University of Twente

A Spatio-Temporal Point Process Model for Firemen Demand in Twente

Bachelor Thesis

Author:
Mike Wendels

Supervisor:
prof. dr. M.N.M. van Lieshout

Stochastic Operations Research
Applied Mathematics

31 March 2017

Contents

A Spatio-Temporal Point Process Model for Firemen Demand in Twente

Mike Wendels[1]

Department of Applied Mathematics, Chair Stochastic Operations Research,
University of Twente, P.O. Box 217 NL-7500 AE Enschede, The Netherlands

31-3-2017

Abstract: In this thesis a spatio-temporal point process will be proposed for modelling firemen demanding emergency calls in the region Twente in the Netherlands. The modelling technique will be described for the level 1a classifications of firemen demanding emergency calls "fire", "service", "accident", "alert" and "environmental". Making accurate expectations for these kinds of emergency calls in the future is very important for emergency services, since it can improve the prevention behaviour and the scheduling of the fire departments and therefore the quality of help. Improvement of the prevention behaviour is made possible because the model describes the influences of the involved covariates on each class of emergency calls. Scheduling could be improved since the number of emergency calls with the corresponding locations and classes can be predicted for future days. In this way it can be predicted for every fire department how many and which kinds of emergency calls they will have to treat the next days. Nowadays these predictions are often made by the industry practice model of partitioning the region of interest in polygons and base the expected number of emergency calls on corresponding information of the past by taking averages. But spatio-temporal point process models have proven to be the more accurate and robust model, since scientific research highly improved the theory for spatial point pattern analysis the last few decades. Spatial point process modelling has also be simplified by the many tools for analysing spatial point patterns available in the `spatstat` package in `R`, available from CRAN (2006). This thesis provides extensions to some of these tools, because a spatio-temporal point process will be developed for the emergency calls rather than a purely spatial point process. This spatio-temporal point process actually involves an ensemble of spatio-temporal point processes for the emergency calls of each level 1a class and each of these models thus have to be modelled separately. These individual spatio-temporal point processes will then be modelled as inhomogeneous Poisson processes for which the intensity function is dependent on spatial and temporal covariates. After modelling, the precise influences of each covariate involved on each kind of emergency calls will be known.

Key words: spatio-temporal point process, spatial point pattern, time series, inhomogeneous Poisson process, maximum pseudolikelihood estimator

[1]Student in Applied Mathematics at the University of Twente, Enschede, The Netherlands, `w.h.m.wendels@student.utwente.nl`

1 Introduction

Nowadays, emergency services base their logistics and prevention behaviour strongly on their expectations for the emergency calls in the future. Making these expectations as accurate as possible has been (and still is) a hot topic in scientific research. Because the better the predictions of emergency calls are, the better emergency services can anticipate their logistics and prevention behaviour on it.

Every emergency call has a time $t \in \mathbb{R}^+$ and a location $\mathbf{x} \in \mathbb{R}^2$ of occurrence. Mostly the description of emergency calls are completed with a classification $c_i, i \in \mathbb{N}$, of the emergency call, for example a description of the emergency call or the priority for serving the emergency call. Emergency services desire to know all the exact times, locations and the potential classifications of the future emergency calls, so they can serve aid on the right time, at the right location and with the right means. Creating a model which predicts these variables exactly for every future emergency call is of course a utopian aim. So each model will in some way involve stochastics and each model will have a horizon for significant prediction.

But how should such a model be built? To give the reader some feeling for these kinds of models, an intuitive and rather simple model is explained first. For this model, the spatial region of interest is partitioned in a set of polygons, and the expected number of emergency calls for each polygon on time t is based on its past information. This is commonly done by taking (weighted) averages of the number of emergency calls in the previous weeks or years for each polygon. This model is the current industry practice (Zhou et al., 2015) and it is not said that applying only these simple statistics provide erroneous expectations. Nonetheless scientific research has developed much more accurate and sophisticated models the last few decades.

The general model for analysing events with a location and time of occurrence is a *spatio-temporal point process (model)*. Such a model is capable of making more accurate predictions in a higher resolution of space and time, since it may take into account detailed distance information in space and time. Modelling a spatio-temporal point process is in general quite complicated, since the causes may also be spatio-temporal next to purely spatial and purely temporal. Often, though, the spatio-temporal causes are not (significantly) present, since the spatial behaviour and temporal behaviour of the events of interest are quite independent. In that case separability may be assumed for the model, in which case the spatial behaviour and the temporal behaviour may be examined individually. Analysing the spatial behaviour involves *spatial point pattern analysis* and analysing the temporal behaviour involves *time series analysis*.

In this thesis, a spatio-temporal point process model will be built for firemen demand, where the region Twente in the Netherlands is the region of interest. This model is based on the data from 1 January 2004 till 31 December 2015. A spatio-temporal point process model seems tailor-made for the problem in this thesis, because emergency calls of each classification all have a specific location and time of occurrence. As a consequence, a spatio-temporal point process can be made for each different class[2]. For each spatio-temporal point process, separability will be assumed. Although there are also no indications of significant spatio-temporal causes for each class, separability is mainly assumed for simplification, since the spatial and temporal behaviour of the firemen demanding emergency calls can as a consequence be examined individually.

[2]The observant reader may note that there could be some dependence between different classes, which makes modelling these classes separately an erroneous choice. If this is the case, a multivariate spatio-temporal point process model is the better option, since they also model the dependencies between different classes.

The aim of this thesis is not only to build a spatio-temporal point process for predicting the future emergency calls of each class[3], but also to get a thorough understanding of the causes of the emergency calls of that class. The former purpose serves for improving the logistics of the fire departments in Twente in both space and time, by anticipating on the predictions of the model. The latter purpose serves for optimizing the prevention behaviour by noting the precise causes of emergency calls of different classes. Both purposes are merged in a spatio-temporal point process, because the predictions of this kind of model are based on the information about the causes of the events.

As a consequence, the emergency calls will be classified by their description rather than by their priority of being served, because the causes of the emergency calls are more dependent on the description of the emergency call than on the priority. The fire departments in Twente classify the description of an emergency call in one of the five following classes: "fire", "service", "accident", "alert" and "environmental". Although there are more kinds of classification, this kind of classification, called the *level 1a classification system*, is the most general and therefore the recommended one. Each of these classes could also be further classified, but these subclasses will not be involved in the modelling described in this thesis.

It will turn out that the dependence of emergency calls on covariates is the most significant cause in Twente, rather than the dependence of emergency calls on other emergency calls. The former cause is called *trend* and the latter cause is called *stochastic interaction*. As a consequence, the models for each class purposed in this thesis will only involve trend as cause. The emergency calls may depend on spatial or temporal covariates[4], for example the population density or a binary variable indicating whether or not the day of interest is 31 December.

Analysing the influences of spatial and temporal covariates on the occurrences of the emergency calls will then be done by comparing the values of these covariates with the number of emergency calls in the region Twente in the time period from 1 January 2004 till 31 December 2015. The relations representing these influences can then be implemented in the model. This will only be done for the covariates that happen to have the most influence, though, for making the model not too complex. After the trends are discovered, a spatio-temporal point process can be built with (extensions of) the `spatstat` package in `R`, which is made available by CRAN (2006).

In which way could optimization with such a model then be achieved? This could be done by adapting the logistics of the fire departments in Twente according to the model of interest, for example by optimizing the scheduling. Although most fire departments in Twente rely on volunteers, optimization could still try to reduce operational costs or to save time. This could implicitely lead to improving the quality of the help. Optimization based on the model could also check whether an extra fire department would be beneficial and where it should be placed. All these kinds of optimization could be done by dynamic programming. This thesis will not involve optimization of logistics according to the model, though.

Next to optimization in logistics, the prevention behaviour can be adapted with the model made, as mentioned earlier. From this model, the fire departments could see influential covariates which cause many emergency calls (of a specific class) in Twente. After finding these hazards, they can try to reduce their influence by reducing or removing the hazards or alerting people for them.

[3]The model of interest is actually an ensemble of spatio-temporal point process models for each class.
[4]As a consequence of separability, spatio-temporal covariates are not allowed

The spatio-temporal point processes proposed in this thesis are inhomogeneous Poisson processes, since the occurrences of emergency calls are assumed to depend on covariates, but not to depend on each other. The important challenge for the modelling is to find the intensity function $\lambda(\mathbf{x}, t), \mathbf{x} \in \mathbb{R}^2, t \in \mathbb{R}^+$, of the inhomogeneous Poisson process for each class of emergency calls. The intensity function is actually the tool for translating the spatial and temporal information of the data to the model. Next to that it has the intuitive property of representing a measure for the expected number of emergency calls for an infinitesimal region around $\mathbf{x}$ and t.

The spatial and temporal information for this intensity function can completely be extracted from the spatial covariate analysis and the temporal covariate analysis, respectively, since the model is assumed to depend only on trend[5]. For the spatial covariate analysis, Twente will be subdivided into a grid of 6291 squares of 500 meter. For the temporal covariate analysis, each year involved will be subdivided in 365 days[6] and so the period of 1 January 2004 till 31 December 2015 will be subdivided into a grid of 4380 days. The extracted information about the influences of the significant covariates will then be translated to the intensity function with help of the `spatstat` package in `R`.

In the following section, the important theory for this thesis will first be summarized, involving spatial point pattern analysis, spatial point process modelling and the extension of the spatial point process to a spatio-temporal one. In section 3, the emergency call data of Twente will be cleaned and analysed. According to this analysis, the spatio-temporal point process model for each class of emergency calls will be chosen. The section concludes by analysing the erroneous data to discover possible trends in the occurrences of such data.

In section 4 the spatial and temporal covariates for this thesis will be selected and will be examined by the spatial and temporal covariate analysis, respectively. In section 5, the method for modelling the discovered covariate information in the intensity function will be explained and the spatio-temporal point process will be built for each class of emergency calls. The complete model, which is the ensemble of the spatio-temporal point processes for each class of emergency calls, will then be validated by comparing them to the available data of emergency calls from 1 January 2016 till 7 December 2016. Section 6 concludes.

The emergency call data is provided by the head fire department in Twente. The cleaning and analysis of it will partially be done by *Microsoft Excel* and partially by the open source program *QGIS*. The remainder of the analysis and the fitting of the involved models of it will be done by `R`.

[5] Even if the model also depends on stochastic interaction, the intensity function for the inhomogeneous Poisson process cannot model the information about these dependencies, since the inhomogeneous Poisson process is only capable of modelling trend.

[6] As will be explained later, leap years are transformed to regular years to make an adequate analysis possible.

2 Literature review

In this section, the reader is given a brief introduction to the theory of spatio-temporal point processes, so he or she will be able to understand the discussions in the remainder of this thesis. The literature review is mostly based on Diggle (1983).

To start with the theory, the two fundamental definitions about spatio-temporal point processes will be given. These definitions will be loosened versions of the formal ones, since these quantitative definitions do the job for this thesis and therefore simplify the discussion in the remainder of this thesis. These definitions are based on Diggle (1983) and Turner (2009). For the formal definitions, the reader is referred to Van Lieshout (2000), Møller and Waagepetersen (2003) and Daley and Vere-Jones (2007).

Definition 2.1 Let $A \subset \mathbb{R}^m, m \in \mathbb{N}$. An *m-dimensional spatial point pattern* S is a data set $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}, \mathbf{x}_i \in A, 1 \leq i \leq n$, in the form of a set of points, distributed within a region of interest A. An *event* $\mathbf{x}_i, 1 \leq i \leq n$, is an element of S. ■

Definition 2.2 Let X_T be a random variable, which takes values in the form of a spatial point pattern out of all possible spatial point patterns for a region of interest $A \subset \mathbb{R}^m, m \in \mathbb{N}$. A *spatio-temporal point process* is a process which generates values for the random variable X_T for a time period of interest $T \subset \mathbb{R}^+$. ■

In most cases, and also in this thesis, $m = 2$, so the region of interest is a bounded subset[7] of the plane $\mathbb{R}^2$. For the general theory ($m \geq 2$), one should read Diggle (1983).

If space and time (covariates) exert influences on the value of the random variable X_T of definition 2.2 independently of each other, the spatio-temporal model becomes *separable*. As mentioned, analysing the spatial and temporal behaviour of the spatio-temporal problem can then be done separately. A spatio-temporal point process is then made by first modelling a spatial point process for the spatial part of the problem, implementing the spatial behaviour analysed. Thereafter this model is extended into time by a time series, which describes the evolution of the spatial point process in time. In this way the separability assumption simplifies the modelling.

For many models, and also for the model proposed in this thesis, separability seems a reasonable assumption. Therefore, the theory about the spatial part and the temporal part of modelling will be examined individually. First, some theory about spatial point pattern analysis will be explained. Then it will be explained how to use the results of this analysis for spatial point process modelling. The section concludes by explaining how to extrapolate the spatial point process in time to a spatio-temporal point process. The time series analysis which gives the information needed for this extrapolation is analogous to the spatial point pattern analysis in this thesis. Therefore extensive theory about time series analysis will not be explained.

2.1 Spatial point pattern analysis

Before any theory about spatial or spatio-temporal point process modelling will be discussed, two very important restrictions have to be made. These restrictions involve that all the spatial point processes (and so spatio-temporal point processes) in this thesis are assumed to be *stationary* and *isotropic*, unless stated otherwise. According to Diggle (1983), the definitions are as follows:

[7] A region of interest involves a bounded subset, since it will never have an infinite area in practice.

Definition 2.3 Let $A \subset \mathbb{R}^2$ the region of interest. A process is *stationary* if all probability statements about the process in any subregion $B \subseteq A$ are invariant under arbitrary translation of B in A. ■

Definition 2.4 Let $A \subset \mathbb{R}^2$ the region of interest. A process is *isotropic* if all probability statements about the process in any subregion $B \subseteq A$ are invariant under arbitrary rotation of B in A. ■

How these assumptions simplify the problem will be explained later. For now it is important to think about the fundamental idea of spatio-temporal point processes. Every model should predict values for the random variable X_T defined as in definition 2.2. But special attention should be paid to the fact that the random variable X_T will not be described by its probability density function, as commonly is the case with random variables, since the probability density function in the context of the random variable X_T is hard to understand and thus hard to work with. A more intuitive description is given by the *intensity function* $\lambda(\mathbf{x}, t), \mathbf{x} \in \mathbb{R}^2, t \in \mathbb{R}^+$.

The reason why the intensity function $\lambda(\mathbf{x}, t)$ is so intuitive is that $\lambda(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$ describes approximately the probability of an event in the subregion $\mathrm{d}\mathbf{x}$ of the region of interest and in the time interval $\mathrm{d}t$. This can also be extended to greater subregions and time intervals. Let $B \subseteq A$ be a subregion of the region of interest $A \subset \mathbb{R}^2$ and $U \subseteq T$ a subperiod of the time period of interest $T \subset \mathbb{R}^+$ for a spatio-temporal point process. Then the expected value $\mathbb{E}[\cdot]$ of (the random variable representing) the number of events $N(B \times U)$ in subregion B and subperiod U is given by:

$$\mathbb{E}\big[N(B \times U)\big] = \int_U \int_{\mathrm{B}} \lambda(\mathbf{x}, t)\ \mathrm{d}\mathbf{x}\ \mathrm{d}t, \qquad \mathbf{x} \in \mathbb{R}^2, t \in \mathbb{R}^+ \tag{1}$$

This result follows from the definition of the intensity function. Before giving a definition, it must be remarked that there does not exist a general kind of intensity function, since different kinds of intensity functions are needed for the different ways a spatial point pattern can be analysed.

In this thesis, only the first-order and second-order properties of the involved spatial point patterns are taken into account, described by the *first-order intensity function* and *second-order intensity function*, respectively. In general, and also in this thesis, the first-order and second-order properties give sufficient information about the random variable X_T and so higher-order properties do not have to be described. For a probability density function as descriptor of X_T, the first-order and second-order properties are described by the first moment and second moment of the probability density function, respectively. So the first-order intensity function and the second-order intensity function are the analogous versions of the first moment and second moment of the probability density function.

One last remark should be made before the first-order intensity function and the second-order intensity function are defined. Since the focus in this part of the thesis lies on modelling spatial point processes, the temporal aspect will be disregarded for the moment and so the period $T \subset \mathbb{R}^+$ for which values for X_T are generated will be fixed for now. Therefore, the values for temporal variable t in the intensity functions are fixed for now and the intensity functions will therefore only be denoted by the spatial variable $\mathbf{x}$, until the discussion of extending the spatial point process to a spatio-temporal one. Now this is mentioned, the first-order intensity function and the second-order intensity function will be defined, according to Diggle (1983).

Definition 2.5 Let $\mathbb{E}[\cdot]$ be the expected value of a random variable, $b_{\mathbf{x}}(s)$ be an open disc with centre $\mathbf{x} \in \mathbb{R}^2$ and radius s, $N(b_{\mathbf{x}}(s))$ be the number of events of the spatial point pattern of interest in this disc and $|\cdot|$ be the operator giving the area of a polygon. Then

$$\lambda(\mathbf{x}) = \lim_{s \to 0} \frac{\mathbb{E}[N(b_{\mathbf{x}}(s))]}{|b_{\mathbf{x}}(s)|}, \qquad \mathbf{x} \in \mathbb{R}^2 \tag{2}$$

is called the *first-order intensity function* and

$$\lambda_2(\mathbf{x}, \mathbf{y}) = \lim_{s_1, s_2 \to 0} \frac{\mathbb{E}[N(b_{\mathbf{x}}(s_1))N(b_{\mathbf{y}}(s_2))]}{|b_{\mathbf{x}}(s_1)||b_{\mathbf{y}}(s_2)|}, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^2 \tag{3}$$

is called the *second-order intensity function.* ■

And now the benefits of the stationarity and isotropy assumptions can be shown. For a stationary and isotropic process:

$$\lambda_2(\mathbf{x}, \mathbf{y}) = \lambda_2(r), \qquad r \in \mathbb{R}^+ \tag{4}$$

where $r \in \mathbb{R}^+$ is the distance between $\mathbf{x}$ and $\mathbf{y}$. Without the stationarity and isotropy assumptions, the second-order intensity function would have been much more complicated.

Now one can clearly see from the definitions that intensity functions give measures for the intensities of occurrences of events. But to let the intensity function represent this measure for a spatial point pattern of interest, the spatial information of this spatial point pattern has to be translated to an intensity function. So it is important to know how spatial information is classified. There are roughly speaking three classifications, according to Diggle (1983):

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

Examples of CSR, aggregated and regular spatial point patterns are given in figure 1.

But how discover which classification fits a spatial point pattern S over the region of interest A? As a first step in answering this question, the following hypothesis test will be executed:

$$\begin{aligned} &H_0\text{: } S \text{ is CSR distributed over } A. \\ &H_1\text{: } S \text{ is not CSR distributed over } A. \end{aligned} \tag{5}$$

The reason why hypothesis test (5) is executed first, is that modelling will be greatly simplified (although it will also make less sense) if H_0 is accepted, since events then tend to occur at each place in A with the same probability. As later will be explained, the (homogeneous) Poisson process model fits S if H_0 is accepted and this process can be modelled very easily. But if H_0 is rejected, things start to get interesting. In that case, the spatial point pattern is aggregated or regularly distributed and modelling makes more sense, but is also more difficult. So actually rejecting H_0 can be seen as a threshold condition for spatially modelling the data.

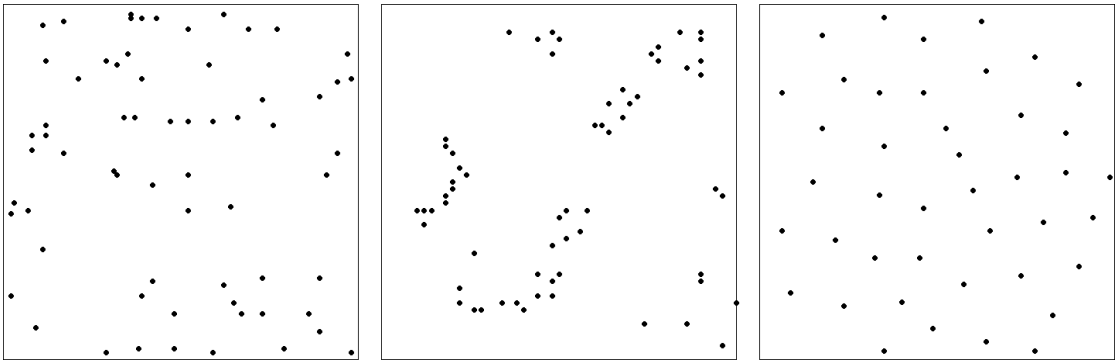

Figure 1: CSR spatial point pattern (left), aggregated spatial point pattern (middle), regular spatial point pattern (right), which respectively represent the datasets `japanesepines`, `redwood` and `cells` from the `spatstat` package in R, available from CRAN (2006).

To execute hypothesis test (5), it is important to know how spatial point patterns are analysed in general. The two ways to analyse spatial point patterns are analysis by *quadrat counts* and analysis by *distance measures*. In quadrat count analysis the region of interest A is partitioned into a number of quadrats and the number of events of the spatial point pattern of interest S is counted in each quadrat. By applying specific statistics, the information of all quadrats can be compared with each other and the spatial point pattern can be classified as CSR, aggregated or regular.

Because the old-fashioned quadrat count analysis is not very accurate and sensitive to errors, the most preferred analysis is the distance measure analysis. This kind of analysis is based on the continuous distance measure $r \in \mathbb{R}, r \geq 0$, between events of the spatial point pattern of interest S, as defined in equation (4). An empirical distribution functions $\hat{f}(r)$ then describes the spatial information of S and thereafter $\hat{f}(r)$ will be compared with the theoretical probability density function $f(r)$ for a CSR spatial point pattern. If $\hat{f}(r_0)$ is significantly different from $f(r_0)$ for some predetermined value r_0 for r and for some significance level α, the hypothesis test (5) is decided in favour of H_1 and otherwise in favour of H_0.

How could it be decided whether $\hat{f}(r_0)$ is significantly different from $f(r_0)$? This can be done by analysing $\hat{f}(r)$ for $r = r_0$ against the upper and lower critical envelopes $U(r)$ and $L(r)$, respectively. According to Diggle (1983), these are defined as follows.

Definition 2.6 Let S be the spatial point pattern of interest in region $A \subset \mathbb{R}^2$, $\tilde{S}_1, \tilde{S}_2, \ldots, \tilde{S}_n$ be n newly sampled CSR spatial point patterns in A and $\hat{f}_i(r)$ be the empirical distribution function representing the 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

$$U(r) = \max_{i=1,2,\ldots,n} \hat{f}_i(r) \tag{6}$$

$$L(r) = \min_{i=1,2,\ldots,n} \hat{f}_i(r) \tag{7}$$

respectively. ■

The critical envelopes $U(r)$ and $L(r)$ actually set boundaries for the region in which a spatial point pattern is still CSR[8]. So the critical envelopes actually form the boundaries between accepting and rejecting H_0: If $\hat{f}(r_0)$ lies between the critical envelopes, so $L(r_0) \leq \hat{f}(r_0) \leq U(r_0)$, H_0 is accepted. Otherwise H_1 is accepted. The significance level α of hypothesis test (5) is dependent on the number of simulations for the critical envelope n. For instance, $n = 39$ implies $\alpha = 0.05$ (Turner, 2009). If H_0 is rejected, the plot of $\hat{f}(r)$ against $U(r)$ and $L(r)$ also reveals whether S is aggregated or regular. Depending on whether $\hat{f}(r_0) > U(r_0)$ or $\hat{f}(r_0) < L(r_0)$, S is aggregated or regular. Which distribution of S requires which condition is dependent on the distance measure analysis used.

One may now ask how different distance measure analyses are possible. The idea is that each distance measure analysis method describes the spatial information of S from a different point of view. Still each distance measure analysis method depends on the distance measure r, but the methods differ in the context in which they implement this distance measure. This also means that the theoretical probability density function differs per method and therefore also the estimator for it, which is the empirical distribution function.

The most popular distance measure analyses are executed by estimating and analysing *Ripley's reduced second moment function* $K(r)$, the *nearest neighbour distance distribution function* $G(r)$, the *empty space function* $F(r)$ or the *summary function* $J(r)$. Before the analysis methods corresponding to these theoretical probability density functions are discussed, an accurate estimator $\hat{\lambda}$ of the intensity λ is needed, since the analysis methods require such an estimator.

For defining this estimator, let S be the spatial point pattern of interest and A the region of interest. Partition A in m polygons $B_i, 1 \leq i \leq m$, of the same area $|B|$, so $|B_1| = |B_2| = \ldots = |B_m| = |B|$. Further let N_i be the random variable, representing the number of events in polygon B_i, and let n_i be the realisations of this random variable for S. Then an estimator for λ is given by:

$$\hat{\lambda}(m) = \frac{\sum_{i=1}^{m} n_i}{m|B|} \tag{8}$$

The strength of this estimator is that it is unbiased for a CSR distribution of S in A. As will be explained more thoroughly later in this section, a CSR distribution for S means that N_i is Poisson distributed with mean $\lambda|B|$ for every $i, 1 \leq i \leq m$. Because the mean is $\lambda|B|$ for all the m polygons in the partition for a CSR distribution of S, it can easily be derived that $\mathbb{E}\big[\hat{\lambda}\big] = \lambda$.

The estimator of equation (8) is also valid for $m = 1$, so it reduces to $\hat{\lambda} = |S||A|^{-1}$ in this case[9]. With this estimator, the distance measure analysis methods can be explained. For the discussion of these methods, let again S be the spatial point pattern of interest and A be the region of interest. Further, let $\mathbf{x}_i \in S$ and $\mathbf{x}_j \in S, j \neq i$, be two arbitrary different events of S with $e(\mathbf{x}_i, \mathbf{x}_j)$ the edge correction weights[10] for these events, $d(\mathbf{x}_i, \mathbf{x}_j)$ the distance between these events and $\mathbb{I}(d(\mathbf{x}_i, \mathbf{x}_j) \leq r)$ an indicator function which equals 1 if the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is smaller than or equal to r and 0 otherwise. Some distance measure analyses may also be based on the distance between an event $\mathbf{x}_i$ and a set $R \subseteq S \setminus \mathbf{x}_i$, denoted by $d(\mathbf{x}_i, R)$. The operators $\mathbb{E}[\cdot]$ and $\mathbb{P}[\cdot]$ give the expected value and the probability of the argument, respectively.

[8] Note that $U(r)$ and $L(r)$ are in practice not equal to the theoretical probability density function $f(r)$.

[9] Be aware that the operator $|\cdot|$ on S represents the cardinality of S, since S is a set, and that this operator on A represents the area of A, since A is a polygon.

[10] Edge correction weights temper the biasing influences of events in the neighbourhood of the border in A. For a more thoroughly discussion about these weights, see Diggle (1983).

Ripley's reduced second moment function
Ripley's reduced second moment function (or Ripley's K-function) $K(r)$ is defined as:

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

So this function describes S by the number of events contained in the circular neighbourhood of radius r for each of the $|S|$ events in S. The empirical distribution function $\hat{K}(r)$ can then be expressed as:

$$\hat{K}(r) = \hat{\lambda}^{-1}|S|^{-1}\sum_{i\neq j} e(\mathbf{x}_i, \mathbf{x}_j)\mathbb{I}(d(\mathbf{x}_i, \mathbf{x}_j) \leq r), \qquad r \geq 0 \tag{9}$$

where the estimator $\hat{\lambda} = |S||A|^{-1}$ can thus be applied, if λ is unknown. Further, the theoretical probability density function under CSR assumption becomes:

$$K(r) = \pi r^2, \qquad r \geq 0. \tag{10}$$

Deviations $K(r) > \pi r^2$ and $K(r) < \pi r^2$ indicate aggregation and regularity, respectively.

Nearest neighbour distance distribution function
The nearest neighbour distance distribution function $G(r)$ is defined as:

$$G(r) = \mathbb{P}\big[\text{distance from an arbitrary event of } S \text{ to the nearest other event of } S \text{ is at most } r\big]$$

So this function describes S by the distances r between the events of the nearest neighbouring event pairs. Let $\mathbf{x}_j \in S \setminus \mathbf{x}_i$ denote the nearest neighbour of event $\mathbf{x}_i \in S$. The empirical distribution function $\hat{G}(r)$ can then be expressed as:

$$\hat{G}(r) = |S|^{-1}\sum_{i\neq j} e(\mathbf{x}_i, \mathbf{x}_j)\mathbb{I}(d(\mathbf{x}_i, S \setminus \mathbf{x}_i) \leq r), \qquad r \geq 0 \tag{11}$$

and the theoretical probability density function under CSR assumption becomes:

$$G(r) = 1 - e^{-\lambda\pi r^2}, \qquad r \geq 0. \tag{12}$$

where the intensity function λ can again be estimated by equation 8, if it is unknown. Deviations $G(r) > 1 - e^{-\lambda\pi r^2}$ and $G(r) < 1 - e^{-\lambda\pi r^2}$ indicate aggregation and regularity, respectively.

Empty space function
The empty space function $F(r)$ is defined as:

$$F(r) = \mathbb{P}\big[\text{distance from an arbitrary point in } A \text{ to the nearest event of } S \text{ is at most } r\big]$$

This function seems similar to the $G(r)$ function. The difference, though, is that this time only one event in the pair is part of S. The other one is an event of a newly sampled CSR spatial point pattern $\tilde{S}$ and such an event is called a point, denoted as $\tilde{\mathbf{x}}_i \in \tilde{S}$. Let $\mathbf{x}_i \in S$ denote the nearest neighbour of the point $\tilde{\mathbf{x}}_i \in \tilde{S}$. The empirical distribution function $\hat{F}(r)$ can then be expressed as:

$$\hat{F}(r) = |\tilde{S}|^{-1}\sum_{\tilde{S}} e(\tilde{\mathbf{x}}_i, \mathbf{x}_i)\mathbb{I}(d(\tilde{\mathbf{x}}_i, \mathbf{x}_i) \leq r), \qquad r \geq 0 \tag{13}$$

and the theoretical probability density function under CSR assumption becomes:

$$F(r) = 1 - e^{-\lambda\pi r^2}, \qquad r \geq 0. \tag{14}$$

where the intensity function λ can again be estimated by equation 8, if it is unknown. Deviations $F(r) < 1-e^{-\lambda\pi r^2}$ and $F(r) > 1-e^{-\lambda\pi r^2}$ indicate aggregation and regularity, respectively.

Summary function
The summary function $J(r)$ is defined as:

$$J(r) = \frac{1 - G(r)}{1 - F(r)}, \qquad r \geq 0. \tag{15}$$

So this function is based on the combination of nearest neighbour distance analysis and empty space analysis for describing S. The benefit of $\hat{J}(r)$ in comparison with $\hat{G}(r)$ and $\hat{F}(r)$ is that it can be computed explicitly for a wide range of spatial point patterns. The empirical distribution function $\hat{J}(r)$ simply becomes:

$$\hat{J}(r) = \frac{1 - \hat{G}(r)}{1 - \hat{F}(r)}, \qquad r \geq 0. \tag{16}$$

and the theoretical probability density function under CSR assumption becomes:

$$J(r) = 1, \qquad r \geq 0. \tag{17}$$

Deviations $J(r) < 1$ and $J(r) > 1$ indicate aggregation and regularity, respectively.

For a more thorough discussion about these distance measure analysis methods, for example how these theoretical probability density functions are found, see for instance Diggle (1983) for the case of $K(r), G(r), F(r)$ and Van Lieshout and Baddeley (1996) for $J(r)$.

So now a spatial point pattern can be classified in a CSR, an aggregated or a regular spatial point pattern. But still little is said about what causes a spatial point pattern to be distributed as one of these classes. One may remember from section 1 that there are two different causes for a distribution of a spatial point pattern. The distribution of the events may be caused by the influences of covariates, called *trend*, or by the influences of other events, called *(stochastic) interaction*. By investigating these causes in more detail, further classification of the spatial point pattern is possible and therefore a more accurate description of the spatial information of the spatial point pattern. In this thesis, only the cause of an aggregated distribution will be examined more thoroughly, since the problem appears to involve an aggregated spatial point pattern (see section 3).

First, the concepts of trend and interaction will first be illustrated in the context of the aggregated spatial point pattern for redwood seedlings from figure 1b, to give the reader a better understanding of these concepts. Suppose the cause of this aggregation is a tendency for the redwood seedlings to grow close to their parents. In this case, the locations of seedlings of the same parent are mutually stochastically dependent on each other, since they all aggregate around this parent, and the cause of the aggregation is stochastic interaction. Another cause of the aggregation may be that some places in the region of interest for this spatial point pattern are more fertile than other places. And since the variation of soil fertility over the region of interest may be introduced as a covariate for the model, the aggregation is now caused by trend. Of course, a combination of the two causes may also be possible.

But how to determine whether trend or interaction is the cause of an aggregated distribution for spatial point pattern S in region A and time period T? Bartlett (1964) mentions this can only

be concluded from n multiple independent and identically distributed realisations $S_1, S_2, \ldots, S_n$ of the random variable X_T from definition 2.2, where X_T represents the (aggregated) spatial point pattern of interest S in the fixed time period $T \subset \mathbb{R}^+$. There are now two different kinds of aggregation possible: the aggregation of the events are around the same points for all the n realisations $S_1, S_2, \ldots, S_n$ in A or are around different points for different realisations. The former kind is caused by trend and the latter kind by stochastic interaction, as one may reason intuitively. Aggregation caused by trend is often called *inhomogeneity* or *heterogeneity*.

In theory, n realisations $S_1, S_2, \ldots, S_n$ of S are not possible, though, since S is the unique spatial point pattern for period T in region A. To solve this problem, time invariance is often assumed for the distribution of S and period T is partitioned in n subperiods $U_i, 1 \leq i \leq n$ of the same length. Then a set of n multiple independent and identically distributed realisations for X_T can be represented by $S_{U_1}, S_{U_2}, \ldots, S_{U_n}$, where $S_{U_i}, 1 \leq i \leq n$ represents the spatial point pattern consisting of the events of S occurred in subperiod T_i. Nonetheless, time invariance is a strong and therefore an often erroneous assumption, causing the conclusions to be interpreted with much caution.

Analysing the cause of aggregation by comparing the clustering points of each spatial point pattern $S_{U_i}, 1 \leq i \leq n$ is a very global and informal analysis. A more formal analysis depends on extensions to the distance measure analyses earlier described. These analysis methods will be explained in section 3.

2.2 Spatial point process modelling

Now some different classifications and their causes for spatial point patterns are known, a start can be made to model each classification and cause in an appropriate way. As earlier mentioned, this thesis focuses on spatial point process models for aggregation. The focus will be even further specified to aggregation caused merely by trend, so inhomogeneity. This because (aggregation caused by) stochastic interaction requires that each event of S must be considered in relation to every other event in S, what makes such models horribly complicated and intractable from an analytical point of view. The verification of this choice, so the assumption that there is (approximately) no stochastic interaction between the events, will later in this thesis be discussed.

One may further question why a complete spatial point process model is required, when the intensity function describes all the spatial information? Well, such a model is the stochastic mechanism for translating this spatial information (packed in the intensity function) into simulations of new spatial point patterns, which therefore have the same distribution[11] as S. Such a spatial point process model is the actual aim and although the intensity function is a great part of it, it is not the whole model.

There are different spatial point process models, made for different kinds of spatial information and each implementing the intensity function in a different way. The search in this thesis concerns finding a spatial point process model for inhomogeneity. For finding such a model, a start will be made by first modelling a CSR spatial point pattern. This will be done by the most fundamental spatial point process model, the *(homogeneous) Poisson spatial point process*, because many applicable spatial point process models are based on this model.

[11] The simulations will never be distributed exactly the same. Part of this arises from the stochastics of the models, so in some way a lack of knowledge. The other part may arise from a wrong or incomplete model used for the modelling.

Definition 2.7 Let $A \subset \mathbb{R}^2$ be the region of interest and $N(A)$ be the number of events in A. Then a point process is *(homogeneous) Poisson* if it satisfies the following conditions:

1. For some $\lambda > 0$, $N(A)$ is Poisson distributed with mean $\lambda|A|$.
2. Given $N(A) = s$, the s events in A form an independent random sample from the uniform distribution on A.

In this process, the intensity function is implemented in the parameter λ. ■

As one can see, λ does not depend on even a single variable. This is what should be expected for a CSR distributed spatial point pattern, because in such a pattern there is no tendency for events to occur at specific places. This intensity function λ can be estimated by the estimator $\hat{\lambda}(m)$ defined as in equation (8), for each $m \in \mathbb{N}$.

Another powerful property of the homogeneous Poisson point process is the following result:

Theorem 2.1 Let $S \in \mathbb{R}^2$ be a CSR spatial point pattern, B_1 and B_2 be two arbitrary subregions of the region of interest $A \subset \mathbb{R}^2$ and $N(B_1)$ and $N(B_2)$ be the number of events in B_1 and B_2, respectively. If $B_1 \cap B_2 = \emptyset$, $N(B_1)$ and $N(B_2)$ are independent.

Proof. Define $B = B_1 \cup B_2$, $p = |B_1||B|^{-1}$ and $q = 1 - p = |B_2||B|^{-1}$. By the conditions, S can be modeled as a Poisson point process in both regions B_1 and B_2. Using the second property of a Poisson point process (Definition 1.6) gives:

$$\mathbb{P}\big[N(B_1) = x, N(B_2) = y \;\big|\; N(B) = x + y\big] = \binom{x+y}{x} p^x q^y, \qquad x \geq 0, y \geq 0$$

and the first property gives the unconditional joint probability distribution of $N(B_1)$ and $N(B_2)$:

$$\begin{aligned}
\mathbb{P}\big[N(B_1) = x, N(B_2) = y\big] &= \binom{x+y}{x} p^x q^y \left[e^{-\lambda|B|} \frac{(\lambda|B|)^{x+y}}{(x+y)!}\right], && x \geq 0, y \geq 0 \\
&= \frac{(x+y)!}{x!\,y!} \frac{|B_1|^x}{|B|^x} \frac{|B_2|^y}{|B|^y} e^{-\lambda(|B_1|+|B_2|)} \frac{\lambda^{x+y}|B|^{x+y}}{(x+y)!}, && x \geq 0, y \geq 0 \\
&= \left[e^{-\lambda|B_1|} \frac{(\lambda|B_1|)^x}{x!}\right] \left[e^{-\lambda|B_2|} \frac{(\lambda|B_2|)^y}{y!}\right], && x \geq 0, y \geq 0 \\
&= \mathbb{P}\big[N(B_1) = x\big] \mathbb{P}\big[N(B_2) = y\big], && x \geq 0, y \geq 0 \;\square
\end{aligned}$$

Theorem 2.1 is the cornerstone of Poisson point processes. It mentions that the events of S are mutually stochastic independent if modelled as a Poisson process. This is the reason why Poisson processes (or extensions to it) are mostly used to model spatial point patterns which (are assumed to) have no stochastic interaction between their events.

Theorem 2.1 also simplifies the second-order intensity function of equation (4) even further to (Diggle, 1983):

$$\lambda_2(r) = \lambda^2, \qquad r \in \mathbb{R}^+. \tag{18}$$

Although such a second-order intensity function is not unique for the homogeneous Poisson process (Baddeley and Silverman, 1984), the Poisson process is a very intuitive and fundamental stochastic mechanism with this characteristic. But this simplicity of the homogeneous Poisson

process has a price, because it could only well be fitted to a CSR spatial point pattern. So now an extension to the homogeneous Poisson process model will be described, which also is capable of modelling inhomogeneity. This extension is called the *inhomogeneous Poisson point process.*

Definition 2.8 Let $A \subset \mathbb{R}^2$ be the region of interest and $N(A)$ be the number of events in A. Then a point process is *inhomogeneous Poisson* if it satisfies the following conditions:

1. For some function $\lambda(\mathbf{x}) > 0, \forall \mathbf{x} \in A$, $N(A)$ is Poisson distributed with mean $\int_A \lambda(\mathbf{x}) d\mathbf{x}$.
2. Given $N(A) = s$, the s events in A form an independent random sample from the distribution on A with a probability density function proportional to $\lambda(\mathbf{x})$.

In this process, the intensity function is implemented in the parameter $\lambda(\mathbf{x})$. ■

Because the intensity function now is variable in $\mathbf{x}$, this model is able to represent differences between locations in the region of interest A. Note that every spatial point pattern generated by the inhomogeneous Poisson point process is expected to have the highest number of events around the same places, so this stochastic mechanism models inhomogeneity, but not aggregation in general. The trend causing the inhomogeneity is described by a number of p covariates $C_i(\mathbf{x}), 1 \leq i \leq p$, in the model, which are represented in the model by the intensity function[12]:

$$\lambda(\mathbf{x}) = \lambda(C_1(\mathbf{x}), C_2(\mathbf{x}), \ldots, C_p(\mathbf{x})) \tag{19}$$

How should this intensity function $\lambda(\mathbf{x})$ then be determined? There is no easy estimator for this process, in contrast to the estimator for the intensity function of the homogeneous Poisson process. Estimating $\lambda(\mathbf{x})$ is complicated because it depends on covariates $C_1, C_2, \ldots, C_p$, whose influences also have to be expressed. So the intensity function should describe how different quantities of different covariates influence the occurrence of an event.

A start for estimating the intensity function can be made by determining the global relation $\hat{\lambda}_\theta(C_1(\mathbf{x}), C_2(\mathbf{x}), \ldots, C_p(\mathbf{x}))$ between the occurrence of an event and each of the p covariates $i, 1 \leq i \leq p$, for example by analysing these relations in the past. The coefficients θ of this relation, called the *influence coefficients*, should then be found in a way as to fit the intensity function, what thus means that the following equation should hold:

$$\mathbb{E}\big[N(B)\big] = \int_{\mathrm{B}} \hat{\lambda}_\theta(C_1(\mathbf{x}), C_2(\mathbf{x}), \ldots, C_p(\mathbf{x}))\ \mathrm{d}\mathbf{x} \tag{20}$$

for any subregion $B \subseteq A$ of the region of interest $A \subset \mathbb{R}^2$. In this way, the function $\hat{\lambda}_\theta(\mathbf{x}) = \hat{\lambda}_\theta(C_1(\mathbf{x}), C_2(\mathbf{x}), \ldots, C_p(\mathbf{x}))$ becomes an estimator for the actual intensity function $\lambda(\mathbf{x})$ of equation (19). The relations can be found by regression analysis and the influence coefficients by estimating the maximum pseudolikelihood. The regression analysis will be discussed in section 4 and the maximum pseudolikelihood estimating technique will be discussed in section 5.

The inhomogeneous Poisson point process is also capable of modelling a CSR spatial point pattern, because the inhomogeneous Poisson point process simplifies to the homogeneous one when $\lambda(\mathbf{x})$ is constant. This is no surprise, because the inhomogeneous Poisson process is an extension to the homogeneous one.

[12] Note that even when the intensity function only depends on the location $\mathbf{x}$ in A, this can also be seen as a covariate $C_i(\mathbf{x}), 1 \leq i \leq p$, so there will always be at least one covariate involved in the model.

The models discussed are part of the class of spatial point process models called *Poisson process models*, which are well-fitted for independently occurring events. To conclude the discussion about spatial point processes, some extensions and other modelling techniques based on interaction are explained. An extension to the inhomogenenous Poisson process which makes it also capable of modelling stochastic interaction between events, is the *Cox process*. This extension is based on making a random variable $\Lambda(\mathbf{x})$ for the intensity function. In this way, different simulations of the spatial point pattern of interest have different points in the region of interest around which the events aggregate. But modelling only based on stochastic interaction is also possible, for example by *pairwise interaction processes* like the *Strauss process* and the *Geyer model*. The former is based on modelling regularity and the latter is the extension for also modelling aggregation based only on interaction.

The spatial point processes discussed did not take into account eventual dependencies between events of distinguishable classes. If events of different classes depend (significantly) on each other, these dependencies can be implemented in a *multivariate spatial point process*. This process consists of the univariate spatial point processes for each class and the dependencies between them.

Except for the Cox process, all models previously discussed are part of the so called *Gibbs point processes* (also known as *Markov point processes*). But spatial point pattern modelling can also be done by *mixture models*, such as *Bayesian semiparametric mixture models* like the *Dirichlet process* with beta or normal densities or the *finite Gaussian mixture model* with a fixed number of components. These models serve as nonparametric models for spatial point patterns. In other words, these models need no thorough information about the classifications and causes of the spatial point patterns involved, since they are not made to model these information. For a more detailed description of Gibbs point process models, the reader is referred to Van Lieshout (2000), Møller and Waagepetersen (2003) and Turner (2009) and for a more detailed description about mixture models, the reader is referred to Zhou et al. (2015).

2.3 Spatio-temporal point process modelling

Spatio-temporal point process modelling involves modelling the spatial and temporal behaviour of one or several classes of events. Because the spatio-temporal point processes discussed in this thesis are univariate, only one class of events is modelled for each spatio-temporal point process. The information about spatial and temporal behaviour of this class of events is translated to a function, well-known as the intensity function $\lambda(\mathbf{x}, t), \mathbf{x} \in \mathbb{R}^2, t \in \mathbb{R}^+$.

Now a spatial point process model can be made, it can be extrapolated in the time to make it a spatio-temporal point process model. A spatial point process is just a spatio-temporal point process for a fixed time period $T \subset \mathbb{R}^+$, as can clearly be seen from definition 2.2. To extrapolate a spatial point process in time, the temporal behaviour of the stochastic variable X_T from definition 2.2 has to be analysed. The intensity of the occurrences of events may for example depend on the time of the day, the season of the year or on the weather. The intensity function of the spatial point process has to be completed with this temporal information. The temporal information can be filtered by time series analysis. For analysing time series, it is useful to first know the exact definition of a time series, though.

Definition 2.9 Let $T \subset \mathbb{R}^+$. A *time series* R is a data set $\{t_1, t_2, ..., t_n\}, t_i \in T, 1 \leq i \leq n$, in the form of a set of points, distributed within a time period of interest T. An *event* $t_i, 1 \leq i \leq n$ is an element of R. ■

Note that this definition is merely definition 2.1 with $m = 1$ and some slightly adapted notation, because the dimension is now time instead of space[13]. So a time series is mathematically a one-dimensional spatial point pattern in time and therefore the analysis of time series is just the one-dimensional version of the two-dimensional spatial point pattern analysis discussed earlier.

In time series analysis, events can also be CSR, aggregated or regular distributed and these distributions can also be caused by trend or stochastic interaction. So the one-dimensional versions of the spatial point pattern analyses discussed could be used to examine the behaviour of events in time. In this thesis, though, a more quantitative analysis method will be used, as explained in section 3. Also for the temporal part of the modelling, it will turn out that modelling trend as the only cause for the distributions involved will be a reasonable assumption.

Because time series analysis is the one-dimensional version of spatial point pattern analysis, time series modelling is also the one-dimensional version of spatial point process modelling. Partition the time period of interest $T \subset \mathbb{R}^+$ in subperiods $U_i, 1 \leq i \leq m$ of the same length and let N_i represent the number of events in subperiod U_i. Further let R be the time series of interest for period T. For a CSR distributed time series R, N_i is Poisson distributed with intensity function λ, which now represents the expected number of events occurring per time unit. In the same way, for an inhomogeneous distributed time series R, N_i is inhomogeneous Poisson distributed with intensity function $\lambda(t)$. The dependence of the intensity function on t makes the intensity function able to model different rates of occurrences of events in time. In this way, accumulations of events in time can be modelled for the inhomogeneous distribution.

For an inhomogeneous Poisson process, adding q temporal covariates $C_i, 1 \leq i \leq q$, to this function and estimating this function can be done in a similar, one-dimensional way for estimating the intensity function for a two-dimensional spatial point process. So note that all the previous discussion about two-dimensional spatial point patterns can be used for time series, if it is reduced to one dimension. This makes time series analysis a lot easier to execute.

So the three-dimensional spatio-temporal problem is reduced to a two-dimensional spatial problem, involving spatial point patterns, and a one-dimensional temporal problem, involving time series. Note that this is only made possible, because the spatio-temporal point process is assumed to be separable. For a nonseparable spatio-temporal point process, the spatio-temporal point patterns may not be analysed by a separate two-dimensional spatial point pattern analysis and a time series analysis. This because the distributions of the spatial point patterns are then different for different times. In that case the spatial point pattern analysis extends to $m = 3$, where two dimensions represent space and one dimension represents time. As a consequence, the model should also be an extension of the spatial point process to $m = 3$.

Also for time series models, the (temporal) information is packed in an intensity function $\lambda(t), t \in \mathbb{R}^+$, analogous to $\lambda(\mathbf{x}), \mathbf{x} \in \mathbb{R}^2$, for spatial point processes. A general spatio-temporal point process has the spatio-temporal information of interest packed in the intensity function $\lambda(\mathbf{x}, t), \mathbf{x} \in \mathbb{R}^2, t \in \mathbb{R}^+$, but the assumed separability reduces this expression to:

$$\lambda(\mathbf{x}, t) = \lambda_\sigma(\mathbf{x})\lambda_\tau(t), \qquad \mathbf{x} \in \mathbb{R}^2, t \in \mathbb{R}^+ \tag{21}$$

and the behaviour of $\lambda_\sigma(\mathbf{x})$ and $\lambda_\tau(t)$ can thus be analysed by two-dimensional spatial point pattern analysis and time series analysis, respectively.

[13] Purely mathematically, new notation is not needed, but practically it makes sense and it avoids confusion.

How are covariates then modelled? If there are p spatial covariates $C_{\sigma,i}, 1 \leq i \leq p$, and q temporal covariates $C_{\tau,i}, 1 \leq i \leq q$, of interest, the separate intensity functions $\lambda_\sigma(\mathbf{x})$ and $\lambda_\tau(t)$ are respectively able to represent the p spatial covariates and q temporal covariates:

$$\lambda_\sigma(\mathbf{x}) = \lambda_\sigma(C_{\sigma,1}(\mathbf{x}), C_{\sigma,2}(\mathbf{x}), \ldots, C_{\sigma,n}(\mathbf{x})) \tag{22}$$

$$\lambda_\tau(t) = \lambda_\tau(C_{\tau,1}(t), C_{\tau,2}(t), \ldots, C_{\tau,m}(t)) \tag{23}$$

Note that because of the separability assumption, spatio-temporal covariates $C_i(\mathbf{x}, t), 1 \leq i \leq n$, cannot be modelled anymore. Also remember that the spatial and temporal covariates can only be modelled for models based on trend, such as the inhomogeneous Poisson process. For this thesis, an inhomogeneous distribution of the events will be accepted and the influences of covariates will appear to have a significant effect on this distribution and therefore the spatio-temporal version of the inhomogeneous Poisson process will be chosen as the model to represent the spatio-temporal point process. The spatio-temporal inhomogeneous Poisson process is defined as follows:

Definition 2.10 Let $A \subset \mathbb{R}^2$ be the region of interest, $T \subset \mathbb{R}^+$ be the time period of interest and $N(A \times T)$ be the number of events in the space-time region $A \times T$. Then a point process is *spatio-temporal inhomogeneous Poisson* if it satisfies the following conditions:

1. For some function $\lambda(\mathbf{x}, t) > 0, \forall(\mathbf{x}, t) \in A \times T$, $N(A \times T)$ is Poisson distributed with mean $\int_T \int_A \lambda(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t$.
2. Given $N(A \times T) = s$, the s events in $A \times T$ form an independent random sample from the distribution on $A \times T$ with a probability density function proportional to $\lambda(\mathbf{x}, t)$.

In this process, the intensity function is implemented in the parameter $\lambda(\mathbf{x}, t)$. ■

Of course, $\lambda(\mathbf{x}, t) = \lambda_\sigma(\mathbf{x})\lambda_\tau(t)$ is the intensity function for the spatio-temporal inhomogeneous Poisson processes to be modelled in this thesis. Note that property 1 is exactly the property of equation (1) and property 2 is exactly the independence property for the spatio-temporal point process to be modelled. So it can clearly be seen from definition 2.10 that a spatio-temporal inhomogeneous Poisson process is suitable as a spatio-temporal point process.

It is important to remark that the intensity functions $\lambda_\sigma(\mathbf{x})$ and $\lambda_\tau(t)$ of the spatio-temporal point process are in general not the same intensity functions as $\lambda(\mathbf{x})$ for the analogous (purely) spatial point process and $\lambda(t)$ for the analogous (purely temporal) time series model, respectively.

Theorem 2.2 Let $\lambda_\sigma(\mathbf{x})$ and $\lambda_\tau(t)$ be the intensity functions for the spatial and the temporal part of a separable spatio-temporal point process and let $\lambda(\mathbf{x})$ and $\lambda(t)$ be the intensity functions of the spatial point process and the time series model corresponding to the spatio-temporal point process. Further, let $B \subseteq A$ be a subregion of the region of interest $A \subset \mathbb{R}^2$, $U \subseteq T$ a subperiod of the time period of interest $T \subset \mathbb{R}^+$ for the spatio-temporal point process and let $N(\cdot)$ be the operator which gives the expected number of events in a spatio-temporal region. Then $\lambda_\sigma(\mathbf{x}) \neq \lambda(\mathbf{x})$ and $\lambda_\tau(t) \neq \lambda(t)$ in general.

Proof. A proof by contradiction. The expected number of events in B according to the spatial point process is:

$$\mathbb{E}\big[N(B)\big] = \int_{\mathrm{B}} \lambda(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

According to the spatio-temporal point process, the expected number of events in B becomes:

$$\begin{aligned}\mathbb{E}\big[N(B)\big] &= \mathbb{E}\big[N(B \times T)\big] \\ &= \int_T \int_{\mathrm{B}} \lambda(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t \\ &= \int_T \int_{\mathrm{B}} \lambda_\sigma(\mathbf{x})\lambda_\tau(t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t \\ &= \int_{\mathrm{B}} \left[\lambda_\sigma(\mathbf{x}) \int_T \lambda_\tau(t)\, \mathrm{d}t\right] \mathrm{d}\mathbf{x}\end{aligned}$$

where $\mathbb{E}\big[N(B)\big] = \mathbb{E}\big[N(B \times T)\big]$ according to Møller and Ghorbani (2012). As a consequence, the following relation holds for every subregion $B \subseteq A$:

$$\lambda(\mathbf{x}) = \lambda_\sigma(\mathbf{x}) \int_T \lambda_\tau(t)\, \mathrm{d}t \tag{24}$$

But this relation is not true in general, since the integral $\int_T \lambda_\tau(t)\, \mathrm{d}t$ depends on the subperiod U chosen and therefore does not equal 1 in general. So a contradiction occurs. In an analogous way, this contradiction can be found for $\mathbb{E}\big[N(U)\big]$. $\square$

Why is separability then such a helpful assumption? The answer is that the relations between the occurrences of events and the spatial and temporal covariates stay the same. The only aspect that changes in estimating the intensity function $\lambda(\mathbf{x}, t)$ are the influence coefficients θ_σ and θ_τ for the spatial and temporal covariates, respectively. These coefficients have to be adapted in such a way that equation (1) stays valid and the contradiction above does not occur for the intensity function $\lambda(\mathbf{x}, t)$. The influence coefficients θ_σ and θ_τ can be found by a spatio-temporal extension to the maximum pseudolikelihood estimating technique used for estimating the influence coefficients θ for a spatial point process. This extension will be described in section 5.

3 Exploratory data analysis

In this section, the data set of emergency calls for firemen demand in Twente will be analysed. This data set is kindly provided by the head fire department in Twente and involves data from period T_t, which is the period from 1 January 2004 till 7 December 2016. The model in this thesis, though, will be based on (the data from) period T_m, which is the period from 1 January 2004 till 31 December 2015. The data from $T_v = T_t \setminus T_m$, so from 1 January 2016 till 7 December 2016, will only be used to validate the model. The region of interest for the model will of course be Twente.

The data consist of all emergency calls with their unique ID number and their information about the time, location and classification of each occurred emergency call. These times, locations and classifications are described by many different aspects, which are sometimes redundant for the analysis executed in this thesis. Next to that, some aspects should even be completed to make the analysis in this thesis possible. So the data set will first be filtered and completed at some points to make it amenable for analysis.

After that, the data set will be analysed. In this way, the behaviour for the occurrences of emergency calls will be classified, to choose a model which is capable of accurately representing this classification. Separability will be assumed, what means that analysing the spatio-temporal behaviour can be divided in analysing the spatial behaviour and analysing the temporal behaviour. Analysing the spatial behaviour will be done by spatial point pattern analysis and analysing the temporal behaviour will be done by time series analysis.

Since a spatio-temporal point process will be made per level 1a class, the spatial and temporal information should be analysed for each different class. The methods for the analyses will only be shown for the emergency calls of class "fire", though, since showing the same analysis methods five times is quite meaningless and tedious. So after showing the methods once for the emergency calls with "fire", the results of the other four classes will only be shortly mentioned. The reason why "fire" is chosen as the leading class for the explanation is that this class happens to be the most interesting one, since most emergency calls of this class have high priority and this class is (presumed to be) very dependent on the influences of covariates.

To conclude the section, the discarded data will be inspected. This discarded data involves erroneous data and data deleted to simplify the modelling. By examining these data, possible trends for them may be discovered.

3.1 Filtering and completion of the emergency call data

As mentioned, the data set of emergency calls for firemen demand in Twente will first be manipulated to make it amenable for analysis. Both filtering and completion is needed for the data. The data can be divided in three kinds of data: the temporal data, the spatial data and the classification data which respectively describe the time, location and classification of the occurred emergency calls. These three kinds of data will be examined individually. All the filtering and completion described can be done by *Microsoft Excel* and *QGIS*.

The temporal data is expressed by many columns in the data set, each describing the temporal information from a different angle (quartile, quartile number, month and year combined). All this information is reduced to the point of time and the date, where the point of time is described in hour, minute and second and the date is described in day, month and year.

But these representations of time and date are not useful for the temporal analysis later in this thesis, since this analysis requires an accurate measure for comparing the same times of different years with each other. For example, 13 June 2005 at 10:00 AM and 13 June 2015 at 10:00 AM need the same time description with respect to their relative years. Such a description is made by representing the temporal information in the day $d_i, 1 \leq d_i \leq 365$ and the second $s_i, 1 \leq s_i \leq 31.536 \cdot 10^6$ of the year $i, 2004 \leq i \leq 2016$ with respect to the beginning of that year (1 January, 0:00 AM).

Leap years though are quite problematic in the descriptions of d_i and s_i, because leap years are one day longer than the other "regular" years and therefore cause a biased comparison with regular years. Comparing data of 2012 with data of 2013 for example, 29 February 2012 at 6:30 AM would be compared with 1 March 2013 at 6:30 AM and 31 December 2012 3:00 PM does not even have a day and a time in 2013 to be compared with. To evade the problems with leap years, the description for leap years is adapted as if the leap year was a regular year (so February 29 is omitted in every leap year involved). For these adapted leap years and of course also for the regular years, the information of d_i and s_i are calculated and added to the data set.

For this thesis, though, the exact point of time is not used for analysis, since the period T_m will later in this thesis be discretized in days. The information of $s_i, 1 \leq s_i \leq 31.536 \cdot 10^6$ is added to the data, though, for possible future extensions to the model in this thesis. The information which will be used is the date of the occurred emergency call, so the year of interest $i, 2004 \leq i \leq 2016$, the day of that year $d_i, 1 \leq d_i \leq 365$ and the month of that year $m_i, 1 \leq m_i \leq 12$. Next to that, the day $d_m, 1 \leq d_m \leq 4380$ of the period T_m is also needed as measure for the discretization of this period in days, later in this thesis. The information of d_m and m_i can easily be derived from the already determined information of the year i and the day d_i of each emergency call. So the temporal information in the adapted data set consists of i, d_i, m_i, d_m and (the in this thesis not used) s_i.

The spatial data of the occurred emergency calls is represented by their longitude and latitude coordinates $(x_{\text{lon}}, x_{\text{lat}})$ and by their X and Y coordinates (x_X, x_Y). Longitude and latitude represent the data in the spatial reference system of EPSG:4326, which is commonly known by the name *WGS84*. X and Y represent the data in a special spatial reference system of EPSG:28992, which is a spatial reference system only used in the Netherlands and is known by the name *RD New*. Note that the longitude is related to X and the latitude to Y. Spatial information is also given in the form of the ID number of the neighbourhood of the emergency call, but this measure is of no use for this thesis and this ID number can always be derived from the spatial coordinates.

To choose the spatial reference system to work with, RD New and WGS84 will first be compared with each other. The main difference is that RD New is a projected coordinate system and WGS84 is a geographic coordinate system. This means that RD New models the earth as a plane, so in $\mathbb{R}^2$, while WGS84 respects the curvature of the earth and models it in $\mathbb{R}^3$. Coordinates expressed in RD New are therefore expressed in two-dimensional Cartesian coordinates $\mathbf{x} = (x_X, x_Y)$, where x_X and x_Y are in meters[14]. Coordinates in WGS84 are expressed in polar coordinates $\mathbf{x} = (r_{\text{earth}}, x_{\text{lon}}, x_{\text{lat}})$, where r_{earth} is the radius of the earth with respect to its center, x_{lon} is the azimuthal angle and x_{lat} is the polar angle. But since the radius of the earth is often assumed constant, the coordinates are commonly, and also in this thesis, expressed as $\mathbf{x} = (x_{\text{lon}}, x_{\text{lat}})$.

[14] These distances are with respect to an origin 120 kilometers south of Paris. With this origin, each location in the Netherlands can be expressed with a positive value for x_X and x_Y.

The reason why no measure for the emergency calls is taken into account that describe their height, for example the radius of the earth for the WGS84 coordinates, is that Twente is very flat and therefore may be regarded as a plain (and of course since the data set does not possess descriptions of such a measure for the emergency calls). If an analogous model would be made for a very mountainous region, like Austria, the coordinate describing the height of each emergency call would be of interest and should be taken into account, since the arrival times of the firemen then also depend on whether they have to drive up a mountain or not.

Although both WGS84 and RD New describe the locations of emergency calls in the plain, RD New is the useful spatial reference system for this thesis, since it is a projected coordinate system rather than a geographic coordinate system like WGS84. Therefore, RD New gives the precise length in meters the firemen should drive, in contrast to WGS84 which gives the angles in the longitudinal and latitudinal directions they cover. For this reason, and because of the better familiarity of the fire departments with RD New in stead of with WGS84, the spatial reference system of RD New is chosen for the model in this thesis[15].

Nonetheless, both the RD New coordinates $\mathbf{x} = (x_X, x_Y)$ as the WGS84 coordinates $\mathbf{x} = (x_{\text{lon}}, x_{\text{lat}})$ will be added to the data set. The WGS84 coordinates are added for the same reason as why s_i was added, namely to make future extensions to the model possible. Some emergency call data only have their spatial information in RD New or in WGS84, though. For these data, completion is needed. Conversion from RD New to WGS84 and vice versa is made possible by a publicly available *Microsoft Excel* file, which is built by Rinus Luijmes and based on an algorithm from dr. ir. E.J.O. Schrama.

The classification data is described in the level 1a, level 2a and level 3a classes of the general classification system of the Netherlands, the level 1, level 2 and level 3 classes of another, less often used informal classification system, the priority for treating the emergency call and the description of the fire if the emergency call of interest is a fire. Because the classification system involving the level 1a, level 2a and level 3a classes is (slightly) more general than the classification system involving the level 1, level 2 and level 3 classes, the former system is preferred to classify the emergency calls by the fire departments in Twente.

In this thesis, the modelling will be based on the level 1a classes $c_{1\text{a}}$ of the emergency calls, on request of the policy and strategics team of the head fire department in Twente. As mentioned in section 1, level 1a involves five descriptions for the firemen demanding emergency calls, represented by the set $C_{1\text{a}} = \{\text{fire}, \text{service}, \text{accident}, \text{alert}, \text{environmental}\}$. So the classification information in the adapted data set consists of $c_{1\text{a}} \in C_{1\text{a}}$.

Also the level 2a and level 3a classifications $c_{2\text{a}}$, $c_{3\text{a}}$ are added to this data set, because they provide useful information specifying the information of $c_{1\text{a}}$, for example specifying the objects of interest as "building" or "car" for emergency calls of the class "fire". But $c_{2\text{a}}$ and $c_{3\text{a}}$ will because of simplification not be taken into account in the modelling in this thesis. They are again only added to the data set for possible future extensions to the model made in this thesis.

The reason why the classification of events is done by description (in level 1a, 2a and 3a) rather than by priority is that classification by description is more useful to express the causes in covari-

[15] As a consequence, all the spatial point patterns and the spatial information in this thesis are in RD New, unless stated otherwise.

ates and thus to help optimizing the prevention behaviour, as mentioned in section 1. Next to that, the data representing the priority of emergency calls is sometimes missing. Also no attention will be paid to the classification of the fire. Although this is actually a subclass of the level 1a class "fire", the subclassifications in level 2a and level 3a are more descriptive. Nonetheless, the priority and the classification of the fire are also added to the data set for possible future extensions to the model.

A number of emergency calls, though, does not possess temporal, spatial or classification information. These emergency calls are deleted from the data set. It is also possible that emergency calls have data entry errors or that the emergency calls occurred outside of Twente[16]. All such emergency calls are deleted from the data, because the modelling in this thesis will be based on reliable data in the region of interest Twente. This filtering is done by *QGIS*.

A remark has to be made, since some of the emergency calls just beyond the borders of Twente are also taken into account in the modelling. This is caused by a discretization of Twente in squares of 500 meter, since the squares of this discretization form the actual region of interest in this thesis. The reason for this will be explained in more detail in section 4. For now, it is important to know that this region is the actual region of interest, unless stated otherwise. Nevertheless, the region of interest will still be denoted by Twente for the moment, to keep the discussions about the analyses intuitive.

In period T_m, $66.707 \cdot 10^3$ firemen demanding emergency calls are received by the fire departments in Twente and 447 of these emergency calls contained such erroneous data. So the fraction deleted data is approximately 0.67 percent, which is already very small. Nonetheless, these data errors will be examined more in detail, to discover whether erroneous data have a certain trend of occurrence and therefore to discover whether they will influence the predictions of the model at specific times in the future or at specific places. Therefore, the erroneous data of the period T_m will be analysed. Next to the erroneous data, the discarded data of 29 February for the involved leap years in T_m will also be analysed on possible trends. The analyses for these discarded data will be done later in this section, after the analysis methods for spatial and temporal analysis are explained for the (correct and important) emergency call data.

3.2 Spatial exploratory data analysis

Now the useful data is filtered, it can be analysed. For modelling the spatial part of the spatio-temporal point process, this spatial point process should be chosen in such a way that it represents the spatial point pattern of interest well. But what is the spatial point pattern of interest? For the spatial point process representing a specific class c_{1a}, the spatial point pattern of interest is the spatial point pattern representing all the emergency calls of this class in Twente for the period on which the model is based, so period T_m. Let S_m denote this spatial point pattern for the class c_{1a} = fire, then figure 2 shows the spatial point pattern of interest S_m[17].

Analysing the distribution of S_m gives precisely the spatial information for the spatio-temporal point process of interest. A first and quantitative examination of figure 2 indicates already that the distribution of the emergency calls seems aggregated. This will now be formally analysed, following the discussion of section 2.

[16] For a very serious emergency call, firemen of different districts could be summoned. This causes the firemen of Twente to go sometimes to an emergency call outside of Twente.

[17] The spatial point pattern plots are made in *QGIS* and the data of the borders of the municipalities in Twente is provided by the following source: BRK bestuurlijke grenzen and CBS buurten.

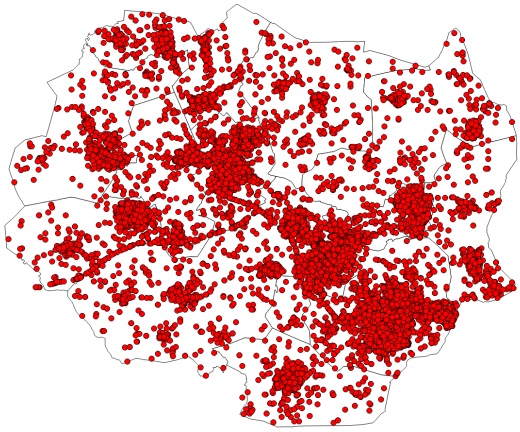

Figure 2: Spatial point pattern for the class c_{1a} = fire in the region Twente and in period T_m.

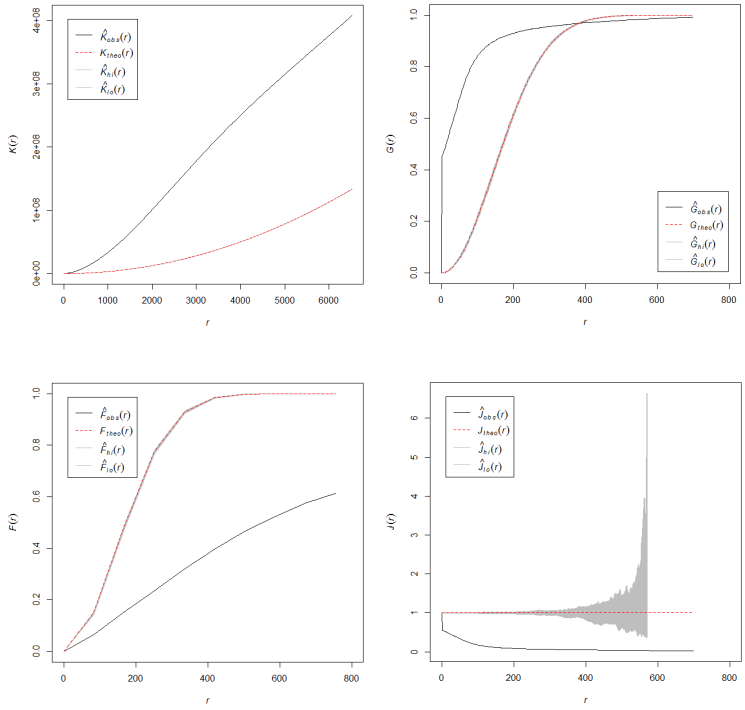

Figure 3: Distance analyses with the estimated functions $\hat{K}(r)$ (figure a, left above), $\hat{G}(r)$ (figure b, right above), $\hat{F}(r)$ (figure c, left below) and $\hat{J}(r)$ (figure d, right below) applied on the data of S_m, plotted against the corresponding theoretical functions and critical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

To start with the formal analysis, hypothesis test (5) will be executed, to analyse whether the distribution of the events in S_m is CSR or not. Hypothesis test (5) can be examined by the distance analysis functions $K(r)$, $G(r)$, $F(r)$ and $J(r)$ described in section 2. The significance level α is set to 0.05. The plots of the empirical distribution functions $\hat{K}(r)$, $\hat{G}(r)$, $\hat{F}(r)$ and $\hat{J}(r)$ of S_m against the corresponding theoretical functions and the critical envelopes are shown in figures 3a, 3b, 3c and 3d, respectively.

Aggregation can quantitatively be seen from these plots (and so CSR can be quantitatively rejected), but how should these plots be examined precisely? Executing hypothesis test (5) formally would mean that a fixed value r_0 should be chosen for r before executing the hypothesis test and that for this value it should be determined whether the empirical distribution function lies inside or outside the critical envelopes. If it lies inside the critical envelopes, H_0 is accepted and otherwise it is rejected. But nothing is till now said about how to choose the value for r_0.

To understand how to choose r_0, it is first important to understand how values for r should be interpreted. Remember that r represents the distance measure for events. For the fixed value r_i for r, it examines how many events there are in the neighbourhood with this distance r_i as radius for the $K(r)$ function and it examines how many neighbouring events have a distance of at most this distance r_i between them for the $G(r)$, $F(r)$ and $J(r)$ function. One can thus see that r_i determines an upper bound for which the interaction between points is examined. It is for this reason that r_i will be called the *interaction distance*.

Which value does this interaction distance r_i have? In theory, all events in the region of interest A may be influences by an earlier occurred event and so the interaction distance would be the largest distance between two events in A. But in practice, the interaction distance r_i is chosen to be smaller, since the events of the specific problem do not have such a large interaction distance (or at least they are assumed to not having such a large value for r_i). To illustrate this idea, an emergency call of the class $c_{1\text{a}}$ = fire may cause another emergency call a few 100 meters further, but probably not many kilometers further.

But how to conclude the value r_i for the interaction distance mathematically? In general, there is no strict mathematical method to conclude this value. But this thesis proposes a method which may give a small indication for it (at least a better indication than choosing the value on the basis of intuition). For this method, the probability distribution is needed for the random variable D, which represents the distance to a newly occurred event with respect to a conditional (earlier occurred) event. The probability distribution for D is not known, but it is assumed to be normal distributed with mean $\mu_D = 0$ and variance σ_D^2. As can be seen, the future event then has the highest probability of occurring on the same location as the conditional event and it has a low probability of occurring far away from is, but still that probability is nonzero. So choosing a normal distribution for D covers all the constraints[18].

How should the value for standard deviation σ_D be found? One may remember that the values in the range $[-\sigma_D, \sigma_D]$ cover 68% of the sample space for a normal distribution. So σ_D should be chosen in such a way that 68% of the emergency calls caused by a conditional emergency call lie within a radius σ_D around this conditional emergency call. Of course, expert judgement and

[18]Note that the distance $D = d$ may be negative in for the normal distribution. This then implies a distance $D = |d|$ for the future event. Although a one-tailed normal distribution would be mathematically cleaner, this two-tailed distribution is easier to work with.

intuition again have to help to conclude such a value for the average emergency call[19], since it is not known which emergency calls are caused by which emergency calls. But there is a method to argue a chosen value for σ_D. A Gaussian mixture model can be made for the spatial point pattern of interest, representing a nonparametric intensity function. Such a model can be made, since values for μ_D and σ_D are now determined for each emergency call. This Gaussian mixture model then represents a nonparametric intensity function intensity, which should resemble the spatial point pattern of interest. If this is (globally) the case, the value of σ_D is chosen well.

Now the distribution for D is concluded, the interaction distance r_i can be concluded from it. The interaction distance is now defined as the range $[-r_i, r_i]$ which covers a faction $1 - \beta$ of all the events caused by the conditional event. So if β is chosen to equal 0.05, $[-r_i, r_i]$ should cover 95% of the sample space for the distribution for D of each emergency call. But according to the properties of σ_D, the range $[-2\sigma_D, 2\sigma_D]$ also covers 95% of these sample spaces. For this thesis, $\beta = 0.05$ and so $r_i = 2\sigma_D$. In analogous ways, the interaction distances r_i can be concluded from the values for σ_D for different values for β.

One may question why this method is better than determining the value for r_i based on expert judgement and intuition directly, since r_i is now derived from σ_D which was determined based on expert judgement and intuition. The reason is that the value for σ_D could be (roughly) validated by comparing the corresponding Gaussian mixture model to the spatial point pattern of interest. In this way, the value for r_i is expected to be more accurate.

For the emergency calls of the class c_{1a} = fire, it was mentioned that the probability of a fire causing another fire many kilometers away from it is expected to be small, but the probability of a fire causing another fire a few 100 meters further is expected to be high. Reasoning in this way, σ_D is chosen to equal 1000 meter. For this value, a Gaussian mixture model is made, which represents a nonparametric intensity function. This intensity function is shown in figure 4 and it can be seen that it resembles S_m nicely, so $\sigma_D = 1000$ seems very reasonable. In a similar way it can be shown that $\sigma_D = 1500$ implies an intensity function where only the hot spots at the (ostensibly) less attractive clustering points are quite faded away and that $\sigma_D = 500$ implies an intensity function where the normal distributions representing D for each occurred event get already quite peaked around these events. So for c_{1a} = fire, $\sigma_D = 1000$ is chosen as standard deviation for D and therefore $r_i = 2000$, since $\beta = 0.05$.

There is one point of attention for the concept of interaction distances, since involving an interaction distance r_i does immediately imply that S_m has stochastic interaction as cause of the distribution. The interaction distance r_i is purely the radius for the neighbourhood in which each event should be examined. Distance analyses may then accept or reject stochastic interaction as cause of the distribution. If stochastic interaction is accepted as cause, it would indeed be sensible to choose r_i also as interaction distance in the intensity function.

Now the interaction distance $r_i = 2000$ is determined, the plots in distance analysis methods can be interpreted. The idea is that not only the value $r_0 = r_i$ is tested for r, but the whole range $[0, r_i]$ will be tested[20]. In the plots for the analysis methods involving $G(r)$, $F(r)$ and $J(r)$, the whole range $[0, 2000]$ is not given, though, since `R` does not calculate the functions for values

[19] Not every (conditional) emergency call may cause the same number of future emergency calls, so the concluded number is an average.

[20] Of course, not infinitely many values can be tested, but by this it is meant that the plots are analysed in this range.

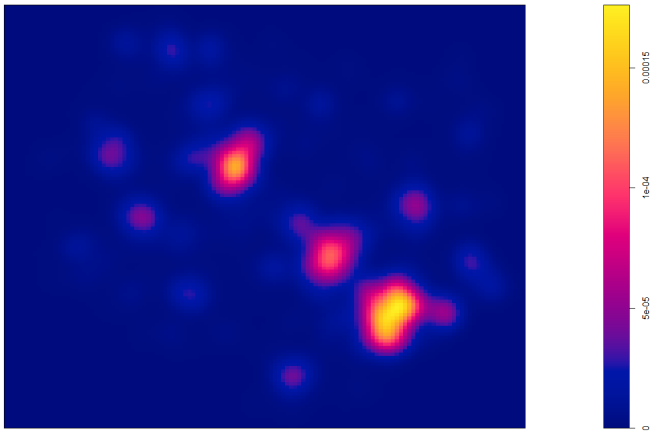

Figure 4: Gaussian mixture model with $\sigma_D = 1000$ for the events of S_m. The plot is provided by the package `spatstat` in `R`.

higher than 700, approximately. This is probably caused by the large amount of emergency call data, although Ripley's K-function is able to deal with this amount. For the range in which the $G(r)$, $F(r)$ and $J(r)$ functions can be examined, they will be interpreted for the hypothesis test of interest. But for the remainder of the values, so approximately $[700, 2000]$, the conclusions for the hypothesis test only rely on the distance analysis method involving Ripley's K-function.

Examining then the plots of figure 3 in this way formally indicates that a CSR distribution is rejected for all plots and that the $K(r)$, $F(r)$ and $J(r)$ functions clearly indicate aggregation. The function $G(r)$ also indicates aggregation for $r < 350$, but after that it indicates regularity. This is exactly the reason why several distance analysis methods were used for the hypothesis test, since each of them focuses on a different aspect of S_m. But basing on all plots together, the distribution seems clearly aggregated. Note that this aggregation is this clear, because of the large amount of data. This also causes the critical envelopes to barely differ from the theoretical function, since a large amount of data reduces the variance.

It is desirable, though, to specify this aggregation even further, that means to check whether inhomogeneity is present. But before examining inhomogeneity, it is important to remember the idea of examining inhomogeneity discussed in section 2. The observant reader may remember from section 2 that testing inhomogeneity for the distribution of S is based on n multiple independent and identically distributed realizations $S_1, S_2, \ldots, S_n$ of S. But since n such realizations were not possible for the same time period of interest T and the same region of interest A, the realizations were based on the spatial point patterns $S_{U_1}, S_{U_2}, \ldots, S_{U_n}$ corresponding to the different subperiods $U_i \subset T, 1 \leq i \leq n$, where $U_1, U_2, \ldots, U_n$ form a partition for T in subperiods of the same length. This method was made possible by the assumption of time invariance. In this way, testing inhomogeneity became actually testing significant difference between spatial point patterns of different subperiods U_i.

In this thesis $n = 12$, since T_m will be partitioned in the twelve years $U_{m,i} \subset T_m, 1 \leq i \leq 12$ involved, where $U_{m,i}$ represents the year $2003 + i$. These subperiods $U_{m,i}$ are therefore mutually exclusive and collectively exhaustive. Further, $S_{m,i}, 1 \leq i \leq 12$, represents the spatial point

pattern for all the emergency calls of the class c_{1a} = fire occurred in period $U_{m,i}$. In this way $S_{m,1}, S_{m,2}, \ldots, S_{m,12}$ represent the twelve realisations of S_m.

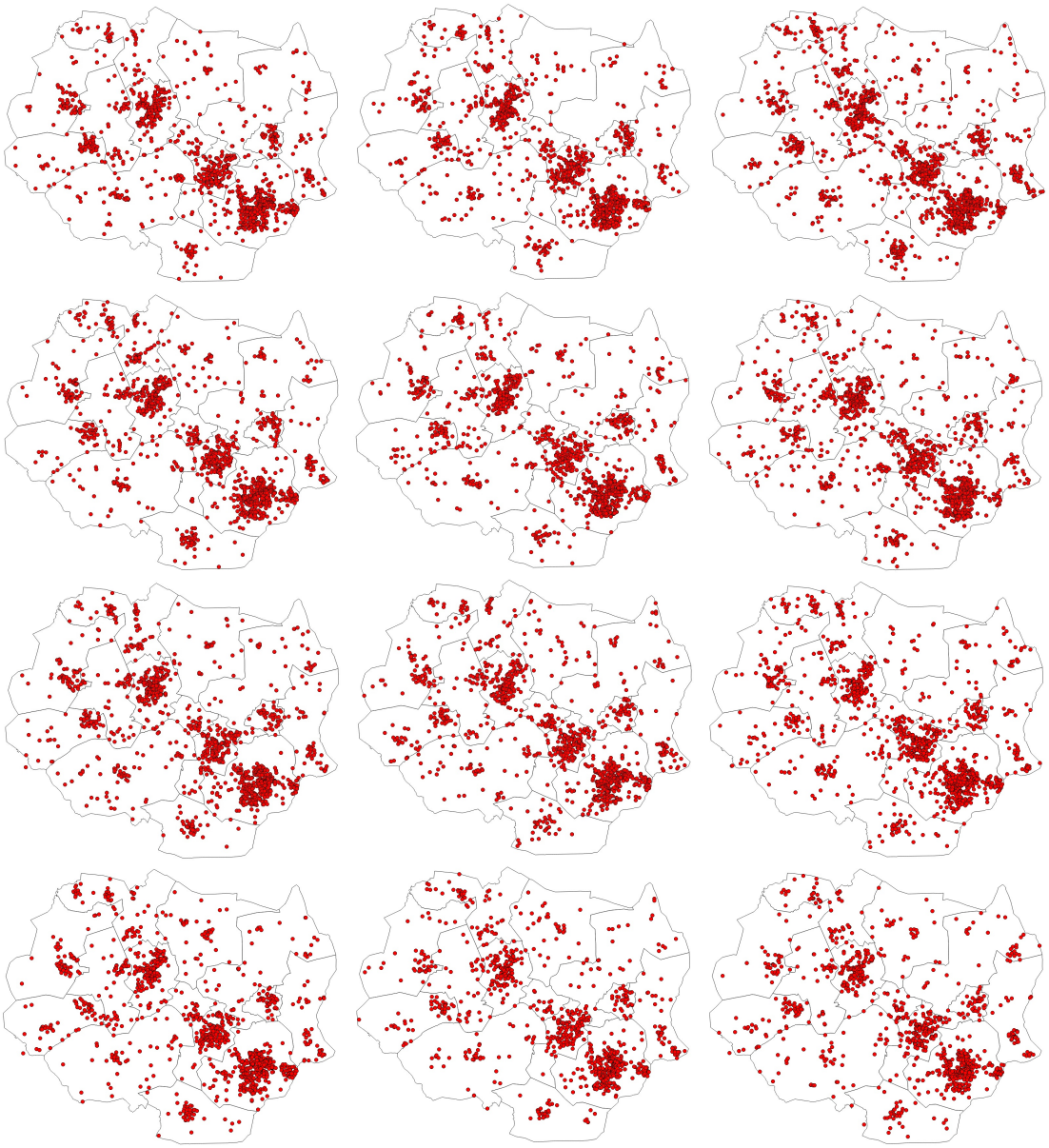

Figure 5: Emergency calls of class c_{1a} = fire for 2004 (figure a, first row left), 2005 (figure b, first row centre), 2006 (figure c, first row right), 2007 (figure d, second row left), 2008 (figure e, second row centre), 2009 (figure f, second row right), 2010 (figure g, third row left), 2011 (figure h, third row centre), 2012 (figure i, third row right), 2013 (figure j, fourth row left), 2014 (figure k, fourth row centre) and 2015 (figure l, fourth row right).

The spatial point patterns $S_{m,1}, S_{m,2}, \ldots, S_{m,12}$ are shown in figure 5. By inspecting each of these spatial point patterns, one can see that the spatial point patterns of each year seem to be aggregated distributed, what also can be tested in the same way as the testing aggregation for S_m. This aggregation even happens to be around the same points in Twente for the spatial point patterns of each year. As a consequence, the distributions for $S_{m,i}, 1 \leq i \leq n$, do not seem to change as years pass by. The time invariance assumption required seems also reasonable for $S_{m,1}, S_{m,2}, \ldots, S_{m,12}$, apart from the fact that the number of emergency calls of class $c_{1\text{a}} = \text{fire}$ decreases when the years pass by. So by this quantitative analysis of $S_{m,1}, S_{m,2}, \ldots, S_{m,12}$, an inhomogeneous distribution seems present.

But inhomogeneity will also be tested in a formal way by the following hypothesis test:

$$\begin{aligned} H_0&\text{: } S \text{ is inhomogeneous distributed over } A. \\ H_1&\text{: } S \text{ is not inhomogeneous distributed over } A. \end{aligned} \tag{25}$$

The earlier discussed distance analysis functions $K(r)$, $G(r)$, $F(r)$ and $J(r)$ cannot be used anymore for testing hypothesis test (25), since these functions are actually only built for testing the hypotheses of (5) against each other. In other words, they are only suited for determining whether a homogeneous Poisson process could be fitted to a spatial point pattern of interest, which requires a CSR distribution to be accepted for it. Therefore the (ordinary) $K(r)$, $G(r)$, $F(r)$ and $J(r)$ are called *homogeneous* or *stationary*.

The homogeneous distance analysis functions can be extended to execute hypothesis test (25) and thus to determine if an inhomogeneous Poisson process could be fitted to the spatial point pattern of interest. Such distance analysis functions are called *inhomogeneous*. The inhomogeneous $K(r)$, $G(r)$, $F(r)$ and $J(r)$ functions are commonly denoted by $K_{\text{inhom}}(r)$, $G_{\text{inhom}}(r)$, $F_{\text{inhom}}(r)$ and $J_{\text{inhom}}(r)$, respectively. In this thesis, only the inhomogeneous K-function is used for executing hypothesis test (25), since the inhomogeneous versions of $G(r)$, $F(r)$ and $J(r)$ gave again no (clear) output and since the earlier executed hypothesis test (25) also mainly relied on Ripley's K-function. For a general discussion of the derivation and the working of the inhomogeneous K-function, the reader is referred to Gabriel and Diggle (2009).

The inhomogeneous K-function is implemented as `Kinhom` in the `spatstat` package in `R`. The plot of the empirical distribution function $\hat{K}_{\text{inhom}}(r)$ for S_m against the corresponding theoretical function $K_{\text{inhom}}(r)$ and the critical envelopes $U(r)$ and $L(r)$ is shown in figure 6. This plot should be interpreted as follows. If the empirical distribution function lies between the critical envelopes, so $L(r) \leq \hat{K}_{\text{inhom}}(r) \leq U(r)$, H_0 of hypothesis test (25) is accepted and an inhomogeneous Poisson process could be (directly) fitted to the data of S_m. Otherwise H_1 is accepted and an inhomogeneous Poisson process seems not suited for the data. If this is the case, deviations $\hat{K}_{\text{inhom}}(r) > U(r)$ again indicate aggregation and $\hat{K}_{\text{inhom}}(r) < L(r)$ again regularity. Further, $n = 39$ implies again $\alpha = 0.05$.

Knowing this, figure 6 can be examined, again for the range [0,2000] as interval of interest. But it can be seen that inhomogeneity is clearly rejected. Still, this does not mean that an inhomogeneous Poisson process is a wrong model for the data, it only means that it is a wrong model for the data implemented with only the locations of the earlier occurred events implemented as covariates. One may remember that events are caused by influences of covariates or influences of earlier occurred events. Although an inhomogeneous Poisson process cannot model the stochastic interaction (as a consequence of theorem 2.1), it can model several more covariates $C_i, 1 \leq i \leq p$, and this will thus be tried first.

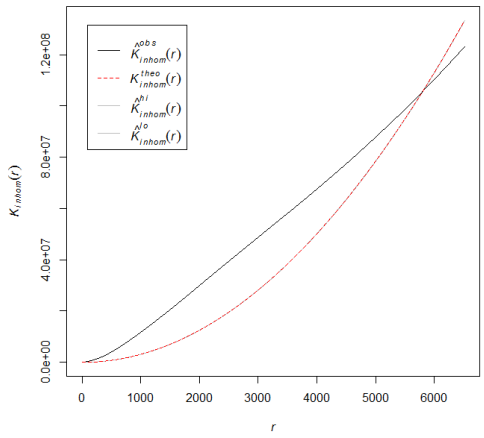

Figure 6: Distance analysis for testing inhomogeneity with the estimated function $\hat{K}_{\text{inhom}}(r)$ applied on the data of S_m, plotted against the corresponding theoretical functions and critical envelopes. The plot is provided by the package `spatstat` in `R`.

If p important covariates are then modelled, (simulated spatial point patterns of) the inhomogeneous Poisson process with these covariates $C_i, 1 \leq i \leq p$, could be validated by the inhomogeneous K-function by examining the empirical density function $\hat{K}_{\text{inhom}}(r)$ for this model to the corresponding theoretical distribution function $K_{\text{inhom}}(r)$ and the critical envelopes. If $\hat{K}_{\text{inhom}}(r)$ lies between the critical envelopes, the inhomogeneous Poisson process model with the covariates $C_i, 1 \leq i \leq p$, is suited for the spatial point pattern of interest. If $\hat{K}_{\text{inhom}}(r)$ lies outside of the envelope, interaction should (also) be modelled, by for example the Strauss process.

So although inhomogeneity is rejected for now for the class $c_{1\text{a}} =$ fire, still an inhomogeneous Poisson process will first be fitted to S_m. But how about the other classes? For analysing them, the interaction distance r_i for each of these classes has again to be determined. But there is no clear clue for these interaction distances, so they are set on the same interaction distance as the one earlier used for the class $c_{1\text{a}} =$ fire, so $r_i = 2000$. This interaction distance can again be checked by comparing the nonparametric intensity function with the spatial point pattern of interest for each class. Also this analysis indicates that $r_i = 2000$ seems very reasonable.

The same spatial exploratory data analysis will then be executed again for the classes $c_{1\text{a}} =$ service, $c_{1\text{a}} =$ accident, $c_{1\text{a}} =$ alert and $c_{1\text{a}} =$ environmental. For each class a CSR distribution is rejected and an aggregated one is indicated for hypothesis test (5). Again, mostly is relied on Ripley's K-function, since the $G(r)$, $F(r)$ and $J(r)$ functions are also in these cases not defined for r (approximately) greater than 700. After that, inhomogeneity is analysed for these classes. Since the $G_{\text{inhom}}(r)$, $F_{\text{inhom}}(r)$ and $J_{\text{inhom}}(r)$ functions do not work for analysing inhomogeneity, this analysis again relies fully on the inhomogeneous K-function. For each of these classes, the empirical distribution functions lie above the upper critical envelope for the whole range $[0, 2000]$ and so inhomogeneity is rejected. Despite this conclusion, an inhomogeneous Poisson process will also be fitted to these classes, for the same reasons as for the class $c_{1\text{a}} =$ fire.

Note that for the data of each level 1a class, inhomogeneity is rejected at least for the range [0,5500]. So even if $r_i = 2000$ may not be the correct interaction distance, rejecting inhomogeneity still seems reasonable, since deviations till $r_i = 5500$ may be possible. And because $r_i = 5500$ is a very large interaction distance which seems not plausible, the conclusion drawn seems strong.

So for each level 1a class $c_{1a} \in C_{1a}$, inhomogeneity is clearly rejected, although an inhomogeneous Poisson process for modelling (the spatial parts of) the spatio-temporal point processes for all these classes is not rejected yet. This proposed model will be examined further in this thesis. Further, the main results of the executed spatial exploratory data analyses for testing CSR and inhomogeneity are shown in the appendices A, B, C and D for the classes c_{1a} = service, c_{1a} = accident, c_{1a} = alert and c_{1a} = environmental, respectively.

3.3 Temporal exploratory data analysis

For modelling the temporal part of the spatio-temporal point process, the evaluation of the spatial point processes in time should be examined. Again the region of interest is Twente and the time period of interest is T_m. The change in the distribution of the occurred emergency calls per year will be examined. This analysis will be done by a quadrat count analysis method, which will be called the *temporal test*.

But why is a quadrat count analysis method used, while the distance analysis methods are more accurate? This is done, since such a quantitative analysis for the temporal behaviour does the job and therefore simplifies the analysis. Since the spatial exploratory data analysis concluded an inhomogeneous Poisson process, a (spatio-temporal) inhomogeneous Poisson process will be proposed for the whole problem rather than a hybrid process with different kinds of models for the spatial and temporal part. The temporal test has therefore only to examine how the temporal part of the data behaves and so how it should be implemented. This test will be explained simultaneously with analysing the temporal behaviour of the occurrences of emergency calls. The significance level α is again set to 0.05.

Before the temporal exploratory data analysis will be explained and executed, the time series of interest R_m for the class c_{1a} = fire will first be described. This time series will be expressed in the number of occurred emergency calls of this class in each subperiod $U_i \subset U_m, 1 \leq i \leq n$, where these subperiods are defined in the same way as for the spatial exploratory data analysis, so U_i represents year $2003 + i$. This time series R_m is given in figure 7.

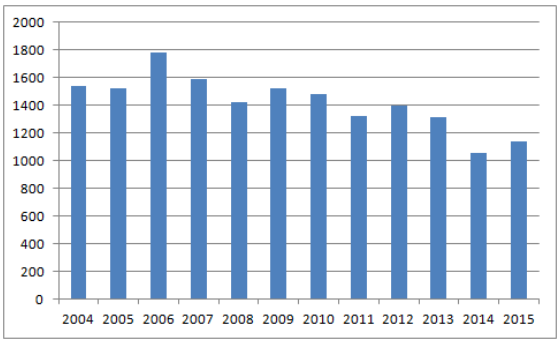

Figure 7: Time series for the amount of emergency calls of the class c_{1a} = fire per year $U_i \subset T_m$.

A slight decrease in the number of emergency calls of the class c_{1a} = fire per year can already be seen from figure 7. In terms of aggregation, the emergency calls seem more aggregated for the early years of time period T_m. To test whether this aggregation is strong enough to reject coincidence (so to reject a CSR distribution), the temporal test will be executed. This test then indicates the years with a significantly different number of occurred emergency calls and so the years in which the emergency calls are aggregated.

If such aggregation is concluded from the temporal test for some years, the inhomogeneous Poisson process seems suited for the emergency calls of interest. Although regularity and stochastic interaction between events cannot be concluded by the temporal test, this is also not needed, since an inhomogeneous Poisson process is the model aimed for in this analysis. And as one may remember, the inhomogeneous Poisson process is not suited for modelling regularity or stochastic interaction.

For the temporal test, the temporal information of the emergency calls of class c_{1a} = fire will thus be described in the year $y, 2004 \leq y \leq 2015$ of occurrence. But to indicate the distribution of each year y, each emergency call will also be described in the day $d, 1 \leq d \leq 365$, of occurrence in that year. To give an indication of expressing the distributions per year by the days, a part of these distributions is given in table 1. By the same reasoning, the distribution for a day d can be indicated by the occurred emergency calls on that day for the years $y, 2004 \leq y \leq 2015$.

Let the 365×12 matrix Ω_T count the number of emergency calls of interest for year y and day d, where the twelve columns $i, 1 \leq i \leq 12$, represent the twelve years $y(i) = i + 2003$ of interest and the 365 rows $j, 1 \leq j \leq 365$, the 365 days $d(j) = j$ of interest. The matrix Ω_T will as a consequence be analogous to table 1, only without the row and column of the table which give respectively the indices j and i for the days and years. Further let the set of indices of the years and days be given by $Y = \{1, 2, \ldots, 12\}$ and $D = \{1, 2, \ldots, 365\}$, respectively, where the elements $i \in Y$ thus represent the year $i + 2003$ and the elements $j \in D$ the day j.

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
1	69	42	37	48	47	94	38	39	37	28	70	29
2	11	4	9	7	10	11	6	5	1	8	6	6
3	7	4	6	2	5	3	5	7	5	7	2	5
4	4	5	8	3	2	6	3	2	4	5	2	2
5	2	4	2	5	3	9	1	2	2	3	3	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
364	9	12	13	16	14	4	7	12	6	7	7	5
365	83	64	134	103	103	58	81	76	130	105	43	86

Table 1: Distribution of the emergency call data of class c_{1a} = fire per year $y, 2004 \leq y \leq 2015$ and per day $d, 1 \leq d \leq 365$.

The temporal test then compares two time periods with each other: the time period $\tau_{k,1}$ in which the emergency calls of interest may have a significant tendency to occur and the reference period $\tau_{k,2}$ for which no such tendency is assumed. Let $k = 1$ if distributions of years are compared to each other and let $k = 2$ if distributions of days are compared to each other. So for $k = 1$, $\tau_{k,1} \subset Y$ and $\tau_{k,2} \subseteq Y \setminus \tau_{k,1}$ and for $k = 2$, $\tau_{k,1} \subset D$ and $\tau_{k,2} \subseteq D \setminus \tau_{k,1}$.

The distributions of the emergency calls of interest for the periods $\tau_{k,1}$ and $\tau_{k,2}$, $k = 1, 2$, can be filtered out of the matrix Ω_T. Let V_i be the 365-tuple[21] representing all elements of column i of matrix Ω_T and W_j be the 12-tuple representing all elements of row j of matrix Ω_T. Define

$$\omega_{1,l} = \bigcup_{i \in \tau_{1,l}} V_i, \qquad l = 1, 2 \tag{26}$$

$$\omega_{2,l} = \bigcup_{j \in \tau_{2,l}} W_j, \qquad l = 1, 2 \tag{27}$$

where the union operator $\cup$ is defined as transforming the elements of all involved tuples to one new tuple $\omega_{k,l}$ where $k, l \in \{1, 2\}$. This new tuple then represents the distribution of the emergency calls of interest in period $\tau_{k,l}$, so for $k = 1$ the distribution in the context of years and for $k = 2$ the distribution in the context of days. With these tuples $\omega_{k,l}$, the temporal (hypothesis) test can be formulated as:

$$\begin{aligned} &H_0\text{: The elements of } \omega_{k,1} \text{ have the same distribution as the elements of } \omega_{k,2}. \\ &H_1\text{: The elements of } \omega_{k,1} \text{ have a different distribution than the elements of } \omega_{k,2}. \end{aligned} \tag{28}$$

where $k = 1, 2$. Testing these hypotheses against each other can be done by the Student's t-test. Let $\bar{X}$ be the sample mean of the elements of $\omega_{k,1}$, s be the sample standard deviance of the elements of $\omega_{k,1}$, n be the number of elements of $\omega_{k,1}$ (so the amount of data in period $\tau_{k,1}$) and μ be the sample mean of the elements of $\omega_{k,2}$, which serves as reference mean. Then Student's t-statistic

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \tag{29}$$

gives a measure how much the distribution of $\omega_{k,1}$ differs from the distribution of $\omega_{k,2}$. Student's t-statistic has a t-distribution with $n - 1$ degrees of freedom. If the value of t lies outside of the $(1 - \alpha)$-confidence interval, H_0 of hypothesis test (28) is rejected, otherwise it is accepted.

One may remark that for executing Student's t-test, a normal distribution is assumed for the occurrences of the data in the tuples $\omega_{k,1}$ and $\omega_{k,2}$, $k = 1, 2$. Although this assumption is quite reasonable for ordinary days and years, since for them no great differences in the mean or the standard deviation of the normal distribution are expected, this assumption is made mainly for simplicity. There are indeed non-parametric tests, like the Mann-Whitney U test, which are also suited for executing hypothesis test (28), but these are a bit more cumbersome.

The temporal test will be done for testing the distribution of each year against that of all the other years, so $\tau_{1,1} = \{y\}, 2004 \leq y \leq 2015$ and $\tau_{1,2} = \{2004, 2005, \ldots, 2015\} \setminus \{y\}$. The results are shown in table 2. One can see already from this table that it possesses the behaviour of the time series R_m, since the values of the t-statistics also seem to decrease globally. Also the largest number of emergency calls in a year is 2006, since the t-statistic has the highest value for this year. This again agrees with R_m. In a similar way, 2014 can be indicated as the year with the smallest number of emergency calls.

Now table 2 will be formally analysed. Since $|\tau_{1,1}| = n = 365$ for each year y, each t-statistic has a t-distribution with 364 degrees of freedom. Therefore the corresponding 95%-confidence interval for each value of y becomes $[-1.975, 1.975]$. It can then be concluded that 2006 has a

[21] Tuples are chosen to represent the data rather than sets, since non-disjoint elements should be taken into account every time they occur.

y	2004	2005	2006	2007	2008	2009
t	1.149	1.259	2.690	1.494	0.009	0.900
y	2010	2011	2012	2013	2014	2015
t	0.612	-1.165	-0.230	-1.035	-4.627	-3.277

Table 2: The values of the Student's t-statistics for the temporal tests for class c_{1a} = fire, which compares each year with the other years.

significant higher amount of occurred emergency calls of class c_{1a} = fire and for 2014 and 2015, this amount is significantly lower. This amount for all the other years of time period T_m is not significantly different from their assumed means μ for $\alpha = 0.05$.

So there are indeed aggregations of events and there seems indeed a decreasing trend of occurring emergency calls of class c_{1a} = fire. Based on this, an inhomogeneous Poisson process will also be fitted for the temporal part of the problem, so for the class c_{1a} = fire, a spatio-temporal inhomogeneous Poisson process will be modelled. Although nothing is concluded about stochastic interaction between the emergency calls in time, it is assumed to be absent. If the spatio-temporal inhomogeneous Poisson process happens to have predictions, the model may be extended by also introducing interaction as cause of the distributions.

Again, the same exploratory data analysis can be executed to analyse the temporal behaviour of the emergency calls of the other level 1a classes. For each of these classes c_{1a} = service, c_{1a} = accident, c_{1a} = alert and c_{1a} = environmental, the time series of interest can also be examined by the temporal test. The years which appear to have a significantly different distribution are shown in table 3, where $t < \mu$ means that the value for the t-statistic for those years is significantly smaller than the assumed mean μ and $t > \mu$ means that this value is significantly larger than the mean μ. The complete results of the executed temporal exploratory data analyses are shown in the appendices A, B, C and D for the classes c_{1a} = service, c_{1a} = accident, c_{1a} = alert and c_{1a} = environmental, respectively.

	$t < \mu$	$t > \mu$
c_{1a} = service	2004 till 2006	2010, 2012, 2014
c_{1a} = accident	2014, 2015	2004
c_{1a} = alert	2004, 2013, 2014, 2015	2006 till 2011
c_{1a} = environmental	2004, 2008, 2009, 2011, 2012, 2014	

Table 3: The years which significantly differ for the temporal tests for the level 1a classes c_{1a} = service, c_{1a} = accident, c_{1a} = alert and c_{1a} = environmental. The years which do not significantly differ are just the remaining years in period T_m.

As a consequence, the conclusion made for the emergency calls of class c_{1a} = fire can be made for emergency calls of the other level 1a classes too. This because for these classes there are also years with a significant different amount of emergency calls and because for these classes independence between the occurrences of emergency calls in time may also be assumed based on the independence concluded for them in the spatial exploratory data analysis. So for emergency calls of each level 1a class $c_{1a} \in C_{1a}$, where $C_{1a} = \{\text{fire}, \text{service}, \text{accident}, \text{alert}, \text{environmental}\}$, the spatio-temporal point process can be modelled as a (spatio-temporal) inhomogeneous Poisson process.

3.4 Analysis of discarded emergency call data

As mentioned, the data set contained some erroneous data, which was deleted from the data set when this set was filtered and completed. Although only 0.67 percent of the data set was erroneous, it is still important to check whether there is a significant tendency for the erroneous data to occur at specific locations or at specific times. This can quite easily be verified by the same techniques as the spatial and temporal exploratory data analyses, as will soon become clear. The region of interest is again Twente and the period of interest is again T_m, as explained earlier.

Before analysing the erroneous data, these data first have to be inspected. To start with this, the erroneous data are classified in four kinds of errors:

1. Data with no, an incorrect or an invalid location of occurrence, called *spatial errors*.
2. Data with no or an incorrect time of occurrence, called *temporal errors*.
3. Data with no or an incorrect classification, called *classification errors*.
4. Data with both a spatial and a temporal error, called *spatio-temporal errors*.

So, in the context of the period of interest T_m and the region of interest Twente for the model, data with spatial errors or spatio-temporal errors have no (correct) spatial information, in contrast to temporal errors and classification errors. In the same way, data with temporal errors or spatio-temporal errors have no (correct) temporal information, in contrast to spatial errors and classification errors.

Remark that for all kinds of temporal errors, it is assumed that the corresponding emergency call has occurred in period T_t, so the period from 1 January 2004 till 7 December 2016. This is also plausible, since the offered data set of emergency calls consisted only of data of the period T_t. But the temporal information of these emergency calls should be estimated even more accurately, since the erroneous data analysis focuses on the period T_m instead of T_t. A quantitative estimation for this can be made by inspecting the ID numbers of the emergency calls containing temporal or spatio-temporal errors, since the ID number is higher for a later occurred emergency call.

Using such a quantitative estimation, the distribution per year of temporal errors and spatial and temporal errors can be determined quantitatively. Such an estimation seems even quite accurate for a discretization in years, since the ID numbers can be classified accurately per year. Applying this estimation on the erroneous data gives the distributions of every type of error per year of period T_t, shown in table 4. With these distributions, the data of period T_v (the period from 1 January 2016 till 7 December 2016) can easily be filtered by removing the column of year 2016.

So table 4 gives the global behaviour of the erroneous data per year. This behaviour will now be analysed in more detail. But how should such an analysis be done? Note that spatial errors can only be investigated in time, temporal errors can only be investigated in space and investigating spatio-temporal errors can neither be done in time nor space. As a consequence, two kinds of analysis will be done: analysis for a tuple ϵ_τ, which contains the correct temporal information of erroneous data without temporal errors and analysis for a tuple ϵ_σ, which contains the correct spatial information of erroneous data without spatial errors. The tuple ϵ_τ will be analysed in the period T_m. The tuple ϵ_σ will be analysed in Twente, because all the locations of the corresponding data happen to be in Twente.

Year(s)	2004	2005	2006	2007	2008	2009	2010
Spatial error	99	98	28	15	20	14	27
Temporal error	4	1	0	0	0	0	0
Classification error	0	0	0	0	0	1	8
Spatial and temporal error	1	4	0	0	0	0	0
Total	104	103	28	15	20	15	35
Year(s)	2011	2012	2013	2014	2015	2016	All
Spatial error	29	19	33	18	27	28	455
Temporal error	0	0	1	0	0	0	6
Classification error	0	0	0	0	0	2	11
Spatial and temporal error	0	0	0	0	0	0	5
Total	29	19	34	18	27	30	477

Table 4: The amount of deleted data for each type of error. "All" refers to 2004 till 2016.

The tuple ϵ_τ will not only consist of the spatial errors of the period T_m, but also of the classification errors of that period. In the same way, the tuple ϵ_σ will also consist of both the temporal errors and classification errors that occurred in Twente. By the definition of the tuples ϵ_τ and ϵ_σ, the classification errors can also be covered by these tuples. The classification errors are analysed in both their spatial and temporal aspects, since these kind of errors still have (the correct) spatial and temporal information about the occurred emergency call.

A further remark for the tuple ϵ_σ is that the temporal and classification errors involved are only from the period T_m, which can be filtered with help from table 4. Although the estimation technique described earlier for filtering the temporal errors is quantitative, there seems in general a small probability that many temporal errors have occurred in 2016, since there are only six temporal errors for T_t (of which five already occurred in 2004 and 2005 according to the quantitative estimation technique).

So the spatial, temporal and classification errors can be analysed by analysing the tuples ϵ_σ and ϵ_τ. But how about the spatio-temporal errors? The only analysis that can be done for these errors is simply counting the number of these errors. The number of occurred spatio-temporal errors is five in the period T_m, where the quantitative filtering technique based on ID numbers is used to classify these errors per year, as in table 4. A more detailed analysis for these errors is not possible, because they cannot be analysed in neither time nor space.

The temporal and classification errors involved in the tuple ϵ_σ manifest themselves as a spatial point pattern in Twente. Therefore, these errors can easily be analysed by the same spatial point pattern analysis techniques as used in the spatial exploratory data analysis. Hypothesis test (5) is the only test that should be executed in this case, since the errors should be CSR distributed if they have no tendency to occur in specific subregions of Twente. This can again be analysed by the distance analysis functions $K(r)$, $G(r)$, $F(r)$ and $J(r)$ for a range [0;2000] for r. Again, $\alpha = 0.05$.

The plots of the estimated empirical distribution functions $\hat{K}(r)$, $\hat{G}(r)$, $\hat{F}(r)$ and $\hat{J}(r)$ for the data of ϵ_σ against the corresponding theoretical functions and the critical envelopes are shown in figures 8a, 8b, 8c and 8d, respectively. It can be seen that accepting a CSR distribution would

be the best conclusion, since the functions stay between the critical envelopes for most values in the range $[0, 2000]$ for r. Although the $\hat{K}(r)$ function does not stay between the envelopes for higher values, this does not affect the conclusion of CSR, since the interaction distance is 2000 meters.

It can also be seen that the small amount of data involved causes tentative critical envelopes and therefore to interpret the results with much caution. But this only strengthens the assumption that there is less probability of occurrences of spatial errors in the future and if they then happen, there seems to be no tendency for a specific subregion of Twente by the acceptation of the CSR distribution.

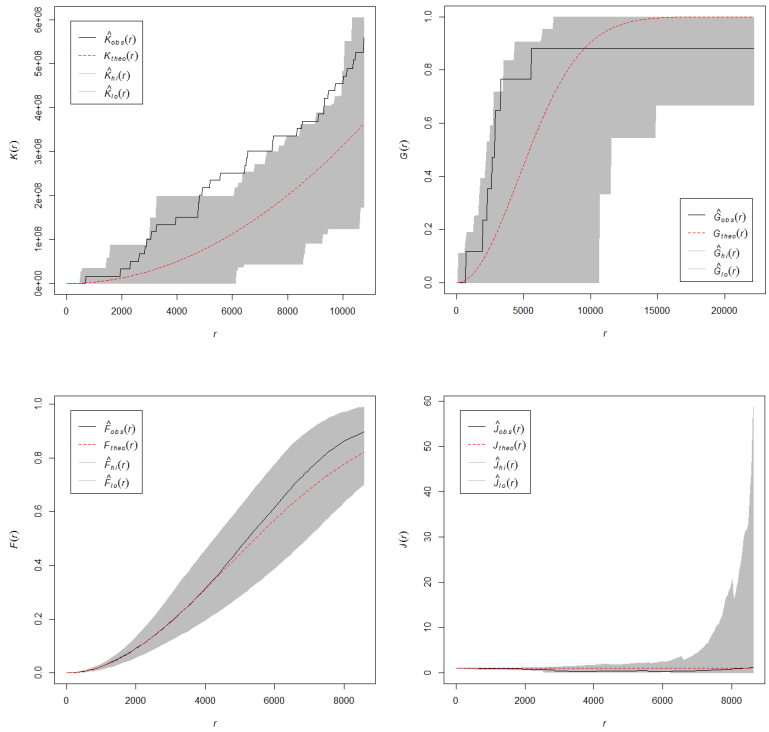

Figure 8: Distance analyses with the estimated functions $\hat{K}(r)$ (figure a, left above), $\hat{G}(r)$ (figure b, right above), $\hat{F}(r)$ (figure c, left below) and $\hat{J}(r)$ (figure d, right below) applied on the erroneous data of ϵ_σ, plotted against the corresponding theoretical functions and critical envelopes. The plots are provided by the package `spatstat` in `R`.

The spatial and classification errors involved in the tuple ϵ_τ manifest themselves as a time series in T_m, which indicates the amount of these errors per year $T_i \subset T_m, 1 \leq i \leq 12$. Analysing this time series could again be done by the temporal test. But this time, the distribution per year will not be expressed per day $d, 1 \leq d \leq 365$, but per month $m, 1 \leq m \leq 12$, since the tuple ϵ_τ is too small to give an accurate distribution of each year per day and vice versa. The distribution of these errors per month $m, 1 \leq m \leq 12$, and per year $y, 2004 \leq y \leq 2015$, is given in table 5.

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
1	7	12	0	4	0	1	1	3	3	2	1	1
2	12	8	3	1	1	2	1	1	0	3	1	2
3	10	13	2	1	0	2	3	3	2	3	2	0
4	13	13	0	2	4	1	2	4	0	3	0	1
5	19	6	4	0	0	2	3	4	0	1	1	2
6	11	9	2	2	3	1	2	1	1	3	2	3
7	6	12	3	2	2	2	2	3	1	2	4	5
8	1	3	4	1	2	0	6	3	3	4	2	1
9	5	6	2	2	2	1	4	2	2	4	1	6
10	5	8	3	0	3	1	1	1	2	4	1	4
11	6	5	5	0	1	1	5	1	3	2	0	0
12	4	3	0	0	2	1	5	3	2	2	3	2

Table 5: Distribution of the erroneous data of ϵ_τ per year $y, 2004 \leq y \leq 2015$, and per month $m, 1 \leq m \leq 12$.

The temporal test will be done for testing the distribution of each year against that of all the other years, so $\tau_{1,1} = \{y\}, 2004 \leq y \leq 2015$ and $\tau_{1,2} = \{2004, 2005, \ldots, 2015\} \setminus \{y\}$. The results are shown in table 6. Since $|\tau_{1,1}| = n = 12$ for each year y, each t-statistic has a t-distribution with 11 degrees of freedom, so the corresponding 95%-confidence interval for each value of y becomes $[-2.201, 2.201]$. It can then be concluded that the erroneous data occurred in 2004 and 2005, what already could be expected from the tables 4 and 5. It can also be concluded from table 6 that significantly less erroneous data occurred in the years 2007, 2008, 2009, 2012 and 2014. For the remainder of the years in period T_m, the null hypothesis of hypothesis test 28 is accepted.

y	2004	2005	2006	2007	2008	2009
t	4.011	5.265	-1.572	-5.527	-3.949	-10.808
y	2010	2011	2012	2013	2014	2015
t	-0.243	-1.983	-4.687	-1.087	-4.944	-1.537

Table 6: The values of the Student's t-statistics for the temporal tests for the erroneous data of ϵ_τ, which compares each year with the other years.

Next to the temporal tests for testing tendencies for erroneous data to occur in specific years, these tendencies will also be examined for specific months. This is made possible, since the distribution per month is also available. For these distributions per month, the rows of table 5 (and so Ω_T) will be examined instead of the columns. In this test, $\tau_{2,1} = \{m\}, 1 \leq m \leq 12$ and $\tau_{2,2} = \{1, 2, \ldots, 12\} \setminus \{m\}$. As a consequence, $|\tau_{2,1}| = n = 12$ for each year m and so again the 95%-confidence interval for each value of y becomes $[-2.201, 2.201]$, since each t-statistic has again a t-distribution with 11 degrees of freedom. The results of these temporal tests are shown in table 7 and it can be concluded that all the temporal tests accept the null hypothesis of hypothesis test (28). In other words, for no month a significant different amount of erroneous data has occurred for $\alpha = 0.05$.

So the temporal tests for the erroneous data of ϵ_τ conclude that there is a tendency for the number of occurring erroneous data to be significantly lower in the years 2007, 2008, 2009, 2012

m	1	2	3	4	5	6
t	-0.121	-0.119	0.371	0.454	0.341	0.358
m	7	8	9	10	11	12
t	0.806	-1.188	0.115	-0.464	-1.015	-1.980

Table 7: The values of the Student's t-statistics for the temporal tests for the erroneous data of ϵ_τ, which compares each month with the other months.

and 2014, which can be explained by an improving system for saving emergency call data. Still this system has its flaws, because the erroneous data in 2015 for example was again higher than in 2014.

The temporal tests also concluded that the number of occurring erroneous data is significantly higher in 2004 and 2005 than for the other years in period T_m. This could be explained by an bad working system for saving data of emergency calls in 2004 and 2005. Firemen in those days namely did not always require specific coordinates for the place of the emergency call, because they knew the neighbourhood well and were sufficiently informed with a quantitative description of the location. Nowadays, the specific coordinates are directly sent to the navigation systems of the fire trucks. Nonetheless, in the future there is also expected a tendency for erroneous data to be significantly lower, by improved systems like the navigation systems for example.

The temporal tests examining the distributions of each month did conclude that there was no such tendency present. In the same way, the spatial point pattern analysis concluded no tendency for erroneous data to occur in specific subregions in Twente. For both cases, it is therefore assumed that this tendency will also stay absent in the future. All the mentioned conclusions are again drawn for a significance level of $\alpha = 0.05$.

Although the erroneous data is now analysed on tendencies in their occurrences, the analysis of discarded emergency call data is not finished. One may remember that the emergency call data of the "leap days" 29 February 2004, 29 February 2008, 29 February 2012 and 29 February 2016 was also discarded, to transform the leap years 2004, 2008, 2012 and 2016 into regular years for the analyses and modelling. This discarded data will also be examined, although only for the emergency call data occurred in period T_m. Since only three days are involved in T_m, there is only a small amount of observations and the analyses for these data are expected to be very tentative. So the results should be interpreted with much caution.

First, the spatial analysis will be done for the data of the leap days. Although the data set is small, an aggregated distribution can clearly be concluded for $\alpha = 0.05$ from the plots of the estimated empirical distribution functions $\hat{K}(r)$, $\hat{G}(r)$, $\hat{F}(r)$ and $\hat{J}(r)$ for this data against the corresponding theoretical functions and the critical envelopes. Again, the interaction distance is assumed to be 2000 meters. A plot of the spatial point pattern representing these data agrees with the conclusion that the distribution is aggregated.

Now the temporal analysis will be done for the data of the leap years. On the leap days in 2004, 2008 and 2012, respectively 11, 10 and 11 emergency calls occurred. For these observations, a temporal test can be executed which compares them to the occurrences of emergency calls on the days in the regular years. So this temporal test compares the distributions of days against each other, where $\tau_{2,1}$ consists of the data of the leap days in period T_m and $\tau_{2,2}$ consists

of the data of the other days in period T_m. As a consequence, $|\tau_{2,1}| = n = 3$ and so again the 95%-confidence interval becomes $[-4.303, 4.303]$, since each t-statistic has a t-distribution with 2 degrees of freedom.

For this temporal test, the t-statistic has the value -13.362 and therefore the alternative hypothesis of hypothesis test (28) is accepted. In other words, there seem to occur significant different emergency calls on the leap days for $\alpha = 0.05$. Since the value of the t-statistic is negative, there occur significant less emergency calls. The value for this t-statistic is quite extreme, though, what is caused by the fact that $\tau_{2,1}$ consists of three elements with approximately the same values. In this way, the sample variance is low and the value for this t-statistic extreme. So exact conclusions cannot be made, but it indeed seems that leap days possess less emergency calls and so that they require a different model than the other days. But to build such a model, more data is required about the number of emergency calls on leap days.

4 Covariate analysis

Since for the emergency calls of all level 1a classes c_{1a} an inhomogeneous Poisson process is modelled, their distributions are assumed to be caused by trend rather than by stochastic interaction. In this section this trend will be examined. As one may remember, trend is actually the influence of covariates $C_i, 1 \leq i \leq n$. A set of covariates should therefore be chosen as the covariates of interest for the spatio-temporal point process of a specific class. The covariate analysis then gives the precise relation between the involved covariates and the occurrences of emergency calls.

The covariates of interest will first be selected and made amenable for the covariate analysis. Since a spatio-temporal point process is made for each class c_{1a}, both spatial covariates and temporal covariates will be involved in the covariate analysis. The spatial and temporal covariates will respectively be compared to the spatial and temporal information about the emergency calls. Spatio-temporal covariates are not involved, since separability was assumed for each model.

After the covariate data are made amenable for analysis, the covariate analysis can be executed. This analysis consists of two parts: the correlation analysis and the regression analysis. The correlation analysis examines the influences of the spatial and temporal covariates on the emergency calls in a global way. After that, the spatial and temporal covariates with the most influence on the emergency calls of interest will be examined further in the regression analysis on their precise relation with the emergency calls of interest. These relations will be implemented in modelling the spatio-temporal point process for each class in section 5.

4.1 Filtering and manipulation of the covariate data

To relate occurrences of emergency calls to covariates, covariate data have first to be obtained. There is no specific guideline to choose useful covariates, it is merely a matter of expert judgement. For this reason, aspects which the involved covariates should describe for this thesis are first chosen globally, in consultation with the policy and strategics team of the head fire department in Twente. Examples of these aspects involve the presence of buildings at a place or the weather of the day. According to these aspects, covariate data sets involving specific covariates for that aspect are selected. Many such covariate data (sets) for the Netherlands are available for public use. The data sets used for this thesis and their sources are:

1. The data of the locations of all the buildings, source: *BAG, het Kadaster.*
2. The data of the locations of national highways and freeways, source: *NWB, Rijkswaterstaat.*
3. The data of the locations of railways and railwaystations, source: *BRT, het Kadaster.*
4. The data of the number of residents, addresses and corresponding information per square of 500 meter, source: *Centraal Bureau voor de Statistiek.*
5. The data of the borders of the towns in Twente, source: *BRT, het Kadaster.*
6. The data of the locations of rivers and lakes, source: *Imergis, made available by J.W. van Aalst.*
7. The data of the locations of canals, source: *Rijkswaterstaat.*
8. The data of (many aspects of) the weather measured by the weather station "Twenthe" in the region Twente per day, source: *Koninklijk Nederlands Meteorologisch Instituut.*

These data sets are for the entire Netherlands and are quite recent, since the years of appearances of the versions vary between 2014 and 2016. One may also see that all these covariate data sets seem to have a certain relationship with the occurrences of emergency calls.

These covariate data sets thus involve many specific covariates for the aspect they represent. For example, the data of the locations of all the buildings is specified for many different types of buildings. Since not every single covariate of these sets is important for the analyses in this thesis, the covariates which are important for this thesis, which means the ones which will indeed be analysed, first have to be filtered from the covariate data sets. Again, this is done in consultation with the policy and strategics team of the head fire department in Twente. The precise covariates which will be analysed are given in table 8, where the spatial covariates are indicated with a subscript σ and the temporal covariates with a subscript τ. The number of the data set from which each specific covariate is extracted is indicated in the square brackets after the description of the specific covariate.

Some covariates need some further explanation. For $C_{\sigma,10}$, $C_{\sigma,11}$ and $C_{\sigma,12}$, the difference between a highway and a freeway has to be made clear for example. Highway are in this thesis assumed to be the roads in the Netherlands which have several lanes for each direction and where a velocity of more than 60 kilometers per hour is necessary for entering these roads. Freeways are in this thesis assumed to be the roads in the Netherlands which have only one lane for each direction and where a velocity of more than 50 kilometers per hour is necessary for entering these ways.

A remark has also to be made about the covariate information involving the number of buildings with a residential function (so $C_{\sigma,22}$), since it is available from both data sets 1 and 4. The reason why data set 4 is used as the source, is that data set 4 gives the information for this covariate per year in the period 2004 till 2014, in contradiction to data set 1, which has measured it only in 2014. Therefore, the information of data set 4 seems more accurate for this covariate.

Another remark has to be made about the covariates of table 8 for the covariates $C_{\tau,3}$ and $C_{\tau,6}$. If the daily precipitation amount was less than 0.05 millimeter or the sunshine duration is less than 0.05 hour, the value of the involved covariate has -1 as value in data set 9. For this thesis, this value is adapted to 0, since it is then assumed for this thesis that $C_{\tau,3} = 0$ if $C_{\tau,3} < 0.5$ and that $C_{\tau,6} = 0$ if $C_{\tau,6} < 0.5$. This can be done by *Microsoft Excel*.

Next to that, a remark should be made about the covariates $C_{\tau,7}$ and $C_{\tau,8}$. It may seem very curious why these covariates are involved. But the temporal tests executed in section 3 for the temporal exploratory data analysis per year are also executed per day. It can then be seen that seen over all level 1a classes of emergency calls, 31 December and 1 January are often rejected in the temporal test, since the number of emergency calls are significantly higher then. This can be explained, since these are the days around new year's eve and therefore a lot of problems involving fireworks occur. The policy and strategics team of the head fire department in Twente agree with this hypothesis.

The seasons of the year are further taken into account as a general measure for the weather by $C_{\tau,9}$, $C_{\tau,10}$, $C_{\tau,11}$ and $C_{\tau,12}$, since each season has its characteristic weather. The dates at which each season begins and ends are based on those for meteorological seasons. Next to that, time in general is also represented by the covariate $C_{\tau,13}$, which involves the day d_m of period T_m. This data was completed in section 3. Since the analysis in section 3 indicated for example for c_{1a} = fire a decreasing trend in time, a general measure for time is introduced as a covariate.

$C_{\sigma,1}$	The total number of buildings [1]
$C_{\sigma,2}$	The number of buildings with an assembly function [1]
$C_{\sigma,3}$	The number of buildings with a healthcare function [1]
$C_{\sigma,4}$	The number of buildings with an industrial or agricultural function [1]
$C_{\sigma,5}$	The number of buildings with an office function [1]
$C_{\sigma,6}$	The number of buildings with an hotel function [1]
$C_{\sigma,7}$	The number of buildings with an educational function [1]
$C_{\sigma,8}$	The number of buildings with a sports function [1]
$C_{\sigma,9}$	The number of buildings with a retail function [1]
$C_{\sigma,10}$	The length in meters of highways and freeways present [2]
$C_{\sigma,11}$	The length in meters of highways present [2]
$C_{\sigma,12}$	The length in meters of freeways present [2]
$C_{\sigma,13}$	The length in meters of railway present [3]
$C_{\sigma,14}$	The number of residents [4]
$C_{\sigma,15}$	The number of residents with an age in the range of 0 till 14 [4]
$C_{\sigma,16}$	The number of residents with an age in the range of 15 till 24 [4]
$C_{\sigma,17}$	The number of residents with an age in the range of 25 till 44 [4]
$C_{\sigma,18}$	The number of residents with an age in the range of 45 till 64 [4]
$C_{\sigma,19}$	The number of residents with an age of 65 or higher [4]
$C_{\sigma,20}$	The number of male residents [4]
$C_{\sigma,21}$	The number of female residents [4]
$C_{\sigma,22}$	The number of buildings with a residential function [4]
$C_{\sigma,23}$	The density of addresses in the neighbourhood [4]
$C_{\sigma,24}$	The urbanity of the neighbourhood [4]
$C_{\sigma,25}$	Boolean variable indicating the presence of a town [5]
$C_{\sigma,26}$	Boolean variable indicating the presence of a pond or canal [6,7]
$C_{\sigma,27}$	Boolean variable indicating the presence of a pond [6]
$C_{\sigma,28}$	Boolean variable indicating the presence of a canal [7]
$C_{\tau,1}$	Daily mean windspeed (in 0.1 meter per second) [8]
$C_{\tau,2}$	Daily mean temperature (in 0.1 degrees Celsius) [8]
$C_{\tau,3}$	Daily precipitation amount (in 0.1 millimeter) [8]
$C_{\tau,4}$	Daily mean sea level pressure (in 0.1 hectopascal) [8]
$C_{\tau,5}$	Daily mean relative atmospheric humidity (in percents) [8]
$C_{\tau,6}$	Sunshine duration calculated from global radiation (in 0.1 hour) [8]
$C_{\tau,7}$	Boolean variable indicating whether or not the day is 1 January
$C_{\tau,8}$	Boolean variable indicating whether or not the day is 31 December
$C_{\tau,9}$	Boolean variable indicating whether or not it is spring (1 March till 31 May)
$C_{\tau,10}$	Boolean variable indicating whether or not it is summer (1 June till 31 August)
$C_{\tau,11}$	Boolean variable indicating whether or not it is autumn (1 September till 31 November)
$C_{\tau,12}$	Boolean variable indicating whether or not it is winter (1 December till 28 February)
$C_{\tau,13}$	The day d_m of period T_m, $1 \leq d_m \leq 4380$

Table 8: The spatial covariates $C_{\sigma,k}, 1 \leq k \leq 28$ and the temporal covariates $C_{\tau,j}, 1 \leq l \leq 13$ involved in the covariate analysis. The number in the square brackets indicates the source covariate data set of that covariate.

A last remark will be made about covariate $C_{\sigma,24}$, since the urbanity of the neighbourhood will be expressed by factors, varying from 1 till 5. These factors should be interpreted as measures for the urbanity, where 1 represents the highest urbanity and 5 the lowest.

The reason that no more covariates are involved than the covariates of table 8 is that all the previous mentioned covariates seem to describe the main aspects of Twente and of the period T_m already. Although it may affect the quality of the model that many covariates are not taken into account, this is the cost for keeping the model (quite) tractable. Note therefore the importance of expert judgement in selecting the right covariates, since this can make the difference between a good and a bad model. If some covariate data happen to be very important later in the analysis but are not implemented in the model, the analysis and modelling can be done again involving these covariates too.

After the ostensibly important covariates are selected in this way, they have to be made amenable for analysis. First, the covariate data should be filtered as to represent only the data for Twente in the period T_m. Then the covariate data should be counted in a grid for space and time to prepare it for analysis. The same should be done for the emergency call data. The reason for this is that emergency calls and covariates can only be compared with each other when there is a similar measure for both of them. Therefore, a discretized grid for the region of interest and the time period of interest will first be defined. The region of interest is again Twente, which will be denoted by A. The time period of interest is again T_m, so the period from 1 January 2004 till 31 December 2015, since the influences of the covariates only have to be examined for the time period on which the model is based.

Since the CBS data is restricted to squares of 500 meter in which it is available and the remainder of the spatial data is described more accurately, it is straightforward to use a grid consisting of 6291 squares of 500 meter as discretization for the spatial region of interest A. In this way a grid is made where the least possible information is lost. To formally define this grid, let $P_\sigma = \{P_{\sigma,1}, P_{\sigma,2}, \ldots, P_{\sigma,6291}\}$ be the partition of region A, so with $P_{\sigma,1} \cup P_{\sigma,2} \cup \cdots \cup P_{\sigma,6291} = A$ and $P_{\sigma,1} \cap P_{\sigma,2} \cap \cdots \cap P_{\sigma,6291} = \emptyset$. All these subregions $P_{\sigma,i}, 1 \leq i \leq 6291$, represent thus squares of 500 meter, so $|P_{\sigma,1}| = |P_{\sigma,2}| = |P_{\sigma,6291}| = 2.5 \cdot 10^5$ squared meter. Then P_σ denotes the spatial grid for A consisting of squares of 500 meter.

For time, all the covariate data is known per day and the emergency call data (in the finest measure) per second, so arguing in the same way as for the spatial discretization, a grid consisting of 4380 days would be the most straightforward time discretization for the time period of interest T_m. Also defining this grid formally, let $P_\tau = \{P_{\tau,1}, P_{\tau,2}, \ldots, P_{\tau,4380}\}$ be the partition of the period T, so with $P_{\tau,1} \cup P_{\tau,2} \cup \cdots \cup P_{\tau,4380} = T$ and $P_{\tau,1} \cap P_{\tau,2} \cap \cdots \cap P_{\tau,4380} = \emptyset$. All these subregions $P_{\tau,j}, 1 \leq j \leq 4380$, represent days, so $|P_{\tau,1}| = |P_{\tau,2}| = |P_{\tau,6291}| = 1$ day. Then P_τ denotes the temporal grid for T_m consisting of time intervals of days.

It is important to identify each grid cell of P_σ and P_τ. For P_σ, each square will be identified by $i, 1 \leq i \leq 6291$, in the sequence going from the smallest value of X to the largest value of X for each value of Y, where X and Y are the coordinates of RD New. This will be done in the sequence going from the smallest value of Y to the largest one. For P_τ, each day will be identified by $j, 1 \leq j \leq 4380$, in a chronological way, so that $j = 1$ represents 1 January 2004, $j = 2$ represents 2 January 2004 and so on till $j = 4380$, which represents 31 December 2015. Remember that all years in period T_m are assumed regular, so 29 February for 2004, 2008 and 2012 are not taken into account.

So now the space-time region $A \times T_m$ is discretized in $P_\sigma \times P_\tau$, so in 6291×4380 grid cells. By the partition P_σ, the spatial part of the emergency call data and the spatial covariates can be compared. In this way the spatial part of the covariate analysis for the spatio-temporal point process of each class is made possible. Analogously by the partition P_τ, the temporal part of the emergency call data and the temporal covariates can be compared. In this way the temporal part of the covariate analysis for the spatio-temporal point processes of each class is made possible. Note again the simplifying property of separability, since it makes the spatial and temporal analysis able in the grids P_σ and P_τ, respectively. If separability was not assumed, both analyses should have been done simultaneously in the $P_\sigma \times P_\tau$ grid, what would have been much more complicated.

To make these covariate analysis possible, the covariates first have to be filtered for the space-time region $A \times T_m$. The temporal covariate data will be filtered for the time period T_m. This can be done by *Microsoft Excel*. Filtering the spatial covariate data will be done so that the data is available in the whole grid P_σ, rather than in the region A alone. The region covered by the grid P_σ, which will be denoted by $\tilde{A}$, differs slightly from the region A, since $\tilde{A}$ also takes into account small regions outside of Twente because of the discretization of Twente by P_σ in squares of 500 meter. The temporal discretization P_τ did not have this problem, since a discretization in days fitted exactly for the time period T_m.

The reason why the region $\tilde{A}$ is assumed the region of interest rather than just the region A is that the analysis for the borders of Twente would be improved if the region $\tilde{A}$ is used as region of interest. This because the squares of 500 meter of P_σ lying on the borders of A otherwise contain the information only for a fraction of these squares, which is the fraction lying within the borders of A. By taking into account all the spatial covariate data and also all the emergency call data for these border squares, the spatial part of the covariate analysis will become more accurate.

One may remark, though, that the emergency call data outside the region A are not involved in the emergency call data set and so it would make no difference whether A or $\tilde{A}$ is chosen as the region of interest. But the observant reader may remember that the original emergency call data set also involved data outside of Twente, since the firemen of Twente also offer their help in other regions sometimes. They only help in regions far away from Twente in case of very serious emergency calls and these emergency calls were of no interest for the models in this thesis. But in regions just across the border of Twente, firemen may also offer help. This because help may not depend on the firemen of the specific region, but on the firemen who can arrive first at the location of the emergency call. Therefore there is indeed emergency call data available near the border of Twente which can be taken into account.

But attention should be paid to the fact that there may still be emergency calls occurred in $\tilde{A}$ which are not involved in the data set. This because firemen of other districts may also help in Twente if they can arrive earlier at the location of the emergency call than firemen of the region Twente. The data of these emergency calls are not available, though, since these data are involved in the data sets of the fire departments of other regions. In this way, the spatial part of the covariate analysis for the squares of the spatial grid P_σ around the border still will not be perfectly accurate, although all the emergency calls for the fire departments of the region Twente in $\tilde{A}$ are taken into account. Nonetheless, the fraction emergency calls for which help is offered by fire departments of other regions is assumed small, based on the experience of the policy and strategics team of the head fire department in Twente.

Based on the previous discussion, the spatial covariate data will thus be filtered for the region $\tilde{A}$. This can be done by *QGIS*. The filtering of the emergency calls for the region $\tilde{A}$ is already done in section 3, since the methods in section 3 already anticipated on the previous discussion and thus used $\tilde{A}$ as the actual region of interest. This explains why some emergency call data laid across the borders of Twente, for example some emergency calls in figure 2. So the models made in this thesis are actually for the region $\tilde{A}$ instead of for A.

Although the useful emergency call data and covariate data now is available for the region $\tilde{A}$, it can still not be analysed. To compare these two kinds of data with each other, a general measure is needed. This measure will be the amount of these data in each grid cell. The spatial covariate data will be counted in the P_σ grid and the temporal covariate data will be counted in the P_τ grid. The emergency call data will of course be counted in both these grids. The counting is not needed for the spatial covariates of covariate data set 4, since this information is already available in the measure of the spatial grid P_σ. In the same way, the temporal covariates of covariate data set 8 are already available in the measure of the temporal grid P_τ and thus neither have to be counted.

Counting the spatial covariate data (that still have to be counted) and the emergency call data in the spatial grid P_σ can be done by the `count points in polygon` function in *QGIS*. This algorithm does not count events which have their locations on the boundaries of the squares of P_σ. Therefore these events will first be translated. Events with a location on the lower boundary of a square of P_σ will be translated a meter to the north, so it is located in the internal region of this square. In a similar way, events with a location on the left boundary of a square of P_σ will be translated a meter to the east, again so it is located in the internal region of this square. In this way no event will be located on the boundary anymore.

After this, the counting algorithm can be run. By the way the events are translated, all events are counted[22]. Moreover, they are even counted only once, in contradiction to some counting algorithms counting an event twice if it is located at the border of these two grid cells. Note that the squares of the grid P_σ do not involve the information for their upper and right boundary, as a consequence of this counting method. So if information about an arbitrary square of the grid P_σ is gives, it involves information about the internal region and the lower and left boundaries of this square.

The counting for the emergency calls in the spatial grid P_σ will further be done per year of the period T_m instead of only for the whole period T_m. This since some covariate information is specified for several years in T_m and therefore the comparision between the occurrence of emergency calls and these covariates can be examined more accurate. If the covariate information is not known for some years in T_m, the unknown years will be based on the year chronologically closest to it. For the covariate analyses, though, the distribution for the whole period T_m is needed. But this distibution can then easily be made from the distributions for each year, by appending these distributions in a chronological way. In this way the information of covariates which are specified per year can accurately be used in the covariate analysis.

Counting the temporal covariate data (that still have to be counted) and the emergency call data in the temporal grid P_τ can be done more easily, since the grid P_τ represents a discrete

[22]Note that it might have been possible that some events had their locations on the right or upper edge of the whole partition P_σ. In that case the translation would make them disappear from the region $\tilde{A}$ and so they would not be counted. But this was not the case for any event.

measure (days) and so has no borders. As a consequence, the data involved is also not able to lie on a border and no adapted counting methods are needed. Next to that, the data of the occurrences of emergency calls is already available in the measure of the day of the year, which was completed to the emergency call data set in section 3. So the counting process for the temporal grid P_τ is very straightforward. It can be done by the `COUNTIF` function in *Microsoft Excel*.

Now the counted information will be summarized. For the spatial part of the covariate analysis, all the spatial covariate data and emergency calls counted in the spatial grid P_σ will be summarized in a table, which will be called the *spatial counting table*, abbreviated as SCT. The rows in the SCT represent each specific square $i, 1 \leq i \leq 6291$, of P_σ. The columns of the SCT represent the level 1a class of interest and the year $y, 2004 \leq y \leq 2015$, of interest in case of emergency calls and the covariate of interest and eventual specified year $y, 2004 \leq y \leq 2015$, for it in case of covariate information.

A same table will be made for the temporal covariate analysis, where all the temporal covariate data and emergency calls counted in P_τ will be summarized. This table will be called the *temporal counting table*, abbreviated as TCT. In the TCT, the rows represent each day $j, 1 \leq j \leq 4380$, of P_τ. The columns of the TCT represent the level 1a class of interestof interest in case of emergency calls and the covariate of interest in case of covariate information.

4.2 Correlation analysis

Now all the data for the spatial part of the covariate analysis is summarized in the SCT and all the data for the temporal part of the covariate analysis is summarized in the TCT, a start can be made with the covariate analysis. As mentioned, this analysis will be done in two steps. The first step is the correlation analysis, which will now be explained.

The correlation analysis examines the influences of the covariates on the emergency calls of each class c_{1a} in a global way. The reason why this is done globally, is that this analysis is meant to make a coarse estimation of this influence. According to these estimations, the covariates with a relative high influence can be distinguished from the other covariates. Since not every covariate of table 8 will be modelled, only the covariates with such a high influence will be filtered from the covariate data set. In this way, the probably most important covariates for each class of emergency calls will be modelled in the corresponding spatio-temporal point process.

The correlation analysis will be executed separately for spatial covariates and temporal covariates. This since spatial covariates have to be related to the spatial information of the emergency calls of interest and the temporal covariates to the temporal information of those emergency calls. For the spatial part of the correlation analysis, the SCT will therefore be used and for the temporal part of the correlation analysis, the TCT will be used.

Both analysis methods are based on the same correlation coefficient for relating the covariates to the emergency calls, though, which is Pearson's correlation coefficient. Let Φ and Ψ be two random variables, $\mathrm{Var}(\cdot)$ be the operator calculating the variance of the involved random variable and $\mathrm{Cov}(\cdot)$ be the operator calculating the covariance between the two involved random variables. Then Pearson's correlation coefficient $\rho_{\Phi,\Psi}$ is defined as follows:

$$\rho_{\Phi,\Psi} = \frac{\mathrm{Cov}(\Phi, \Psi)}{\sqrt{\mathrm{Var}(\Phi)}\sqrt{\mathrm{Var}(\Psi)}} \tag{30}$$

For the spatial part of the correlation analysis, the correlation coefficient $\rho(C_{\sigma,k}, c_{1\mathrm{a}})$ is desired between the spatial covariate of interest $C_{\sigma,k}, 1 \leq k \leq 28$, and the emergency calls of the class of interest $c_{1\mathrm{a}} \in C_{1\mathrm{a}}$. But the correlation coefficient of equation (30) cannot be (immediately) calculated, though, since the covariance between the involved variables and the variances of each of them are not known. The covariance and variance will therefore be replaced by the sample covariance and the sample variance, respectively. These will be calculated from the data of the SCT.

Let $\hat{C}_{\sigma,k,i}$ represent the counted data for covariate $C_{\sigma,k}$ in square $i, 1 \leq i \leq 6291$, of P_σ and let $\hat{c}_{1\mathrm{a},i}$ represent the counted data for the emergency calls of class $c_{1\mathrm{a}}$ in square $i, 1 \leq i \leq 6291$, of P_σ. Implementing the sample covariance and the sample variances in equation (30), the estimated correlation coefficient $\hat{\rho}(C_{\sigma,k}, c_{1\mathrm{a}})$ between the spatial covariate of interest $C_{\sigma,k}, 1 \leq k \leq 28$, and the emergency calls of the class of interest $c_{1\mathrm{a}} \in C_{1\mathrm{a}}$ becomes:

$$\hat{\rho}(C_{\sigma,k}, c_{1\mathrm{a}}) = \frac{\sum_{i=1}^{6291}(\hat{C}_{\sigma,k,i} - \hat{\mathbb{E}}[C_{\sigma,k}])(\hat{c}_{1\mathrm{a},i} - \hat{\mathbb{E}}[c_{1\mathrm{a}}])}{\sqrt{\sum_{i=1}^{6291}(\hat{C}_{\sigma,k,i} - \hat{\mathbb{E}}[C_{\sigma,k}])^2}\sqrt{\sum_{i=1}^{6291}(\hat{c}_{1\mathrm{a},i} - \hat{\mathbb{E}}[c_{1\mathrm{a}}])^2}} \tag{31}$$

where $\hat{\mathbb{E}}[C_{\sigma,k}]$ is the sample mean of covariate $C_{\sigma,k}$, so

$$\hat{\mathbb{E}}[C_{\sigma,k}] = \frac{1}{6291}\sum_{i=1}^{6291}\hat{C}_{\sigma,k,i} \tag{32}$$

and $\hat{\mathbb{E}}[c_{1\mathrm{a}}]$ is the sample mean of the emergency calls of class $c_{1\mathrm{a}}$, so

$$\hat{\mathbb{E}}[c_{1\mathrm{a}}] = \frac{1}{6291}\sum_{i=1}^{6291}\hat{c}_{1\mathrm{a},i} \tag{33}$$

For the temporal part of the correlation analysis, the correlation coefficient $\rho(C_{\tau,l}, c_{1\mathrm{a}})$ is desired between each temporal covariate of interest $C_{\tau,l}, 1 \leq l \leq 13$, and the emergency calls of each class $c_{1\mathrm{a}} \in C_{1\mathrm{a}}$. The same problem arises for calculating the covariance and variances of equation (30). Therefore, the same method for solving this problem can be applied in this case. So let $\hat{C}_{\tau,l,j}$ represent the counted data for covariate $C_{\tau,l}$ in day $j, 1 \leq j \leq 4380$, of P_τ and let $\hat{c}_{1\mathrm{a},j}$ represent the counted data for the emergency calls of class $c_{1\mathrm{a}}$ in day $j, 1 \leq j \leq 4380$ of P_τ,. Then the estimated correlation coefficient $\hat{\rho}(C_{\tau,l}, c_{1\mathrm{a}})$ between the temporal covariate of interest $C_{\tau,l}, 1 \leq l \leq 13$, and the emergency calls of the class of interest $c_{1\mathrm{a}} \in C_{1\mathrm{a}}$ becomes:

$$\hat{\rho}(C_{\tau,l}, c_{1\mathrm{a}}) = \frac{\sum_{j=1}^{4380}(\hat{C}_{\tau,l,j} - \hat{\mathbb{E}}[C_{\tau,l}])(\hat{c}_{1\mathrm{a},j} - \hat{\mathbb{E}}[c_{1\mathrm{a}}])}{\sqrt{\sum_{j=1}^{4380}(\hat{C}_{\tau,l,j} - \hat{\mathbb{E}}[C_{\tau,l}])^2}\sqrt{\sum_{j=1}^{4380}(\hat{c}_{1\mathrm{a},j} - \hat{\mathbb{E}}[c_{1\mathrm{a}}])^2}} \tag{34}$$

where $\hat{\mathbb{E}}[C_{\tau,l}]$ is the sample mean of covariate $C_{\tau,l}$, so

$$\hat{\mathbb{E}}[C_{\tau,j}] = \frac{1}{4380}\sum_{j=1}^{4380}\hat{C}_{\tau,l,j} \tag{35}$$

and $\hat{\mathbb{E}}[c_{1\mathrm{a}}]$ the sample mean of the emergency calls of class $c_{1\mathrm{a}}$, so

$$\hat{\mathbb{E}}[c_{1\mathrm{a}}] = \frac{1}{4380}\sum_{j=1}^{4380}\hat{c}_{1\mathrm{a},j} \tag{36}$$

	"fire"	"service"	"accident"	"alert"	"environmental"
$C_{\sigma,1}$	0.674	0.571	0.147	0.400	0.419
$C_{\sigma,2}$	0.459	0.556	0.124	0.398	0.326
$C_{\sigma,3}$	0.178	0.168	0.057	0.310	0.146
$C_{\sigma,4}$	0.370	0.287	0.117	0.197	0.221
$C_{\sigma,5}$	0.326	0.408	0.132	0.373	0.213
$C_{\sigma,6}$	-0.002	0.012	-0.005	0.011	0.009
$C_{\sigma,7}$	0.385	0.293	0.086	0.231	0.226
$C_{\sigma,8}$	0.267	0.203	0.071	0.126	0.171
$C_{\sigma,9}$	0.431	0.527	0.104	0.408	0.289
$C_{\sigma,10}$	-0.010	-0.014	0.116	-0.006	0.001
$C_{\sigma,11}$	0.027	0.002	0.138	-0.002	-0.007
$C_{\sigma,12}$	-0.025	-0.016	0.060	-0.006	0.005
$C_{\sigma,13}$	0.175	0.159	0.063	0.106	0.115
$C_{\sigma,14}$	0.634	0.489	0.127	0.319	0.389
$C_{\sigma,15}$	0.547	0.376	0.104	0.217	0.330
$C_{\sigma,16}$	0.629	0.499	0.122	0.315	0.366
$C_{\sigma,17}$	0.641	0.499	0.128	0.315	0.376
$C_{\sigma,18}$	0.586	0.438	0.120	0.273	0.376
$C_{\sigma,19}$	0.571	0.503	0.127	0.409	0.379
$C_{\sigma,20}$	0.632	0.489	0.128	0.313	0.388
$C_{\sigma,21}$	0.631	0.491	0.128	0.331	0.390
$C_{\sigma,22}$	0.664	0.535	0.134	0.350	0.407
$C_{\sigma,23}$	0.639	0.538	0.174	0.391	0.393
$C_{\sigma,24}$	-0.085	-0.072	0.007	-0.052	-0.032
$C_{\sigma,25}$	0.369	0.288	0.146	0.224	0.252
$C_{\sigma,26}$	0.019	0.024	0.047	0.007	0.001
$C_{\sigma,27}$	0.002	-0.005	0.005	-0.016	-0.006
$C_{\sigma,28}$	0.024	0.037	0.059	0.024	0.007
$C_{\tau,1}$	0.000	0.009	0.017	-0.035	0.116
$C_{\tau,2}$	-0.045	0.078	-0.044	0.044	0.028
$C_{\tau,3}$	-0.048	0.106	0.058	0.057	0.290
$C_{\tau,4}$	0.061	-0.045	-0.012	0.027	-0.095
$C_{\tau,5}$	-0.065	-0.051	0.014	0.072	0.013
$C_{\tau,6}$	0.040	0.042	0.000	-0.026	-0.029
$C_{\tau,7}$	0.398	0.022	-0.008	0.084	-0.002
$C_{\tau,8}$	0.763	0.070	0.010	0.079	0.005
$C_{\tau,9}$	0.018	-0.017	0.022	-0.075	-0.025
$C_{\tau,10}$	-0.047	0.064	-0.038	0.002	0.027
$C_{\tau,11}$	-0.076	-0.014	-0.007	0.089	-0.011
$C_{\tau,12}$	0.106	-0.033	0.024	-0.015	0.009
$C_{\tau,13}$	-0.073	0.092	-0.066	0.013	-0.002

Table 9: Correlation coefficients between each covariate $C_{\sigma,i}, 1 \leq i \leq 28$, $C_{\tau,l}, 1 \leq l \leq 13$, and each level 1a class of emergency calls $c_{1a} \in C_{1a}$.

The resulting correlation coefficients for both the spatial part and the temporal part of the correlation analysis are shown in table 9. Now these correlation coefficients are calculated, the most influent covariates for each level 1a class of emergency calls can be filtered. The six most important covariates for each class will be filtered from tabel 9. The results of this filtering is shown in table 10, where 1 marks the covariate with the highest value for the correlation coefficient between this covariate and the class of emergency calls of interest, 2 marks the covariate with the second highest value for this correlation coefficient, and so on. The covariates of table 10 will then be further analysed by regression analysis.

	"fire"	"service"	"accident"	"alert"	"environmental"
1	$C_{\tau,8}$	$C_{\sigma,1}$	$C_{\sigma,23}$	$C_{\sigma,19}$	$C_{\sigma,1}$
2	$C_{\sigma,1}$	$C_{\sigma,2}$	$C_{\sigma,1}$	$C_{\sigma,9}$	$C_{\sigma,22}$
3	$C_{\sigma,22}$	$C_{\sigma,23}$	$C_{\sigma,25}$	$C_{\sigma,1}$	$C_{\sigma,23}$
4	$C_{\sigma,17}$	$C_{\sigma,22}$	$C_{\sigma,11}$	$C_{\sigma,2}$	$C_{\sigma,21}$
5	$C_{\sigma,23}$	$C_{\sigma,9}$	$C_{\sigma,22}$	$C_{\sigma,23}$	$C_{\sigma,14}$
6	$C_{\sigma,14}$	$C_{\sigma,19}$	$C_{\sigma,5}$	$C_{\sigma,5}$	$C_{\sigma,20}$

Table 10: The 6 most influent covariates per level 1a class of emergency calls $c_{1a} \in C_{1a}$.

It can be seen that the spatial covariates seem much more influent than the temporal ones in general. This may be since the focus is more laid on spatial covariates than on temporal covariates, since 28 spatial covariates are involved and 12 temporal covariates. As a consequence, a spatial point process model for the classes c_{1a} = service, c_{1a} = accident, c_{1a} = alert and c_{1a} = environmental then seems sufficient and time invariance per day seems assumed. This seems intuitively wrong, but also temporal tests for the distributions of the days for each class of emergency calls mentioned prove that there are days with a significantly different distributions (although not so strongly different). So this means that other covariates are needed to describe the temporal part of the emergency calls of the earlier mentioned classes. This will not be done anymore in this thesis, though.

Next to that, many spatial covariates in table 10 are again (strongly) correlated with each other. For example, $\rho(C_{\sigma,14}, C_{\sigma,20}) = 0.843$, which is also very logical, since the number of male residents depends (intuitively) on the total number of residents. This will be no problem, though, since the spatio-temporal point process modelling for a specific level 1a class is a kind of multivariate regression and therefore it keeps an eye on the covariances between covariates and it then filters the covariates that clearify the occurrences of the emergency calls of interest the best. So it will say whether $C_{\sigma,14}$ or $C_{\sigma,20}$ is a better descriptor of the emergency calls of the class "environmental", for example. This is the main reason why six covariates for each class are filtered, since many of these six may be strongly correlated with each other and therefore the number of significant covariates may be reduced a lot in the modelling.

4.3 Regression analysis

To conclude the covariate analysis, regression analysis will be done. This analysis will relate each selected covariate in table 10 to the emergency calls of the corresponding level 1a class. These relations will be needed for the spatial point process modelling, involving the classes $c_{1a} \in \tilde{C}_{1a}$, $\tilde{C}_{1a} = \{\text{service}, \text{accident}, \text{alert}, \text{environmental}\}$, and for the spatio-temporal point process modelling, involving the class c_{1a} = fire. Regression analysis also has to be done separately for the spatial covariates and the temporal covariates, for the same reasons as for the correlation analysis.

Let $C_{c_{1a},\sigma,k}, 1 \leq k \leq p$, represent the selected spatial covariates of table 10 for the emergency calls of the class c_{1a}. So for $c_{1a} = \text{fire}$, $p = 5$ and for the other classes, $p = 6$. The spatial part of the regression analysis then tries to find the global relation $g(c_{1a}, C_{c_{1a},\sigma,1}, C_{c_{1a},\sigma,2}, \ldots, C_{c_{1a},\sigma,p})$ between the occurrences of emergency calls of this class c_{1a} and each spatial covariate $C_{c_{1a},\sigma,k}$. For this relation, it is assumed that each covariate influences the emergency calls of interest independently from each other, so cross terms of different covariates will not appear in the function representing the relation $g(c_{1a}, C_{c_{1a},\sigma,1}, C_{c_{1a},\sigma,2}, \ldots, C_{c_{1a},\sigma,p})$.

The exact relation $g(c_{1a}, C_{c_{1a},\sigma,1}, C_{c_{1a},\sigma,2}, \ldots, C_{c_{1a},\sigma,p})$ is unknown and therefore will be estimated. This relation will be estimated by determining the global relations $g_k(c_{1a}, C_{c_{1a},\sigma,k})$ between the occurrences of emergency calls of the class c_{1a} and each individual spatial covariate $C_{c_{1a},\sigma,k}$. This is made possible by the assumption that each covariate exerts influences independently from each other. In this way finding a $(p+1)$-dimensional function representing the relation $g(c_{1a}, C_{c_{1a},\sigma,1}, C_{c_{1a},\sigma,2}, \ldots, C_{c_{1a},\sigma,p})$ is reduced to finding p two-dimensional functions representing the relations $g_k(c_{1a}, C_{c_{1a},\sigma,k})$.

The estimates $\hat{g}_k(c_{1a}, C_{c_{1a},\sigma,k})$ of these relations $g_k(c_{1a}, C_{c_{1a},\sigma,k})$ will be based on the data $\hat{C}_{c_{1a},\sigma,k,i}$ and $\hat{c}_{1a,i}$ from the SCT, where $\hat{C}_{c_{1a},\sigma,k,i}$ represents the counted data for covariate $C_{c_{1a},\sigma,k}$ in square $i, 1 \leq i \leq 6291$, of P_σ and $\hat{c}_{1a,i}$ represents again the counted data for the emergency calls of class c_{1a} in square $i, 1 \leq i \leq 6291$, of P_σ. Each estimate $\hat{g}_k(c_{1a}, C_{c_{1a},\sigma,k})$ can then be found by examining which kind of function fits the best to the data $\hat{C}_{c_{1a},\sigma,k,i}$ and $\hat{c}_{1a,i}$. Comparing the quality of these fits for different kinds of functions can be done by univariate regression. The less error the function of interest has with respect to the data, the better this kind of function represents the relation.

Recall that the estimates $\hat{g}_k(c_{1a}, C_{c_{1a},\sigma,k})$ found in this way only represent the global relations, since the found coefficients for the fitted functions are not the (influence) coefficients for the model. The reason for this is that the univariate regression assigns all emergency calls of the class of interest to the covariate of interest in the regression, while some of these emergency calls may have no relation with the covariate of interest, since different emergency calls may be caused by different covariates.

A multivariate regression with the found global functions for each covariate involved would indeed solve this problem and give the true coefficients. Remark that in this way, the function $\hat{g}(c_{1a}, C_{c_{1a},\sigma,1}, C_{c_{1a},\sigma,2}, \ldots, C_{c_{1a},\sigma,p})$ is an estimator for the intensity function for each square $i, 1 \leq i \leq 6291$, of P_σ. Spatial point process modelling then uses the global relation of this estimate for determining the influence coefficients of the intensity function for each location $\mathbf{x} \in A$, where A is the region of interest, in stead of for each square i of P_σ. This will be done in section 5.

Since there are infinitely many kinds of functions, some kinds of functions have to be chosen that will be examined. This can be done by analysing the plots of $\hat{C}_{c_{1a},\sigma,k,i}$ against $\hat{c}_{1a,i}$ and by restricting the function from getting too complex. In this way, the following three families are selected to analyse the relations for the regression analysis in this thesis:

1. $Y = \zeta_4 X^4 + \zeta_3 X^3 + \zeta_2 X^2 + \zeta_1 X + \zeta_0, \qquad \zeta_l \in \mathbb{R}, 0 \leq l \leq 4$
2. $Y = {\eta_0}^X, \qquad \eta_0 \in \mathbb{R}$
3. $Y = X^{(\theta_0 + \theta_1 X + \theta_2 X^2)} e^{(\theta_3 + \theta_4 X + \theta_5 X^2)}, \qquad \theta_l \in \mathbb{R}, 0 \leq l \leq 5$

where $Y = c_{1a}$ and $X = C_{c_{1a},\sigma,k}$ in this thesis. Remark that a lot of functions are covered with these three families of functions, since coefficients $\zeta_0, \zeta_1, \ldots, \zeta_4, \eta_0, \theta_0, \theta_1, \ldots, \theta_5$ may be zero. For example, if the real relation would be $y = x^3 + 1$, family 1 will appear to have the best fitting regression for the data, where $\zeta_3 = \zeta_0 = 1$ and $\zeta_4 = \zeta_2 = \zeta_1 = 0$.

The above families are only allowed to fit continuous relations and such a relation is the case for almost any selected spatial covariate from table 10. The only spatial covariate from this table for which the relation is discrete is $C_{\sigma,25}$, which is a Boolean variable equaling 1 if a town is present in a specific square $i, 1 \leq i \leq 6291$, of P_σ and equaling 0 if a town is not present in that square. The function

4. $Y = \kappa_0 + \kappa_1 X, \qquad \kappa_l \in \mathbb{R}, l \in \{0, 1\}$

can be used to fit such a Boolean variable $X \in \{0, 1\}$, where κ_0 represents the fitted value for $X = 0$ and $\kappa_0 + \kappa_1$ represents the fitted value for $X = 1$. In the case of examining the influence of $C_{\sigma,25}$, $Y = c_{1a}$ and $X = C_{\sigma,25}$.

A last remark about the fitted functions is that they should be nonnegative, since a negative number of emergency calls cannot occur. If a function has negative values, though, it often has a bad fit for the regression as a consequence. If although this fit still happens to have the best fit, it could be adapted to be nonnegative by changing all negative values to the value 0. In this way, the fit will even become better. Further, all involved covariates have nonnegative values, so the function will live in the first quadrant.

Now the spatial part of the regression analysis will be executed for each $\hat{C}_{c_{1a},\sigma,k}, 1 \leq k \leq p$, by `R`. By examining the regression plots, residual plots and the summary of the regression analysis in `R`, one can see how well each kind of function fits the data. For example, the regression analysis plot for data of $C_{\sigma,14}$ against the emergency calls of c_{1a} = fire is given in figure 10, in which it can be seen that family 3 seems to be the best fitting relation between the data of $C_{\sigma,14}$ and the emergency calls of c_{1a} = fire. Further, it must be noted that family 1 fits very badly, since the fitting for it does not even converge in `R`. The residual plots for each of the involved families agrees with this. The residual plot for familiy 3 is shown in figure 10.

Even for every continuous spatial covariate of any class c_{1a}, spatial regression analysis indicates family 3 as the best fitting family. This could have been expected in some way, since this family has the most coefficients and so the most degrees of freedom for fitting to the data. Nonetheless, all three families have been tested, since there could have been a change that a polynomial of degree 3 or 4 or a power function would have been the better fit. The discrete spatial covariate $C_{\sigma,25}$ is further fitted by family 4, which also results in a good fit.

A same approach can be taken for the temporal part of regression analysis. As earlier mentioned, $C_{\tau,8}$ was the only temporal covariate in table 10 and the corresponding emergency calls were of the class c_{1a} = fire. Let $\hat{C}_{\tau,8,j}$ represent the counted data for covariate $C_{\tau,8}$ on day $j, 1 \leq j \leq 4380$, of P_τ and let $\hat{c}_{\text{fire},j}$ represent the counted data for the emergency calls of class c_{1a} = fire in day $j, 1 \leq j \leq 4380$, of P_τ. Since $C_{\tau,8}$ is a Boolean variable, family 4 will again be used for fitting, as was also the case for $C_{\sigma,25}$. Also in this case, the residual plots and summary of the regression indicate that the fit is good. The fitting and analysing is again done by `R`.

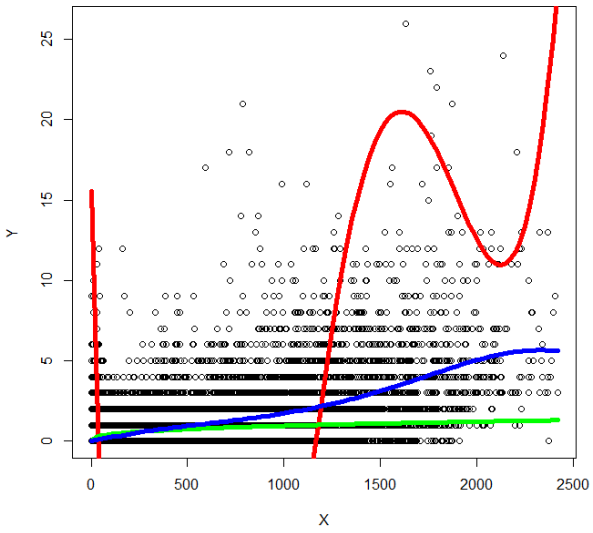

Figure 9: The data of $C_{\sigma,14}$ plotted against the emergency calls of $c_{1\mathrm{a}} = \text{fire}$ for the described regression analysis involving the functions of family 1 (the red graph), 2 (the green graph) and 3 (the blue graph).

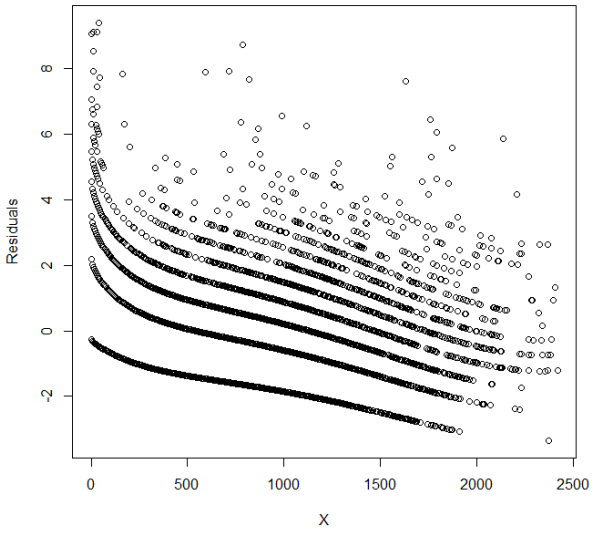

Figure 10: The residual plot of the described regression analysis of the data of $C_{\sigma,14}$ against the emergency calls of $c_{1\mathrm{a}} = \text{fire}$ for family 3.

5 Spatio-temporal point process fitting

So the emergency calls of the class c_{1a} = fire will be modelled as a spatio-temporal inhomogeneous Poisson process and the emergency calls of the remaining level 1a classes will be modelled as a (purely) spatial inhomogeneous Poisson process. Further, in the modelling of each process (a maximum of) six covariates and their corresponding relations found in section 4 will be taken into account. The data of the occurred emergency calls of each class will be related to this corresponding covariate information by the intensity function, as described in section 2. These intensity functions are the main ingredient for the inhomogeneous Poisson processes to be made.

But how should this intensity function now be estimated from the analyses earlier executed? In section 2 it was only mentioned that the relations between the occurrences of emergency calls of interest and each involved covariate can be examined globally by regression analysis and that the influence coefficients corresponding to these relations can be found by the method of maximum pseudolikelihood estimation. The reason why this estimation technique is used in stead of for example maximum likelihood estimation will be explained in this section. If these influence coefficients are then estimated, they can be implemented in the intensity function for each class together with the (global) relations found in the regression analysis. In this way, the intensity functions for the spatial point processes and spatio-temporal point processes can be made.

In this section, the described method will be explained more thoroughly and will be executed. First, the maximum pseudolikelihood estimate for a spatial point process and for a spatio-temporal point process will be examined. Then it will be explained how the former estimate is implemented in the functions of the `spatstat` package in `R` for the modelling of spatial point processes. Since `R` has no immediate function for modelling spatio-temporal point processes, it will also be explained how the maximum pseudolikelihood estimate for a spatio-temporal point process can be implemented for this extension. This implementation is already done in `R` by prof. dr. M.N.M. van Lieshout and Adina Iftimi and their algorithm will be used in this thesis for the spatio-temporal point process modelling. The section then concludes by validating the modelled spatial or spatio-temporal point processes for each class of emergency calls. In this way, the quality of the models can be checked and so whether or not they have to be extended further (by for example also modelling stochastic interaction).

A last remark has to be made, since the methods in this section are only applicable to (a time series of) spatial point patterns where there are no two events with the same $(\mathbf{x}, t)$ coordinates, since spatial point patterns are in a formal way a set and not a tuple, as can be seen from definition 2.1. This means that the techniques made for analysing them are actually not meant for analysing also the duplicated events. The analyses for the spatial point patterns in section 3 is actually still fine, since the distance analysis methods only look at nonzero distances.

The reason why duplicated events were not removed in the filtering executed in section 3 is that for the covariate analyses, each of the several duplicated events (so with the same $(\mathbf{x}, t)$ coordinates) contains information about the occurrences of the emergency calls. But for the modelling, these duplicated events have to be removed, since building a whole new theory for spatial point patterns with duplicated events is "a Herculean task", according to Turner (2009). Duplicated events could be taken into account by slightly translating them with a similar method as for translating the events on the boundaries of the spatial grid P_σ in section 4, but even this is quite complicated to do. So this thesis involves models where duplicated events are removed.

5.1 Estimation of the intensity function

The emergency calls of the classes $c_{1a} \in \tilde{C}_{1a}$, $\tilde{C}_{1a} = \{\text{service}, \text{accident}, \text{alert}, \text{environmental}\}$ will be modelled as a purely spatial inhomogeneous Poisson process, so the one formulated in definition 2.8. This because the six most influential covariates for these classes happen to be all spatial. The intensity function for these classes is as a consequence the one from equation (19). Since the global relations $g_k(c_{1a}, C_{c_{1a},\sigma,k})$ between the emergency calls of each of the involved classes $c_{1a} \in \tilde{C}_{1a}$ and the six most influential covariates for these classes $C_{c_{1a},\sigma,k}, 1 \leq k \leq 6$ were concluded from the regression analysis in section 4, equation (19) can be specified further as:

$$\lambda_{\theta_\sigma}(\mathbf{x}) = \prod_{k=1}^{6} g_k(c_{1a}, C_{c_{1a},\sigma,k}) \tag{37}$$

Now let $h(\mathbf{x})$ be defined as:

$$h_k(\mathbf{x}) = \mathbf{x}^{(\theta_{k,0}(\mathbf{x})+\theta_{k,1}C_k(\mathbf{x})+\theta_{k,2}(C_k^2(\mathbf{x})^2))} e^{(\theta_{k,3}+\theta_{k,4}C_k(\mathbf{x})+\theta_{k,5}(C_k(\mathbf{x})^2))} \tag{38}$$

Then the intensity function for the classes $c_{1a} = \text{service}$, $c_{1a} = \text{alert}$ and $c_{1a} = \text{environmental}$ can be formulated as:

$$\lambda_{\theta_\sigma}(\mathbf{x}) = \prod_{k=1}^{6} h_k(\mathbf{x}) \tag{39}$$

and the intensity function for the class $c_{1a} = \text{accident}$ as:

$$\lambda_{\theta_\sigma}(\mathbf{x}) = \left(\prod_{k=1}^{2} h_k(\mathbf{x}) \right) (\theta_{3,0} + \theta_{3,1} C_{\sigma,25}) \left(\prod_{k=4}^{6} h_k(\mathbf{x}) \right) \tag{40}$$

In equations (38), (39) and (40), $C_k(\mathbf{x}), 1 \leq k \leq 6$, represents the k^{th} most important covariate for the class $c_{1a} \in \tilde{C}_{1a}$ of interest and $\theta_\sigma = (\theta_{1,0}, \theta_{1,1}, \ldots, \theta_{6,5})$ is the vector representing all the influence coefficients, which are related to the six spatial covariates involved.

Now it will be explained how the influence coefficients of θ_σ in the intensity functions of equation (39) and (40) will be discovered for each level 1a class $c_{1a} \in \tilde{C}_{1a}$. As mentioned, this will be done by the maximum pseudolikelihood estimating technique. The discussion will be based on Baddeley and Turner (2000). To maximise the pseudolikelihood, the pseudolikelihood first has to be determined. For the (purely) spatial inhomogeneous Poisson process, the pseudolikelihood PL is defined as follows (Baddeley and Turner, 2000):

$$\text{PL}(\theta_\sigma) = \prod_{i=1}^{|S|} \lambda_{\theta_\sigma}(\mathbf{x}_i) e^{-\int_A \lambda_{\theta_\sigma}(u)\, \mathrm{d}u} \tag{41}$$

where S is the spatial point pattern consisting of the emergency calls for the level 1a class of interest $c_{1a} \in \tilde{C}_{1a}$ and A is the region of interest. The idea now is to find the values $\hat{\theta}_\sigma$ of θ_σ corresponding to the maximum of $\text{PL}(\theta_\sigma)$. These estimates $\hat{\theta}_\sigma$ are the values to complete the estimated intensity functions for the spatial inhomogeneous Poisson processes of interest. Estimating $\hat{\theta}_\sigma$ is done by maximising the logarithm of the pseudolikelihood rather than just the pseudolikelihood, since taking the logarithm of the right-hand side of equation (41) has a much simpler expression[23]. The logarithm of the pseudolikelihood, abbreviated as the "log

[23]The reason that maximising the logarithm of the pseudolikelihood has the same answer as maximising the pseudolikelihood directly is that the logarithm is an increasing function.

pseudolikelihood", for each spatial inhomogeneous Poisson process becomes:

$$\log\big[\mathrm{PL}(\theta_\sigma)\big] = \sum_{i=1}^{|S|} \log\big[\lambda_{\theta_\sigma}(\mathbf{x}_i)\big] - \int_A \lambda_{\theta_\sigma}(u)\,\mathrm{d}u \tag{42}$$

The difficulty in maximizing the (log) pseudolikelihood is that the integral $\int_A \lambda_{\theta_\sigma}(u)\,\mathrm{d}u$ is quite difficult to calculate. To approximate this integral, region A will be partitioned in p polygons $B_k \subset A, 1 \leq k \leq p$. Further, some new concepts are needed. Let $Q = \{u_1, u_2, \ldots, u_m\}, u_i \in A$ be a set of random points distributed over the polygons $B_1, B_2, \ldots, B_p$ of the partition, respectively. Then these points $u_i \in A, 1 \leq i \leq m$ are called the *quadrature points*. Each quadrature point u_i also has an associated weight $w_i \geq 0$ and these weights are called the *quadrature weights*. With these concepts, the integral of interest will be approximated by the Berman-Turner method (Berman and Turner, 1992) as follows:

$$\int_A \lambda_{\theta_\sigma}(u)\mathrm{d}u \approx \sum_{i=1}^{m} \lambda_{\theta_\sigma}(u_i) w_i \tag{43}$$

The Berman-Turner method then chooses the quadrature points $u_i \in Q$ to include all the (locations of the) emergency calls of interest $\mathbf{x}_j \in S$ and some other "dummy" points in the region of the grid. Let D denote the set of dummy points, so $Q = S \cup D$, and let z_i be a Boolean variable equaling 1 if $u_i \in S$ and 0 if $u_i \in D$ and $y_i = z_i/w_i$. Then equation (42) reduces further to:

$$\log\big[\mathrm{PL}(\theta)\big] \approx \sum_{i=1}^{m} \big(z_i \log[\lambda_{\theta_\sigma}(u_i)] - w_i \lambda_{\theta_\sigma}(u_i)\big) \tag{44}$$

$$= \sum_{i=1}^{m} w_i \big(y_i \log[\lambda_{\theta_\sigma}(u_i)] - \lambda_{\theta_\sigma}(u_i)\big) \tag{45}$$

Now one can see the benefits of the Berman-Turner device, since equation (45) has the same form as the log likelihood of multiple independent Poisson random variables with weights w_i, means $\lambda_{\theta_\sigma}(u_i)$ and responses y_i. Therefore the log pseudolikelihood for an inhomogeneous Poisson process representing a spatial point process can be maximised by the standard software for fitting generalized linear models.

The previous discussion involved finding an estimation $\hat{\theta}_\sigma$ for the influence coefficients θ_σ for spatial point processes representing emergency calls of the classes $c_{1\mathrm{a}} \in \tilde{C}_{1\mathrm{a}}$. These spatial point processes were modelled as an ordinary spatial inhomogeneous Poisson processes. The emergency calls of the class $c_{1\mathrm{a}} =$ fire cannot be modelled in such a way, since it has to be modelled as a spatio-temporal point process. This was a consequence of the fact that the six most influent covariates for $c_{1\mathrm{a}} =$ fire not did only involve spatial covariates, but also a temporal one. This spatio-temporal point process will be modelled as a spatio-temporal inhomogeneous Poisson process with intensity function $\lambda(\mathbf{x}, t)$.

First, the intensity function $\lambda_\theta(\mathbf{x}, t)$ for this process will be described, where $\theta = (\theta_\tau, \theta_\sigma)$ is the vector representing the influence coefficients θ_τ and θ_σ for the temporal and spatial covariates, respectively. Remember that separability was assumed, so $\lambda_\theta(\mathbf{x}, t) = \lambda_{\sigma,\theta_\sigma}(\mathbf{x})\lambda_{\tau,\theta_\tau}(t)$. The global relations between the emergency calls of class $c_{1\mathrm{a}} =$ fire and the spatial covariates, which are concluded from the regression analysis in section 4, can be used to express the global behaviour of $\lambda_{\sigma,\theta_\sigma}(\mathbf{x})$:

$$\lambda_{\sigma,\theta_\sigma}(\mathbf{x}) = \prod_{k=2}^{6} h_k(\mathbf{x}) \tag{46}$$

and in the same way the global behaviour of $\lambda_{\tau,\theta_\tau}(t)$ can be expressed based on the global relation between the emergency calls of class $c_{1\text{a}}$ = fire and the only temporal covariate $C_{\tau,8}$:

$$\lambda_{\tau,\theta_\tau}(\mathbf{x}) = (\theta_{1,0} + \theta_{1,1}C_{\tau,8}) \tag{47}$$

In equation (46), $C_k(\mathbf{x})$ is the k^{th} most important covariate of the level class $c_{1\text{a}}$ = fire where $2 \leq k \leq 6$ and $C_1(\mathbf{x}) = C_{\tau,8}$, since the temporal covariate $C_{\tau,8}$ in equation (47) is the most important covariate. Further, $\theta = (\theta_{1,0}, \theta_{1,1}, \theta_{2,0} \ldots, \theta_{6,5})$ is the vector representing all the influence coefficients, where $\theta_\tau = (\theta_{1,0}, \theta_{1,1})$ and $\theta_\sigma = (\theta_{2,0}, \theta_{2,1}, \ldots, \theta_{6,5})$.

The vector $\theta = (\theta_\tau, \theta_\sigma)$ consisting of the influence coefficients of the intensity function $\lambda_\theta(\mathbf{x}, t)$ will be estimated by an extension to the method for estimating the purely spatial intensity function $\lambda(\mathbf{x})$. It will also involve the maximum pseudolikelihood method for estimating θ. First the pseudolikelihood for a spatio-temporal inhomogeneous Poisson process will be defined:

$$\text{PL}(\theta_\sigma) = \prod_{i=1}^{|S|} \lambda_{\theta_\sigma}(\mathbf{x}_i, t_i) e^{-\int_T \int_A \lambda_{\theta_\sigma}(u,v)\ \mathrm{d}u\ \mathrm{d}v} \tag{48}$$

where S is the spatial point pattern consisting of emergency calls for the level 1a class $c_{1\text{a}}$ = fire, T is the time period of interest and A is the region of interest. Again the log likelihood will be maximised instead of the ordinary likelihood. The log likelihood for a spatio-temporal inhomogeneous Poisson process becomes:

$$\log\big[\text{PL}(\theta)\big] = \sum_{i=1}^{|S|} \log\big[\lambda_\theta(\mathbf{x}_i, t_i)\big] - \int_T \int_A \lambda_{\theta_\sigma}(u, v)\ \mathrm{d}u\ \mathrm{d}v \tag{49}$$

Again the integral from equation (49) will be approximated by a partition, this time for the space-time region $A \times T$. The spatial region A will again be partitioned in p polygons $B_k \subseteq A, 1 \leq k \leq p$ and the time period T will be partitioned in q subperiods $U_l \subseteq T, 1 \leq l \leq q$. Now define the m quadrature points by the set $Q = \{(u_1, v_1), (u_2, v_2), \ldots, (u_m, v_m)\}, (u_i, v_i) \in (A \times T)$, where $u_i \in A$ is the spatial part of these quadrature points and $v_i \in T$ is the temporal part. The weight $w_i \geq 0$ is associated with the quadrature point (u_i, v_i). The integral of interest is then again approximated by the Berman-Turner method:

$$\int_T \int_A \lambda_{\theta_\sigma}(u, v)\ \mathrm{d}u\ \mathrm{d}v \approx \sum_{i=1}^{m} \lambda_\theta(u_i, v_i) w_i \tag{50}$$

Let S now denote the data set consisting of the space-time coordinates $(\mathbf{x}, t)$ of all the occurred events (so actually a spatio-temporal point pattern) and let D again denote the set of dummy points. According to the Berman-Turner method, $Q = S \cup D$ again and z_i is defined as the Boolean variable equaling 1 if $(u_i, v_i) \in S$ and 0 if $(u_i, v_i) \in D$. Again defining $y_i = z_i / w_i$, equation (49) reduces to:

$$\log\big[\text{PL}(\theta)\big] \approx \sum_{i=1}^{m} \big(z_i \log[\lambda_\theta((u_i, v_i))] - w_i \lambda_\theta(u_i, v_i)\big) \tag{51}$$

$$= \sum_{i=1}^{m} w_i \big(y_i \log[\lambda_\theta(u_i, v_i)] - \lambda_\theta(u_i, v_i)\big) \tag{52}$$

and again equation (52) can be approximated by software for fitting generalized linear models, since it is also similar to the log likelihood of multiple independent Poisson variables Y_i with weights w_i, means $\lambda_{\theta_\sigma}(u_i)$ and responses y_i.

5.2 Model fitting and validation

Now the method for estimating the influence coefficients by maximising (log) pseudolikelihoods is explained, it will be executed for each level 1a class of emergency calls. First, the spatial point process for the classes $c_{1a} \in \tilde{C}_{1a}$ will be fitted and after that the spatio-temporal point process for the class $c_{1a} =$ fire will be fitted. After the models are fitted, their quality will be checked by diagnostic plots[24] and by comparing the model for each class to the industry model for the emergency calls of the respective class. The idea of the current industry model was already explained in section 1, based on the discussion in Zhou et al. (2015).

For modelling the spatial point processes corresponding to the classes $c_{1a} \in \tilde{C}_{1a}$, A is again partitioned in the spatial grid P_σ (as defined in section 4) for the Berman-Turner method. In the center of each square of 500 meter involved in P_σ a dummy point is placed. The counting weights concept of Baddeley and Turner (2000) is further used to define the weights w_i, which are therefore defined as $w_i = 2.5 \cdot 10^5 / n_i$, where $2.5 \cdot 10^5$ represents the area of a square of P_σ in squared meter and n_i represents the number of quadrature points in square $i, 1 \leq i \leq 6291$.

As a next step, each quadrature point u_i should be completed with the corresponding values of the covariates of interest. This covariate information is needed for each quadrature point $u_i \in Q_{c_{1a}}$, where $Q_{c_{1a}}$ is the set containing all quadrature points for the level 1a class of interest, since fitting the spatial point process actually involves a kind of multivariate regression between the location $\mathbf{x} \in \mathbb{R}^2$ (with realizations u_i) as dependent variable and the covariates $C_{c_{1a},\sigma,k}, 1 \leq k \leq 6$, as explanatory variables. Even more specific, this regression is done for both the X and Y coordinates, and these coordinates of quadrature points u_i will be denoted by $u_{i,X}$, $u_{i,Y}$, respectively. So a table involving the values of $u_{i,X}$, $u_{i,Y}$ and $C_{c_{1a},\sigma,k}, 1 \leq k \leq 6$, for all quadrature points $u_i, 1 \leq i \leq |Q|$, is needed for fitting a spatial point process for the classes $c_{1a} \in \tilde{C}_{1a}$. This table will be denoted by TQP and is partially given for the class $c_{1a} =$ service in table 11. The covariate information can be filtered for each quadrature point by *QGIS*.

i	type u_i	$u_{i,X}$	$u_{i,Y}$	$C_{\sigma,1}$	$C_{\sigma,2}$	$C_{\sigma,23}$	$C_{\sigma,22}$	$C_{\sigma,9}$	$C_{\sigma,19}$
1	data	242923	487246	900	5	1434	648	185	198
2	data	241799	486175	956	13	2046	658	194	396
3	data	240858	485049	886	2	1593	719	91	272
4	data	258574	470923	1196	3	3401	1075	23	434
5	data	243595	487947	436	3	1004	411	14	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
6366	dummy	241750	501750	0	0	150	193	0	70
6367	dummy	241250	501250	0	0	88	0	0	0
6368	dummy	241750	501250	2	0	141	10	2	3
6369	dummy	242250	501750	2	0	157	180	0	75
6370	dummy	242250	501250	18	0	150	20	1	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 11: A part of the table containing the information about the quadrature points (TQP) for the class $c_{1a} =$ service, which consists of 6365 data points and 6291 dummy points.

[24] The diagnostic plots are only used for validation of the purely spatial point processes, so for the classes $c_{1a} \in \tilde{C}_{1a}$, since `spatstat` does not have the tools yet to make diagnose plots for spatio-temporal point processes.

A remark should be made for making the TQP needed for fitting the corresponding spatial point process. It may be possible that covariates are involved for which the values are known for several years in T_m. But for which year should these values be involved in the TQP? In this thesis, the average is taken over the values of all years in T_m. Even when these values are not known for a specific year, the values of the year chronologically closest to it are assigned to that specific year. In this way, the whole period T_m is represented well, which is required since the model is based on this period. This averaging method is for example applied to $C_{\sigma,23}$ of table 11.

Now it is explained how the TQP can be made, it will be used to fit a spatial (inhomogeneous Poisson) point process to the emergency calls of each level 1a class $c_{1a} \in \tilde{C}_{1a}$. For fitting the spatial point processes, the `ppm` function of the `spatstat` package in `R` will be used, which estimates the influence coefficients θ_σ corresponding to the (global) intensity functions for each level 1a class found in section 4. These estimated influence coefficients $\hat{\theta}_\sigma$ are found for the fitted Poisson processes by the maximum pseudolikelihood estimating technique for spatial (inhomogeneous Poisson) point processes earlier described, so by finding the values for w_i and y_i for each quadrature point $u_i \in Q_{c_{1a}}$ and fitting the generalized linear model with weights w_i, means $\log(\lambda_{\theta_\sigma}(u_i))$ and responses y_i and of course Poisson distributed errors.

Before using the `ppm` function, the required arguments for this function will be explained. The first argument involves a data frame of the two columns of TQP which respectively contain the values for $u_{i,X}$ and $u_{i,Y}$. The second argument involves the global relations between the events of the spatial point pattern of interest and the covariates of interest in a log linear form. The third argument involves the kind of interaction between the events of the spatial point pattern of interest. A data frame containing the covariate information of the TQP is finally involved as the `covariates` argument.

Since the required data for the model fitting by the `ppm` function is contained in the TQP, this table is first imported a data frame in `R` (by the command `data.frame`). Let `uX` and `uY` represent the columns of the TQP involving the values for $u_{i,X}$ and $u_{i,Y}$, respectively, and let `C1`, `C2`, ..., `C6` represent the columns of the TQP involving the values for $C_{c_{1a},\sigma,1}, C_{c_{1a},\sigma,2}, \ldots, C_{c_{1a},\sigma,6}$, respectively. For the classes c_{1a} = service, c_{1a} = alert and c_{1a} = environmental, the influence coefficients θ_σ can then be fitted as follows:

```
ppm(cbind(uX,uY), ~ (log(C1+1e-14)*polynom(C1,2) + log(C2+1e-14)*polynom(C2,2) +
log(C3+1e-14)*polynom(C3,2) + log(C4+1e-14)*polynom(C4,2) +
log(C5+1e-14)*polynom(C5,2) + log(C6+1e-14)*polynom(C6,2)),
Poisson(), covariates=cbind(C1,C2,C3,C4,C5,C6))
```

where the first argument is the data frame `cbind(uX,uY)`, the second argument is the log linear form of equation (39), the third argument is the command `Poisson()`, which indicates that there is no stochastic interaction (assumed) between the emergency calls of the level 1a class $c_{1a} \in \tilde{C}_{1a}$ of interest and the `covariates` argument is the data frame `cbind(C1,C2,C3,C4,C5,C6))`. For the class c_{1a} = accident, an eye must be kept on the Boolean covariate $C_{\sigma,25}$ involved. To implement this covariate in `R`, the command `factor()` must be used. The influence coefficients θ_σ for equation (40) can then be fitted as follows:

```
ppm(cbind(uX,uY), ~ (log(C1+1e-14)*polynom(C1,2) + log(C2+1e-14)*polynom(C2,2) +
factor(C3) + log(C4+1e-14)*polynom(C4,2) + log(C5+1e-14)*polynom(C5,2) +
log(C6+1e-14)*polynom(C6,2)), Poisson(), covariates=cbind(C1,C2,C3,C4,C5,C6))
```

In this way, the influence coefficients θ_σ can be estimated for all classes $c_{1a} \in \tilde{C}_{1a}$. But how could the quality of the fitted influence coefficients $\hat{\theta}_\sigma$, and so of the fitted spatial point process, be checked? This could be done in the context of the inhomogeneous Poisson process by again examining the empirical distribution function $\hat{K}_{\text{inhom}}(r)$, as explained in section 2, but there are even more adequate verification tools implemented in `R`. These tools involve the `diagnose.ppm` function, analogous to a residual plot for ordinary regression analysis, and the `qqplot.ppm` function, analogous to a Q-Q plot of the residuals in a linear model (Baddeley and Turner, 2006). In this thesis, only the `diagnose.ppm` function will be used to examine the quality of the fits.

It can be concluded that only the fits for the classes c_{1a} = environmental and c_{1a} = accident are reasonable, but that the fits for the classes c_{1a} = service and c_{1a} = alert are bad. It can also be concluded that the fit for the class c_{1a} = service does not differ significantly for implementing the five most important covariates or the six most important ones. To reduce the complexity of the fit, only the five most important covariates $C_{\sigma,k}, 1 \leq k \leq 5$ are therefore taken into account for the corresponding model. For the same reason, only the four most important covariates $C_{\sigma,k}, 1 \leq k \leq 4$ are taken into account for the model of the class c_{1a} = environmental. The reason that some covariates may not exert a great influence anymore is that it may have a high correlation with the already modelled covariates and so, for the reasons explained in section 4, this covariate may not have a significant influence anymore. This can be clearly seen for the class c_{1a} = environmental, where $C_{\sigma,5}$ and $C_{\sigma,6}$ have a high correlation with $C_{\sigma,4}$.

Let p denote the number of most important covariates modelled for a specific level 1a class, then for c_{1a} = service, $p = 5$, for c_{1a} = environmental, $p = 4$ and for c_{1a} = alert and c_{1a} = accident, $p = 6$. The fitted influence coefficients $\hat{\theta}_\sigma$ corresponding to these classes c_{1a} = service, c_{1a} = alert, c_{1a} = environmental and c_{1a} = accident are given in the tables 12, 13, 14 and 15, respectively, and the corresponding diagnostic plots are given in figure 11.

	$C_{\sigma,1}$	$C_{\sigma,2}$	$C_{\sigma,23}$	$C_{\sigma,22}$	$C_{\sigma,9}$
$\theta_{k,0}$	-14.02717				
$\theta_{k,1}$	0.02548	0.00495	0.00072	0.00801	0.00390
$\theta_{k,2}$	0.08491	-0.05457	0.02006	-0.06203	0.00199
$\theta_{k,3}$	0.00013	-0.00856	0.00001	-0.00015	-0.00006
$\theta_{k,4}$	-0.01642	0.04944	-0.00307	0.01330	0.00018
$\theta_{k,5}$	-0.00001	0.00149	0.00000	0.00002	0.00001

Table 12: The influence coefficients $\hat{\theta}_\sigma$ for c_{1a} = service, fitted by the `ppm` function in `R`.

	$C_{\sigma,19}$	$C_{\sigma,9}$	$C_{\sigma,1}$	$C_{\sigma,2}$	$C_{\sigma,23}$	$C_{\sigma,5}$
$\theta_{k,0}$	-14.49913					
$\theta_{k,1}$	-0.00608	0.00134	0.05336	0.02037	0.01296	0.02077
$\theta_{k,2}$	-0.08477	-0.05795	0.10299	-0.27645	0.03224	0.10277
$\theta_{k,3}$	-0.00040	-0.00050	0.00018	-0.02343	0.00002	0.00452
$\theta_{k,4}$	0.02076	0.01649	-0.02032	0.16863	-0.00517	-0.04200
$\theta_{k,5}$	0.00005	0.00007	-0.00002	0.00401	0.00000	-0.00076

Table 13: The influence coefficients $\hat{\theta}_\sigma$ for c_{1a} = alert, fitted by the `ppm` function in `R`.

	$C_{\sigma,1}$	$C_{\sigma,22}$	$C_{\sigma,23}$	$C_{\sigma,21}$
$\theta_{k,0}$	-13.98994			
$\theta_{k,1}$	0.03744	0.01035	-0.01191	-0.00112
$\theta_{k,2}$	0.04710	-0.03365	0.01947	0.01800
$\theta_{k,3}$	0.00005	-0.00012	0.00001	0.00008
$\theta_{k,4}$	-0.00847	0.00845	-0.00303	-0.00501
$\theta_{k,5}$	-0.00001	0.00001	0.00000	-0.00001

Table 14: The influence coefficients $\hat{\theta}_\sigma$ for $c_{1\mathrm{a}}$ = environmental, fitted by the `ppm` function in `R`.

	$C_{\sigma,23}$	$C_{\sigma,1}$	$C_{\sigma,25}$	$C_{\sigma,11}$	$C_{\sigma,22}$	$C_{\sigma,5}$
$\theta_{k,0}$	-13.30804					
$\theta_{k,1}$	0.00273	0.01430	0.36379	0.03113	0.01006	0.00307
$\theta_{k,2}$	0.01016	0.03424		-0.00256	-0.03645	0.28036
$\theta_{k,3}$	0.00001	0.00005		-0.00001	-0.00007	0.01611
$\theta_{k,4}$	-0.00171	-0.00655		0.00071	0.00732	-0.14048
$\theta_{k,5}$	0.00000	-0.00001		0.00000	0.00001	-0.00268

Table 15: The influence coefficients $\hat{\theta}_\sigma$ for $c_{1\mathrm{a}}$ = accident, fitted by the `ppm` function in `R`.

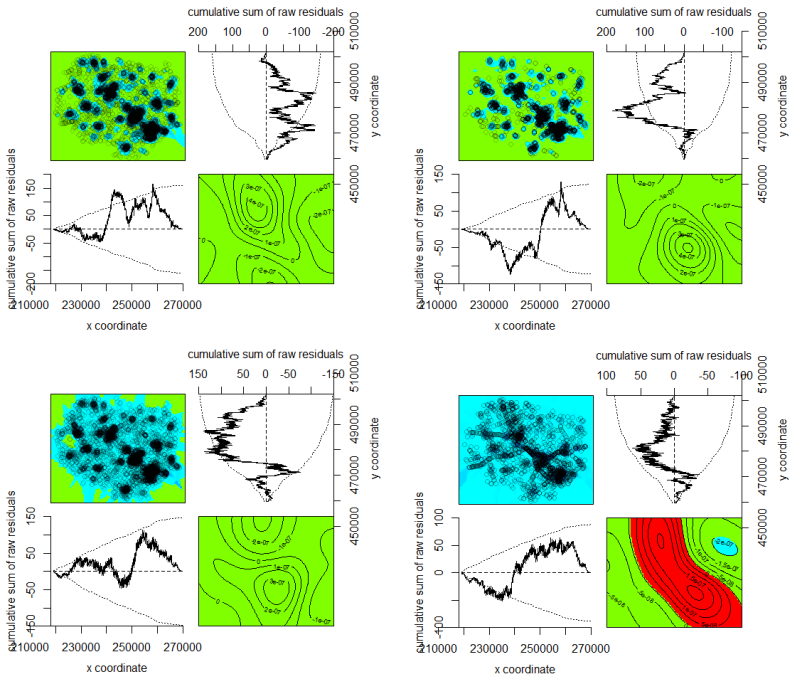

Figure 11: Diagnostic plots for the fitted influence coefficients $\hat{\theta}_\sigma$ for $c_{1\mathrm{a}}$ = service (figure a, left above), $c_{1\mathrm{a}}$ = alert (figure b, right above), $c_{1\mathrm{a}}$ = environmental (left below) and $c_{1\mathrm{a}}$ = accident (right below). These plots are made by the `diagnose.ppm` function.

Some remarks have to be made for these tables. For each of these tables, there is only one value of $\theta_{k,0}$ involved, since the output of the `ppm` function gives not the values of $\theta_{1,0}$, $\theta_{2,0}$, ..., $\theta_{p,0}$ individually, but it multiplicates these values and gives this single value as output. This value is listed as $\theta_{1,0}$ in each table. Next to that, there values 0.00000 may occur in the tables, which not directly implies that the corresponding coefficient has a value of zero, but it does have a value smaller than 0.000005. Further for the tables, the covariates in the upper row are ordered from the most important one (left) to less important ones (right). Concluding the discussion of the tables, empty cells indicate that the influence coefficient of interest is not defined.

Also some remarks have to be made for the diagnostic plots in figure 11, since it was not explained how to interpret these plots precisely. The subplot in the upper left part of the plot further gives an indication of the fitted intensity function (in blue) and a simulated point pattern according to it (in black). The subplot in the lower right part gives the residuals in the plain, which should be zero for a perfectly fitting model. If the residuals are high, though, the fitted model represents the data badly and another kind of model should be chosen to represent the data. This test is although quite quantitative and a more qualitative test can be executed by examining the lower left subplot and the upper right subplot.

The subplots in the lower left part and the upper right part further give the residuals of the lower right subplot for the X and Y coordinates, respectively. These two subplots can be interpreted in an analogous way as the plots for the distance analysis methods described in section 2 and therefore examining these subplots give a more qualitative test whether the fitted kind of model is the right one or not. In the mentioned subplots, the empirical distribution function (which is thus based on the fitted model) is denoted by the solid graph and the critical envelopes (which are thus based on simulations of the theoretical model that was fitted to the data, so in this case, an inhomogeneous Poisson process) are denoted by the dashed graphs. If the empirical distribution function lies outside of the critical envelopes, the fitted model is rejected as the right model, and otherwise it is accepted.

By examining the diagnose plots in figure 11, it can indeed be seen what was earlier mentioned, namely that an inhomogeneous Poisson process is surely not the right model for the classes $c_{1\text{a}}$ = service and $c_{1\text{a}}$ = alert and that this model is reasonable for the classes $c_{1\text{a}}$ = environmental and $c_{1\text{a}}$ = accident, but not very suitable. Despite, the models will be compared to the emergency call data of the corresponding classes occurred in period T_v (so the period from 1 January 2016 till 7 December 2016) later in this section, to make sure that it does not give accurate predictions.

Now the spatial point processes are fitted, the spatio-temporal point process for the class $c_{1\text{a}}$ = fire will be fitted. The TQP must slightly be extended in this case, since it should also possess a column which describes the day of each quadrature point $u_i \in Q_{\text{fire}}$. For Q_{fire}, 1300 dummy points are generated rather than 6291, to reduce the complexity of the model fitting.

To generate these 1300 dummy points, 100 points are evenly distributed over the region A_{ext}, which is defined as the region in the square in the range [218000,271000] in the X direction and in the range [459000,502000] in the Y direction (a square which completely includes the region $\tilde{A}$). The reason why the points were not generated in $\tilde{A}$ itself is that the algorithm (by prof. dr. M.N.M. van Lieshout and Adina Iftimi) used requires a square as spatial region for generating the dummy points. For each of these 100 points, 13 points are generated in time period T_m in an regular way, so on days 168, 505, 842, and so on till day 4212. The corresponding grid in $A_{\text{ext}} \times T$ with the dummy points as centres will be denoted by $P_{(10,10,13)}$.

After this, the values of the six most important covariates can be assigned to each quadrature point $u_i \in Q_{\text{fire}}$. Assigning the values for the spatial covariates is then based on the X and Y coordinates of the quadrature points u_i, denoted by $u_{i,X}$ and $u_{i,Y}$, respectively. In an analogous way, assigning the values for the temporal covariates is based on the t coordinates (so on the day d of interest) of the quadrature points u_i, denoted by $u_{i,t}$. The values for the spatial covariates can again be assigned by *QGIS* and the values for the temporal covariates by *Microsoft Excel*.

Now, all the information is collected to build the extended TQP, denoted by ETQP. The ETQP involves three columns which describe the values of $u_{i,X}$, $u_{i,Y}$ and $u_{i,t}$ for each value of $i, 1 \leq i \leq Q_{\text{fire}}$. The ETQP also involves six columns describing the six most influent covariates $C_{\tau,1}$, $C_{\sigma,k}, 2 \leq k \leq 6$ for the class $c_{1\text{a}} = \text{fire}$ concluded from section 4. With this ETQP for the emergency calls of class $c_{1\text{a}} = \text{fire}$, a spatio-temporal point process can be made for this class.

To fit the spatio-temporal point process, the ETQP is imported in `R` as a data frame. Let `uX`, `uY` and `ut` represent the columns of of the TQP involving the values of $u_{i,X}$, $u_{i,Y}$ and $u_{i,t}$, respectively, and let $\texttt{C1}, \texttt{C2}, \ldots, \texttt{C6}$ represent the columns of the ETQP involving the values for $C_{\tau,1}, C_{\sigma,2}, \ldots, C_{\sigma,6}$, respectively. For fitting the spatio-temporal inhomogeneous Poisson process, the `ppm` function from the `spatstat` package cannot be used now anymore, since this function is only designed for modelling spatial point patterns to the data. So a new algorithm should be made, implementing the method for maximising the pseudolikelihood for a spatio-temporal inhomogeneous Poisson process[25].

As mentioned in the discussion about the maximum pseudolikelihood estimating technique for spatio-temporal inhomogeneous Poisson processes, a generalized linear model can be used to (approximately) maximise the expression in equation (52) (and so the log likelihood), since this expression was similar to the log likelihood of multiple independent Poisson variables Y_i with weights w_i, means $\lambda_{\theta_\sigma}(u_i)$ and responses y_i. So first w_i and y_i will be calculated from the data in the ETQP. The weights can be calculated by $w_i = 7.68 \cdot 10^{11}/n_i$, where $7.68 \cdot 10^{11}$ represents the area of a cell of the earlier mentioned grid $P_{(10,10,13)}$ in the unit squared meter times days and n_i represents the number of quadrature points in cell i of this grid, where $1 \leq i \leq 1300$. The values for y_i can then be calculated by $y_i = z_i/w_i$, where z_i is a Boolean variable equaling 1 if u_i is a data point (event) and 0 if u_i is a dummy point.

After calculating the values for w_i and y_i, they are also added to the ETQP in the columns `w` and `yy`, respectively. With this completion, the ETQP can be used to estimate the values for the influence coefficients $\theta = (\theta_\sigma, \theta_\tau)$ corresponding to the (global) intensity function for the spatio-temporal point process for $c_{1\text{a}} = \text{fire}$ found in section 4. In other words, it can be used to fit an spatio-temporal inhomogeneous Poisson process to the data of class $c_{1\text{a}} = \text{fire}$. For completing the ETQP in `R` with the dummy points and the values for w_i and y_i, the `pseudolikelihood` function of the mentioned script made by prof. dr. M.N.M. van Lieshout and Adina Iftimi is used.

Now, the estimates $\hat{\theta} = (\hat{\theta}_\sigma, \hat{\theta}_\tau)$ for the influence coefficients $\theta = (\theta_\sigma, \theta_\tau)$ can be find by the `glm` function in `R`, which fits a generalized linear model, of course in this case with (quasi)poisson errors:

[25] Remark that this new algorithm is not the spatio-temporal version of the whole `ppm` function, but only of the part of the `ppm` function that fits Poisson spatial point processes, since the earlier discussion about the maximum pseudolikelihood estimating technique was only valid for Poisson point processes.

```
glm(formula = yy ~ (factor(C1) + log(C2+1e-14)*polynom(C2,2) +
log(C3+1e-14)*polynom(C3,2) + log(C4+1e-14)*polynom(C4,2) +
log(C5+1e-14)*polynom(C5,2) + log(C6+1e-14)*polynom(C6,2)),
family = "quasipoisson", weights = w, data = Q, start = c(0,0,0,0,0,0))
```

Since `R` was not able to calculate the influence coefficients θ for the (global) intensity function implemented in the above function, so for the product of the equations (46) and (47), the fitted (global) intensity function is simplified to the following one:

$$\lambda_\theta = e^{\theta_0+\theta_1 C_1(\mathbf{x})+\theta_2 C_2(\mathbf{x})+\theta_3 C_3(\mathbf{x})+\theta_4 C_4(\mathbf{x})+\theta_5 C_5(\mathbf{x})+\theta_6 C_6(\mathbf{x})} \tag{53}$$

which can be fitted as follows:

```
glm(formula = yy ~ (factor(C1)+C2+C3+C4+C5), family = "quasipoisson",
weights = w, data = Q, start = c(0,0,0,0,0,0))
```

The reason why the relation of equation (53) can be fitted, but the relation consisting of equations (46) and (47) cannot, is probably the complexity of the algorithm and the huge amount of data involved. It would be needed to reduce the complexity of the algorithm or, if the complexity of the algorithm is not too high, to use a more powerful computer, to find the influence coefficients θ for the (global) intensity function consisting of the product of the equations (46) and (47). For the simplified fit involving the relation of equation (53), the influence coefficients are given in table 16.

	intercept	$C_{\tau,8}$	$C_{\sigma,1}$	$C_{\sigma,22}$	$C_{\sigma,17}$	$C_{\sigma,23}$	$C_{\sigma,14}$
θ_k	-21.39236	-0.05956	-0.00125	0.00015	-0.00062	0.00120	0.00092

Table 16: The influence coefficients θ for the class c_{1a} = fire, fitted by the `glm` function in `R`.

The quality of this fit cannot be tested by the `diagnose.ppm` function, though, since this function is also only suited for (purely) spatial point processes. Instead, the quality of the newly made spatio-temporal point process will be tested by making predictions according to this model and compare these to the (real) emergency call data of the class c_{1a} = fire occurred in period T_v (the verification period, involving 1 January 2016 till 7 December 2016). This kind of verification will, as earlier announced, also be done for the models representing the other level 1a classes, $c_{1a} \in \tilde{C}_{1a}$.

Before simulating the point process models earlier made can be done, the intensity function of each model has to be evaluated in every 500 meter square of the spatial grid P_σ for Twente. For this evaluation, the value for each covariate in 2016 is assumed the same as the corresponding value in 2015. Of course, each value for the intensity function is then multiplied by the area of each square, so 25000 squared meters, as a corollary of equation (1). Further, to produce a model for the period T_v, consisting of 341 days, the values are also multiplied by $\frac{341}{4380}$, since the intensity function is made for 4380 days.

After this, all values for each level 1a class $c_{1a} \in C_{1a}$ are summarized in a 101×85 matrix $I_{c_{1a}}$, which represents the region A_{ext} discretized in 500 meter squares, where each square corresponds to a cell in $I_{c_{1a}}$. If a cell is inside the region A_{ext}, but outside the region $\tilde{A}$, the value for the intensity function is zero. Determining and summarizing the values for the intensity function is done by *QGIS* and *Microsoft Excel*.

After this is done, the matrix $I_{c_{1a}}$ for each level 1a class $c_{1a} \in C_{1a}$ is imported as an image in `R` (by the command `im`). Simulating a spatial point process for each level 1a class $c_{1a} \in C_{1a}$ in period T_v can then be done by applying the `rpoispp` function to the corresponding (image of) matrix $I_{c_{1a}}$.

So now it is explained how spatial point patterns can be simulated for the verification analysis according the found intensity functions. But next to predicting the number of emergency calls for the period T_v according to the made point processes, this number will also be predicted according to the current industry practice model, mentioned in section 1 and based on the discussion in Zhou et al. (2015). This is in this thesis done by taking the average number of occurred events in each year 2004, 2005, ..., 2015 in T_m for each level 1a class, although there are of course many other ways possible for applying this model.

Let n_v be the real number of occurred emergency calls of the level 1a class of interest $c_{1a} \in C_{1a}$ in period T_v, $\tilde{n}_v(k)$ be the predicted number of emergency calls for period T_v according to model $k, k = 1, 2$, $E_v(k)$ be the error corresponding to $\tilde{n}_v(k)$ and further let model 1 represent the fitted inhomogeneous Poisson processes and let model 2 represent the current industry practice with the average taken over all the years in period T_m. Then the results of the model verification analysis are shown in table 17.

	n_v	$\tilde{n}_v$ model 1	E_v model 1	$\tilde{n}_v$ model 2	E_v model 2
fire	1187	490	-697	1424	+237
service	802	483	-319	688	-114
accident	146	161	+15	191	+45
alert	1878	279	-1599	2639	+761
environmental	452	401	-51	577	+125

Table 17: The predictions of the fitted inhomogeneous Poisson processes and the current industry practice for period T_v for each level 1a class $c_{1a} \in C_{1a}$.

It can thus be seen that the inhomogeneous Poisson process models representing the level 1a classes c_{1a} = fire, c_{1a} = service and c_{1a} = alert give bad predictions, and although the current industry practice model applied to the earlier mentioned classes gives better predictions, these predictions are still bad. The result that the inhomogeneous Poisson process models for c_{1a} = service and c_{1a} = alert fit bad agrees with the corresponding diagnosis plots earlier made. It can thus be concluded that another model should be fitted to the data, which is able to represent these classes more accurately.

The inhomogeneous Poisson process models for c_{1a} = accident and c_{1a} = environmental fit quite well, though. This was also earlier concluded from the corresponding diagnosis plots. These models fit even better than the current industry practice model applied to these level 1a classes. Still some improvements have to be made for the model representing the class c_{1a} = environmental, since the prediction has a deviation of 50 emergency calls from the occurred number of events of this class in period T_v.

6 Conclusion and discussion

In this thesis, a spatio-temporal point process was only fitted to the class c_{1a} = fire, since for the other classes $c_{1a} \in \tilde{C}_{1a}$, $\tilde{C}_{1a} = \{\text{service}, \text{accident}, \text{alert}, \text{environmental}\}$, temporal influences seemed not present and thus a (purely) spatial point process was modelled. For all these point processes, an inhomogeneous Poisson process was proposed as model, since trend seemed the only kind of cause for the spatial point patterns of interest and an inhomogeneous Poisson process is the easiest kind of model to fit data involving trend as cause. Fitting the inhomogeneous Poisson processes for the classes $c_{1a} \in \tilde{C}_{1a}$ and c_{1a} = fire was respectively done by the `ppm` function (from the `spatstat` package) and the `glm` function in `R`

As a consequence of choosing inhomogeneous Poisson processes as the models of interest, regularity could not be modelled anymore. This was not that problematic, since the exploratory data analyses in section 3 clearly indicated an aggregated distribution for the emergency calls of each level 1a class in the region $\tilde{A}$ (defined as in section 4) and in the time period T_m (defined as in section 3). It was more problematic that an inhomogeneous Poisson process has not the ability to model eventual stochastic interaction between the events of the involved spatial point patterns. But the reward for this shortcoming was that the inhomogeneous Poisson process is tractable and quite easily understandable.

In section 3, it was already indicated that the data for neither class fitted an inhomogeneous Poisson process directly. Therefore covariates were introduced. The inhomogeneous Poisson processes for the classes c_{1a} = accident and c_{1a} = environmental fit quite well with the respectively six and four most important covariates, which were found in the covariate analysis in section 4. But the inhomogeneous Poisson processes with the most important covariates involved for the classes c_{1a} = fire, c_{1a} = service and c_{1a} = alert did not predict the future emergency calls well. These conclusions were made based on the diagnosis plots for the classes $c_{1a} \in \tilde{C}_{1a}$ and on the verification analysis for all level 1a classes.

The bad fitting spatio-temporal point process model for the class c_{1a} = fire is very likely to be caused by the implementation of the wrong (global) intensity function in the fitting process, which involved the intensity function of equation (53). The reason for fitting this intensity function instead of the intensity function concluded from the regression analysis in section 4 (so the product of equations (46) and (47)), is that the initial concluded intensity function could not be fitted by `R`. This is probably caused by the large amount of data for this intensity function in combination with the complexity of this intensity function. Therefore the algorithms in `R` may have problems in calculating the maximum pseudolikelihood estimate for this fit and therefore the influence coefficients may not be estimated well. So a more efficient algorithm would be desired to model the intensity function initially concluded from the regression analysis.

Further, the bad fitting model for the classes c_{1a} = service and c_{1a} = alert (and the model for the class c_{1a} = fire may also cope with this problem) is probably caused by the restriction to only modelling trend and so not taking into account eventual stochastic interaction between the emergency calls of these classes. So extending the current model to a Cox process seems a good proposition for a following research, since in that way both trend and stochastic interaction can be modelled. The reason why a Strauss model seems not the best option is that trend also seems to have significant influence, as one could see in section 4. Of course, modelling interaction may also be an improvement to the models representing the classes c_{1a} = accident and c_{1a} = environmental.

It may also be the case that the bad fits for the classes c_{1a} = service and c_{1a} = alert are caused by the fact that the models were univariate, or in other words that the models did not take into account eventual dependencies between them and other classes. It may be the case that a place with less emergency calls of the level 1a class service also has less emergency calls of the level 1a class alert, for example. Of course, this extension may also be a good one for the classes c_{1a} = fire, c_{1a} = accident and c_{1a} = environmental. So extending the model to a mulitvariate spatio-temporal point process involving all the level 1a classes would also be a good proposition for a following research.

Further, the bad fits for the classes c_{1a} = service and c_{1a} = alert may be caused by an erroneous implementation of the temporal behaviour (and again the quality of the fitted models for the classes c_{1a} = fire, c_{1a} = accident and c_{1a} = environmental may also suffer from this aspect). It may be expected from an initial inspection of the corresponding time series that the behaviour of the occurrences of emergency calls differs per year and the models made in this thesis did not take these differences into account thoroughly. So it would be a good extension to the models made to take a closer look at the behaviour of the occurrences of emergency calls in time, which can be done by modelling the temporal aspect by a Fourier series. It may also be a good extension to choose the period on which the model is based, so T_m, to represent a shorter period, so for example 2010 till 2015. In this way, the models are based on more recent data.

Nonetheless, this thesis did take into account the temporal behaviour by the temporal covariates, but these seemed to have little influence in general, since only the model representing the class c_{1a} = fire involved a temporal covariate. This is also a bit curious, since in practice time appears to have influence. It would also be a good extension to this research to take a closer look on these temporal covariates and maybe the covariate analysis in general, since there may be some methods involved in it which cause the results to be erroneous.

There will be one other extension proposed to the ensemble of models made, since the fire departments in Twente also desire to know the precise causes of level 2a and level 3a emergency calls. And although not all the level 1a classes are well modelled in this thesis, it is not a bad idea to first try the modelling for the level 2a and level 3a emergency calls, since they involve less data, which reduces the complexity of the modelling and causes the algorithms to diverge not as quickly as for the emergency calls of the level 1a classes.

Concluding, this thesis did only find quite reasonable models for the classes c_{1a} = accident and c_{1a} = environmental. It did although explain the basic theory of making a spatio-temporal point process, it made the emergency call data amenable for analysis, it analysed the (global) spatial and temporal properties of the emergency call data, it examined the erroneous data and the discarded data for the "leap days" (as defined in section 3), it introduced a reasonable amount of covariates which could be examined for their influences on the emergency calls of each involved level 1a class, it examined the relations between these covariates and the emergency call data for each level 1a class, it modelled the spatial and spatio-temporal inhomogeneous Poisson processes which were proposed for the classes $c_{1a} \in \tilde{C}_{1a}$ and c_{1a} = fire, respectively, and finally it analysed the quality of these models. And although the fits did not appear to be that good for the most level 1a classes, this thesis gave much information for a follow-up research of these data and gives a good introduction to the modelling of spatio-temporal point processes.

Acknowledgements

My most sincere thanks go to my supervisor, prof. dr. M.N.M. van Lieshout, for the many help and good advice she provided me during this research. I also want to thank her for introducing me to the field of spatio-temporal point process modelling. I also wish to thank Emiel Sanders from the policy and strategics team of the head fire department of Twente for all the help and practical information he gave me about the functioning of the firemen in Twente, but also for the hospitality to execute the research partially in the head fire department of Twente.

Bibliography

Adrian Baddeley and Rolf Turner. Practical maximum pseudolikelihood for spatial point patterns. *Australian & New Zealand Journal of Statistics*, 42(3):283–322, 2000.

Adrian Baddeley and Rolf Turner. Modelling spatial point patterns in r. In *Case studies in spatial point process modeling*, pages 23–74. Springer, 2006.

AJ Baddeley and BW Silverman. A cautionary example on the use of second-order methods for analyzing point patterns. *Biometrics*, pages 1089–1093, 1984.

MS Bartlett. The spectral analysis of two-dimensional point processes. *Biometrika*, 51(3/4): 299–311, 1964.

Mark Berman and T Rolf Turner. Approximating point process likelihoods with glim. *Applied Statistics*, pages 31–38, 1992.

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

Peter J Diggle. Statistical analysis of spatial point processes. *Academic, London*, 1983.

Edith Gabriel and Peter J Diggle. Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica*, 63(1):43–51, 2009.

MNM van Lieshout and AJ Baddeley. A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica*, 50(3):344–361, 1996.

Jesper Møller and Mohammad Ghorbani. Aspects of second-order analysis of structured inhomogeneous spatio-temporal point processes. *Statistica Neerlandica*, 66(4):472–491, 2012.

Jesper Møller and Rasmus Plenge Waagepetersen. *Statistical inference and simulation for spatial point processes*. CRC Press, 2003.

Rolf Turner. Point patterns of forest fire locations. *Environmental and ecological statistics*, 16 (2):197–223, 2009.

MNM Van Lieshout. *Markov point processes and their applications*. World Scientific, 2000.

Zhengyi Zhou, David S Matteson, Dawn B Woodard, Shane G Henderson, and Athanasios C Micheas. A spatio-temporal point process model for ambulance demand. *Journal of the American Statistical Association*, 110(509):6–15, 2015.

A Results exploratory data analysis for c_{1a} = service

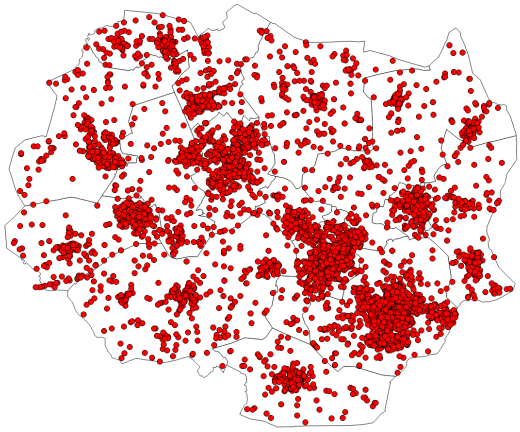

Figure 12: Spatial point pattern for the class c_{1a} = service in the period T_m.

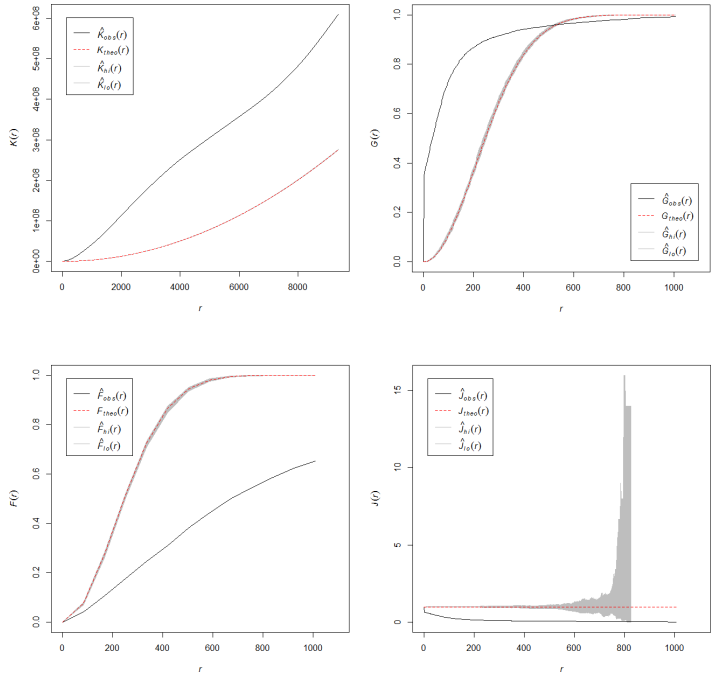

Figure 13: Homogeneous distance analyses with the estimated functions $\hat{K}(r)$ (figure a, left above), $\hat{G}(r)$ (figure b, right above), $\hat{F}(r)$ (figure c, left below) and $\hat{J}(r)$ (figure d, right below) applied on the data of the class c_{1a} = service, plotted against the corresponding theoretical functions and critical envelopes. The plots are provided by the package `spatstat` in `R`. The $\hat{G}(r)$, $\hat{F}(r)$ and $\hat{J}(r)$ functions are defined till $r = 1000$, approximately.

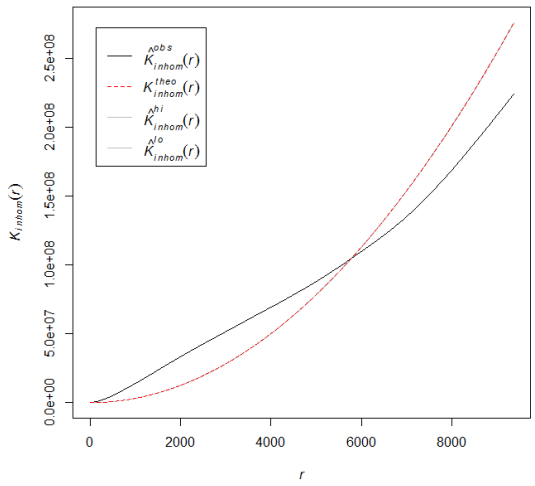

Figure 14: Inhomogeneous distance analysis for the estimated function $\hat{K}_{\text{inhom}}(r)$ applied on the data of the class $c_{1\text{a}}$ = service, plotted against the corresponding theoretical functions and critical envelopes. The plot is provided by the package `spatstat` in `R`.

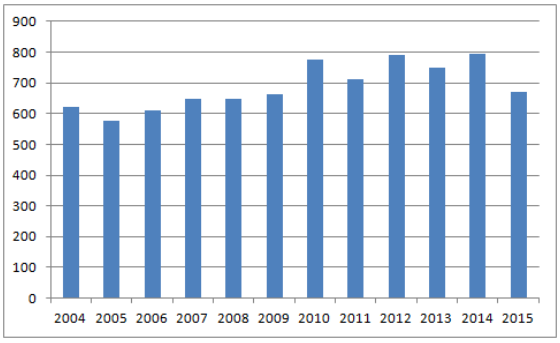

Figure 15: Time series for the amount of emergency calls of the class $c_{1\text{a}}$ = service per year $U_i \subset T_m$.

y	2004	2005	2006	2007	2008	2009
t	-2.120	-4.100	-2.920	-1.039	-1.518	-1.135
y	2010	2011	2012	2013	2014	2015
t	2.643	0.971	3.817	2.026	3.919	-0.666

Table 18: The values of the Student's t-statistics for the temporal tests for class $c_{1\text{a}}$ = service, which compares each year with the other years. The corresponding 95%-confidence interval for each value of y is $[-1.975, 1.975]$.

B Results exploratory data analysis for c_{1a} = accident

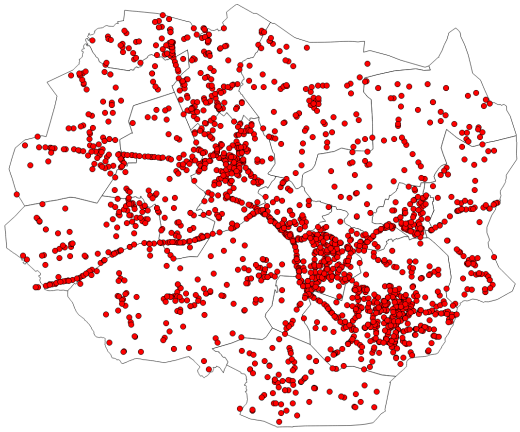

Figure 16: Spatial point pattern for the class c_{1a} = accident in the period T_m.

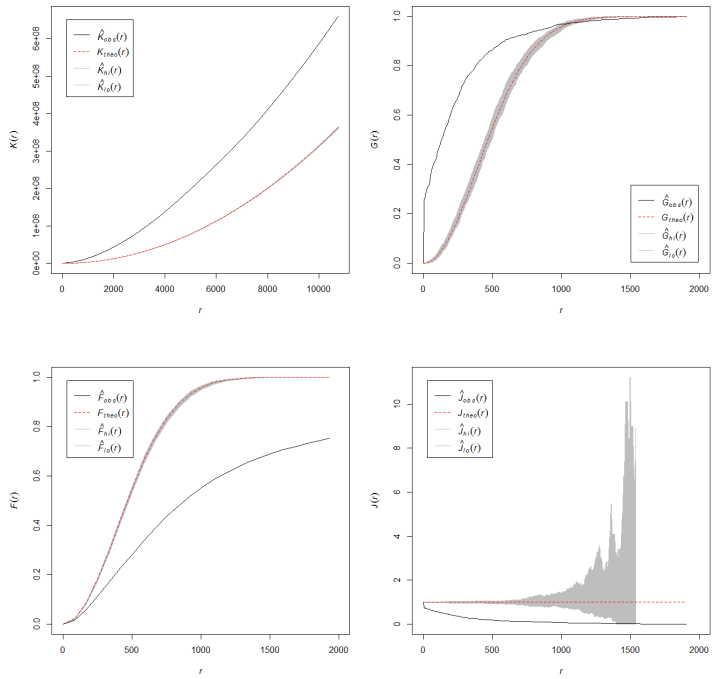

Figure 17: Homogeneous distance analyses with the estimated functions $\hat{K}(r)$ (figure a, left above), $\hat{G}(r)$ (figure b, right above), $\hat{F}(r)$ (figure c, left below) and $\hat{J}(r)$ (figure d, right below) applied on the data of the class c_{1a} = accident, plotted against the corresponding theoretical functions and critical envelopes. The plots are provided by the package `spatstat` in `R`. The $\hat{G}(r)$, $\hat{F}(r)$ and $\hat{J}(r)$ functions are defined till $r = 1900$, approximately.

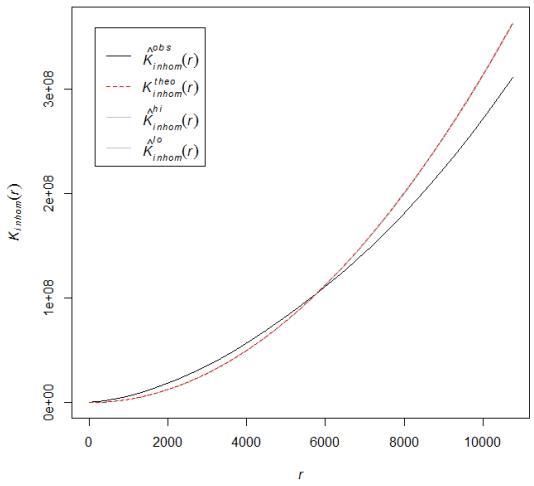

Figure 18: Inhomogeneous distance analysis for the estimated function $\hat{K}_{\text{inhom}}(r)$ applied on the data of the class $c_{1\text{a}}$ = accident, plotted against the corresponding theoretical functions and critical envelopes. The plot is provided by the package `spatstat` in `R`.

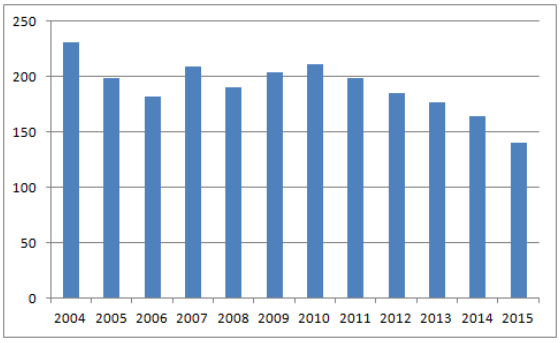

Figure 19: Time series for the amount of emergency calls of the class $c_{1\text{a}}$ = accident per year $U_i \subset T_m$.

y	2004	2005	2006	2007	2008	2009
t	2.694	0.578	-0.738	1.279	-0.072	1.068
y	2010	2011	2012	2013	2014	2015
t	1.354	0.631	-0.474	-1.146	-2.331	-4.683

Table 19: The values of the Student's t-statistics for the temporal tests for class $c_{1\text{a}}$ = accident, which compares each year with the other years. The corresponding 95%-confidence interval for each value of y is $[-1.975, 1.975]$.

C Results exploratory data analysis for c_{1a} = alert

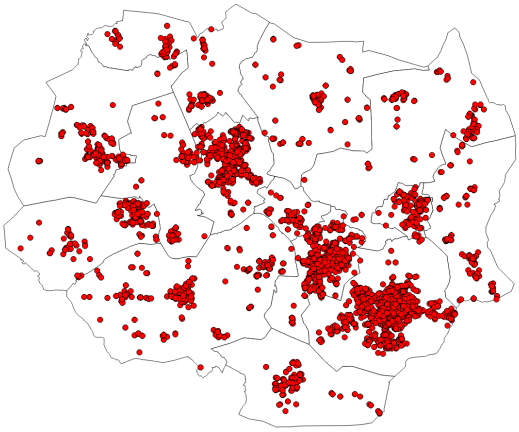

Figure 20: Spatial point pattern for the class c_{1a} = alert in the period T_m.

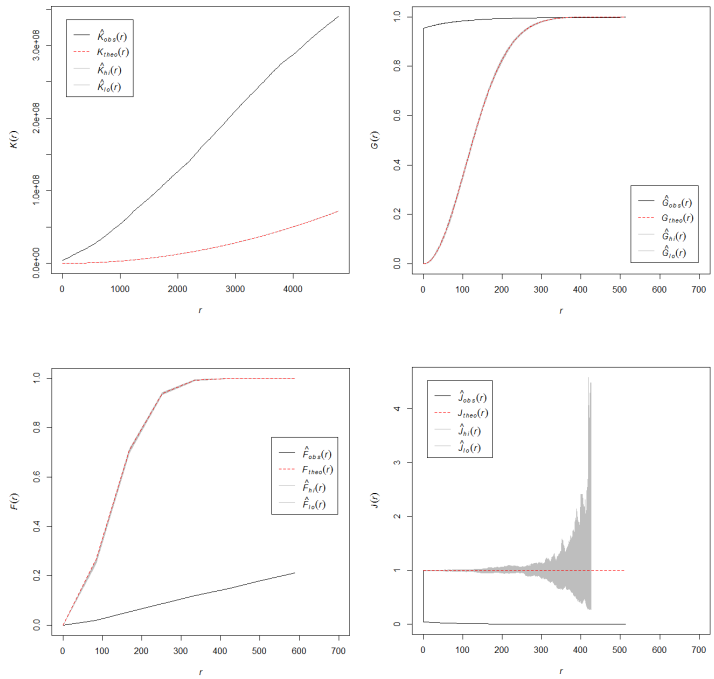

Figure 21: Homogeneous distance analyses with the estimated functions $\hat{K}(r)$ (figure a, left above), $\hat{G}(r)$ (figure b, right above), $\hat{F}(r)$ (figure c, left below) and $\hat{J}(r)$ (figure d, right below) applied on the data of the class c_{1a} = alert, plotted against the corresponding theoretical functions and critical envelopes. The plots are provided by the package `spatstat` in `R`. The $\hat{G}(r)$, $\hat{F}(r)$ and $\hat{J}(r)$ functions are defined till $r = 500$, approximately.

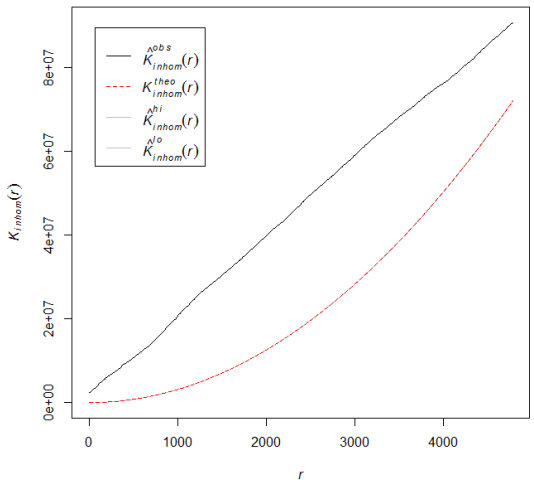

Figure 22: Inhomogeneous distance analysis for the estimated function $\hat{K}_{\text{inhom}}(r)$ applied on the data of the class c_{1a} = alert, plotted against the corresponding theoretical functions and critical envelopes. The plot is provided by the package `spatstat` in `R`.

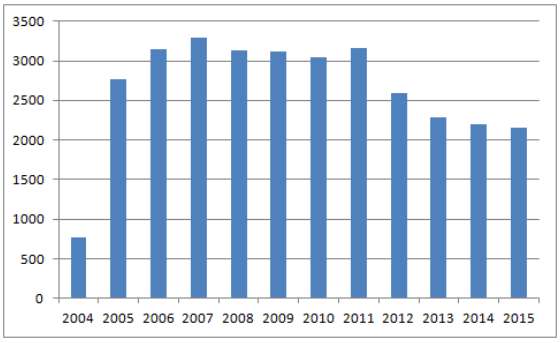

Figure 23: Time series for the amount of emergency calls of the class c_{1a} = alert per year $U_i \subset T_m$.

y	2004	2005	2006	2007	2008	2009
t	-64.371	2.005	8.327	11.040	8.154	7.536
y	2010	2011	2012	2013	2014	2015
t	7.328	7.545	-0.776	-7.470	-8.892	-9.126

Table 20: The values of the Student's t-statistics for the temporal tests for class c_{1a} = alert, which compares each year with the other years. The corresponding 95%-confidence interval for each value of y is $[-1.975, 1.975]$.

D Results exploratory data analysis for c_{1a} = environmental

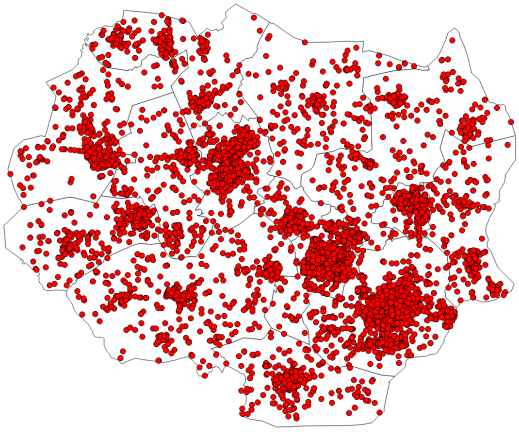

Figure 24: Spatial point pattern for the class c_{1a} = environmental in the period T_m.

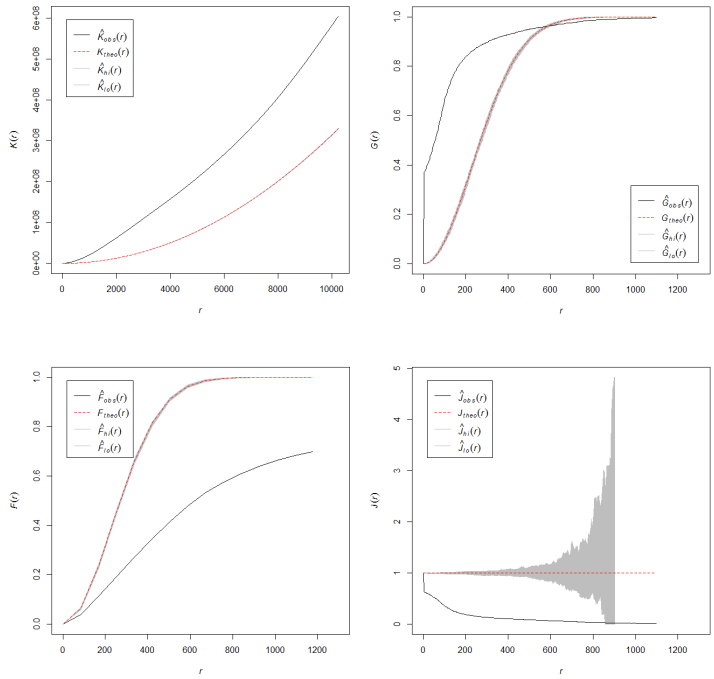

Figure 25: Homogeneous distance analyses with the estimated functions $\hat{K}(r)$ (figure a, left above), $\hat{G}(r)$ (figure b, right above), $\hat{F}(r)$ (figure c, left below) and $\hat{J}(r)$ (figure d, right below) applied on the data of the class c_{1a} = environmental, plotted against the corresponding theoretical functions and critical envelopes. The plots are provided by the package `spatstat` in `R`. The $\hat{G}(r)$, $\hat{F}(r)$ and $\hat{J}(r)$ functions are defined till $r = 1100$, approximately.

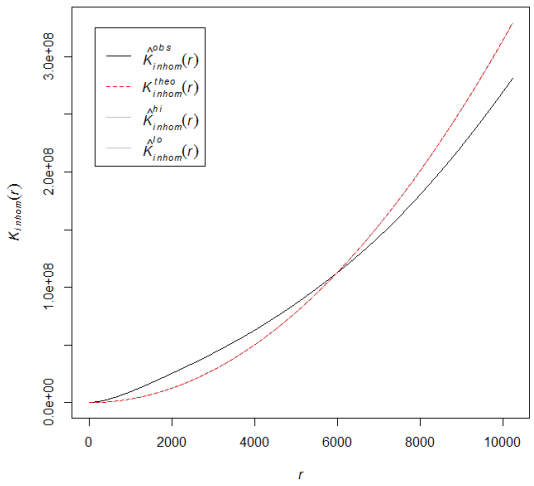

Figure 26: Inhomogeneous distance analysis for the estimated function $\hat{K}_{\text{inhom}}(r)$ applied on the data of the class $c_{1\text{a}}$ = environmental, plotted against the corresponding theoretical functions and critical envelopes. The plot is provided by the package `spatstat` in `R`.

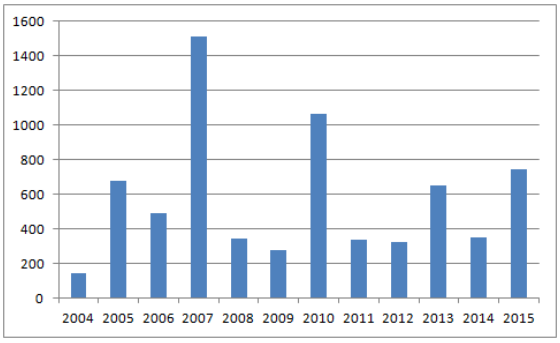

Figure 27: Time series for the amount of emergency calls of the class $c_{1\text{a}}$ = environmental per year $U_i \subset T_m$.

y	2004	2005	2006	2007	2008	2009
t	-11.762	0.366	-0.537	1.135	-2.406	-7.946
y	2010	2011	2012	2013	2014	2015
t	1.0155	-5.515	-5.356	0.306	-3.363	0.705

Table 21: The values of the Student's t-statistics for the temporal tests for class $c_{1\text{a}}$ = environmental, which compares each year with the other years. The corresponding 95%-confidence interval for each value of y is $[-1.975, 1.975]$.