Determining high-risk zones using point process methodology

Realization by building an R package

BACHELOR THESIS

Heidi Seibold

July 11, 2012

Department of Statistics, LMU

Supervision: Prof. Dr. Helmut Küchenhoff

Dipl.-Stat. Monia Mahling

Abstract

The determination of high-risk zones is an important part in finding unexploded bombs that have been lying under the ground in German properties since the Second World War. This thesis gives an overview on possible methods to determine high-risk zones: on the one hand, methods that draw a circle around each bomb crater with a certain radius, which can be fixed or data-driven, to determine the high-risk zone. Those methods are together classified as distance-based methods. On the other hand, there is a method that is based on the intensity of the crater point pattern.
Furthermore, the thesis describes the implementation of the different methods which is realized in the R package `highriskzone`. In oder to give an overview on the package, example usages are demonstrated.
Additionally, simulation studies were conducted in which different point patterns were simulated: homogeneous Poisson processes and two-dimensional normal mixture distributions. The study showed that the distance-based method and the intensity-based method deliver comparably good results.

Heidi Seibold

Contents

1 Motivation

It is now 67 years since the official end of the Second World War and still unexploded bombs from that time are an issue. For instance, during construction work yet duds are being found. To provide evacuations during construction works and to ensure the safety of the workers, it is neccesary to scan high risk areas for bombs before starting any other works. Since the search for unexploded bombs is expensive, it is reasonable to consider carefully where to search.
In the course of this thesis, different ways of computing high-risk zones for duds are being discussed and implemented in an R package, the package is illustrated in example analysis and a simulation study compares the two most relevant methods for specific point patterns.
Methods have been developed in order to determine high-risk zones for unexploded bombs. Nevertheless they are highly useful for other matters than bombs and so is the R package. In this thesis, the bombs topic will be used to facilitate understanding.

The project of determining high-risk zones for unexploded bombs from the Second World War was initiated by the *Oberfinanzdirektion Niedersachsen* (OFD) in order to remove duds in federal properties in Germany. This thesis is based on investigations on this field by Monia Mahling et al. [Mahling et al. 2013].

2 Methods for computing high-risk zones

The following sections describe three methods to determine a high-risk zone. The first method, the method of fixed radius, was formerly used at the OFD. The other two methods are the nowadays competing methods in calculating zones to be searched for unexploded bombs. The quantile-based method is connatural to the method of fixed radius, but more advanced. The third method is using the intensity for the computation. This method was developed in the past years by Monia Mahling and colleagues in order to achieve an improved approach. The upcoming explanations are only basics. For further information see Mahling et al. [2013].

2.1 Notation

Let X be the spatial point process, which in the case here is the location of all bombs and Y is a subset of X describing the observed process, i.e. the bomb crater of exploded bombs. The process of unexploded bombs (unobserved events) then is $Z = X \setminus Y$, meaning that Z and Y are disjoint and together forming X. Further W ist the observation window of X. A clear observation window is needed since there is no information about bombs outside the observed area.

2.2 Distance-based methods

The two following methods use the distance to the next event to determine the high-risk zone. The first assumes that unexploded bombs lie within a certain radius around the bomb crater. The second uses a quantile of the nearest-neighbour-distance to designate the zone.

2.2.1 Method of fixed radius

The method of fixed radius is a simple approach. Here, high-risk zones are appointed by drawing a circle around each bomb crater with a fixed radius r—the region within the circles defines the zone.

$$R_r = \{s \in W : \min_j ||s - y_j|| \leq r\} \tag{2.1}$$

where s is a location in the observation window W and y_j an event of the observed process.

2.2.2 Quantile-based method

This technique is a data-based development of the method of fixed radius. For each exploded bomb y_i in the point process Y, the distance to the nearest other exploded bomb in Y is calculated by the nearest-neighbour-distance

$$t_i = \min_{j \neq i} \|y_i - y_j\|. \tag{2.2}$$

From the empirical distribution function of the nearest-neighbour-distances of the point pattern

$$G(r) = \frac{1}{n_Y} \sum_i \mathbb{1}\{t_i \leq r\}, \tag{2.3}$$

the p-quantile $Q(p)$ can be calculated. The radius of the circles around each observed event is assesed the p-quantile, so the radius in this method depends on the distances between the observed events. The value of p is a real number between 0 and 1. Is p set near 1, there should be few unexploded bombs outside the searched zone, since the relative numer of unexploded bombs should be around $1 - p$.
On these grounds the high-risk zone R_r is defined as

$$R_r = \{s \in W : \min_j ||s - y_j|| \leq Q(p)\} \tag{2.4}$$

[Mahling et al. 2013].

Comparing the formulas of the high-risk zone of the two distance-based methods shows that the only difference between them, is the way the radius is specified.

2.3 Intensity-based method

The intensity-based method is built on spatial point process methods, more precisely on Poisson point process methodology.

2.3.1 Spatial point processes

Patterns of exploded bombs can be described by spatial point processes. According to Illian et al. [2008], point processes "are *stochastic models* of irregluar point patterns" in which a point pattern is an accumulation of points in a set or area. Usually point patterns are seen as samples or realisations of point processes [Illian et al. 2008].
In the following $N_X(A)$ will designate the number of bombs in a region $A \subseteq W$. The expected number of bombs in region A, also called intensity measure $E(N_X(A)) = \Lambda_X(A)$, is defined as

$$\Lambda_X(A) = \int_A \lambda_X(\mathbf{x}) d\mathbf{x} \tag{2.5}$$

with $\lambda_X(s)$ as the intensity function in location $s \in W$ which is proportional to the point density around the location [Mahling et al. 2013; Illian et al. 2008, p. 28].

Inhomogeneous Poisson process

The spatial point process X, the process of exploded bombs, is assumed to be an inhomogeneous Poisson point process. In contrary to a homogeneous Poisson process, an inhomogeneous Poisson process does not have a constant intensity $\lambda_X(s)$, but one that depends on the location. Other properties of the Poisson process are that the number of bombs in one area $A \subset W$ is Poisson distributed with the mean $\Lambda_X(A)$ and the number of bombs in two disjoint areas $A \subset W$ and $B \subset W$ are independent [Illian et al. 2008, p. 118].

Cluster process

It is difficult to distinguish an inhomogeneous Poisson process from a cluster process [Illian et al. 2008, p. 372]. We can not say with absolute certainty that the point pattern of exploded bombs is an inhomogenous Poisson process and not a cluster process. Thus the cluster process has to be mentioned here. Illian et al. [2008] define

clusters as "groups of points with an inter-point distance that is below the average distance in the pattern".
A special cluster process is the Neyman-Scott process, in which the parent points form a Poisson process. The parent points stand for the cluster centre and are not part of the actual process which is formed by the daughter points. The daughter points are scattered around the parent points. The Neyman-Scott process used in this work has cluster centres that form an inhomogeneous Poisson point process. The purpose of the Neyman-Scott process here will be explained in detail in Sections 3.2.2 and 3.2.3. For further informations on cluster processes see Illian et al. [2008, Chap. 6.3].

2.3.2 Estimation of the intensity function

The initial step of determining the high-risk zone by the intensity-based method is to estimate the intensity of exploded bombs $\lambda_Y(s)$. The estimation is accomplished by using a nonparametric method that works with a anisotropic Gaussian kernel $K_H(\cdot)$:

$$\hat{\lambda}_Y(s) = e(s) \cdot \sum_{i=1}^{n_Y} K_H(s - y_i). \tag{2.6}$$

where $e(s)$ is an edge effect bias correction

$$e(s) = \left(\int_W K_H(s - v) dv \right)^{-1} \tag{2.7}$$

which is needed because there is no information about bombs outside the observation window W and without it the bias at the margins of W would be negative. H in K_H stands for the covariance matrix of the kernel and is chosen by the cross-validation technique.

The probability q for a bomb not to explode is assumed to be homogeneous in the observation window [Mahling et al. 2013] . Therefore it is possible to calculate the estimator for the intensity of unexploded bombs directly from $\hat{\lambda}_Y(s)$:

$$\hat{\lambda}_Z(s) = \frac{q}{1-q} \cdot \hat{\lambda}_Y(s) \tag{2.8}$$

[Mahling et al. 2013]

2.3.3 Determination of the threshold c

Given the calculated intensity, the boundaries of the high-risk zone are not yet clear. A critical value $c > 0$, for which areas with intensity $\hat{\lambda}_Z$ larger than or equal to c should be searched, is to be set so that the high-risk zone is

$$R_c = \{s \in W : \hat{\lambda}_Z(s) \geq c\} \tag{2.9}$$

which means in words that the high-risk zone is all locations s of the observation window for which the intensity of Z is larger than or equal to c. The difficulty is the election of c. Goal in the determination of the high-risk zone is that the probability to have an unexploded bomb outside this zone $P\{N_z(W \setminus R_c) > 0\}$ should be small. This failure propability will therefore be set to $0 \leq \alpha \leq 1$. The number of unexploded bombs $N_Z(W)$ is unknown and so is the number of unexploded bombs outside the high-risk zone. Therefore, the failure probability needs to be estimated. What can be used, is the given probability of non-explosion q, the estimated intensity function and the assumption that Z is an inhomogeneous Poisson point process, so

$$N_Z(A) \sim \text{Po}\{\Lambda_Z(A)\} \text{ with } \Lambda_Z(A) = q\Lambda_X(A) = \frac{q}{1-q}\Lambda_Y(A). \tag{2.10}$$

The threshold c is the smallest value that applies to

$$\begin{aligned}
\hat{P}\{N_Z(W \setminus R_c) > 0\} &= 1 - \hat{P}\{N_Z(W \setminus R_c) = 0\} \\
&= 1 - \exp\{-\hat{\Lambda}_Z(W \setminus R_c)\} \cdot \overbrace{\{\hat{\Lambda}_Z(W \setminus R_c)\}^0 \cdot \frac{1}{0!}}^{=1} \\
&= 1 - \exp\left[-\left\{\frac{q}{1-q}\hat{\Lambda}_Y(W \setminus R_c)\right\}\right] \\
&= 1 - \exp\left[-\left\{\frac{q}{1-q}\left(\int_{(W \setminus R_c)} \hat{\lambda}_Y(\mathbf{y})d\mathbf{y}\right)\right\}\right] \\
&\stackrel{!}{=} \alpha
\end{aligned} \tag{2.11}$$

[Mahling et al. 2013] .

3 Implementation

In the previous chapter the different methods to determine high-risk zones were explained. This chapter is about the implementation of these methods and some useful tools for this matter. The implementation is realized in **R**, which is an open source statistical software. Since it is open source, everyone can use and write **R** packages, which are extensions providing utilities for different statistical techniques [R Development Core Team 2012].
In course of this work the `highriskzone`-package is implemented.
This chapter explains the package using the bombs subject to facilitate understanding. However, the package can be used for various other themes.

3.1 Package dependencies

The `highriskzone`-package depends on the two packages `spatstat` [Baddeley and Turner 2005] and `ks` [Duong 2012]. The following explanations are just a short overview on some important utility functions and objects from the packages for the package `highriskzone`. For further information see the manuals of the packages `spatstat` [Baddeley and Turner 2005] and `ks` [Duong 2012].

Package `spatstat`

There exist various packages in **R** that deal with spatial data. One of them is the package `spatstat`. It deals with the analysis of spatial point patterns and supplies several tools which are highly useful for the implementation of the `highriskzone`-package. Next to functions which estimate the density, calculate a distancemap etc., important objects for the matters of the `highriskzone`-package are implemented. The most important of which are:

`ppp` To represent two-dimensional point patterns, objects of class `ppp` were implemented by the authors of `spatstat`. With the function `ppp()` the user can generate such objects. The `highriskzone`-package needs the data used for the analysis in this format.

owin Every object of class `ppp` contains an object of class `owin`, which is the observation window of the point pattern, i.e. the `owin` objects store the coordinates of the border of the observation window. Objects of type `owin` can also exist without being part of a `ppp`-object. For the `highriskzone`-package this type of object is needed for two matters. On the one hand for the actual observation window of the data. On the other hand the zone of high-risk will be of class `owin`.

im Another object class is `im`, which stands for a two-dimensional pixel image. "A pixel image is essentially a matrix of numerical values associated with a rectangular grid of points inside a window in the x, y plane" [Baddeley 2010, Chap. 10, p.63]. A pixel is one unit of the grid.
To create objects of this class, the function `im()` is provided. For changing the number of pixels in the x and y direction set `spatstat.options(npixel)` [Baddeley and Turner 2005]:

```
spatstat.options(npixel = c(250, 250))
```

For details on object-oriented programming in **R** see Matloff [2011, Chap. 9].

Package `ks`

The package `ks` is used for the selection of the smoothing bandwidth while estimating the intensity function. The function, that conducts the selection is `Hscv()` [Duong 2012].

3.2 Package `highriskzone`

The `highriskzone`-package provides a toolbox dealing with the determination of high-risk zones. The two main user functions are `det_hrz()` and `eval_method()`; the first determines the high-risk zone, the second evaluates the zone and accordingly evaluates the methods. To do that, data is either simulated or thinned. Several other functions help dealing with the data, the estimation, the simulation or the evaluation.

3.2.1 Determination of the high-risk zone

The central function of the `highriskzone`-package is `det_hrz`, which determines the high-risk zone using the method the user wants it to. In the R-Code below we see

the arguments that can be set in this function.

```
det_hrz(ppdata, type, criterion, cutoff,
        distancemap, intens, nxprob, covmatrix)
```

The arguments have the following meanings:

ppdata Observed spatial point process of class ppp. In the calculation of high-risk zones for unexploded bombs, this is the data of the point pattern of exploded bombs.

type Method to use for the determination of the high-risk zone (see Chapter 2). Can be one of "dist" or "intens". "dist" is to be used for the methods that calculate the high-risk zoneusing the distance to the next event, i.e. the method of fixed radius and the quantile-based method and as described below, a method that calculates the radius from a fixed area. The methods of type = "dist" can be classified together as distance-based methods. Which one of the three distance-based methods is being carried out depends on the criterion option (see below). If type = "intens", the high-risk zone is determined using the intensity-based method.

criterion Criterion to choose how the high-risk zone should be limited. This can be one of "area", "direct" or "indirect". For criterion = "area", the area is fixed, for criterion = "direct" the radius or the threshold c to cut is fixed and for criterion = "indirect" the radius or the threshold c are calculated indirectly.
More precisely, let type be "dist", then criterion = "area" means the high-risk zone shall be of a certain size and the radius of the circles around the data points is calculated from that. For criterion = "direct" the method of fixed radius is conducted. For criterion = "indirect" the determination is done by a fixed quantile.
If type = "intens" and criterion = "area", the area of the high-risk zone is fixed and the intensity-based method is used. For criterion = "direct", it means the determination is done using a fixed value of the threshold c. The indirect way of determining the high-risk zone by the intensity-based method is to give the failure probability α

and calculate the threshold c from that. That is what happens for `criterion = "indirect"`.

`cutoff` The cutoff value represents the actual fixed value of limitation, i.e. the value of the area, the quantile, the failure probability, the radius or the threshold c, depending on what is set for `type` and `criterion`.

`distancemap` A distance map gives the distance of every pixel to the nearest observation of the point pattern. It is of class `im` [Baddeley 2010, p.85, 115]. The distance map is only needed for the quantile-based method. To put in the distance map is optional for the user. If it is not given, it will be computed.

`intens` The user can specify the intensity. This is optional and only needed for the intensity-based method. The intensity has to be of class `im`. If the user does not specify the intensity and `type = "intens"` it will be estimated.

`nxprob` The argument `nxprob` stands for the probability of having unobserved events, e.g. the probability for a bomb not to explode.

`covmatrix` The covariance matrix of the Gaussian kernel is needed for the determination of the smoothing bandwidth, when the high-risk zone is determined by the intensity-based method (see Section 2.3.2). To set the argument is optional. If it is needed but not given, it will be computed within the function.

The return value of `det_hrz` is an object of class `highriskzone` which basically is a list of the type, criterion and cutoff used, the determined high-risk zone (object of class `owin`), the calculated threshold and cutoff value (`calccutoff`) and the covariance matrix. The threshold is the threshold c if `type = "intens"`. For `type = "dist"`, it is the value of the quantile of the next-neighbour distance for `citerion = "indirect"` or the radius for `criterion = "direct"` and `"area"`. The calculated cutoff `calccutoff` is NA (not available) if the `criterion` is anything else but `"area"`. If `criterion = "area"`, it is the value of the quantile of the next-neighbour distance if the quantile-based method is used or the failure probability α if the intensity-based method is chosen. Further details on the return values can be found in the R Documentation on `det_hrz`.

The following minimal example shows the determination of a high-risk zone using the intensity-based method with a fixed threshold c of 0.15, i.e. regions with estimated intensity higher than 0.15 are in the high-risk zone. The data is read using

the function read_pppdata, which will be explained further in Section 3.2.3.
As we can see cutoff and threshold are equal which is consistent since we give the threshold c as cutoff value. Figure 3.1 shows the resulting high-risk zone in green. plot.highriskzone is a generic plotting function for objects of class highriskzone. It is also executed if a highriskzone-object is given to the function plot().

```
#generate example data
ppdat <- read_pppdata(xppp = c(1, 2, 1, 2, 5, 5.8, 8.5, 1:10),
                      yppp = c(1, 1.5, 1.8, 0.45, 6.5, 7.1, 2.5,
                               9:6, 4:1, 7.5, 8),
                      xwin = c(0.5, 12, 10, 13, 0),
                      ywin = c(-1, 0, 5, 10, 11))


hrz <- det_hrz(ppdat, type = "intens", criterion = "direct",
               cutoff = 0.15, nxprob = 0.1)
hrz

## high-risk zone of type intens
## criterion: direct
## cutoff: 0.15

hrz$threshold

## [1] 0.15

class(hrz)

## [1] "highriskzone"


plot(hrz, zonecol = 3, main = "example hrz", box = FALSE,
     pattern = ppdat, win = ppdat$window, plotpattern = TRUE,
     plotwindow = TRUE)
```

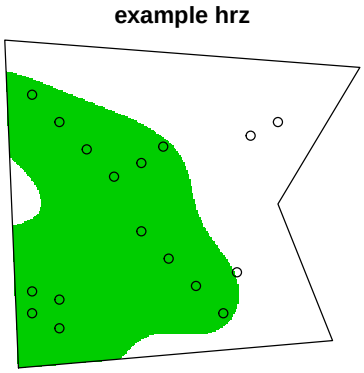


Figure 3.1: High-risk zone for example data

3.2.2 Evaluation of the methods

The high-risk zone is the zone that will be searched for unexploded bombs. Thus a perfect high-risk zone is one which coveres all unexploded bombs and none lie outside and the area to be searched is preferably small. In reality, it is very sumptuous to discover the quality of such a zone. The whole property would have to be searched and the found unexploded bombs inside and outside the high-risk zone would have to be counted.
A way to work around that procedure is to simulate data. The problem is, that one has to know exactly what kind of data is needed, e.g. the type of point process and how high the probability of non-explosion is. But if both type of point process and probability of non-explosion are known, one can simulate observed and unobserved events, determine a high-risk zone based on the observed events and and see how many unobserved events, i.e. unexploded bombs are in- and outside the area to be searched. Another possibility is to use the real data, split it (randomly) into two parts and call one the observed and one the unobserved. With that procedure one does only need the probability of having an unobserved event. The structure will be thinned and the thinning has to be done using assumptions like the amount of thinning, i.e. probability of non-explosion, but no distribution assumptions or similar have to be made for the data structure.

The idea of evaluating the high-risk zone based on a simulation is demonstrated

by the subsequent example. First an inhomogeneous Poisson process is simulated and then randomly split into observed and unobserved events whith a probability to have unobserved events of 0.1. Then the high-risk zone is determined. At last the function `eval_hrz()` is executed, in which the zone of high risk, the observed and the unobserved events have to be set as arguments. It gives back an object of class `hrzeval`, which has a list as basic structure. It contains the number of unobserved events outside the high-risk zone, the number of events in the unobserved point pattern, the fraction of the first two values, the area of the high-risk zone, the number of events in the observed point pattern, a subset of the unobserved events which are outside the high-risk zone and a subset of the unobserved events which are inside the high-risk zone.

```
# simulate a Poisson process
set.seed(123)
lambda <- function(x, y) { 50 * exp( 3 * x) }
simdat <- rpoispp(lambda)

# split data in observed and unobserved
ssimdat <- thin(simdat, nxprob = 0.1)

# determine the high-risk zone
hrzsim <- det_hrz(ssimdat$observed, type = "dist",
                  criterion = "area", cutoff = 0.7,
                  nxprob = 0.1)

# evaluation of the high-risk zone
eval <- eval_hrz(hrzsim$zone, unobspp = ssimdat$unobserved,
                 obspp = ssimdat$observed)
eval

## evaluation of a hig-risk zone based on 279 observed events
## number of unobserved events: 31
## number of unobserved events located outside the high-risk zone: 6

class(eval)

## [1] "hrzeval"
```

As we can see, the high-risk zone determined here is not satisfying because 6 out of the 37 unobserved events do not lie in the high-risk zone, which in the real world would mean that six duds would not be found and stay a hidden danger.

Evaluation gets more solid if more iterations are done, i.e. not only one data set is simulated and the high-risk zone for that data is determined and evaluated. Besides that, it is usually requested that the simulated data is similar to existing real data. For those two tasks the `highriskzone`-package provides a function. The user can choose the number of iterations and also the way the simulation is to be done. What the function does is to simulate data in every iteration and then `det_hrz` and `eval_hrz` are executed on this data.
There are three possible ways of simulating data to evaluate the high-risk zone using the function `eval_method`: the first is that we have a data set and split it randomly with a given probability of having unobserved events into two, as shown above. One we call the observed and one the unobserved. The splitting is done by drawing from a binomial distribution with probability of success equal the probability of having unobserved events. The second kind to simulate data is generating an inhomogeneous Poisson process based on the intensity of the given data set. The third kind is to simulate a cluster process (Neyman-Scott process) also based on the intensity of the given data set and on the maximum radius of a random cluster as well as on the amount of clustering. This last kind of simulation is used to check what happens if the underlying process is not an inhomogeneous Poisson process but a cluster process (see Section 2.3.1).
The possible arguments to set in the function can be seen in the R-Code below:

```
eval_method(ppdata, type, criterion, cutoff,
            numit, nxprob, distancemap,
            intens, covmatrix, simulate,
            radiusClust, clustering)
```

As we can see, this function has mainly the same arguments as `det_hrz` which is consistent, considering that within the function `eval_method` the determination of high-risk zones for the simulated data is executed and the the results are evaluated for each iteration.

`numit` The argument `numit` stands for the number of iterations to be done.

`simulate` The string given for `simulate` represents the way the simulation is to be performed. The option `"thinning"` stands for the random thinning

of the given data, "intens" for the simulation of an inhomogeneous Poisson process via the intensity and thinning afterwards, "clintens" for the simulation of a cluster process via the intensity and thinning of that data.

radiusClust This argument is only needed if a cluster process is simulated and even then it is optional, because it can also be calculated within eval_method. If set, the numeric value of this argument is used as radius of the circles around the parent points in which the cluster points are located.

clustering Has to be a value larger than or equal to 1 which describes the amount of clustering. The adjusted estimated intensity of the observed pattern is divided by this value and it is also the parameter of the Poisson distribution for the number of points per cluster.

The following R code snippet illustrates the use of the function and its arguments. Ten iterations are done and data is simulated by thinning or rather splitting the actual data. The "actual" data here is the simulated data set from above. Usually the number of iterations is higher, but to understand the procedure, it is sufficient.

```
set.seed(123)
ev <- eval_method(simdat, type = "dist", criterion = "area",
                  cutoff = 0.7, nxprob = 0.1, numit = 10,
                  simulate = "thinning", pbar = FALSE)

ev$missingfrac

##  [1] 0.10714 0.03125 0.04167 0.00000 0.00000 0.00000 0.03226 0.00000
##  [9] 0.00000 0.03226
```

The example shows that the fraction of the number of unobserved points outside the high-risk zone and the number of observations in the unobserved point pattern varies over the iterations. There are five high-risk zones that cover all the unobserved events. The highest fraction here is 0.10714.

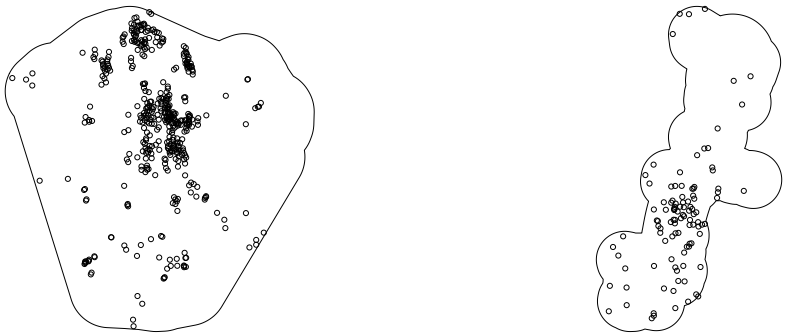
3.2.3 Further functions of the package

There exist several functions in the highriskzone-package that do important work in addition to the mentioned main functions.

In the first example, the function `read_pppdata()` was used reading in the data as a `ppp`-object, so it can be used for analysis concerning the high-risk zone topic.
Another important function we have already learned about is `eval_hrz()`, which is mainly used inside the function `eval_method()`. In cases other point patterns are to be simulated, it is highly useful in direct usage (see Chapter 5). It takes the determined high-risk zone, the observed and unobserved part of the point pattern and tells the user how good the zone is.
For splitting point patterns artificially into observed and unobserved events, the function `thin()` can be used. This function is also part of `eval_method()`.
A highly useful tool for the intensity-based method in the package is `est_intens()`, which estimates the intensity of the given point pattern. It returns a list of the estimated intensity which is an object of class `im` and the covariance matrix. This function is required whenever the intensity-based method is in use.
Last but not least, there is the function `sim_nsppp()`, which simulates a Neyman-Scott process using the intensity of the given data (see Section 2.3.1).
The following chapter shows some examples which illustrate the exploit of the functions.

4 Application

This chapter illustrates the usage of the `highriskzone`-package, drawing on two examples of real data of bomb craters supplied by the *Oberfinanzdirektion Niedersachsen*. To keep data privacy, relative coordinates are used in this example.



(a) Example pattern A (b) Example pattern B

Figure 4.1: Point patterns and observation window of the main examples

The first step of data analysis is always to read in the data. For matters of high-risk zone determination, an object of class `ppp` is needed. Often data is supplied as two data frames: one of the coordinates of the point pattern and one of the coordinates of the observation window. It is, for instance, the point pattern of example A available in the data frame called `patternA`, which stores the x- and y-coordinates of the 443 observations, while the coordinates of the observation window are stored in in `windowA`. In the R-Code below we can see how to read in such type of data.

```
str(patternA)


## 'data.frame': 443 obs. of  2 variables:
##  $ x: num  1088 1103 991 976 1000 ...
##  $ y: num  2413 2400 2373 2368 2341 ...
```

```
str(windowA)

## 'data.frame': 208 obs. of  2 variables:
##  $ x: num  30.94 22.84 15.92 10.22 5.76 ...
##  $ y: num  1948 1931 1912 1893 1874 ...

craterA <- read_pppdata(xppp = patternA$x, yppp = patternA$y,
                        xwin = windowA$x, ywin = windowA$y)
craterA

##  planar point pattern: 443 points
## window: polygonal boundary
## enclosing rectangle: [0, 2334.4] x [0, 2456.4] units
```

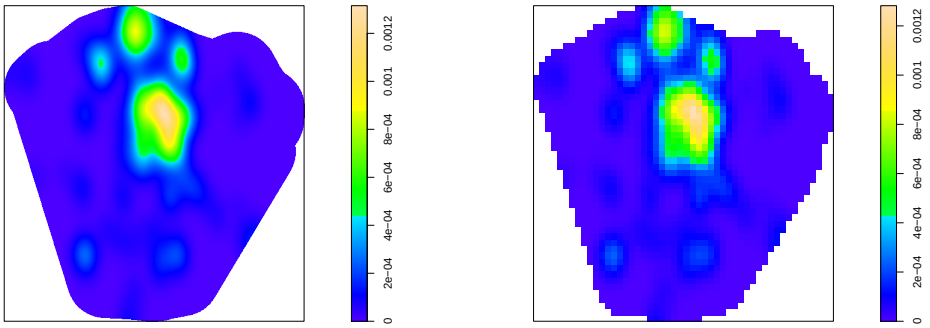
To get a better understanding of the intensity based method, the intensity of example pattern A is estimated and visualized by a colour image, which maps the intensity to a colour. Blue stands for a low intensity, while yellow stands for higher intensity [Baddeley 2010, Chap. 10]. After that, the high-risk zone for this example is determined via the intensity-based method.

```
spatstat.options(npixel = 500)
intensity1 <- est_intens(craterA)
plot(intensity1$intensest, main = "")

spatstat.options(npixel = 50)
intensity2 <- est_intens(craterA)
plot(intensity2$intensest, main = "")
```

As we can see the image of the intensity with 500 pixels is more accurate than the one with 50 pixels. The obvious advantage of less pixels over more is the time of calculation.

There are various ways of determining the high-risk zone using the intensity-based method. The user can choose between different criterions, decide wether to use the already estimated intensity or not and wether to set a covariance matrix or let the function compute it. Here only one way is shown, but in the R description page of the function `det_hrz()` all possibilities can be found. Figure 4.3 shows the resulting high-risk zone.



(a) Estimated intensity 500 pixels (b) Estimated intensity 50 pixels

Figure 4.2: Colour images of the intensity of example A

```
hrzAi <- det_hrz(craterA, type = "intens", criterion = "indirect",
    cutoff = 0.2, intens = intensity1$intensest, nxprob = 0.1)

plot(hrzAi, win = craterA$window, plotwindow = TRUE, zonecol = 3,
    main = "High risk zone", box = FALSE)
```

For the example B in the following R code snippet, a high-risk zone is determined for each method: one with fixed radius of 150 meters, one with the 90%-quantile and one with an α of 0.1. For all three methods the high-risk zone seems to be more or less similar (see Figure 4.4). The method of fixed radius and the quantile-based method have results that are more alike than the result of the intensity-based to those, because they both draw circles aroud the observations. Just that the 90%-quantile is larger that 150 meters.

```
data(craterB)
hrzBdd <- det_hrz(craterB, type = "dist", criterion = "direct",
                  cutoff = 150, nxprob = 0.1)
hrzBdi <- det_hrz(craterB, type = "dist", criterion = "indirect",
                  cutoff = 0.9, nxprob = 0.1)
hrzBi <- det_hrz(craterB, type = "intens", criterion = "indirect",
                 cutoff = 0.1,  nxprob = 0.1)

op <- par(mfrow = c(1, 3), mar=c(0, 4, 3, 2), oma=c(0.1,1,1,1))
```

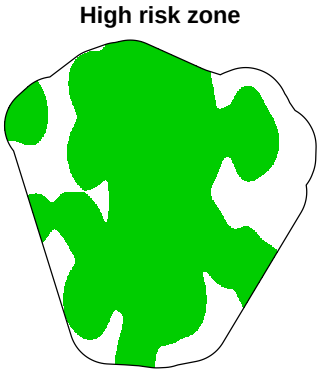


Figure 4.3: High-risk zone for example A determined via the intensity-based method using a fixed failure probability of 0.2

```
plot(hrzBdd, zonecol = 4, main = "hrz by fixed radius\n (r = 150)",
     win = craterB$window, plotwindow = TRUE, box = FALSE)
plot(hrzBdi, zonecol = 4, main = "hrz by quantile\n (q_0.9)",
     win = craterB$window, plotwindow = TRUE, box = FALSE)
plot(hrzBi, zonecol = 4, main = "hrz by intensity\n (alpha = 0.1)",
     win = craterB$window, plotwindow = TRUE, box = FALSE)
```

```
par(op)
```

The next step is to take a closer look at the evaluation prozess. Here we take the data of example B, split it into what we call observed and unobserved events, meaning exploded and unexploded bombs, and determine a high-risk zone giving the observed events. Here determination is conducted using the quantile-based method giving an area of 1.5 million square metres the high-risk zone should have. Further the evaluation is performed for the calculated zone. One third of the unobserved objects lies outside the high-risk zone, which we can also see in Graphic 4.5 produced by the generic function `plot()`, which for an object of class `hrzeval` as input argument calls the function `plot.hrzeval()`. The grey zone is the determined high-risk zone. The blue points resemble the observed events, the magenta ones the unob-

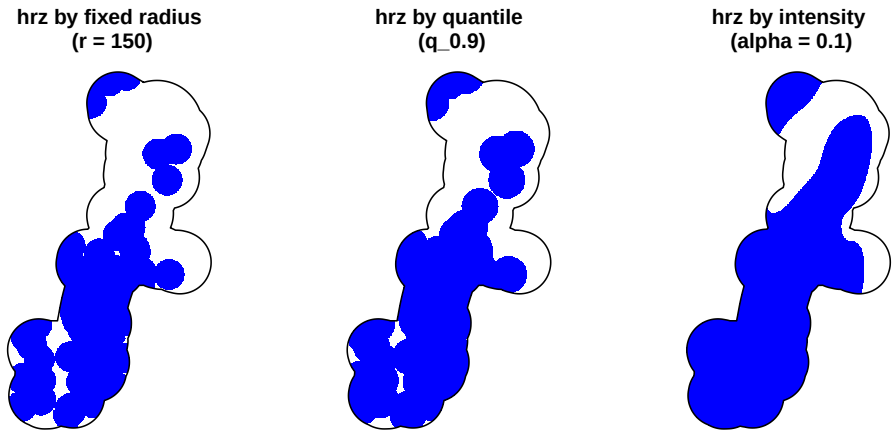


Figure 4.4: High-risk zones for example B determined via the three methods

served. The filled magenta points are the unobserved events outside the high-risk zone.

```
# thin data
set.seed(100)
thdata <- thin(craterB, nxprob=0.1)

# determine hrz for the "observed events"
hrz <- det_hrz(thdata$observed, type = "dist", criterion = "area",
               cutoff = 1500000, nxprob = 0.1)

# evaluate the hrz
evaluation <- eval_hrz(hrz = hrz$zone, unobspp = thdata$unobserved,
                       obspp = thdata$observed)
evaluation$missingfrac

## [1] 0.3333


op <- par(mar=c(1, 4, 1, 6) , xpd=TRUE)
plot(evaluation, hrz = hrz, obspp = thdata$observed, plothrz = TRUE,
     plotobs = TRUE, insidecol = "magenta", outsidecol = "magenta",
```

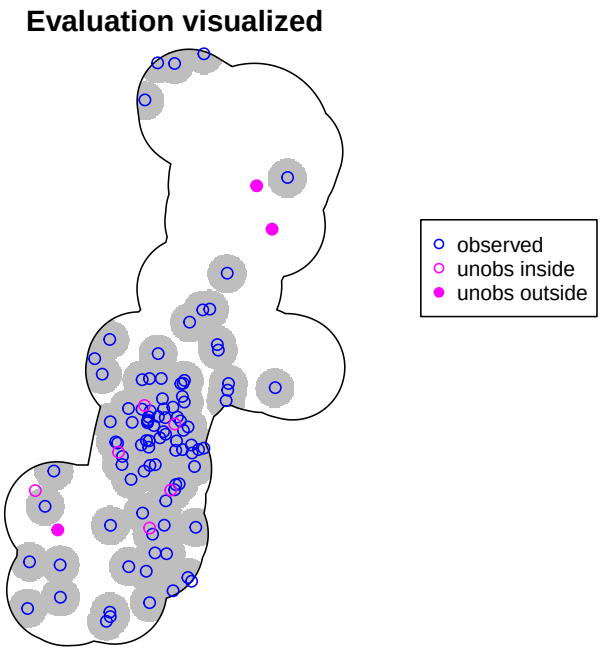


Figure 4.5: Visualisation of what happens in the evaluation of the high-risk zone

```
     obscol = "blue", insidepch = 1, outsidepch = 19,
     main = "Evaluation visualized", box = FALSE)
legend(2400, 2456.4061,
       c("observed", "unobs inside", "unobs outside"),
       col = c("blue", "magenta", "magenta"), yjust=1,
       pch=c(1, 1, 19), cex=0.8)
```

```
par(op)
```

The next application example is for the function `eval_method()`. A cluster process is simulated based on the intensity and the observation window of example B. The maximum possible radius is 300 metres and the paramter of the Poisson distribution for the number of points per cluster is 15. Since it is hard to envision what a cluster process looks like, two example simulated processes are shown first.
For the evaluation, the high-risk zone is determined using on the one hand the quantile-based method and on the other hand the intensity-based method, each giving a fixed area equal to the example before. Ten iterations are done, so ten

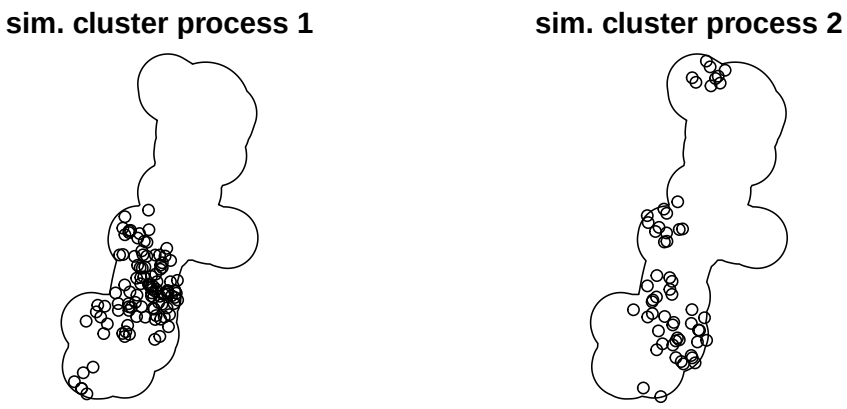


Figure 4.6: Simulated cluster processes

cluster processes are simulated and 20 high-risk zones evaluated, since two zones are determined for each process.

```
set.seed(100)
sim_pp1 <- sim_nsppp(craterB, radius=300, clustering=15,
                     thinning=0.1)
sim_pp2 <- sim_nsppp(craterB, radius=300, clustering=15,
                     thinning=0.1)
op <- par(mfrow = c(1, 2))
plot(sim_pp1, main = "sim. cluster process 1")
plot(sim_pp2, main = "sim. cluster process 2")
```

```
par(op)

evalm <- eval_method(craterB, type = c("dist", "intens"),
         criterion = c("area", "area"), cutoff = c(1500000, 1500000),
         nxprob = 0.1, numit = 10, simulate = "clintens",
         radiusClust = 300, clustering = 15, pbar = FALSE)
evalm_d <- subset(evalm, evalm$Type == "dist")
evalm_i <- subset(evalm, evalm$Type == "intens")
```

```r
data.frame(pmiss_d = mean(evalm_d$missingfrac),
           pmiss_i = mean(evalm_i$missingfrac),
           pout_d = ( sum(evalm_d$numbermiss > 0) / nrow(evalm_d) ),
           pout_i = ( sum(evalm_i$numbermiss > 0) / nrow(evalm_i) ))

##   pmiss_d pmiss_i pout_d pout_i
## 1  0.2113  0.2323    0.8    0.9
```

At an equal area the results of the two methods seems to be quite similar. The mean fraction of duds outside the high-risk zone (`pmiss`) is with 0.2113 a little lower for the quantile-based method than for the intensity based. The fraction of high-risk zones with at least one unexploded bomb outside the zone (`pout`) is again lower for the quantile-based method, which means that for the ten simulated processes, the quantile-based method performed better on average.

5 Simulation study

In the previous chapters different methods of the determination of high-risk zones, the implementation in the `highriskzone`-package and the usage the package were presented.
It is of interest how the methods perform on different point patterns, even ones a method was not meant for in the first place. In this chapter, we will investigate performance of the intensity-based method versus a distance-based method when they are applied to a homogeneous Poisson process or to a process that has the form of a normal mixture distribution. Larger high-risk zones lead in trend to better results in terms of covering unexploded bombs. Therefore it is reasonable to give a fixed area when comparing the methods. Thus, the intensity-based method is used here giving a fixed area (`type = "intens", criterion = "area"`) and so is the distance-based method (`type = "dist", criterion = "area"`) in which the required radius is calculated from the given area (see Section 3.2.1).
The study will be performed by simulating data, splitting it into an "observed" and an "unobserved" process, determining a high-risk zone for the observed part and evaluating the high-risk zone by investigating how many unobserved events lie outside and inside the high-risk zone. The procedure will be conducted 1000 times in each case.

5.1 Homogeneous Poisson process

In Section 2.3.1 we have learned that the difference between a homogeneous and a inhomogeneous Poisson process is the intensity λ. For the inhomogeneous process it depends on the location, whereas it is constant for the homogeneous process.
Figure 5.1 shows a point pattern generated by a homogeneous Poisson process on the unit square with $\lambda = 50$ on the left side and the estimated intensity (by `est_intens`) on the right side. It displays clearly that the estimated intensity does not equal the theoretical intensity 50 on all locations. If the estimated intensity actually was constant on all locations, this would mean for the intensity-based method that either the high-risk zone would equal the observation window or the area of the high-risk

Table 5.1: Mean fraction of unobserved events outside the high-risk zone p_{miss} and fraction of high-risk zones that leave at least one unobserved event uncovered p_{out} for homogeneous Poisson processes; calculation only with data sets that contain unobserved events in brackets; method intens stands for intensity based method, dist for distance-based method

	$\lambda = 50$		$\lambda = 500$	
method	intens	dist	intens	dist
p_{miss}	0.24059	0.2422	0.2503	0.2489
p_{out}	0.7160 (0.7157)	0.7100 (0.7097)	1	1

zone would be zero, depending on the value of the threshold c. That is why the homogeneous Poisson process is interesting for this investigation.

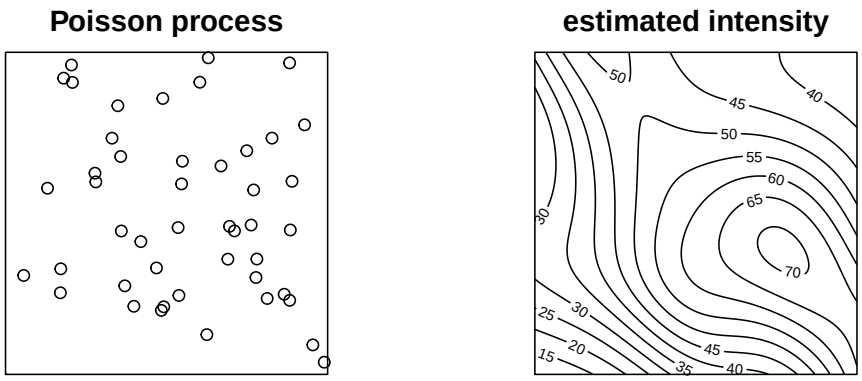


Figure 5.1: Simulated Poisson process with intensity 50 and corresponding estimated intensity

In this section two different types of homogeneous Poisson processes are simulated. One with $\lambda = 50$ and one with $\lambda = 500$, each with the unit square as observation window. The simulated data is randomly split into unobserved and observed events, with a probability to have unobserved events of 10 percent. The procedure described at the beginning of the chapter is iterated 1000 times. To determine the high-risk zones, a fixed area of 0.75 is given.

A summary of the results is shown in Table 5.1. It depicts the mean fraction of unobserved events, i.e. unexploded bombs, outside the high-risk zone in 1000 iterations

which is enlabeled with p_{miss} and the fraction of high-risk zones that leave at least one unobserved event uncovered (p_{out}). Both values are rounded to four decimal places.
For the Poisson processes with intensity 50, the mean fraction of unobserved events outside the high-risk zone is slightly lower, i.e. the high-risk zones are better in average for the intensity-based method. However, there are more high-risk zones that cover all bombs for the distance-based method. 9 out of the 1000 data sets simulated had no unobserved events. These data frames could not be included in the calculation of p_{miss} for dividing by zero is not possible. p_{out} was calculated both for all cases and only for cases with at least one unobserved event. The latter aspect is shown in brackets.
Figure 5.2 shows the empirical cumulative distribution functions of the fraction of unobserved events outside the high-risk zone which will be denominated as missing fraction. Generally, it can be said that the better the method, the larger the area under the curve, hence the curve being left or above the other curve is the one of the method with better outcome.
The first plot shows the results for the Poisson process with intensity 50. The jump in zero shows that there is more than a quarter of high-risk zones covering all unobserved events, which is just the same as $1 - p_{out}$. The green line standing for the intensity-based method and the red line standing for the distance-based method are very similar. None of the methods can visually be rated as the favourable here since no curve is constantly above the other.
The same shows us the boxplot of the differences

$$m_{\text{intens},j} - m_{\text{dist},j} \quad \forall j = \{1, 2, ..., 1000\} \tag{5.1}$$

with $m_{\text{intens},j}$ missing fraction of the intensity-based method and $m_{\text{dist},j}$ missing fraction of the distance-based method in iteration j which is shown in Figure 5.3. It is highly symmetric which means there is no tendency to which method is better.

For the processes with an intensity of 500 the distance-based method in average performs very little better than the intensity-based method. While for both methods no high-risk zone covers all unexploded bombs, the mean fraction of duds outside the high-risk zones takes a value of 25.03 percent for the intensity based method and 24.89 percent for the distance-based method. The curves of the empirical distribution functions are very close. From the plot one can not make out which method is the better one. The corresponding boxplot (Figure 5.3 left box) tells us the same.

Altogether, both methods show similar performances for the homogeneous Poisson processes.

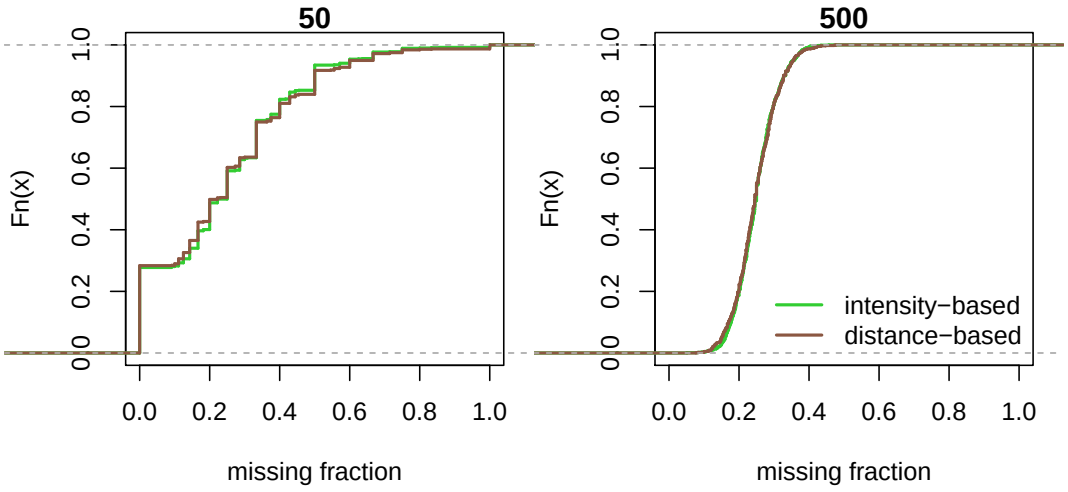


Figure 5.2: Empirical distribution functions of the missing fraction for the Poisson processes with intensity 50 and 500

5.2 Normal mixture distribution

A mixture distribution is a mixture of two or more distributions. Let X be a random variable with

$$X \sim \eta_1 \cdot N_2(\mu_1, \Sigma_1) + \eta_2 \cdot N_2(\mu_2, \Sigma_2) \tag{5.2}$$

then X originates from a mixture distribution of two two-dimensional multivariate normal distributions with weights η_1, η_2. Then the mixture density is

$$f_x = \eta_1 \cdot f_2(y; \mu_1, \Sigma_1) + \eta_2 \cdot f_2(y; \mu_2, \Sigma_2). \tag{5.3}$$

The weights are values between 0 and 1 and sum to 1, so the integral of the density equals 1. In random number generation the weight η_i stands for the mixing proportions, e.g. here η_1 is the probability that the sample variable is part of the distribution $N_2(\mu_1, \Sigma_1)$ [Frühwirth-Schnatter 2006, Chap. 1.2].

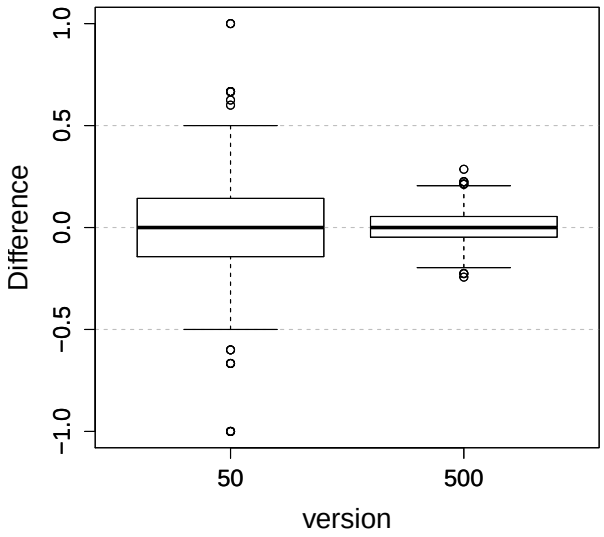


Figure 5.3: Boxplot of the differences $m_{intens,j} - m_{dist,j}$ for the Poisson processes with intensity 50 and 500

Two-dimensional normal mixture distributions with circular shape

In the following, point patterns are simulated from a two-dimensional normal mixture distribution with proportions 0.75 and 0.25. Thus, an average of three quarters of the 250 points to be simulated are expected to be part of the first set of two-dimensional normal distributions and one quarter of the other. Since the number of points in total is fixed to 250, the simulated point patterns are no inhomogeneous Poisson processes [Illian et al. 2008, p. 118].
First, four different types of normal mixture distributions are simulated. Later on two other types will be introduced. All four of the primary versions have the same variances and covariances equal zero for one set of multivariate normal distribution. In all cases, the mean of the first set is the two dimensional null vector and the covariance matrix has 1 on the diagonal and 0 else. Means and covariance matrices of the second sets vary, yet the variances are always larger than in the first set. The window stays the same for all versions: a rectangle with a range of -5 to 30 in x-direction and -10 to 10 in y-direction. It is possible that generated points lie outside this window which does not matter since in the real world sometimes events also lie outside the observed field and are not considered in the analysis.
To be able to compare the results, the high-risk zone is determined fixing the area again. Here the fixed area is 100.

Table 5.2: Overview on mixture distributions used for the simulation

Version 1	$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$; $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\mu_2 = \begin{pmatrix} 20 \\ 0 \end{pmatrix}$; $\Sigma_2 = 10 \cdot \Sigma_1$
Version 2	$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$; $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\mu_2 = \begin{pmatrix} 20 \\ 0 \end{pmatrix}$; $\Sigma_2 = 1.5 \cdot \Sigma_1$
Version 3	$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$; $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\mu_2 = \begin{pmatrix} 5 \\ 0 \end{pmatrix}$; $\Sigma_2 = 10 \cdot \Sigma_1$
Version 4	$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$; $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\mu_2 = \begin{pmatrix} 5 \\ 0 \end{pmatrix}$; $\Sigma_2 = 1.5 \cdot \Sigma_1$

Table 5.2 shows the kinds of normal mixture distributions which will be used. A visual display of the densities of the four variations can be seen in Figure 5.4. The plots display clearly that the first two versions do not cross even at very low level (0.002), but in the other two the densities mix.

For random numbers generated for the mentioned normal mixture distributions, the estimated density or intensity is usually not equal but similar to the shown theoretical densities in Figure 5.4. An example that shows this is Figure 5.5.

Table 5.3: Mean fraction of unobserved events outside the high-risk zone p_{miss} and fraction of high-risk zones that leave at least one unobserved event uncovered p_{out} for normal mixture distributions versions 1 to 4; method intens stands for intensity based method, dist for distance-based method

	version 1		version 2		version 3		version 4	
method	intens	dist	intens	dist	intens	dist	intens	dist
p_{miss}	0.1234	0.1399	0.0073	0.0038	0.0839	0.0989	0.0011	0.0011
p_{out}	0.9470	0.9680	0.1670	0.0920	0.8740	0.9160	0.0280	0.0270

Table 5.3 shows a summary of the results of the simulation study for the four versions of normal mixture distributions with circular shape. The intensity-based method performs better on average of the 1000 iterations for the first and third version. Both the mean fraction of unobserved events outside the high-risk zone and the fraction of high-risk zones that leave at least one unobserved event uncovered is higher for the distance-based method. For version 2, however, the tendency is contrary and for version 4 p_{miss} is the same for both methods while for the intensity-based method slightly more high-risk zones do not cover all unobserved events.
The empirical distribution function curve of the intensity-based method in version 1

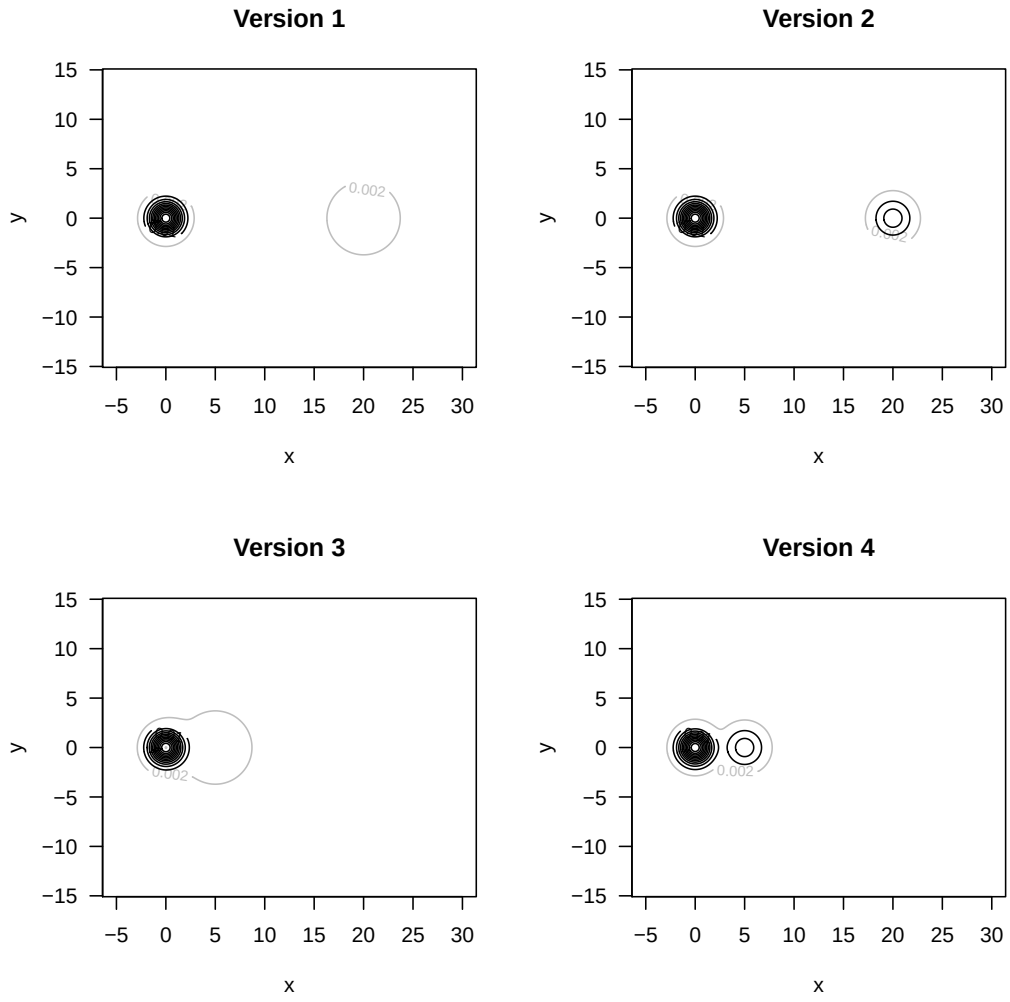


Figure 5.4: Overview on theoretical densities of mixture distributions used for the simulation. The grey line is always on 0.002, the black contour lines are in equal distances.

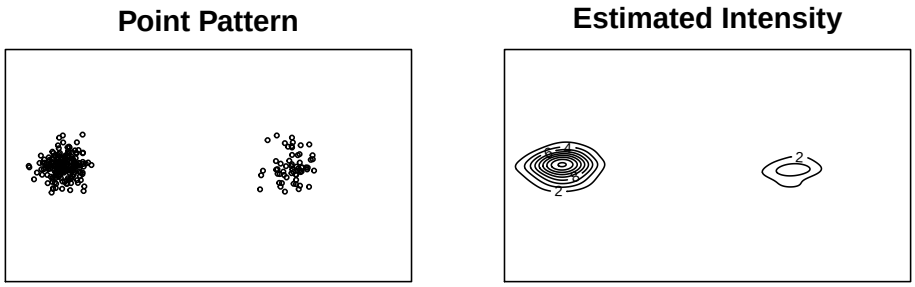


Figure 5.5: Example of randomly generated numbers from the normal mixture distribution version 2

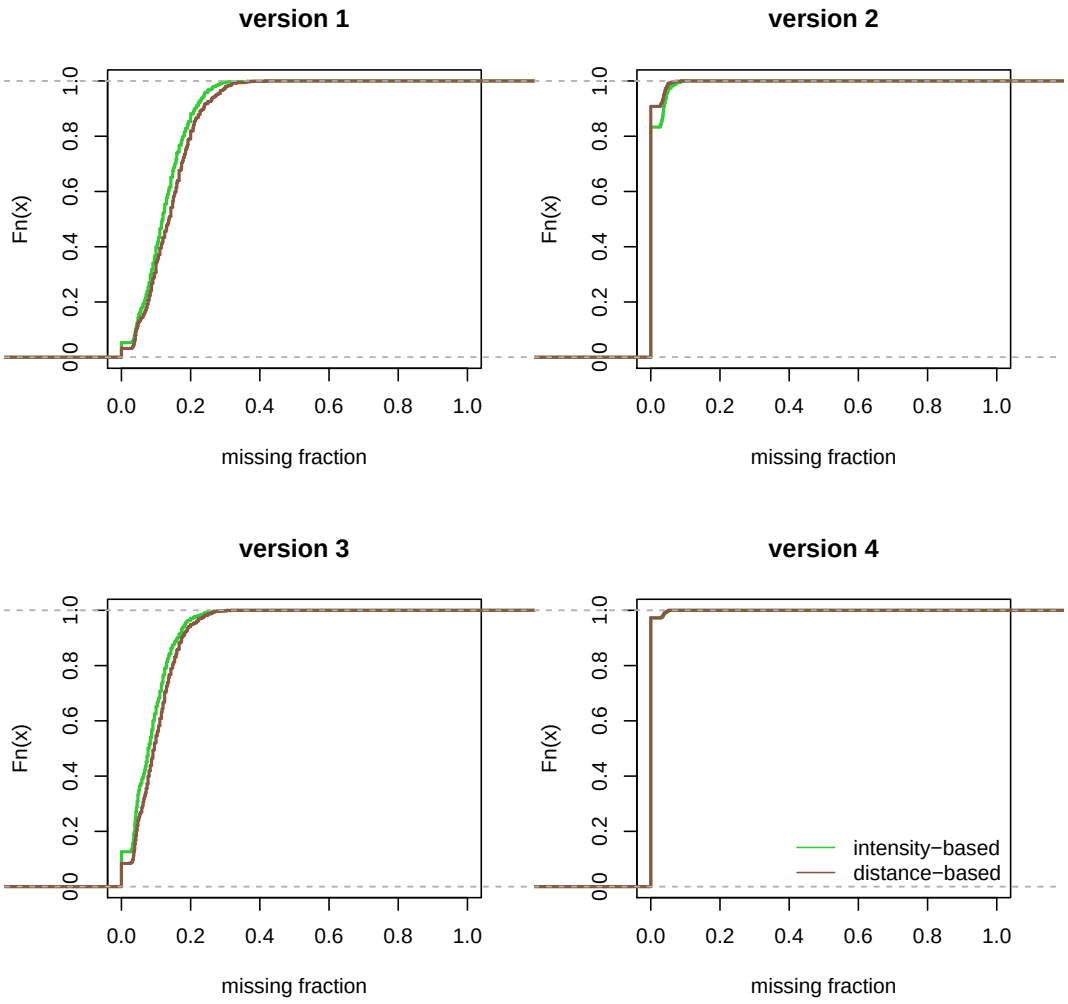


Figure 5.6: Empirical distribution functions of the missing fraction for the normal mixture distributions version 1-4

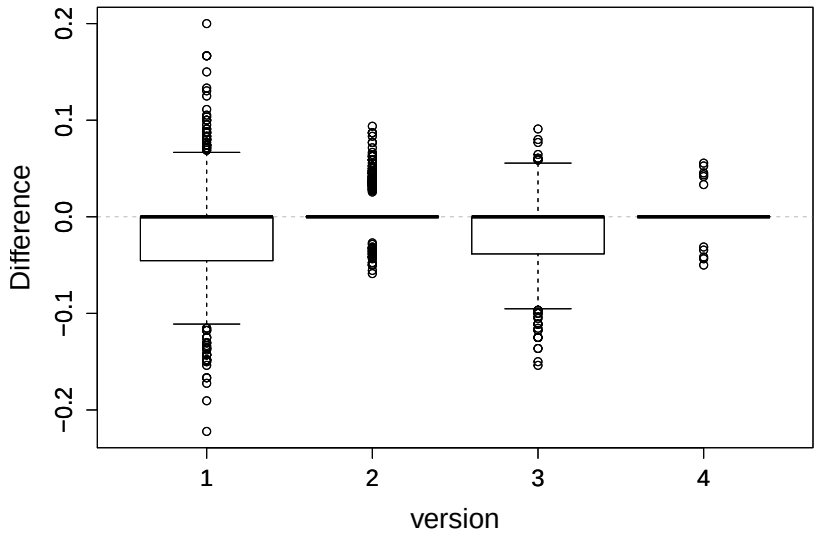


Figure 5.7: Boxplot of the differences $m_{intens,j} - m_{dist,j}$ for the normal mixture distributions version 1-4

and 3 (Figure 5.6) is a little above the curve of the distance-based method which tells us that the intensity-based method operates better on such type of point pattern. The area under the curve for version 2 and 4 is large for both methods which sticks to the small values of p_{out}. Due to the fact that for the distance-based method even more high-risk zones cover all unexploded bombs, the area is even a little smaller than for the intensity-based method. For version 4 the difference can hardly be perceived.
The distribution of the differences $m_{\text{intens},j} - m_{\text{dist},j}$ has a negative skew for methods 1 and 3 which is shown in Figure 5.7. This again delivers that the intensity-based method performes better here. The boxes of version 2 and 4 are reduced to the median line which is clear since for both methods most missing fractions are zero.
In conclusion for the versions with the bigger difference ($\Sigma_2 = 10 \cdot \Sigma_1$) between the variances of the sets, the intensity-based method delivers a better performance; for $\Sigma_2 = 1.5 \cdot \Sigma_1$, the distance-based method performs slightly better.
Furthermore both methods achieve better results for the smaller difference of the variances. With the smaller distances of the means which we have for version 3 and 4 the high-risk zones of both methods cover the unobserved events better than with the distances of means in version 1 and 2.

Table 5.4: Overview on mixture distributions used for the simulation

Version 5	$\mu_1 = \begin{pmatrix} 5 \\ 0 \end{pmatrix}$; $\Sigma_1 = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$	$\mu_2 = \begin{pmatrix} 17.5 \\ 0 \end{pmatrix}$; $\Sigma_2 = 1.5 \cdot \Sigma_1$
Version 6	$\mu_1 = \begin{pmatrix} 5 \\ 0 \end{pmatrix}$; $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\mu_2 = \begin{pmatrix} 17.5 \\ 0 \end{pmatrix}$; $\Sigma_2 = \begin{pmatrix} 1.5 \cdot 1 & 0 \\ 0 & 1.5 \cdot 3 \end{pmatrix}$

Two-dimensional normal mixture distributions with elliptic shape

As supplement to the normal mixture distributions with equal variances in one set, we now consider elliptical sets. The two cases are explained in Table 5.4 and the theoretical densities are shown in Figure 5.8. Observation windows and weights η_1, η_2 are the same as for the normal mixture distributions above. The distance of the means is the average distance of the means for the earlier simulations, 12.5. For the first version with elliptic sets – version 5 – the variance in x-direction of Σ_1 is three times the variance in y-direction and the covariance matrix Σ_2 is 1.5 times Σ_1. In version 6 the variances of x- and y-direction of Σ_2 are swapped in reference to version 5.

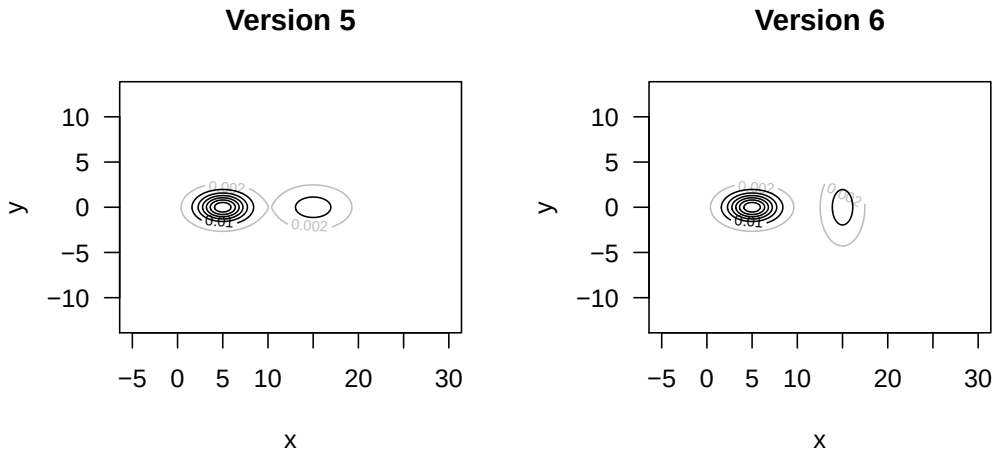


Figure 5.8: Overview on theoretical densities of mixture distributions used for the simulation. The grey line is always on 0.002, the black contour lines are in equal distances.

Both versions show highly similar results. Values p_{miss}, p_{out} as well as the Visualizations 5.9 and 5.10 demonstrate that for version 5 and 6 the intensity-based method delivers better results than the distance-based. Using the intensity-based method, about half of the high-risk zones cover all unobserved events in both version 5 and 6.

Table 5.5: Mean fraction of unobserved events outside the high-risk zone p_{miss} and fraction of high-risk zones that leave at least one unobserved event uncovered p_{out} for normal mixture distributions versions 5 and 6; method intens stands for intensity based method, dist for distance-based method

method	version 5		version 6	
	intens	dist	intens	dist
p_{miss}	0.0268	0.0384	0.0301	0.0387
p_{out}	0.4870	0.6160	0.5290	0.6200

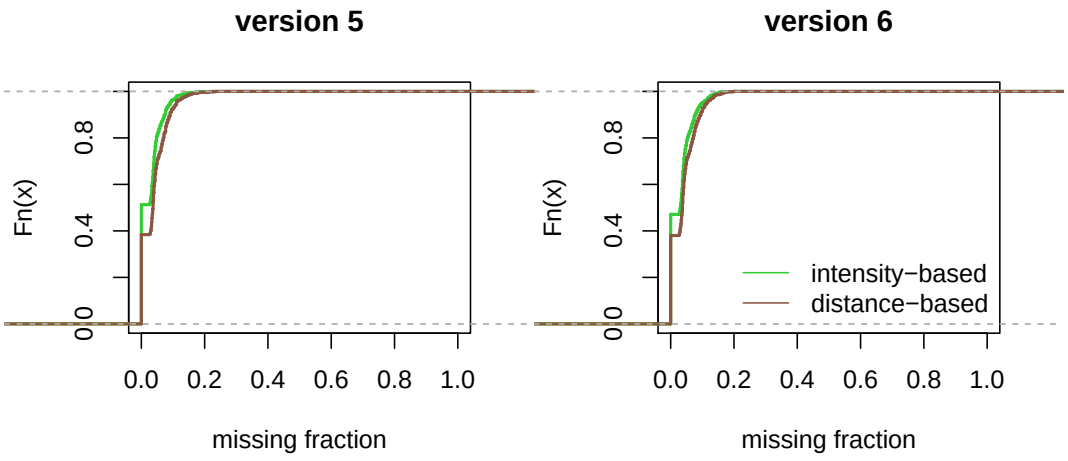


Figure 5.9: Empirical distribution functions of the missing fraction for the normal mixture distributions version 5 and 6

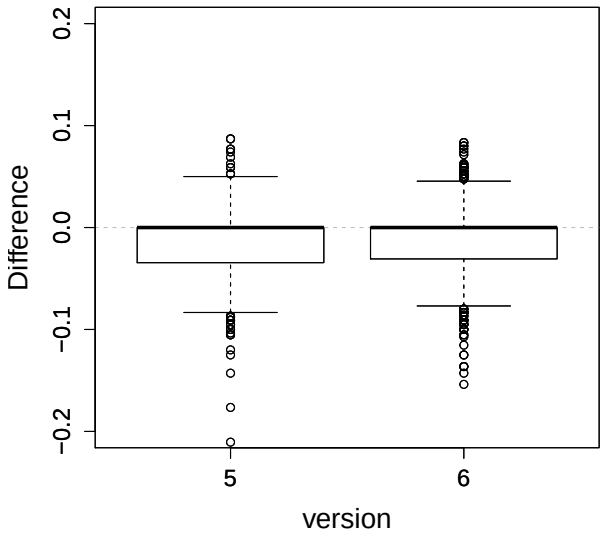


Figure 5.10: Boxplot of the differences $m_{intens,j} - m_{dist,j}$ for the normal mixture distributions version 5 and 6

For the distance-based method, it only is about 40 percent. The curve of the empirical distribution function of the intensity-based method is always left or equal the curve of the distance-based method. The boxplots of the differences $m_{\text{intens},j} - m_{\text{dist},j}$ show similar results as for versions 1 and 3.

One can hardly say that one method performed better in the simulation study than the other. All point patterns simulated were no inhomogeneous Poisson processes, so one would expect the intensity-based method, which bases on the assumption that the underlying process is an inhomogeneous Poisson process, to not perform as decently as it did. Apparently the intensity-based method is quite robust. At least for the point patterns shown here. Still the performance of the distance-based method was not much worse. For version 2 and 4 of the normal mixture distributions it even did slightly better.
There are uncountable further possibilities of point patterns that could be simulated. One idea is to simulate an inhomogeneous Poisson process using the intensity of the normal mixture distributions version 1 to 6 and see if this leads to an improvement towards the normal mixture distributions simulated in this thesis [Baddeley 2010, Chap. 15.1].

6 Conclusion

In this thesis different methods to determine high-risk zones were presented: The distance-based methods and the intensity-based method. These methods and various tools regarding the high-risk zone topic were implemented in an R package called `highriskzone`. Among the tools were a function for data preparation, one for simulating cluster processes, functions to evaluate high-risk zones and generic functions for visualization. The package depends on two other R packages: `ks` [Duong 2012] and the highly useful package `spatstat` [Baddeley and Turner 2005], which deals with the analysis of spatial point patterns.
In Chapter 4 some possible ways of using the `highriskzone`-package are shown for real point patterns of exploded bombs from German properties. Data was provided by the *Oberfinanzdirektion Niedersachsen*.

A simulation study where homogeneous Poisson processes and two-dimensional normal mixture distributions were generated, showed that it is not clear which of the intensity-based and the distance-based method is the one with overall better performance for the chosen types of point patterns. Though, the intensity-based method did slightly better on some process types. The intensity-based method also seems to be quite robust for point patterns which are no inhomogeneous Poisson processes.
To get deeper knowledge about when which method delivers better results, there have to be more simulations with more different types of point patterns and different fixed areas.

List of Figures

List of Tables

Bibliography

A. Baddeley. Analysing spatial point patterns in R. Technical report, CSIRO, 2010. Version 4.1., 2010. URL http://www.spatstat.org/spatstat/. 3.1, 3.2.1, 4, 5.2

A. Baddeley and R. Turner. Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005. URL www.jstatsoft.org. ISSN 1548-7660. 3.1, 3.1, 6

T. Duong. *ks: Kernel smoothing*, 2012. URL http://CRAN.R-project.org/package=ks. R package version 1.8.7. 3.1, 3.1, 6

S. Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Verlag, 2006. 5.2

J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan. *Statistical analysis and modelling of spatial point patterns*. Wiley-Interscience, 2008. 2.3.1, 2.3.1, 2.3.1, 2.3.1, 5.2

M. Mahling, M. Höhle, and H. Küchenhoff. Determining high-risk zones for unexploded World War II bombs by using point process methodology. *expected in Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 2013. 1, 2, 2.2.2, 2.3.1, 2.3.2, 2.3.2, 2.3.3

N. Matloff. The art of R programming. *No Starch Pess*, 2011. 3.1

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org/. 3

Eidesstattliche Erklärung

Hiermit erkläre ich, Heidi Seibold, Matrikel-Nr. 10069365, dass ich meine Bachelorarbeit mit dem Thema

Determining high-risk zones using point process methodology

selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

München, den 11. Juli 2012

Heidi Seibold

Digital appendix

The enclosed CD includes the following contents:

Package The highriskzone-package folder, the PDF file with the documentation and an R script to install all necessary packages

Graphics All graphics shown in the thesis

Thesis The thesis

Literature Literature available as PDF files