Determining high-risk zones by using spatial point process methodology

Monia Mahling

München 2013

Determining high-risk zones by using spatial point process methodology

Monia Mahling

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Monia Mahling
aus Altdorf b. Nürnberg

München, den 28.03.2013

Erstgutachter: Prof. Dr. Helmut Küchenhoff
Zweitgutachter: Dr. Janine B. Illian
Drittgutachter: PD Michael Höhle, Ph.D.
Tag der Disputation: 7. Juni 2013

Contents

Abstract

Methods for constructing high-risk zones, which can be used in situations where a spatial point pattern has been observed incompletely, are introduced and evaluated with regard to unexploded bombs in federal properties in Germany. Unexploded bombs from the Second World War represent a serious problem in Germany. It is desirable to search high-risk zones for unexploded bombs, but this causes high costs, so the search is usually restricted to carefully selected areas. If suitable aerial pictures of the area in question exist, statistical methods can be used to determine such zones by considering patterns of exploded bombs as realisations of spatial point processes. The patterns analysed in this thesis were provided by Oberfinanzdirektion Niedersachsen, which supports the removal of unexploded ordnance in federal properties in Germany. They were derived from aerial pictures taken by the Allies during and after World War II.

The main task consists of finding as small regions as possible containing as many unexploded bombs as possible. In this thesis, an approach based on the intensity function of the process is introduced: The high-risk zones consist of those parts of the observation window where the estimated intensity is largest, i.e. the estimated intensity function exceeds a cut-off value c. The cut-off value can be derived from the risk associated with the high-risk zone. This risk is defined as the probability that there are unexploded bombs outside the zone.

A competing approach for determining high-risk zones consists in using the union of discs around all exploded bombs as high-risk zone. The radius is chosen as a high quantile of the nearest-neighbour distance of the point pattern. In an evaluation procedure, both methods yield comparably good results, but the theoretical properties of the intensity-based high-risk zones are considerably better.

A further goal is to perform a risk assessment of the investigated area by estimating the probability that there are unexploded bombs outside the high-risk zone. This is especially important as the estimation of the intensity function is a crucial issue for the intensity-based method, so the risk cannot be determined exactly in advance. A procedure to calculate the risk is introduced. By using a bootstrap correction, it is possible to decide on acceptable risks and find the optimal, i.e. smallest, high-risk zone for a fixed probability that not all unexploded bombs are located inside the high-risk zone.

The consequences of clustering are investigated in a sensitivity analysis by exploiting the procedure for calculating the risk. Furthermore, different types of models which account for clustering are fitted to the data, classical cluster models as well as a mixture of bivariate normal distributions.

Zusammenfassung

Methoden zur Konstruktion von Risikozonen, die verwendet werden können, wenn ein räumliches Punktmuster unvollständig beobachtet wurde, werden am Beispiel von Blindgängern auf Bundesliegenschaften in Deutschland eingeführt und evaluiert. Blindgänger aus dem Zweiten Weltkrieg stellen in Deutschland ein schwerwiegendes Problem dar. Es ist daher wünschenswert, Risikozonen nach Blindgängern abzusuchen. Da dies jedoch hohe Kosten verursacht, beschränkt sich die Suche normalerweise auf sorgfältig ausgewählte Gebiete. Falls für das fragliche Gebiet geeignete Luftbilder existieren, können zur Bestimmung solcher Zonen statistische Methoden angewandt werden, indem das Muster der detonierten Bomben als Realisation eines räumlichen Punktprozesses betrachtet wird. Die in dieser Arbeit analysierten Muster wurden von der Oberfinanzdirektion Niedersachsen zur Verfügung gestellt, die die Kampfmittelräumung auf deutschen Bundesliegenschaften unterstützt. Sie wurden aus Luftbildern gewonnen, die die Alliierten während und nach dem Zweiten Weltkrieg aufgenommen haben.

Das primäre Ziel besteht darin, möglichst kleine Regionen zu finden, die möglichst viele Blindgänger enthalten. In dieser Arbeit wird ein Ansatz vorgestellt, der auf der Intensitätsfunktion des Prozesses basiert: Die Risikozonen bestehen aus denjenigen Teilen des Beobachtungsfensters, in denen die geschätzte Intensität am höchsten ist, d.h. in der die geschätzte Intensitätsfunktion einen Cutoff-Wert c überschreitet. Der Cutoff-Wert kann vom Restrisiko der entsprechenden Risikozone abgeleitet werden.

Ein konkurrierender Ansatz zur Bestimmung von Risikozonen besteht darin, die Vereinigung aller Kreisscheiben um die detonierten Bomben herum als Risikozone zu definieren. Der Radius ergibt sich als hohes Quantil des Nächste-Nachbarn-Abstandes des Punktmusters. Bei der Evaluation liefern beide Methoden ähnlich gute Ergebnisse, jedoch sind die theoretischen Eigenschaften der intensitätsbasierten Risikozonen deutlich besser.

Ein weiteres Ziel ist eine Risikoabschätzung für das untersuchte Gebiet, wofür die Wahrscheinlichkeit für Blindgänger außerhalb der Risikozone geschätzt wird. Dies ist insbesondere deswegen wichtig, weil sich die Schätzung der Intensität als kritischer Punkt der intensitätsbasierten Methode erwiesen hat und das Risiko im Voraus nicht exakt festgelegt werden kann. Es wird ein Verfahren zur Risikobestimmung vorgestellt. Mit Hilfe einer Bootstrap-Korrektur ist es möglich, das akzeptable Risiko festzulegen und die optimale (d.h. kleinste) Risikozone für eine vorgegebene Wahrscheinlichkeit, dass nicht alle Blindgänger in der Risikozone liegen, zu finden.

Die Auswirkungen von Clustering werden in einer Sensitivitätsanalyse untersucht, wozu das Verfahren zur Risikobestimmung verwendet wird. Darüber hinaus werden verschiedene Arten von Clustermodellen an die Daten angepasst, sowohl klassische Clustermodelle als auch Mischungen von bivariaten Normalverteilungen.

Besonders bedanken möchte ich mich bei

- Helmut Küchenhoff, dem ich nicht nur das spannende Thema zu verdanken habe, sondern von dem ich in mehr als sieben Jahren Mitarbeit im Statistischen Beratungslabor auch sehr viel lernen konnte,

- Michael Höhle für seine Diskussionsfreudigkeit nicht nur im Hinblick auf die Arbeit selbst, das Aufwerfen neuer Blickwinkel, kritische Nachfragen und Zuspruch,

- der Oberfinanzdirektion Niedersachsen, insbesondere Herrn Christian Meinhardt und Herrn Wilfried Möller, für die finanzielle Unterstützung des Projekts und viele interessante Diskussionen,

- Herrn Robert Brosy und Herrn Andreas Bernhardt von der Mull und Partner Ingenieurgesellschaft für die Zusammenarbeit und die Bereitstellung von Daten und Hintergrundinformationen, sowie die kurzfristige Beschaffung des amerikanischen Luftbildes für Kapitel 1,

- Volker Schmid, Torsten Hothorn und Thomas Kneib für ihr Interesse an dem Thema und verschiedenste Hinweise dazu,

- Heidi Seibold für die gute Zusammenarbeit bei der Erstellung des R-Paketes,

- zahlreichen Freunden und Kollegen, insbesondere Felix Heinzl, Verena Guttenberg, Anne Kraus und Cornelia Oberhauser,

- meinen Eltern für ihre Unterstützung vom ersten Tag an bis heute,

- meiner Schwester Marina für unzählige Telefonate und ihre Geduld

- und schließlich Jürgen Grimm für Entlastung, Unterstützung, Verständnis, Ablenkung, Aufmunterung und Rückhalt.

1. Background and motivation

High-risk zones are relevant in all situations which can be characterized as follows: Some kind of event is observed in random locations, but not all locations are known and one would like to find a region where the unobserved events will be discovered with a high probability. Such situations can arise in a variety of applications, such as epidemiology (e.g. if not all cases of a certain type of infectious disease are reported to the authority in charge) or ecology (e.g. for locations of rare plants which are difficult to detect). The focus in this thesis is on one specific application, unexploded World War II bombs.

Even more than 65 years after the end of the Second World War, unexploded bombs still represent a serious problem in Germany. Their clearance usually requires the evacuation of houses and the closing of roads and railway lines. As they are often accidentally found during construction work, these actions have to be taken quickly and often at times which are especially inconvenient. Even worse, unintended detonations have resulted in severe accidents in several cases.

To avoid accidents and render evacuations more foreseeable, it is desirable to search high-risk areas for unexploded bombs before any construction work starts. Depending on several characteristics of the subsoil, the search alone–without clearance or possible reconstruction–costs between 0.20 and 20 € per m^2, so it must be restricted to carefully selected areas.

During and after the Second World War, the Allies took aerial pictures of regions they had bombed. An example is given in Figure 1.1. Nowadays, experts analyse these aerial pictures and derive the locations of bomb craters. In some cases, smaller structures in the aerial pictures, which are difficult to detect, indicate that a bomb may have thuded in this place, but did not detonate. However, this does not necessarily mean that an unexploded bomb is located in this position, as such findings from the aerial pictures are rather vague and it is often impossible to retrace where unexploded ordnance was removed during and in the first years after World War II.

If suitable aerial pictures of the area in question exist, the locations of bomb craters can be used to determine high-risk zones for unexploded bombs. Note that this is usually not possible for properties situated in cities because the bomb craters are mostly covered by ruins of houses and therefore cannot be discerned in the pictures. The high-risk zones determined on the basis of the bomb craters are an important step in deciding where to search for unexploded bombs. Additionally, other aspects such as historical data from archives are also considered before the final decision is taken.

Up to now, high-risk zones have been defined in a way where only very little information from the data is used, namely the coordinates of every single observation, but no characteristics of the pattern in general.

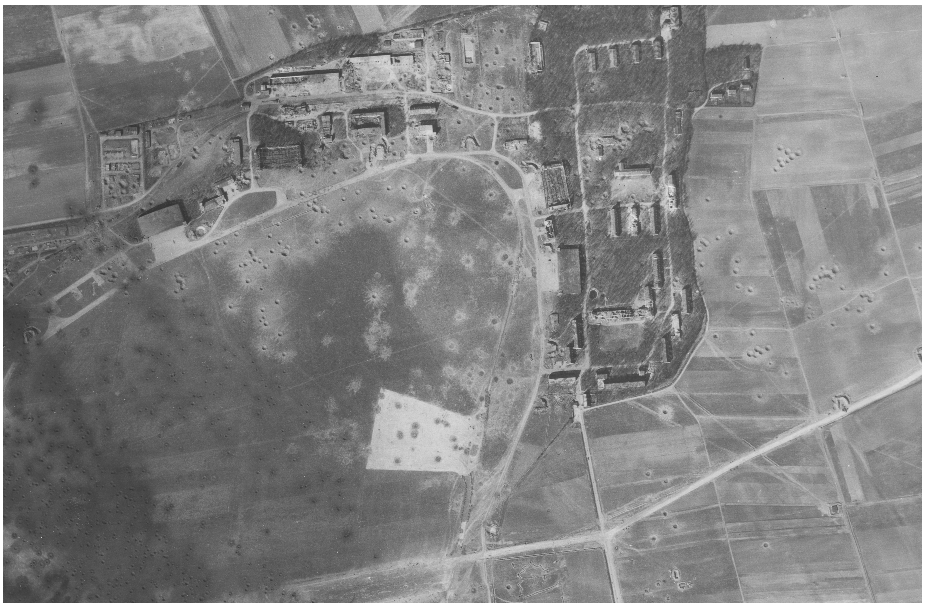
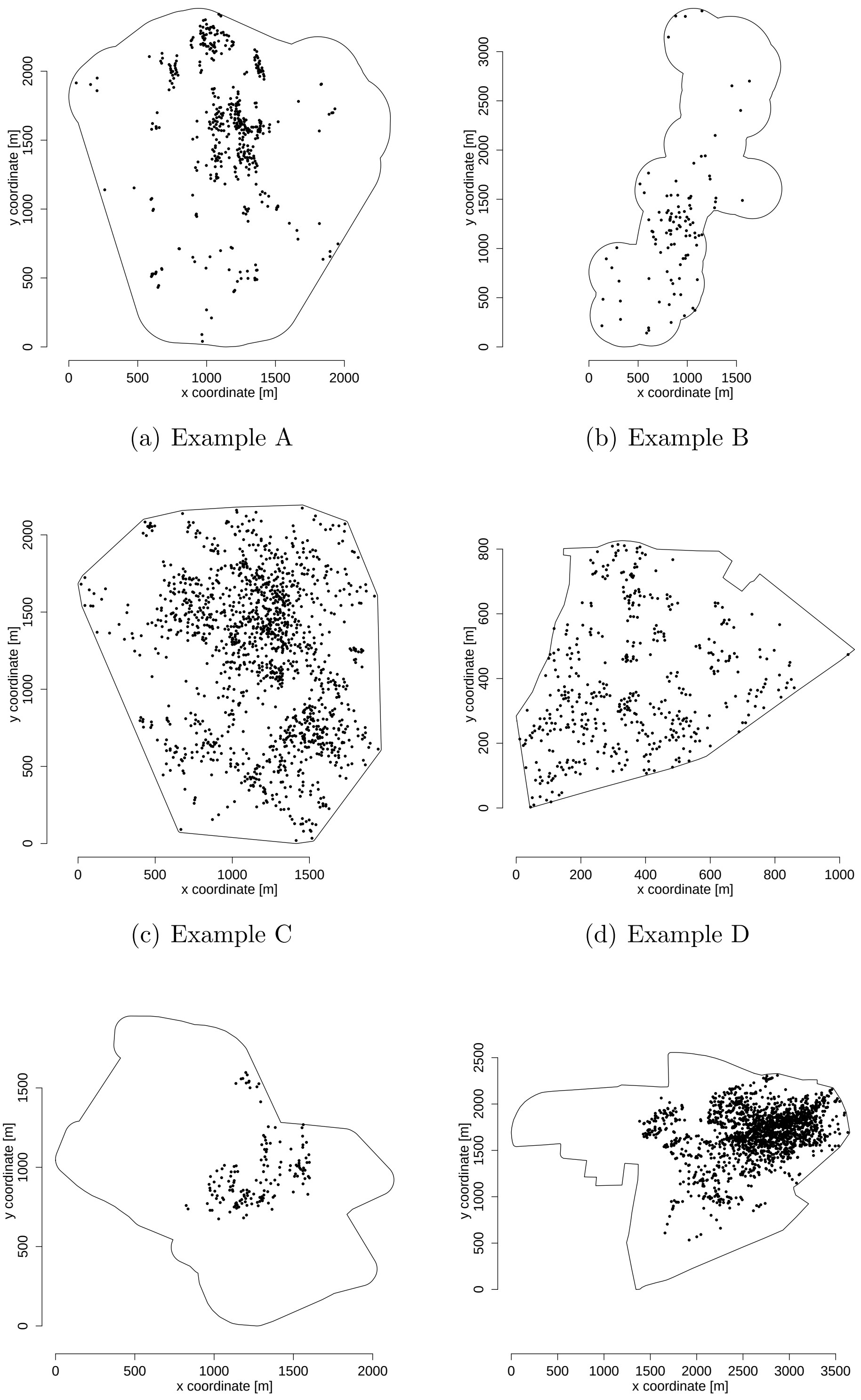


Figure 1.1.: Example of an aerial picture showing bomb craters (source: National Archive - Aerial Photo, Sortie 34-3658, Date 24.03.1945 (WW II Europe)).

Therefore, *Oberfinanzdirektion Niedersachsen* (OFD), which supports the removal of unexploded ordnance in federal properties in Germany, and *Mull und Partner Ingenieurgesellschaft*, who are experts in the analysis of aerial pictures, searched for more sophisticated approaches. In a cooperation project with the Statistical Consulting Unit (*Statistisches Beratungslabor*), Department of Statistics, Ludwig-Maximilians-Universität München, the task of evaluating existing methods and developing a novel approach was addressed. A further aim was to perform a risk assessment of the investigated areas, e.g. by estimating the probability that there are unexploded bombs outside the high-risk zone. The final goal of the cooperation was to develop a procedure which can be applied automatically by users who are not experts in statistics. The cooperation project started in May 2009 and was funded by *Oberfinanzdirektion Niedersachsen.*

Examples of the data are presented in Figure 1.2. The georeferenced locations of the bomb craters have been provided by *Mull und Partner.* As all information is derived from the aerial pictures, no data are available about the locations of unexploded bombs that have been found for the specific areas of interest. Example A comprises 443 observations of bomb craters in an area of approximately 400 *ha*. Example B consists of 104 observations in an area of approximately 350 *ha*. The bomb craters are mainly located in the southern part of the property. The 1369 observations of Example C are scattered over large parts of the property with an area of 334 *ha*. Example D consists of 451 observations on 52 *ha*. They seem to be more dense in the south of the property. The 152 bomb craters of Example E are concentrated on a rather small part of the property, which has an area of 239 *ha*. Example F comprises 1706 observations on 504 *ha*. Most of them are located in the north-east of the property.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 1.2.: Properties to be cleared: The solid lines represent the border of the areas for which data are available, the points illustrate the locations of bomb craters.

The properties to be cleared typically had military importance during war (i.e. barracks, airfields, military training areas) or were important from a strategic or economic point of view (i.e. rivers, floodgates, roads and industrial plants). Since detailed information on specific premise type and location could facilitate the identification of the property and hence have an economic impact, no further information other than relative coordinates on the specific property was provided. In general, it is not justified to assume several targets in a property. In some cases, like for Example B, the target of the attack even seems to be situated outside the property to be cleared.

The probability of non-explosion of every bomb can vary depending for example on the subsoil and the year of the attack. A well-established reference value is 0.1. In some cases, it is possible to estimate the probability of non-explosion for a given property from historical records. It is usually assumed to be constant on the whole property, as there is little data from which more specific assumptions could be derived.

The novel approach for constructing high-risk zones consists in interpreting the observed pattern of bomb craters as a realisation of a spatial point process. This point of view is widespread in the analysis of the locations of lightning strikes (Schabenberger and Gotway, 2005) or earthquakes (Vere-Jones, 1970; Choi and Hall, 1999). McDonald and Small (2006, 2009) used spatial point process methodology for analysing patterns of unexploded ordnance at former air force bombing ranges. Neyman and Scott (1972) mention patterns resulting from bombing as an example for a special type of spatial point process (see Section 2.5 for Neyman-Scott processes). These point process models had been used to optimise the formation bombing strategy for clearing land mines from the landing beaches in Normandy during World War II.

Interpreting the observed bomb crater patterns as realisations of spatial point processes provides a rich methodology to analyse the patterns and develop a construction method for high-risk zones: Various point process characteristics can be considered to investigate the properties of the observed patterns. In particular, the intensity function serves as basis for high-risk zones.

This thesis is organised as follows: Chapter 2 procures selected notation and properties of spatial point processes. The properties of the six real-data examples are investigated in an exploratory analysis via functional summary characteristics in Chapter 3 in order to find an appropriate model for the data. In Chapter 4, three methods for constructing high-risk zones are presented: The traditional method, the quantile-based method, which is not entirely new, but based on the considerations of *Mull und Partner*, and–as a completely novel approach–the intensity-based method. The behaviour of the construction methods is investigated in Chapter 5, which also contains a model check and a comparison of the theoretical properties of intensity-based and quantile-based high-risk zones. The chapter finishes with a recommendation for the intensity-based method. In Chapter 6, the risk associated with an intensity-based high-risk zone is assessed. As this risk does not reliably equal the parameter which is intended to specify it, a correction method is proposed. The consequences of spatial clustering are investigated in Chapter 7. A sensitivity analysis is performed and different types of models which account for clustering are fitted to the data. The R package `highriskzone` comprising an implementation of the main methods of this thesis is introduced in Chapter 8. Finally, a summary of the most important results and a review of open research questions is given in Chapter 9.

All analyses were performed by using the statistical software R (R Development Core Team, 2012). In particular, the R package spatstat (Baddeley and Turner, 2005, 2006) for the analysis of spatial point patterns was employed.

Parts of Chapters 3 to 7 have been published in an article in the Journal of the Royal Statistical Society (Series C) (Mahling et al., 2013). This article contains contributions by Michael Höhle and Helmut Küchenhoff. Most of the ideas are my own. I performed all analyses and wrote the article. Helmut Küchenhoff and Michael Höhle commented on the manuscript.

The R package highriskzone which is introduced in Chapter 8 was created by Heidi Seibold on the basis of my implementation of the methods for constructing and evaluating high-risk zones. The package is the major part of Heidi Seibold's bachelor thesis (Seibold, 2012), which was supervised by Helmut Küchenhoff and me jointly.

2. Spatial point processes

In this chapter, much of the notation that will be needed later on is introduced, as well as the most important concepts regarding spatial point processes. These aspects are discussed in a variety of books, such as Illian et al. (2008); Møller and Waagepetersen (2003); Schabenberger and Gotway (2005); Diggle (2003); Daley and Vere-Jones (1988); Cressie (1993); Gelfand et al. (2010); Ripley (1981, 1988). The history of the theory of point processes in one dimension is summarized in Daley and Vere-Jones (1988).

A particularity of spatial point patterns is that simulation is an important part of the analysis as many important characteristics cannot be determined explicitly, at least not for more complex models. Therefore, Monte Carlo tests as advocated by Diggle (1983, page 7) will frequently be used in later chapters.

Many parts of this chapter are essentially based on Illian et al. (2008), especially on Sections 1.5 and 1.6, 2.1 to 2.4, 3.4 and 6.1 to 6.4.

2.1. Introduction to spatial point processes

2.1.1. Definition and basic properties

Spatial point processes are "stochastic models of irregular point patterns" (Illian et al., 2008, page 23). A spatial point process $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ with $\mathbf{x}_i \in \mathbb{R}^d$ is a finite or infinite random set. The notation implies that all points are different and do not coincide (*assumption of simplicity*). A point pattern $\mathcal{X}$ is a realisation of the point process X. Note that only point processes on the plane will be considered in what follows, i.e. $d = 2$.

Random points in X will be called 'events' (alternatively, 'points' or 'sites'), whereas arbitrary points in $\mathbb{R}^d$, which may be in X or not, will be called 'locations' or 'positions'.

Møller and Waagepetersen (2003) give a more formal definition of point processes taking into account the measure theoretical background.

Counting measure

Let $W \subseteq \mathbb{R}^d$ be the observation window of the spatial point process X, i.e. the area for which data are available. The area of a set $\mathcal{B} \subseteq W$ is denoted by $\nu(\mathcal{B})$.

The counting measure $N_X(\mathcal{B})$ denotes the random number of points in a Borel set $\mathcal{B} \subseteq \mathbb{R}^d$. To keep the notation simple, the index X will usually be omitted. N is locally finite, i.e. $N(\mathcal{B}) < \infty \;\; \forall$ bounded sets $\mathcal{B}$. For disjoint sets $\mathcal{B}_1$ and $\mathcal{B}_2$,

$$N(\mathcal{B}_1 \cup \mathcal{B}_2) = N(\mathcal{B}_1) + N(\mathcal{B}_2)$$

holds (*property of additivity*).

Distributions

Every point process can be described by infinitely many random variables and their distributions. The most important of these are the *number distributions*

$$P(N(\mathcal{B}) = n) \text{ and } P(N(\mathcal{B}_1) = n_1, \dots, N(\mathcal{B}_k) = n_k)$$

and the *void probabilities*

$$P(N(\mathcal{B}) = 0).$$

2.1.2. Intensity

Intensity measure and intensity function

The intensity measure $\Lambda_X(\mathcal{B})$ equals the expected number of events in $\mathcal{B}$, $E\{N_X(\mathcal{B})\}$, and the intensity function $\lambda_X(\mathbf{s})$, which is defined via

$$\Lambda_X(\mathcal{B}) = \int_{\mathcal{B}} \lambda_X(\mathbf{x}) d\mathbf{x},$$

represents the probability of an event in an infinitesimal disc centered at a given location $\mathbf{s} \in W$. The intensity function exists under continuity conditions (for example, the points may not be arranged on a lattice). It is proportional to the point density around the location $\mathbf{s}$.

Papangelou conditional intensity

The Papangelou conditional intensity $\lambda(\mathbf{s}|\mathcal{X})$ is motivated as follows: The conditional probability for a point of X in an infinitesimal sphere containing the deterministic location $\mathbf{s}$, given the realisation $\mathcal{X}$, a point pattern, of X outside the sphere is $\lambda(\mathbf{s}|\mathcal{X})d\mathbf{s}$. Note that $E(\lambda(\mathbf{s}|X)) = \lambda(\mathbf{s})$.

Point density distribution function

The point density distribution function $G(t)$ describes the frequency of values of the intensity function:

$$G(t) = \frac{\nu(W_t)}{\nu(W)},$$

where $\nu(W)$ is the area of the observation window and $W_t = \{\mathbf{s} \in W : \lambda(\mathbf{s}) \leq t\}$.

Campbell theorem

For non-negative functions $f(\mathbf{x})$, the expected value of point process sums

$$S_f = f(\mathbf{x}_1) + f(\mathbf{x}_2) + \ldots = \sum_{(i)} f(\mathbf{x}_i) = \sum_{\mathbf{x}\in X} f(\mathbf{x}) = \int f(\mathbf{x}) N(d\mathbf{x})$$

can be computed as follows:

$$E(S_f) = E\left(\sum_{\mathbf{x}\in X} f(\mathbf{x})\right) = \int f(\mathbf{x})\lambda(\mathbf{x}) d\mathbf{x}.$$

2.1.3. Moments

Variance and Covariance

$$Var(N(\mathcal{B})) = E\left[(N(\mathcal{B}) - \Lambda(\mathcal{B}))^2\right] = E\left[(N(\mathcal{B}))^2\right] - \Lambda(\mathcal{B})^2$$

$$Cov(N(\mathcal{B}_1), N(\mathcal{B}_2)) = E\left[(N(\mathcal{B}_1) - \Lambda(\mathcal{B}_1)) \cdot (N(\mathcal{B}_2) - \Lambda(\mathcal{B}_2))\right]$$

Moment measure

The kth-order moment measure $\mu^{(k)}$ is defined by

$$\int_{\mathbb{R}^{nd}} f(\mathbf{x}_1, \ldots, \mathbf{x}_n) \mu^{(n)}(d(\mathbf{x}_1, \ldots, \mathbf{x}_n)) = E\left(\sum_{\mathbf{x}_1,\ldots,\mathbf{x}_n \in X} f(\mathbf{x}_1, \ldots, \mathbf{x}_n)\right),$$

where $f(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is any non-negative measurable function on $\mathbb{R}^{nd}$. It expresses expected values involving the counting measure:

$$\mu^{(k)}(\mathcal{B}_1 \times \ldots \times \mathcal{B}_k) = E(N(\mathcal{B}_1) \cdots N(\mathcal{B}_k)) \text{ and } \mu^{(k)}(\mathcal{B}^k) = E(N(\mathcal{B})^k)$$

In particular, it represents the kth moment of the real-valued random variable $N(\mathcal{B})$:

$$\mu^{(1)}(\mathcal{B}) = E(N(\mathcal{B})) = \Lambda(\mathcal{B}) \quad \text{and} \quad \mu^{(2)}(\mathcal{B}_1 \times \mathcal{B}_2) = E(N(\mathcal{B}_1) \cdot N(\mathcal{B}_2)), \quad \text{so}$$

$$Var(N(\mathcal{B})) = \mu^{(2)}(\mathcal{B} \times \mathcal{B}) - (\Lambda(\mathcal{B}))^2 \text{ and}$$

$$Cov(N(\mathcal{B}_1), N(\mathcal{B}_2)) = E\left[N(\mathcal{B}_1) \cdot N(\mathcal{B}_2)\right] - E(N(\mathcal{B}_1)) \cdot E(N(\mathcal{B}_2)) = \mu^{(2)}(\mathcal{B}_1 \times \mathcal{B}_2) - \Lambda(\mathcal{B}_1) \cdot \Lambda(\mathcal{B}_2).$$

Factorial moment measure

The kth-order factorial moment measure $\alpha^{(k)}$ is defined by

$$\int_{\mathbb{R}^{nd}} f(\mathbf{x}_1, \ldots, \mathbf{x}_k)\alpha^{(k)}(d(\mathbf{x}_1, \ldots, \mathbf{x}_k)) = E\left(\sum_{\mathbf{x}_1,\ldots,\mathbf{x}_k \in X}^{\neq} f(\mathbf{x}_1, \ldots, \mathbf{x}_k) \right).$$

In contrast to the kth-order moment measure, only all k-tuples of distinct points in X are considered. If $\mathcal{B}_1, \ldots, \mathcal{B}_k$ are pairwise disjoint:

$$\mu^{(k)}(\mathcal{B}_1 \times \ldots \times \mathcal{B}_k) = \alpha^{(k)}(\mathcal{B}_1 \times \ldots \times \mathcal{B}_k)$$

An important relation is obtained for $k = 2$:

$$\mu^{(2)}(\mathcal{B}_1 \times \mathcal{B}_2) = \Lambda(\mathcal{B}_1 \cap \mathcal{B}_2) + \alpha^{(2)}(\mathcal{B}_1 \times \mathcal{B}_2)$$

The kth-order factorial moment of a real-valued random variable $N(\mathcal{B})$ is $\alpha^{(k)}(\mathcal{B}^k)$:

$$\alpha^{(k)}(\mathcal{B}^k) = E\left[N(\mathcal{B}) \cdot (N(\mathcal{B}) - 1) \cdots (N(\mathcal{B}) - n + 1)\right]$$

Product density

The "frequency of possible configurations of k points" (Illian et al., 2008, page 32) is reflected by the product density $\rho^{(k)}$. Let $b_1, \ldots, b_k$ be pairwise disjoint discs with centres $\mathbf{x}_1, \ldots, \mathbf{x}_k$ and infinitesimal areas (or volumes) $dV_1, \ldots, dV_k$. The probability that there is a point of X in each of the discs $b_1, \ldots, b_k$ is

$$\rho^{(k)}(\mathbf{x}_1, \ldots, \mathbf{x}_k)dV_1, \ldots, dV_k.$$

The product density is defined if continuity properties are satisfied for $\alpha^{(k)}$, then

$$\alpha^{(k)}(\mathcal{B}_1 \times \ldots \times \mathcal{B}_k) = \int_{\mathcal{B}_1} \cdots \int_{\mathcal{B}_k} \rho^{(k)}(\mathbf{x}_1, \ldots, \mathbf{x}_k) d\mathbf{x}_1 \cdots d\mathbf{x}_k.$$

It coincides with $\lambda(\mathbf{x})$ for $k = 1$. If $\rho^{(2)}(\mathbf{x}_1, \mathbf{x}_2)$ depends only on the distance r of $\mathbf{x}_1$ and $\mathbf{x}_2$, the simplified notation $\rho(r)$ is used.

2.1.4. Stationarity and isotropy

Consider a point process $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ and the translated point process

$$X_{\mathbf{y}} = \{\mathbf{x}_1 + \mathbf{y}, \mathbf{x}_2 + \mathbf{y}, \ldots\}$$

which is obtained by shifting all points of the process X by the same vector $\mathbf{y}$. A point process X is *stationary* if X and $X_{\mathbf{y}}$ have the same distribution for all translations $\mathbf{y}$, i.e.

$$P(N_X(\mathcal{B}_1) = n_1, \ldots, N_X(\mathcal{B}_k) = n_k) = P(N_{X_\mathbf{y}}(\mathcal{B}_1) = n_1, \ldots, N_{X_\mathbf{y}}(\mathcal{B}_k) = n_k).$$

Such processes are also called *homogeneous*. The intensity measure of a stationary point process can be written in a simple form:

$$\Lambda(\mathcal{B}) = \lambda\nu(\mathcal{B})$$

A point process X is *isotropic* if its distributional properties are not affected by rotations around the origin, i.e. if X and the rotated point process $R_\alpha X$ have the same distribution for all angles α.

2.2. Complete spatial randomness: The Poisson point process

The Poisson point process or, more specifically, the homogeneous Poisson point process, is the most important point process model. It represents the case of complete spatial randomness and is often used as a null model which does not have any systematic structure. In addition, it can be used as a starting point for the construction of more complex point process models. Another important property is that many summary characteristics can be computed explicitly for this reference model (cf. Section 3).

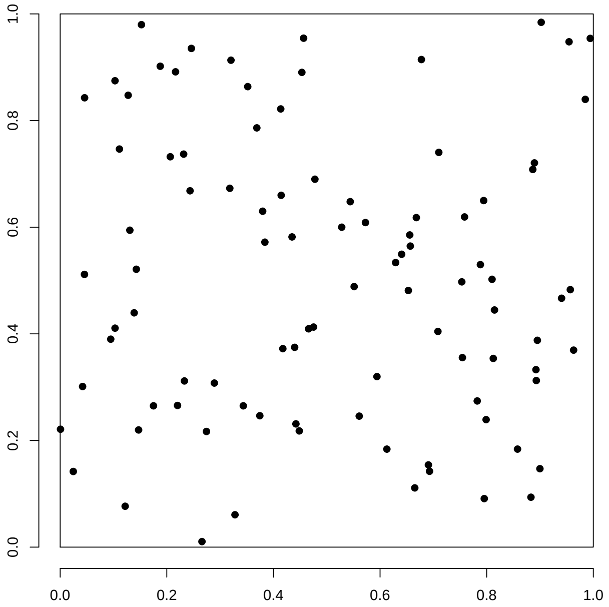


Figure 2.1.: Simulated Poisson process.

The homogeneous Poisson point process is defined as follows:

- The number of points in any bounded set $\mathcal{B}$ follows a Poisson distribution: $N(\mathcal{B}) \sim Po(\lambda\nu(\mathcal{B}))$.
- The numbers of points of X in k disjoint sets are independent for arbitrary k (*property of independent scattering*).

A simulated example of a Poisson point process on a unit square is depicted in Figure 2.1. Some points are very close to the border of the unit square. There are points which are located close to their neighbours and other points which are located quite far away from them.

The homogeneous Poisson point process is stationary and isotropic. Its intensity λ is constant and describes the mean number of point in a unit square. For known intensity, all types of distributions of the process can be determined. In particular, the one-dimensional distributions follow from the Poisson distribution of the point counts:

$$P(N(\mathcal{B}) = n) = \frac{\lambda^n(\nu(\mathcal{B}))^n}{n!}\exp(-\lambda\nu(\mathcal{B}))$$

A closely related model (which, however, does not exactly express complete spatial randomness) is the *binomial point process.* Here, the number n of points in W is fixed. The points $x_1, \dots, x_n$ are uniformly and independently distributed in the bounded observation window.

The number of points in $\mathcal{B} \subset W$ follows a binomial distribution:

$$P(N(\mathcal{B}) = k) = \binom{n}{k} p^k(1-p)^{n-k} \;\text{ with }\; p = \frac{\nu(\mathcal{B})}{\nu(W)} \;\text{ for }\; k = 0, \dots, n.$$

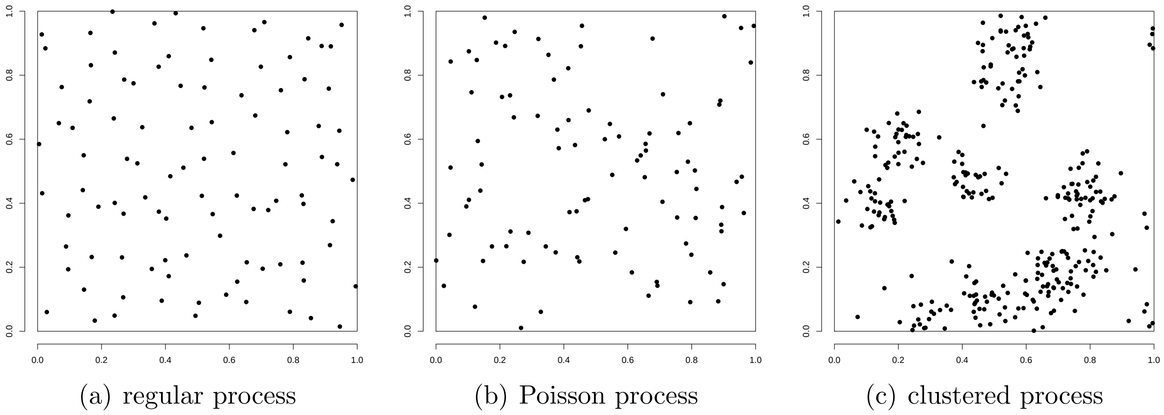
As the total number of points in W is fixed, the numbers of points in different subsets of W are not independent, but there is spatial correlation. This is the reason why the binomial point process–unlike the Poisson point process–is not a perfect model for complete spatial randomness. However, it can be shown that conditioning a Poisson process on a fixed number of points yields a binomial point process.

It is relatively easy to simulate a binomial point processes, especially if the observation window is rectangular. In this case, both the x and y coordinate for each of the n points are drawn from a continuous uniform distribution. If the shape of the observation window is more complicated, a binomial process can be simulated by using rejection sampling: Points are simulated in a rectangle containing the observation window W and are rejected if they are located outside W. In addition or as an alternative approach, the observation window can by approximated by a union of disjoint squares. For more details, see Illian et al. (2008, pages 64/65).

The simulation procedure for binomial point process can be made use of for simulating Poisson point processes. As a first step, a Poisson random number is drawn for the total number of points in the observation window. The second step consists in simulating a binomial point process with the specified number of points.

2.3. Clustering and regularity

A possible deviation from complete spatial randomness consists in interaction between the points. Two types of interaction are distinguished: The points can either repulse or attract each other. Repulsion leads to *regular* patterns. Sometimes the term *inhibition* is used instead of *regularity*. The distance from an arbitrary point to its nearest neighbour is typically large, the distance from an arbitrary location in the observation window to the nearest point of the process is roughly the same. A simulated example of such a regular pattern is depicted in Figure 2.2(a), next to a simulated example of a Poisson process in Figure 2.2(b).



(a) regular process (b) Poisson process (c) clustered process

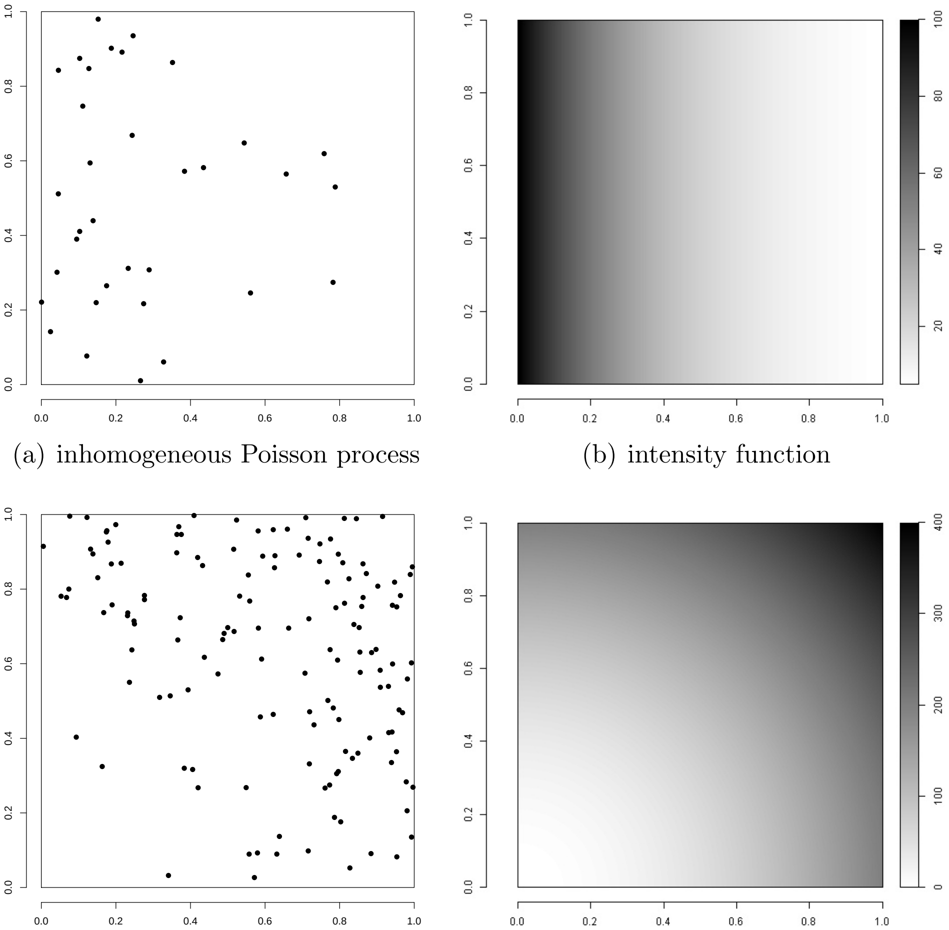
Figure 2.2.: Simulated examples of a regular process, a Poisson process and a clustered process.

The opposite type of interaction between points, attraction, yields *clustered* or *aggregated* patterns. A simulated example (a realisation of the so-called *Thomas process*) is shown in Figure 2.2(c). Here, the distance from an arbitrary point to its nearest neighbour is typically small, whereas the distance between an arbitrary location in the observation window and the nearest point of the process is typically large.

As we will see, regular patterns are not relevant in the bomb crater application and will therefore play a minor role in the following parts of this thesis. Models for clustered patterns are introduced in Section 2.5.2.

2.4. Inhomogeneity

Another possible deviation from complete spatial randomness is varying expectation for the point counts. The result are non-stationary processes, which are usually called *inhomogeneous*. An important model for such patterns is the *inhomogeneous Poisson point process*. It takes the varying expectation into account by means of a spatially varying intensity function $\lambda(\mathbf{s})$. Two simulated examples of inhomogeneous Poisson point processes are depicted in Figure 2.3, as well as the corresponding intensity functions. Inhomogeneous Poisson point processes will be discussed in detail in Section 2.5.1.

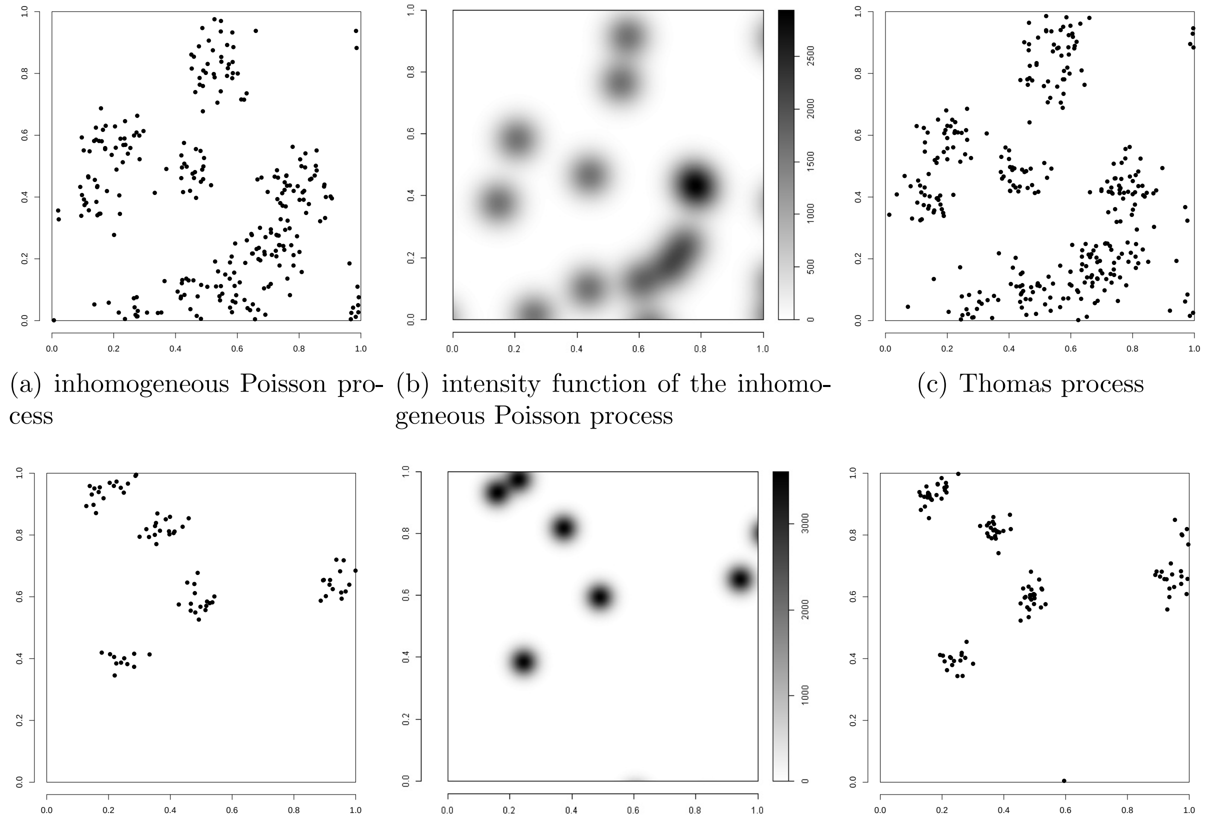


(a) inhomogeneous Poisson process (b) intensity function

(c) inhomogeneous Poisson process (d) intensity function

Figure 2.3.: Simulated examples of inhomogeneous Poisson processes and their intensity functions.

As the intensity function of the inhomogeneous Poisson point process varies in space, the resulting patterns may resemble clustered patterns although the reason for a higher point density in some areas is completely different: It results from a higher intensity in case of the inhomogeneous Poisson process, whereas it is induced by (mutual) attraction of the points for the clustered process. Figure 2.4 illustrates the possible resemblance of the resulting



(a) inhomogeneous Poisson process

(b) intensity function of the inhomogeneous Poisson process

(c) Thomas process

(d) inhomogeneous Poisson process

(e) intensity function of the inhomogeneous Poisson process

(f) Thomas process

Figure 2.4.: Simulated examples of inhomogeneous Poisson processes (with their intensity function) and of Thomas processes.

patterns for two simulated examples. The inhomogeneous Poisson processes have been simulated from the underlying intensity functions in Figures 2.4(b) and 2.4(e). The clustered patterns have been simulated as a so-called *Thomas process*, a popular model for clustered processes, which will be introduced in Section 2.5.2. Moreover, the relation between clustered and inhomogeneous patterns will be discussed in more detail in Section 2.5.3.

2.5. Spatial point process models

In this section, some of the most important point process models are introduced. A useful concept–which is not related to a specific model, but will be applied in Section 4 and can also be used for simulating inhomogeneous Poisson point processes and Cox processes–is thinning of point processes:

The random deletion of points in the process X yields a thinned process $Y \subset X$. There are three ways to delete points:

- *p-thinning:* Every point is deleted with probability $1 - p$ independently of all other points and the specific location. The parameter p is called *retention probability.* In a simulation approach, this kind of thinning can by performed by drawing Bernoulli random numbers.
- $p(\mathbf{s})$*-thinning:* The retention probability is not fixed, but given by a deterministic function $p(\mathbf{s})$ with $0 \leq p(\mathbf{s}) \leq 1$. Thus, the retention probability depends on the location.
- $P(\mathbf{s})$*-thinning:* $p(\mathbf{s})$ is random and based on a random field $P(\mathbf{s})$.

More general approaches include thinning depending on the configuration of the initial process X. As these are mainly relevant for regular patterns, they are not discussed further.

The following properties will be used later on:

- If $\lambda_X(\mathbf{s})$ is the intensity function of the initial process X and $p(\mathbf{s})$-thinning is applied, the intensity function of the thinned process Y is $\lambda_Y(\mathbf{s}) = p(\mathbf{s})\lambda_X(\mathbf{s})$.
- If the initial process X is an inhomogeneous Poisson process, the $p(\mathbf{s})$-thinned process Y is an inhomogeneous Poisson process as well, and so is $Z = X \backslash Y$. Y and Z are independent (Møller and Waagepetersen, 2003, page 23).

2.5.1. Inhomogeneous Poisson point process

An important model for (finite) non-stationary processes is the *inhomogeneous Poisson point process*, a generalisation of the homogeneous Poisson point process introduced in Section 2.2. The inhomogeneous Poisson point process is defined as follows:

- The number of points in any bounded set $\mathcal{B}$ follows a Poisson distribution:

$$N(\mathcal{B}) \sim Po\left(\int_{\mathcal{B}} \lambda(\mathbf{x})d\mathbf{x}\right)$$

- The numbers of points of X in k disjoint sets are independent for arbitrary k (*property of independent scattering*).

While the property of independent scattering remains unchanged compared to the homogeneous Poisson point process, the formerly constant intensity λ is replaced by an intensity function $\lambda(\mathbf{s})$ whose values depend on the location $\mathbf{s} \in W$.

Some of the properties of homogeneous Poisson point processes can be generalised for the inhomogeneous case. In particular, the one-dimensional distributions still follow from the Poisson distribution of the point counts:

$$P(N(\mathcal{B}) = n) = \frac{(\Lambda(\mathcal{B}))^n}{n!} \exp(-\Lambda(\mathcal{B})),$$

where $\Lambda(\mathcal{B}) = \int_{\mathcal{B}} \lambda(\mathbf{x}) d\mathbf{x}$.

The assumption of an inhomogeneous Poisson point process implies that beyond spatial variation in the intensity function, there is no stochastic dependence between observations.

An inhomogeneous Poisson point process Y with intensity function $\lambda_Y(\mathbf{s})$ can be simulated by applying $p(\mathbf{s})$-thinning to a homogeneous Poisson point process X with intensity λ_X, where

$$\lambda_X = \max_{\mathbf{s}} \lambda_Y(\mathbf{s}) \text{ and } p(\mathbf{s}) = \frac{\lambda_Y(\mathbf{s})}{\lambda_X}.$$

2.5.2. Cluster processes

The fundamental idea of *cluster processes* is that every point of a 'parent process' is replaced by a cluster of 'daughter points'. The union of these cluster points is the cluster process, whereas the parent points are usually unobserved (and often fictitious) and describe the cluster centres. They are usually not part of the cluster process.

A *Poisson cluster processes* (Bartlett, 1964; Møller and Waagepetersen, 2007; Illian et al., 2008) is obtained if the cluster centres form a Poisson process. The Poisson cluster process consists of the cluster points only (Schabenberger and Gotway, 2005).

Neyman-Scott processes (Neyman and Scott, 1958) are special Poisson cluster processes: Each cluster centre has a random number of offspring, the cluster points. The number of cluster points is independent and identically distributed with a discrete probability mass function. The positions of the cluster points relative to the cluster centres are independent and identically distributed according to a bivariate distribution function. Note that according to the definition of Cressie (1993), the cluster centres of a Neyman-Scott process may form an inhomogeneous Poisson point process, whereas most other authors postulate a homogeneous Poisson process (e.g. Stoyan et al. (1995)).

Popular Neyman-Scott processes are the Matérn cluster process (Matérn, 1960) and the Thomas process (Thomas, 1949): The cluster points of each cluster of a *Matérn cluster process* are independently uniformly distributed in a disc of radius R around the cluster centre. For the *Thomas process*, the positions of the cluster points relative to the cluster centres are given by an isotropic normal distribution with parameter σ. They will be considered in more detail in Section 7.2.

Cluster processes can be simulated in three steps:

1. The process of cluster centres is generated, e.g. as an (inhomogeneous) Poisson point process.
2. The number of offspring per cluster is simulated using a discrete distribution, e.g. the Poisson distribution.
3. The positions of the cluster points relative to the centres are determined using a bivariate distribution, e.g. a uniform distribution on a disc or a bivariate normal distribution.

2.5.3. Cox processes

An extension of the Poisson process is the *Cox process* (Cox, 1955), where the intensity $\lambda(\mathbf{s})$ is replaced by a non-negative random field $\Phi(\mathbf{s})$, the *intensity field.* Because of this second stochastic component, this process is frequently called 'doubly stochastic process'. Conditional on Φ, the Cox process is a Poisson process with intensity function Φ. If the intensity field $\Phi(\mathbf{s})$ is stationary, the resulting Cox process is stationary as well, whereas an inhomogeneous Poisson point process is not.

As the intensity field $\Phi(\mathbf{s})$ can take various forms, the class of Cox processes is large. Some of the most popular examples are the log-Gaussian Cox process (Møller et al., 1998), where $\Phi(\mathbf{s}) = \exp Z(\mathbf{s})$ and $Z(\mathbf{s})$ is a Gaussian random field, the Poisson-gamma random field Cox process (Ickstadt and Wolpert, 1997; Wolpert and Ickstadt, 1998), the Thomas process and the Matérn cluster process. The three latter models are examples of shot-noise Cox processes (see Møller (2003) and–for a generalisation–Møller and Torrisi (2005)). Other types of Cox processes can be obtained by $P(\mathbf{s})$-thinning of Poisson processes. The class of Cox processes also comprises many cluster processes.

In general, a Cox process can be simulated in two steps: First, a realisation of the random field is generated. The obtained intensity function is then used to simulate an inhomogeneous Poisson point process.

As conditional on Φ, the Cox process is a Poisson process, inhomogeneous Poisson processes and Cox processes cannot be distinguished when only one realisation is available (Møller and Waagepetersen, 2007). This is a fundamental problem which is not only relevant with respect to Cox processes: As Diggle et al. (2007) stated, there is a "fundamental ambiguity" between clustering and inhomogeneity: Both mechanisms generate patterns with aggregation, so they are difficult to distinguish–see also Ripley (1981), who states that inhomogeneous Poisson processes and cluster processes can "have identical distributions and so cannot be distinguished by any amount of data". Bartlett (1964) has shown that a Neyman-Scott process is identical to a Cox process if the number of offspring follows a Poisson distribution, so "no method of statistical analysis could discriminate between the two interpretations".

Given these problems in theory, Schabenberger and Gotway (2005) recommend to use the point process model which fits the subject-matter theory best.

In the following chapter, the bomb crater patterns are examined to find out which model should be used. The subject-matter theory will be considered in Section 3.6.

3. Properties of the bomb crater point pattern data

In this chapter, the properties of the observed patterns are investigated in order to find an appropriate model for the data.

The observation window W of a spatial point process X gives the area for which data are available. In the simplest case, it is identical to the property of interest (for more complex settings, see Chapters A and B in the Appendix). The process X now represents the locations of all bombs, exploded as well as unexploded. However, only a thinned version Y of the full process X, namely the exploded bombs in form of the bomb craters, has been observed. It consists of the $N_Y(W) = n_Y$ observations whose coordinates can be derived from the aerial pictures. The process of unexploded bombs, $Z = X \backslash Y$, is unobserved, i.e. its locations are unknown. The probability of non-explosion for every bomb q is assumed to be homogeneous in W, which means that every $\mathbf{x} \in X$ is element of Z with probability q, regardless of its location $\mathbf{s} \in W$ and independently of the behaviour of the other elements of X.

To find an appropriate model for the data depicted in Figure 1.2, second-order characteristics such as Ripley's K-function and the pair correlation function, as well as nearest-neighbour and empty-space characteristics will be considered. The intensity is a fundamental first-order characteristic. The largest average intensity was observed for Example D (0.00087), the second largest for Example C (0.00041). For Examples B (0.000030) and E (0.000064), the average intensity was much smaller. An intermediate intensity was obtained for Examples A (0.000108) and F (0.000338). The estimation of intensity functions varying in space is discussed in Section 4.3.2.

The following explanations are mainly based on Illian et al. (2008), Chapters 1.7, 4.1–4.3 and 4.10. The general use of summary functions for spatial point processes is discussed in Cressie (1993), Ripley (1981) and Diggle (2003). No numerical summaries such as indices will be considered, but functional summaries. These contain more information and are therefore also useful for fitting models.

Summary characteristics can be classified as location-related (like the empty-space distribution function) or point-related, as the nearest-neighbour distribution function or Ripley's K-function. For the proper definition of point-related characteristics, Palm distributions are needed.

Palm characteristics are probabilities or expected values referring to individual points in the process, for example the expected number of points in a disc of radius r centered at $\mathbf{x}$ (where $\mathbf{x}$ is not counted) or the probability that there is at least one further point in such a disc (Illian et al., 2008, page 177). In the stationary case, it is possible to define these characteristics independently of the particular position of the point $\mathbf{x}$. In consequence, the origin o or a so-called *typical point* is considered and the characteristics are written as

$$P_o\left[N(b(o,r)\backslash\{o\})>0\right]$$

and

$$E_o\left[N(b(o,r)\backslash\{o\})\right].$$

The exact definitions are

$$\lambda\nu(W)P_o(X\in\mathcal{A})=E\left[\sum_{\mathbf{x}\in X\cap W}\mathbb{1}_{\mathcal{A}}(X-\mathbf{x})\right],$$

where W is a 'test set', $X\in\mathcal{A}$ means that the process X has property $\mathcal{A}$ and $X-\mathbf{x}$ is the shifted process, and

$$\lambda\nu(W)E_o\left[\mathcal{S}(X)\right]=E\left[\sum_{\mathbf{x}\in X\cap W}\mathcal{S}(X-\mathbf{x})\right],$$

where $\mathcal{S}(X)$ is a number assigned to X.

Details can be found in Stoyan et al. (1987), Chapter 2.4.3. The main difficulty with Palm characteristics is that probabilities such as $P_o(X\in\mathcal{A})$ can be interpreted as a conditional probability given that there is a point in o, but this conditional probability cannot be defined in the classical way, as the probability that there is a point in o is zero for stationary processes.

The *Campbell-Mecke formula* is a version of the Campbell theorem in which the function that is considered does not only depend on $\mathbf{x}$, but also on other points of the process X:

$$E\left(\sum_{\mathbf{x}\in X}f(\mathbf{x},X)\right)=\lambda\int E_o\left[f(\mathbf{x},X_{-\mathbf{x}})\right]d\mathbf{x}=\lambda E_o\left[\int f(\mathbf{x},X_{-\mathbf{x}})d\mathbf{x}\right].$$

Classical summary characteristics are defined for stationary processes. This implies the assumption that the pattern is infinite and can be continued outside the observation window in the same way. However, only the points inside the window are known, which results in *edge effects.* Illian et al. (2008) summarize edge-correction methods in Chapter 4.2.2. The most important methods for the following sections can be classified as follows:

- The *border method* (also called *minus sampling*) can be illustrated for a situation where only neighbours within a distance r are relevant, which is the case when the expected number of points in a disc of radius r centered at the typical point or the probability that there is a point in this disc is estimated. Then, a possible approach is to use only the points whose distance from the boundary ∂W of the observation window is larger than r. These points are located in

$$W_{\ominus r} = \{\mathbf{s} \in W : b(\mathbf{s}, r) \subseteq W\},$$

where $b(\mathbf{s}, r)$ denotes a disc of radius r centered at $\mathbf{s}$. Only the points in $W_{\ominus r}$, which is a subset of W, are taken into account for the estimation. The points in $W \backslash W_{\ominus r}$, however, are used to determine the correct distances to the nearest neighbours of the points in $W_{\ominus r}$ or the correct numbers of points inside the given discs.

- A more sophisticated approach is the *nearest-neighbour edge-correction*: Only those points are taken into account for estimation whose nearest-neighbour distance $d(\mathbf{x})$ is shorter than their distance $e(\mathbf{x})$ to the boundary ∂W. A weight $1/\nu(W_{\ominus d(\mathbf{x})})$, where $W_{\ominus d(\mathbf{x})} = W \ominus b(o, d(\mathbf{x})) = \{\mathbf{s} \in W : b(\mathbf{s}, d(\mathbf{x})) \subseteq W\}$, is attached to every point.

- The typical structure of estimators for second-order characteristics consists in double sums of pairs of points in W. For *second-order edge-corrections*, pairs of points $(\mathbf{x}_1, \mathbf{x}_2)$ with a large inter-point distance–where both $\mathbf{x}_1$ and $\mathbf{x}_2$ are located in W–are attributed large weights. The first type of weights is called *stationary* or *translational*: $1/\nu(W_{\mathbf{x}_1} \cap W_{\mathbf{x}_2})$, where $W_{\mathbf{x}} = \{\mathbf{z} + \mathbf{x} : \mathbf{z} \in W\}$ denotes the translated window. For the second type of weights, *isotropic* or *rotational edge-correction* is obtained: $1/w(\mathbf{x}_1, \mathbf{x}_2)$, where $w(\mathbf{x}_1, \mathbf{x}_2)$ is the boundary length in W of $b(\mathbf{x}_1, ||\mathbf{x}_1 - \mathbf{x}_2||)$, divided by the circle perimeter length $2\pi||\mathbf{x}_1 - \mathbf{x}_2||$.

The border method was introduced by Ripley (1977). Hanisch (1984) introduced a nearest-neighbour estimator which uses nearest-neighbour edge-correction.

Ripley (1988, page 32) explains the isotropic correction as follows: Pairs $(\mathbf{x}_1, \mathbf{x}_2)$ are counted $k(\mathbf{x}_1, \mathbf{x}_2)$ times, where $\frac{1}{k(\mathbf{x}_1, \mathbf{x}_2)}$ is the proportion of the perimeter of the circle $\partial b(\mathbf{x}_1, d(\mathbf{x}_1, \mathbf{x}_2))$ which is in the window (see also Hanisch (1983), Ohser (1983), Ripley (1976) and Ohser and Stoyan (1981) for second-order edge-corrections).

The problem of spatial censoring and edge correction is discussed in detail in Baddeley (1999) and Ripley (1988, Chapter 3), who presents corrections for nearest-neighbour methods and 'interpoint distance methods' such as the K-function. Moreover, he gives asymptotic variances for edge-corrected estimates and limit theorems for interpoint distances. Baddeley (1999) distinguishes two types of edge effects, sampling bias (which results if size and shape of a certain object influence the probability that this object is included in the sample) and censoring effects (i.e. not the full extent of a geometrical object partially located within the window can be observed). Edge effects are especially severe if the window is small or has a complex shape. Possible strategies for correction include data-dependent weighting and methodology from survival analysis. Concerning sampling bias, the border method can be used. Alternatively, 'real' edge corrections are proposed, weighted empirical distributions of distances between points which can be derived from the Campbell-Mecke formula and are Horvitz-Thompson style estimators (e.g. translation correction and Ripley's isotropic correction; see Horvitz and Thompson (1952) for Horvitz-Thompson estimators in general). Concerning censoring effects, the analogy between censoring and edge effects can be made use of to obtain Kaplan-Meier and Hanisch-type estimators.

The Kaplan-Meier estimator (Baddeley and Gill, 1997) is more efficient than the border method for the empty-space function and for the nearest-neighbour distance distribution function. For Ripley's K-function, however, the Kaplan-Meier estimator is less efficient than sophisticated edge corrections using weights which are reciprocal to the observation probability. The Chiu-Stoyan estimator (Chiu and Stoyan, 1998) is based on the nearest-neighbour estimator proposed by Hanisch (1984). Its adaption to the empty-space function yields a new estimator which is closely related to the Kaplan-Meier estimator, as Chiu and Stoyan (1998) have shown.

Illian et al. (2008) recommend to check the assumption of stationarity before applying the following summary characteristics. Moreover, the window should be adapted (e.g. the pair-correlation function might exhibit a strange behaviour for large arguments if an inappropriate window is used). However, summary characteristics can formally be applied to non-stationary patterns (Illian et al., 2008, page 280). Therefore, the assumption of stationarity will not be checked, but summary characteristics for inhomogeneous processes will be applied additionally. As the observation windows have a special meaning for the bomb crater patterns (they represent the properties to be cleared), they will not be adapted.

Some of the results for Examples A and B have been shown in Mahling et al. (2013).

3.1. Empty-space function

The *empty-space function* (also called *spherical contact distribution function* or *point-to-nearest-event distribution function*)

$$H_s(r) = 1 - P(N(b(o,r)) = 0)$$

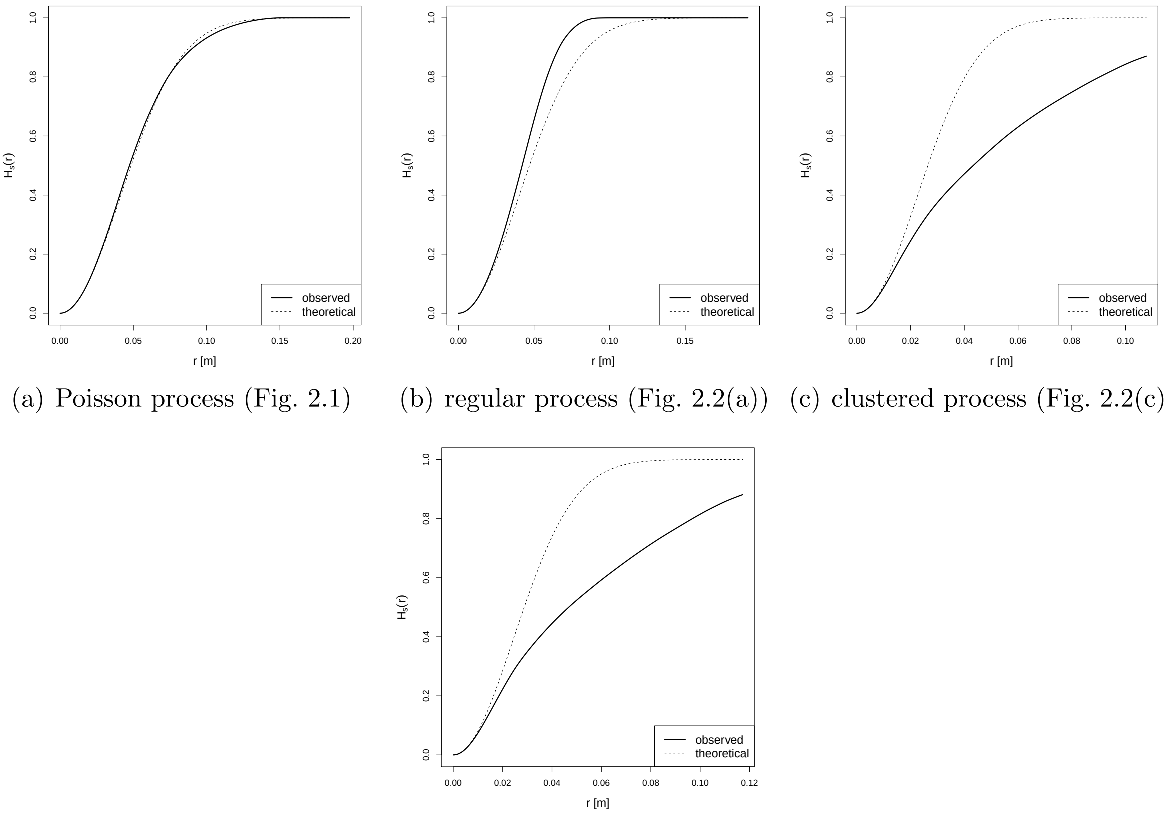
for $r \geq 0$ is a location-related summary characteristic. It is "the distribution of the distance from an 'arbitrary' test location to its nearest neighbour" (Illian et al. (2008), p. 200). It describes the distribution of the smallest radius for a disc at the origin o to a contact point in the spatial point process X.

For a homogeneous Poisson point process, the empty-space function is

$$H_s(r) = 1 - \exp(-\lambda \pi r^2).$$

Smaller values suggest clustering, whereas larger values suggest regularity. This is illustrated in Figure 3.1, where estimated empty-space functions are shown for some of the simulated patterns from Chapter 2. Note that $H_s(r)$ also takes smaller values than $1 - \exp(-\lambda \pi r^2)$ for inhomogeneous Poisson processes.

In `spatstat`, three types of correction are available, namely the border method estimator, the Kaplan-Meier estimator and the Chiu-Stoyan estimator. The Kaplan-Meier estimator is recommended.



(a) Poisson process (Fig. 2.1)

(b) regular process (Fig. 2.2(a))

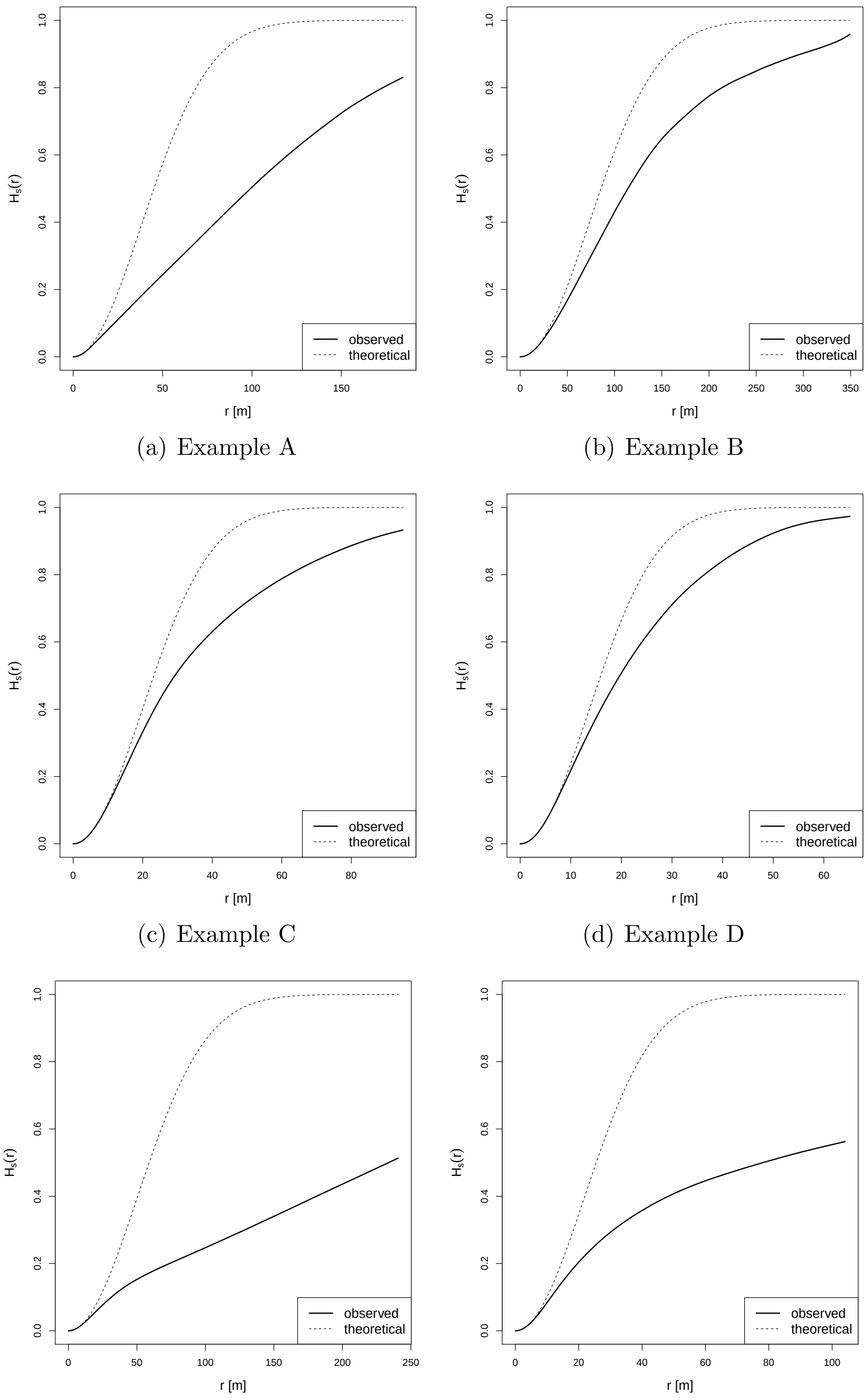
(c) clustered process (Fig. 2.2(c))

(d) inhomogeneous Poisson process (Fig. 2.4(a))

Figure 3.1.: Empty-space function: The solid lines represent the estimated empty-space functions for the simulated patterns, the dashed lines correspond to the theoretical empty-space functions of a homogeneous Poisson process.

Two discrete approximations are used: The observation window W is discretized on a regular lattice and $[0, r]$ is replaced by a large number of values from this interval. Finally, the estimator is obtained by a distance transform algorithm of image processing, where the Euclidean metric is discretely approximated (Borgefors, 1986).

The estimated empty-space functions (Kaplan-Meier estimator) for Examples A to F are depicted in Figure 3.2. For all six examples, the values for the observed pattern are smaller than the theoretical values for a homogeneous Poisson process, especially for Examples E and F, whose points are scattered on a small part of the window.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 3.2.: Empty-space function: The solid lines represent the estimated empty-space functions for the observed patterns, the dashed lines correspond to the theoretical empty-space functions of a homogeneous Poisson process.

3.2. Nearest-neighbour distance distribution function

The nearest-neighbour distance distribution function

$$D(r) = P_o(N(b(o,r)\setminus\{o\}) > 0)$$

for $r \geq 0$ describes the random distance from a typical point to its nearest neighbour and is a point-related summary characteristic. As only the nearest neighbour is considered, it is "short-sighted" and cannot describe the behaviour of the process at large distances (Illian et al., 2008, page 207).

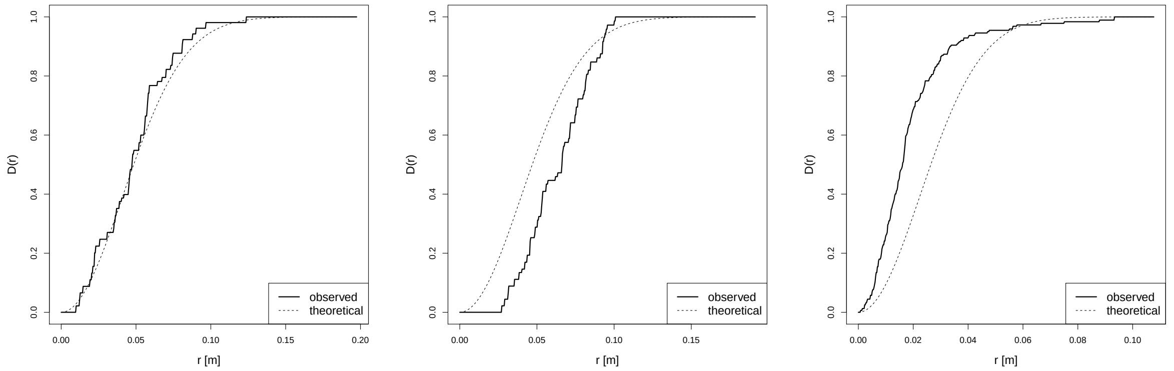
For a homogeneous Poisson point process, the nearest-neighbour distance distribution function is

$$D(r) = 1 - \exp(-\lambda\pi r^2).$$

Larger values suggest clustering, whereas smaller values suggest regularity. As a consequence,

- $D(r) = H_s(r)$ for a Poisson process,
- $D(r) \leq H_s(r)$ for a clustered process and
- $D(r) \geq H_s(r)$ for a regular process.

The estimated nearest-neighbour distribution functions for some of the simulated patterns from Chapter 2 are shown in Figure 3.3.

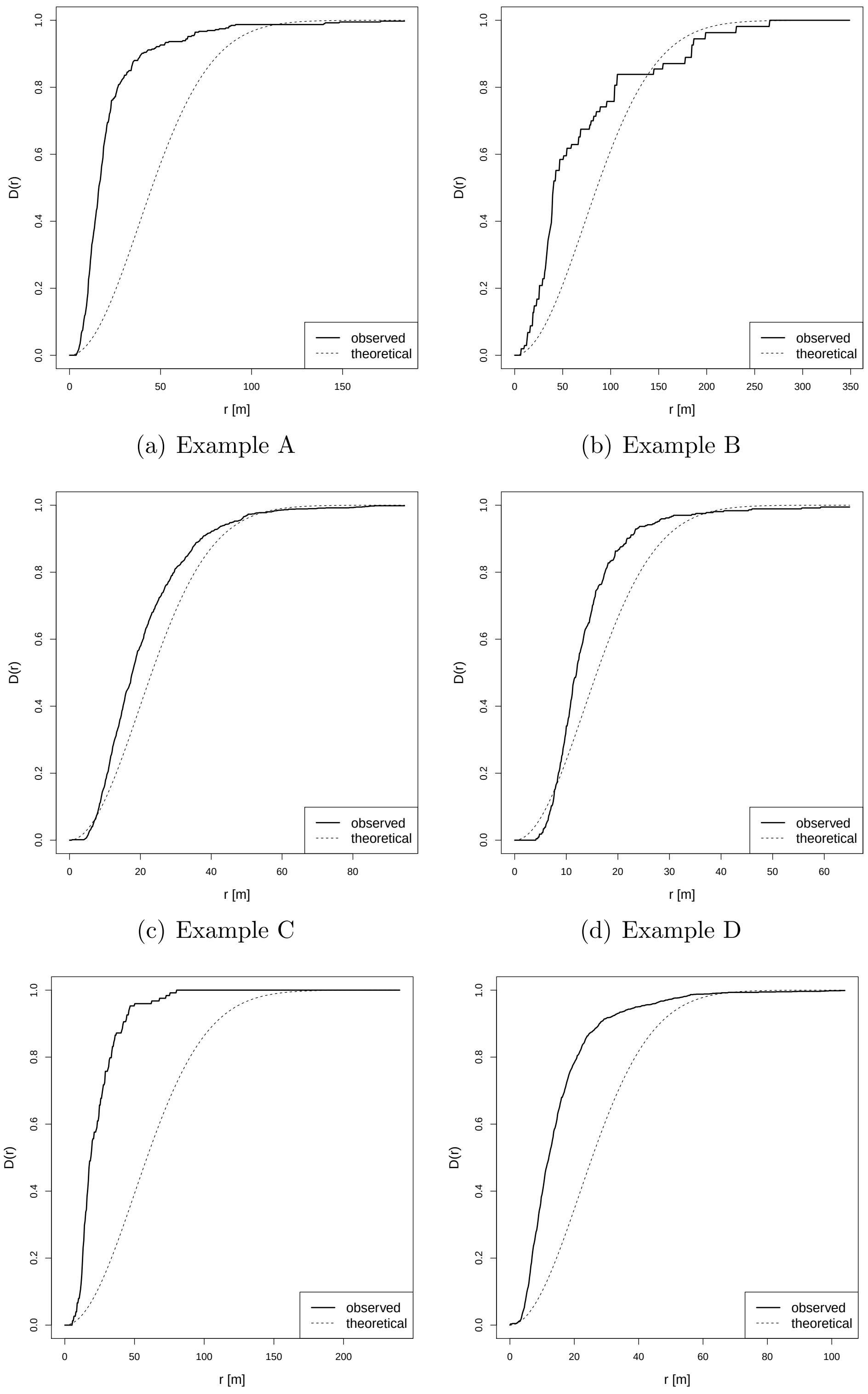


(a) Poisson process (Fig. 2.1) (b) regular process (Fig. 2.2(a)) (c) clustered process (Fig. 2.2(c))

Figure 3.3.: Nearest-neighbour distance distribution function: The solid lines represent the estimated nearest-neighbour distance distribution functions for the simulated patterns, the dashed lines correspond to the theoretical nearest-neighbour distance distribution functions of a homogeneous Poisson process.

In spatstat, three estimators are implemented: The border estimator, the nearest-neighbour estimator and the Kaplan-Meier estimator. Estimation is performed based on histogram counts.

The estimated nearest-neighbour distance distribution functions (Kaplan-Meier estimator) for Examples A to F are depicted in Figure 3.4. For all six examples, the values for the observed pattern are larger than the theoretical values for a homogeneous Poisson process for most values of r.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 3.4.: Nearest-neighbour distribution function: The solid lines represent the estimated nearest-neighbour distribution functions for the observed patterns, the dashed lines correspond to the theoretical nearest-neighbour distribution functions of a homogeneous Poisson process.

3.3. J-function

The J-function

$$J(r) = \frac{1 - D(r)}{1 - H_s(r)} \quad \text{for } r \leq 0 \text{ with } H_s(r) < 1$$

has been introduced by Van Lieshout and Baddeley (1996). It “compares the environment of a typical random point of the process with the environment of a fixed arbitrary point” in form of the ration of the probabilities that there is no further point within the distance r of a given point or location. Van Lieshout and Baddeley (1996) have shown that $J(r)$ is constant for distances r which are larger than the range of spatial interaction for stationary processes whose Papangelou conditional intensity exists. They have also shown that

- $J(r) \equiv 1$ for a Poisson process,
- $J(r) \leq 1$ for a clustered process and
- $J(r) \geq 1$ for a regular process.

The J-function is not invariant under thinning (Van Lieshout and Baddeley, 1996).

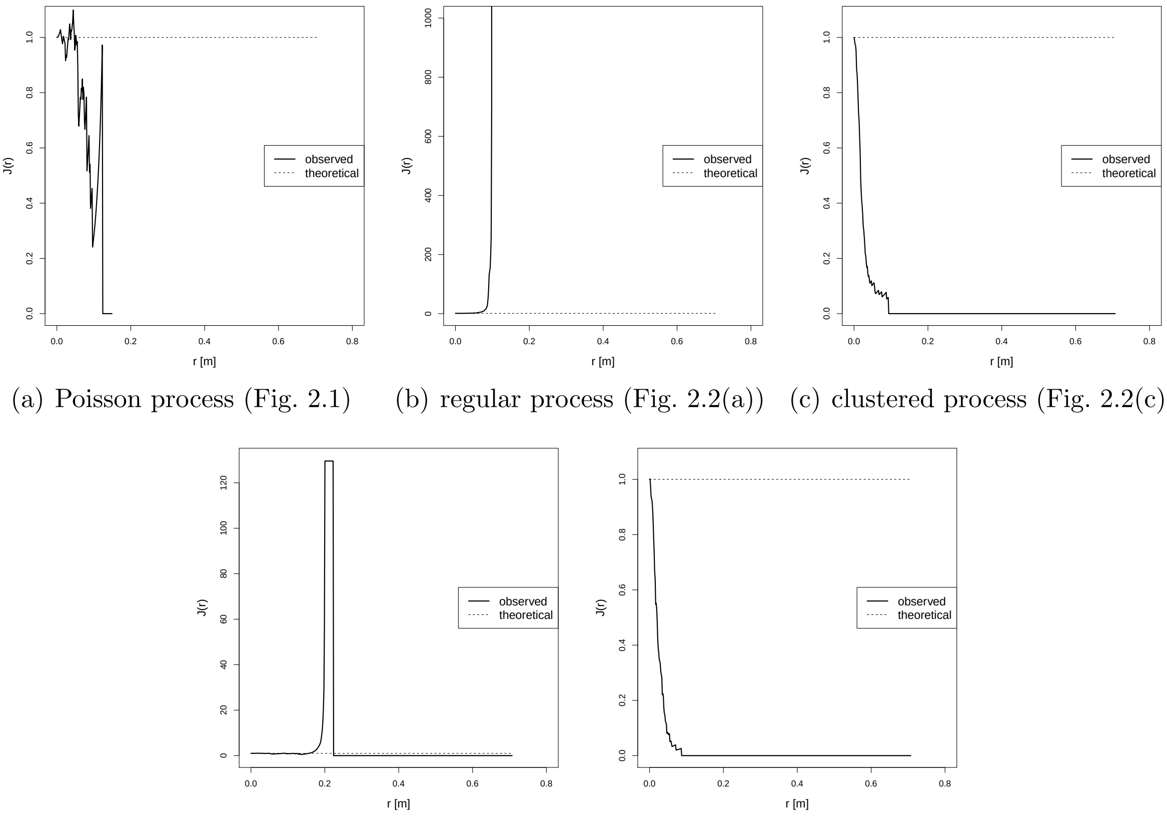
Thönnes and Van Lieshout (1999) found in simulations that the J-function is a “competitive alternative” to the nearest-neighbour distance distribution function and the empty-space function for testing complete spatial randomness. As advantages compared to those, they name that the J-function can be evaluated explicitly for a larger class of models and that it measures type, strength and range of spatial interaction.

Bedford and van den Berg (1997) have shown that there are non-Poisson processes with $J(r) \equiv 1$. They considered the one-dimensional case and showed that there are even processes with $J(r) \equiv 1$ whose interpoint distances are bounded (and hence not exponentially distributed).

Baddeley, Kerscher, Schladitz, and Scott (2000) have shown that the J-function is insensitive to edge effects: The uncorrected estimator is approximately unbaised for Poisson point processes. For testing complete spatial randomness, the uncorrected estimator yields at least as powerful tests as the corrected estimators.

The estimation of the J-function is difficult, as it is a combination of two characteristics of different nature (point-related vs. location-related). The denominator is small for large r, which results in fluctuations of the estimator (Illian et al., 2008, page 213). Van Lieshout and Baddeley (1996) recommend plugging in estimators of $H_s(r)$ and $D(r)$ obtained by comparable methods, especially using Kaplan-Meier estimators for both function as these have the advantage that they are proper distribution functions. As Figure 3.5 shows, even this procedure may result in questionable estimates, as for the Poisson process from Figure 2.1, whose estimated J-function is depicted in Figure 3.5(a) and which should be close to 1, or for the inhomogeneous Poisson process from Figure 2.3(a).

The estimated J-functions (Kaplan-Meier estimator for both nearest-neighbour distance distribution function and empty-space function) for Examples A to F are depicted in Figure 3.6. For all six examples, the values for the observed pattern are smaller than the



(a) Poisson process (Fig. 2.1)

(b) regular process (Fig. 2.2(a))

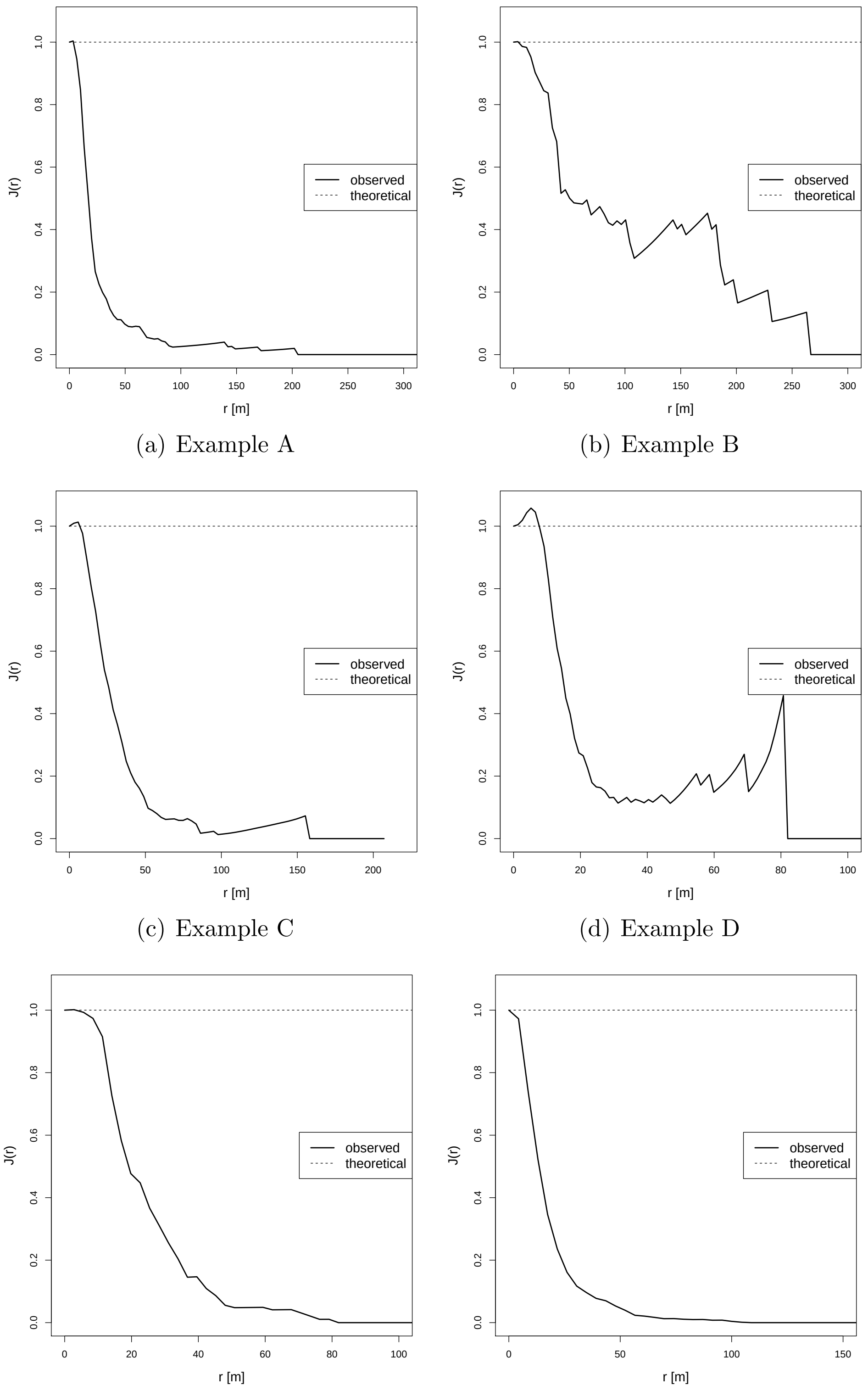
(c) clustered process (Fig. 2.2(c))

(d) inhomogeneous Poisson process (Fig. 2.3(a))

(e) inhomogeneous Poisson process (Fig. 2.4(a))

Figure 3.5.: J-function: The solid lines represent the estimated J-functions for the simulated patterns, the dashed lines correspond to the theoretical J-functions of a homogeneous Poisson process.

theoretical values for a homogeneous Poisson process. Only for Examples C and D, a value larger than 1 is observed for a small value of r. For Example A, the J-function is constant for $r > 210$. The range of spatial interaction is about 270 for Example B, 160 for Example C, 83 for Example D and 82 for Example E. The J-function of Example F is constant for $r > 110$, but it is already almost constant from $r = 70$ on.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 3.6.: J-function: The solid lines represent the estimated J-functions for the observed patterns, the dashed lines correspond to the theoretical J-functions of a homogeneous Poisson process.

3.4. Ripley's K-function

A well-established tool for investigating homogeneous spatial point patterns is Ripley's K-function (Ripley, 1977)

$$K(r) = \frac{1}{\lambda} E_o(N(b(o,r)\backslash\{o\})),$$

defined for $r \geq 0$, where λ is the intensity of the stationary process. Thus, multiplied with λ, the K-function corresponds to the expected number of other points within distance r from the typical point o. The division by λ is performed to separate out the global point density and local point density fluctuations (Illian et al. (2008), p. 215).

Ripley (1977) introduces the K-function as follows: "$\lambda^2 K(t)$ is the expected number of ordered pairs of distinct points less than distance t apart with the first point in a given set of unit area" and "$\lambda K(t)$ is the expected number of further points within t of an arbitrary point of the process".

For a Poisson process, $K(r) = \pi r^2$, larger values are obtained for cluster processes, smaller values for regular processes. Examples for the simulated patterns from Chapter 2 are shown in Figure 3.7.

As the K-function is a cumulative characteristic, its interpretation is complicated. It is rather difficult to see for which values of r deviations from the case of complete spatial randomness are observed.

The K-function is the first second-order characteristic presented in this thesis. Second-order characteristics for stationary processes can be motivated as follows (Illian et al., 2008, pages 223–225): The second-order factorial moment measure $\alpha^{(2)}(\mathcal{B}_1 \times \mathcal{B}_2)$ gives the mean number of pairs $\mathbf{x}_1 \neq \mathbf{x}_2$ with $\mathbf{x}_1 \in \mathcal{B}_1$ and $\mathbf{x}_2 \in \mathcal{B}_2$. If $\mathcal{B}_1$ and $\mathcal{B}_2$ are disjoint, then

$$\alpha^{(2)}(\mathcal{B}_1 \times \mathcal{B}_2) = E\left[N(\mathcal{B}_1) \cdot N(\mathcal{B}_2)\right].$$

If $\mathcal{B}_1 = \mathcal{B}_2 = \mathcal{B}$:

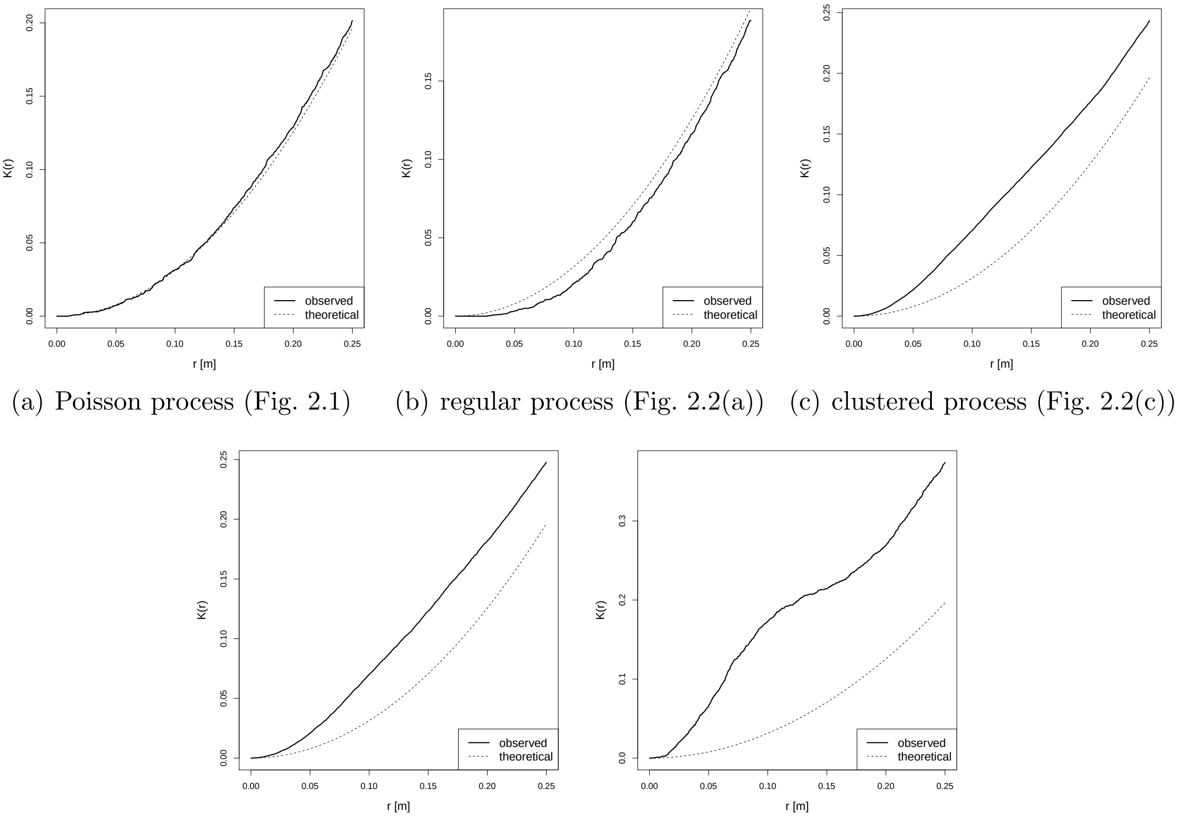
$$\alpha^{(2)}(\mathcal{B} \times \mathcal{B}) = E\left[N(\mathcal{B})(N(\mathcal{B}) - 1)\right] = E\left[N(\mathcal{B})^2\right] - E\left[N(\mathcal{B})\right].$$

So $Var(N(\mathcal{B}))$ can be written as follows:

$$Var(N(\mathcal{B})) = \alpha^{(2)}(\mathcal{B} \times \mathcal{B}) + E\left[N(\mathcal{B})\right] - \left(E\left[N(\mathcal{B})\right]\right)^2 = \alpha^{(2)}(\mathcal{B} \times \mathcal{B}) + \lambda\nu(\mathcal{B}) - (\lambda\nu(\mathcal{B}))^2$$

As $Cov(N(\mathcal{B}_1), N(\mathcal{B}_2))$ can be expressed in terms of $\alpha^{(2)}$ and λ as well, the intensity and the second-order factorial moment measure completely describe the second-order behaviour. Therefore, it is useful to derive simpler expressions for the second-order factorial moment measure containing the second-order product density ρ and the *reduced second-order moment measure* $\mathcal{K}$. It can be shown that

$$\alpha^{(2)}(\mathcal{B}_1 \times \mathcal{B}_2) = \int_{\mathcal{B}_1} \int_{\mathcal{B}_2 - \mathbf{x}} \rho(\mathbf{h}) d\mathbf{h} d\mathbf{x}$$



(a) Poisson process (Fig. 2.1)

(b) regular process (Fig. 2.2(a))

(c) clustered process (Fig. 2.2(c))

(d) inhomogeneous Poisson process (Fig. 2.4(a))

(e) inhomogeneous Poisson process (Fig. 2.4(d))

Figure 3.7.: K-function: The solid lines represent the estimated K-functions for the simulated patterns, the dashed lines correspond to the theoretical K-functions of a homogeneous Poisson process.

with $\mathbf{h} = \mathbf{x}_1 - \mathbf{x}_2$ and

$$\alpha^{(2)}(\mathcal{B}_1 \times \mathcal{B}_2) = \lambda^2 \int_{\mathcal{B}_1} \mathcal{K}(\mathcal{B}_2 - \mathbf{x}) d\mathbf{x},$$

where $\mathcal{K}$ is the *reduced second-order moment measure* defined by

$$\lambda \mathcal{K}(\mathcal{B}) = E_o(N(\mathcal{B} \backslash \{o\})).$$

$$\Rightarrow \lambda^2 \mathcal{K}(\mathcal{B}) = \int_{\mathcal{B}} \rho(\mathbf{h}) d\mathbf{h}$$

As one can see easily, Ripley's K-function is $K(r) = \mathcal{K}(b(o, r))$ for $r \geq 0$.

The typical second-order characteristics are Ripley's K-function and the pair correlation function. The pair correlation function was originally introduced for X-ray experiments of Max von Laue around 1900, the K-function was introduced by Bartlett (1964) and–in its modern form–by Ripley (1977) (see Illian et al., 2008, page 226).

Illian et al. (2008) recommend working with the pair correlation function rather than with Ripley's K-function: Ripley's K-function should be used for very small samples and for goodness-of-fit tests, the pair correlation function should be used for exploratory analysis.

Baddeley and Silverman (1984) show that the second-order description of a pattern does not contain all information. They construct a so-called 'cell process': The rectangular observation window is divided in unit squares with random numbers N_S of points per cell. Every cell contains 0, 1 or 10 points with probability 1/10, 8/9 and 1/90, respectively. They show that the intensity is the same as for a Poisson process with intensity 1 and that the two processes have identical K-functions. So the K-function is not unique, very different patterns may have exactly the same K-function.

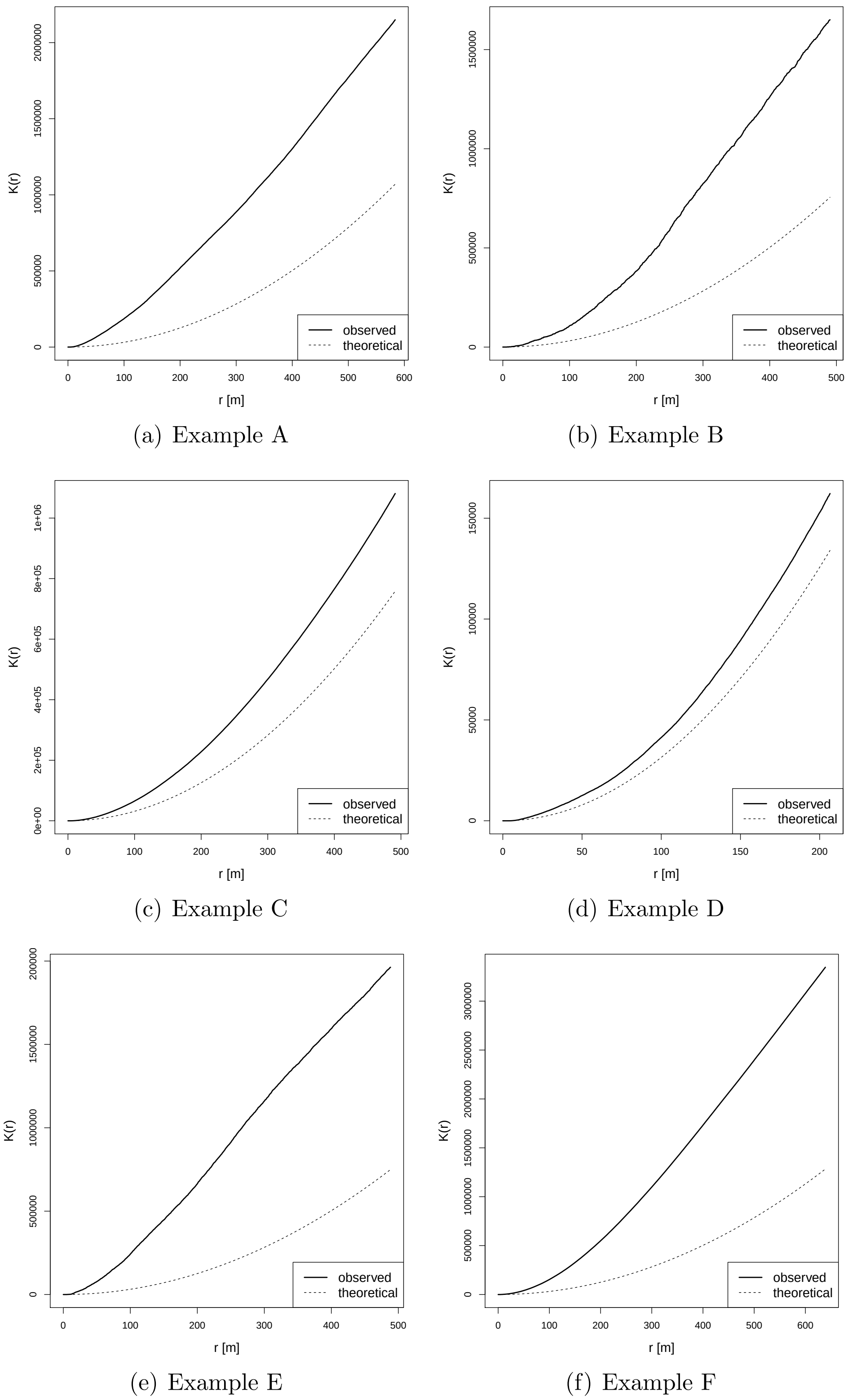
The K-function is invariant under thinning (Stoyan et al., 1987, page 134). Ripley (1977) and Ohser (1983) give unbaised estimators.

In `spatstat`, the K-function is estimated as follows:

$$\hat{K}(r) = \frac{\nu(W)}{(n-1)\pi} \sum_{i<j} \mathbb{1}(d(i,j) \leq r) \cdot e(i,j),$$

where n is the number of points, $d(i,j)$ is the distance between two (ordered) points and $e(i,j)$ the corresponding edge correction weight. The implemented edge correction methods are the border method, Ripley's isotropic correction (which is recommended) and Ohser's translation correction, which is slow for complex windows.

The values of the estimated K-functions (with Ripley's isotropic correction, which is used following the recommendation although anisotropy cannot be ruled out) clearly exceed the theoretical values for all six examples (Fig. 3.8), especially for Examples E and F, not so much for Examples C and D. This may suggest spatial clustering, but the deviation may also be due to inhomogeneity. Note that even larger values would result if the translation correction was applied.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

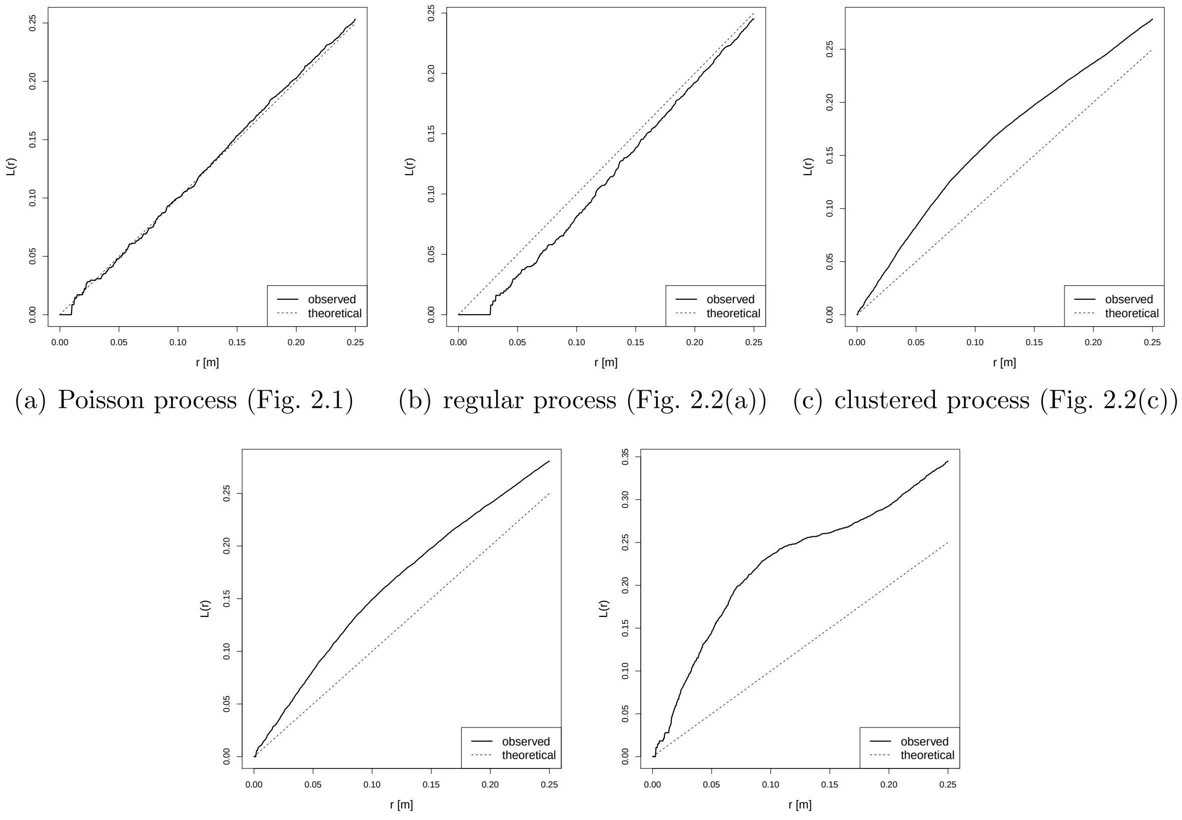
(f) Example F

Figure 3.8.: Ripley's K-function: The solid lines represent the estimated K-functions for the observed patterns, the dashed lines correspond to the theoretical K-functions of a homogeneous Poisson process.

In the discussion to Ripley (1977), Julian Besag suggested "a slight modification to K-plots" and showed a plot of $\sqrt{\frac{K(r)}{\pi}}$ against r. This resulted in Besag's L-function (Besag, 1977):

$$L(r) = \sqrt{\frac{K(r)}{\pi}} \quad \text{for } r \geq 0.$$

For a Poisson point process $L(r) = r$, larger values suggest clustering. Examples for the simulated patterns from Chapter 2 are shown in Figure 3.9.



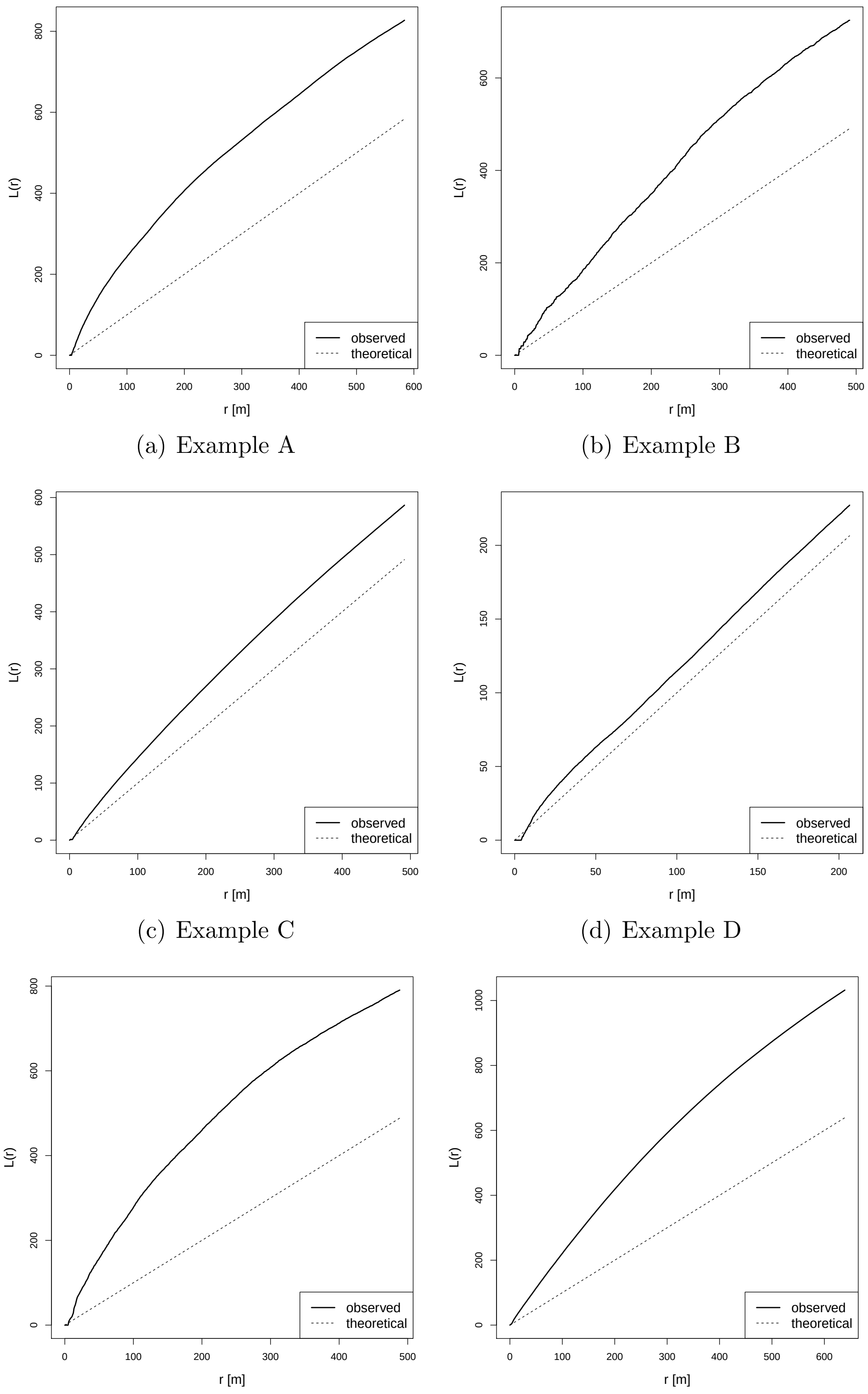
(a) Poisson process (Fig. 2.1) (b) regular process (Fig. 2.2(a)) (c) clustered process (Fig. 2.2(c))

(d) inhomogeneous Poisson process (Fig. 2.4(a)) (e) inhomogeneous Poisson process (Fig. 2.4(d))

Figure 3.9.: L-function: The solid lines represent the estimated L-functions for the simulated patterns, the dashed lines correspond to the theoretical L-functions of a homogeneous Poisson process.

The advantages of the L-function are that the interpretation and visualization are easier as the function is proportional to r. Moreover, the root transformation stabilises fluctuations. A modified version $L^*(r) = \sqrt{\frac{K(r)}{\pi}} - r$ of the L-function is somewhat misleading as the cumulative nature of the function is covered up.

In `spatstat`, the L-function is estimated by transforming the estimate of the K-function. Figure 3.10 shows the estimates for Examples A to F obtained via Ripley's isotropic correction. The values for the observed patterns exceed the theoretical values.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 3.10.: Besag's L-function: The solid lines represent the estimated L-functions for the observed patterns, the dashed lines correspond to the theoretical L-functions of a homogeneous Poisson process.

3.5. Pair correlation function

As we have seen, the cumulative nature of the K-function makes its interpretation difficult. An alternative second-order characteristic which is not cumulative is the pair correlation function $g(r)$. It is proportional to the derivative of the K-function:

$$g(r) = \frac{K'(r)}{2\pi r} \text{ for } r \geq 0.$$

It can be motivated as follows (Illian et al., 2008, page 219): The probability to observe an event in the infinitesimal disc $b(\mathbf{x})$ with area $d\mathbf{x}$ centered on $\mathbf{x}$ is $\lambda d\mathbf{x}$. Consider a second location $\mathbf{y}$ with distance r from $\mathbf{x}$. The probability to observe an event in $b(\mathbf{x})$ as well as in $b(\mathbf{y})$ is

$$p_2(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y},$$

where $\rho(\mathbf{x}, \mathbf{y})$ is the second-order product density.

If this probability only depends on the distance r, but not on the specific location of $\mathbf{x}$ and $\mathbf{y}$, this expression simplifies to

$$p_2(\mathbf{x}, \mathbf{y}) = g(r) \lambda d\mathbf{x} \lambda d\mathbf{y}.$$

The pair correlation function is $g(r) = \rho(r)/\lambda^2$. In case of complete spatial randomness, $p_2(\mathbf{x}, \mathbf{y}) = \lambda d\mathbf{x} \lambda d\mathbf{y}$, so $g(r) \equiv 1$. For cluster processes, larger values are obtained, especially for small radius r. For regular processes, smaller values are obtained for small r. The range of correlation is the finite distance r_{corr} with $g(r) = 1$ for $r \geq r_{corr}$. Examples for some of the simulated patterns from Chapter 2 are shown in Figure 3.11.

The pair correlation function is invariant under p-thinning (Illian et al., 2008, page 366).

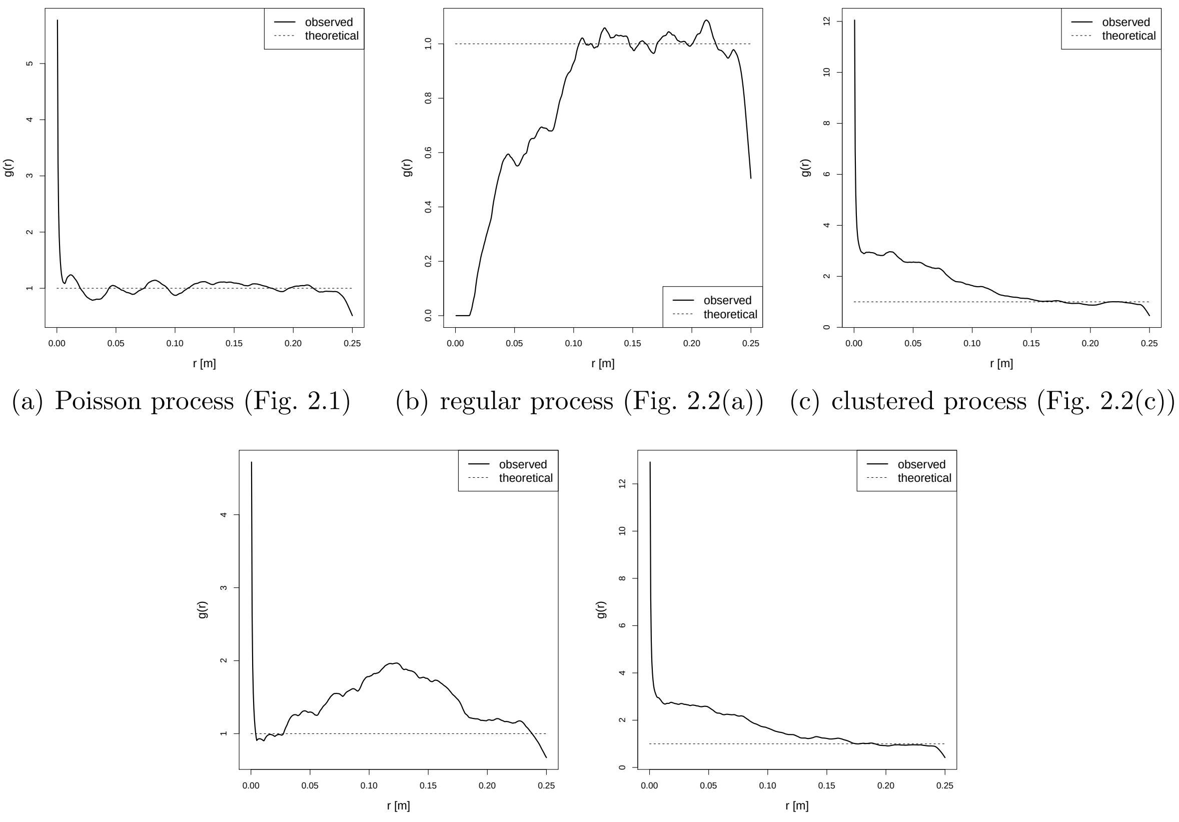
In `spatstat`, the pair correlation function is estimated as recommended in Stoyan and Stoyan (1992, 1994): In a first step, the second-order product density $\rho^{(2)}(r)$ is estimated. The estimate for the pair correlation function is obtained by dividing by an estimate for λ^2. A suitable estimator, which is unbaised for Poisson processes, is $\widehat{\lambda^2} = \frac{n(n-1)}{\nu(W)^2}$ (Stoyan and Stoyan, 1992, page 302).

For the estimation of $\rho^{(2)}(r)$, they use an Epanechnikov kernel $k_h(\cdot)$ with support $[-h, h]$, where $h = c \ / \ \sqrt{\lambda}$, c is a value between 0.1 and 0.2 and λ is a simple estimator for the intensity (Stoyan and Stoyan, 1992, page 310). In `spatstat`, $c = 0.15$ per default. For cluster processes, Guan (2007) recommends smaller values for h.

If Ripley's isotropic correction is used, the resulting estimator is

$$\hat{\rho}_R(r) = \frac{1}{2\pi r} \sum_{i=1}^{n} \sum_{\substack{j=1, \\ j \neq i}}^{n} \frac{k_h(r - ||\mathbf{x}_j - \mathbf{x}_i||) b_{ij}}{\nu(W^{||\mathbf{x}_j - \mathbf{x}_i||})},$$

where $W^r = \{\mathbf{x} \in W : \partial(b(\mathbf{x}, r)) \cap W \neq \emptyset\}$ and $b_{ij} = \frac{2\pi}{\alpha_{ij}}$, where α_{ij} is the sum of all angles of the arcs in W of a circle with radius $||\mathbf{x}_j - \mathbf{x}_i||$ centered at $\mathbf{x}_i$. If $\alpha_{ij} = 0$, then $b_{ij} = 0$



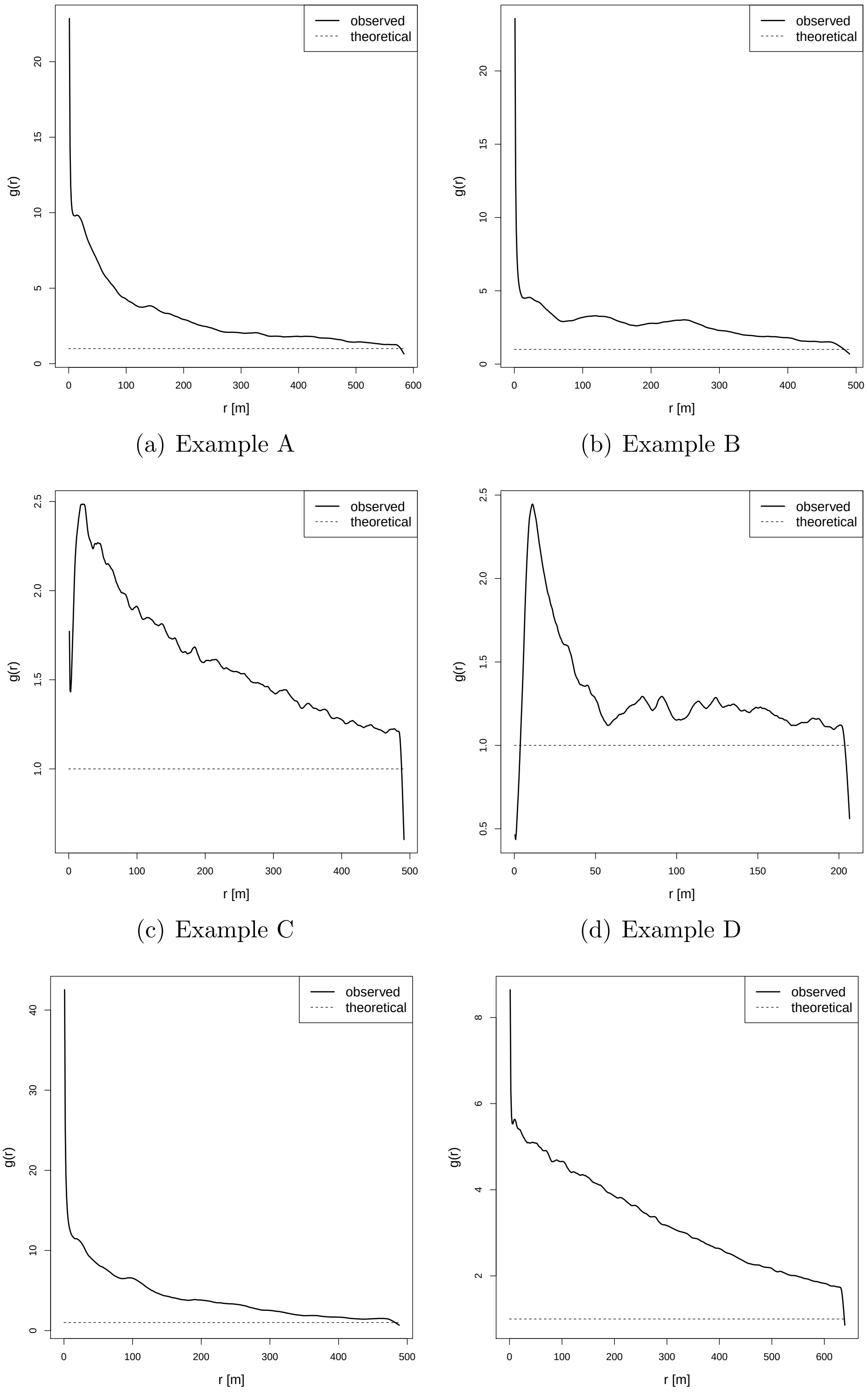
(a) Poisson process (Fig. 2.1) (b) regular process (Fig. 2.2(a)) (c) clustered process (Fig. 2.2(c))

(d) inhomogeneous Poisson process (Fig. 2.3(a)) (e) inhomogeneous Poisson process (Fig. 2.4(a))

Figure 3.11.: Pair correlation function: The solid lines represent the estimated pair correlation functions for the simulated patterns, the dashed lines correspond to the theoretical pair correlation functions of a homogeneous Poisson process.

(Illian et al. (2008), p. 229). The translation correction and Ripley's isotropic correction are available in `spatstat`.

Like for the K-function, Ripley's isotropic correction was applied. Figure 3.12 shows the estimates for Examples A to F. The values for the observed patterns are larger than 1 even for large values of the argument r. In part, this may be due to the choice of the observation window (Examples E and F), but even more to inhomogeneity. For very small values of r, the estimation is difficult (see Illian et al., 2008, page 235). In particular, large values may be obtained for $r < h$. For Examples A, B and E, the pair correlation function takes indeed large values for small r, whereas these values are smaller than 1 for Example D. The following values were obtained for h: 14.4 for Example A, 27.3 for Example B and 7.4 for Example C. For Examples D, E and F, $h = 5.1$, $h = 18.8$ and $h = 8.2$, respectively. This means that the remarkable shape of $\hat{g}(r)$ for Examples A, C and D is due to difficulties with regard to the estimation. Note that larger values would have been obtained with the translation correction.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

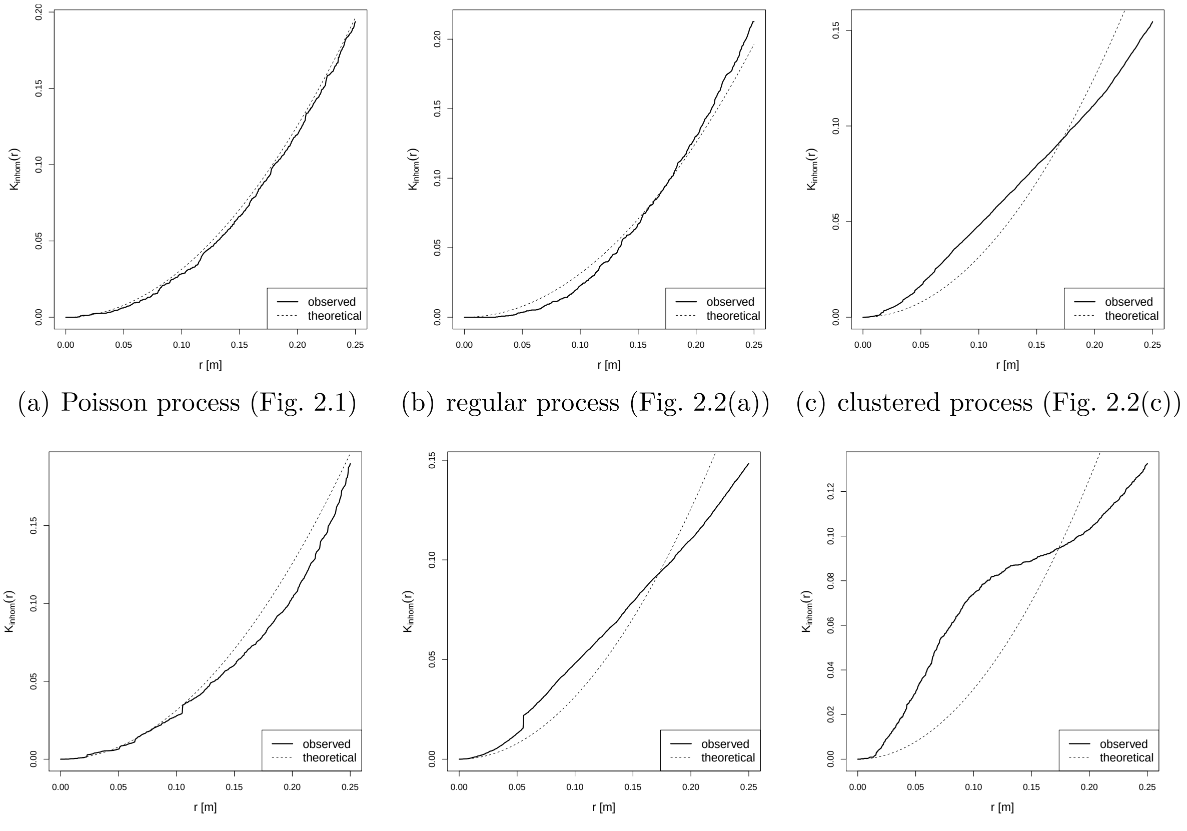
(f) Example F

Figure 3.12.: Pair correlation function: The solid lines represent the estimated pair correlation functions for the observed patterns, the dashed lines correspond to the theoretical pair correlation functions of a homogeneous Poisson process.

3.6. Characteristics for inhomogeneous processes

All summary characteristics considered so far are intended to be used for stationary processes. Although their definition explicitly refers to stationarity (e.g. in terms of a constant intensity λ), formal application to non-stationary processes is possible. It may lead to bi- or multimodal distributions in $\hat{D}(r)$ and $\hat{H}_s(r)$; as observed for the bomb crater patterns, $\hat{g}(r)$ and $\hat{K}(r)$ are similar to estimates which would be obtained for cluster processes with (spurious) large clusters (Illian et al., 2008, page 280).

Baddeley, Møller, and Waagepetersen (2000) introduced an approach for non-stationary processes based on intensity-reweighting, which can be applied for so-called *second-order intensity-reweighted stationary processes* such as inhomogeneous Poisson processes and nonstationary processes resulting from $p(\mathbf{s})$-thinning of stationary processes.



(a) Poisson process (Fig. 2.1) (b) regular process (Fig. 2.2(a)) (c) clustered process (Fig. 2.2(c))

(d) inhomogeneous Poisson process (Fig. 2.3(c)) (e) inhomogeneous Poisson process (Fig. 2.4(a)) (f) inhomogeneous Poisson process (Fig. 2.4(d))

Figure 3.13.: Inhomogeneous K-function: The solid lines represent the estimated inhomogeneous K-functions for the simulated patterns, the dashed lines correspond to the theoretical inhomogeneous K-functions of an (inhomogeneous) Poisson process.

The inhomogeneous K-function is defined as

$$K_{inhom}(r) = \frac{1}{\nu(\mathcal{B})} E \sum_{\mathbf{x}_i \in X \cap \mathcal{B}} \sum_{\mathbf{x}_j \in X \setminus \{\mathbf{x}_i\}} \frac{\mathbb{1}(||\mathbf{x}_i - \mathbf{x}_j|| \leq r)}{\lambda(\mathbf{x}_i)\lambda(\mathbf{x}_j)},$$

where $\mathcal{B}$ is a bounded Borel set in $\mathbb{R}^2$. The inhomogeneous K-function does not depend on the choice of $\mathcal{B}$ and can be interpreted as Palm expectation:

$$K_{inhom}(r) = E_{\mathbf{s}} \sum_{\mathbf{x}_i \in X \setminus \{\mathbf{s}\}} \frac{\mathbb{1}(||\mathbf{x}_i - \mathbf{s}|| \leq r)}{\lambda(\mathbf{x}_i)}.$$

Just like the K-function, the inhomogeneous K-function is not unique: For every intensity function $\lambda(\mathbf{s})$ it is possible to find a non-Poisson process whose inhomogeneous K-function is identical to the inhomogeneous K-function of an inhomogeneous Poisson process with intensity function $\lambda(\mathbf{s})$. The K-function is invariant under random thinning, a similar property holds for the inhomogeneous K-function (Baddeley, Møller, and Waagepetersen, 2000).

The estimation of inhomogeneous summary functions implies the estimation of $\lambda(\mathbf{s})$. To enable distinction of large-scale variation of the intensity and small-scale correlation, a smooth estimate needs to be used. For kernel estimation, the bandwidth should be chosen larger than in the estimation of summary characteristics (Illian et al. (2008), page 281). If a kernel estimator is used, $\lambda(\mathbf{s})$ is overestimated at data points (Baddeley, Møller, and Waagepetersen, 2000).

Examples for the simulated patterns from Chapter 2 are shown in Figure 3.13. For the homogeneous Poisson process as well as for the inhomogeneous Poisson processes, the estimated functions for the simulated patterns should be very close to the theoretical inhomogeneous K-function. However, the estimated functions typically take larger values for small r and smaller values for large r.

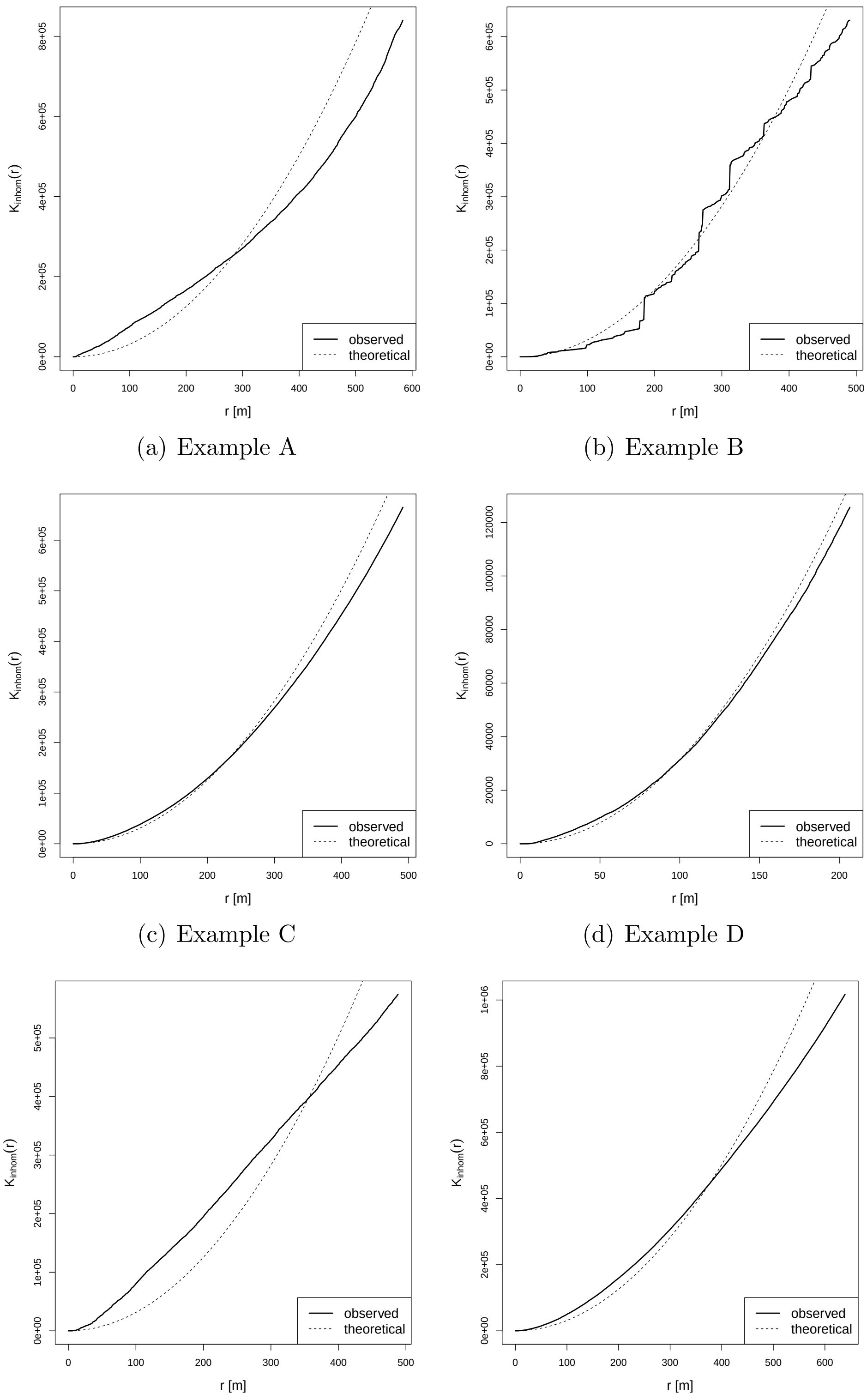
In `spatstat`, the inhomogeneous K-function is estimated as follows:

$$\hat{K}_{inhom}(r) = \sum_i \sum_j \mathbb{1}(d(i,j) \leq r) \cdot \frac{e(\mathbf{x}_i, \mathbf{x}_j, r)}{\hat{\lambda}(\mathbf{x}_i)\hat{\lambda}(\mathbf{x}_j)},$$

where $d(i,j)$ is the distance between two (ordered) points and $e(\mathbf{x}_i, \mathbf{x}_j, r)$ the corresponding edge correction weight, which depends on r. $\hat{\lambda}(\mathbf{x}_i)$ and $\hat{\lambda}(\mathbf{x}_j)$ are the values of the estimated intensity function. A leave-one-out estimator can be used to avoid overestimation at data points. The implemented edge correction methods are the border method, a modified version of the border method, Ripley's isotropic correction and Ohser's translation correction.

To reduce variability and bias in small samples and in case of strong inhomogeneity, the estimator can be rescaled.

In Fig. 3.14, the theoretical K-function for the inhomogeneous Poisson point process together with the estimates for Examples A to F obtained using Ripley's isotropic edge correction is depicted. For Example A, the observed values are a bit higher than the theoretical values for $r < 300$ and smaller for $r > 300$. The results for Examples C, E and F are similar. For Example D, the estimated function is very close to the theoretical function. Because of the irregular shape of the study region, the behaviour is less clear for Example B. In general, the deviations from an inhomogeneous Poisson point process seem to be small.



(a) Example A

(b) Example B

(c) Example C

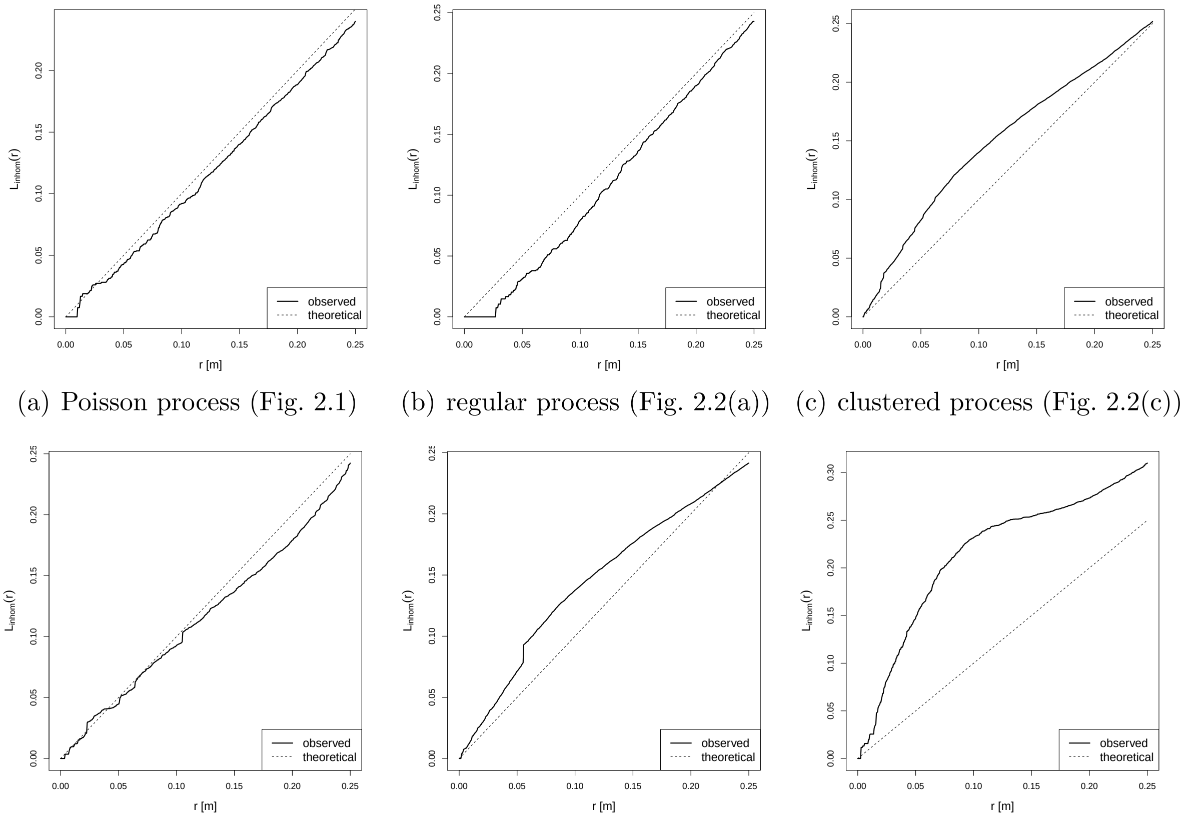
(d) Example D

(e) Example E

(f) Example F

Figure 3.14.: Inhomogeneous K-function: The solid lines represent the estimated inhomogeneous K-functions for the observed patterns, the dashed lines correspond to the theoretical inhomogeneous K-functions of an (inhomogeneous) Poisson process.

An inhomogeneous L-function can be defined analogously. The results are less clear than for the inhomogeneous K-function both for the simulated patterns (Figure 3.15) and the observed patterns (Figure 3.16). In particular, the estimated values for Examples A, E and F are larger than expected for an inhomogeneous Poisson process.



(a) Poisson process (Fig. 2.1)

(b) regular process (Fig. 2.2(a))

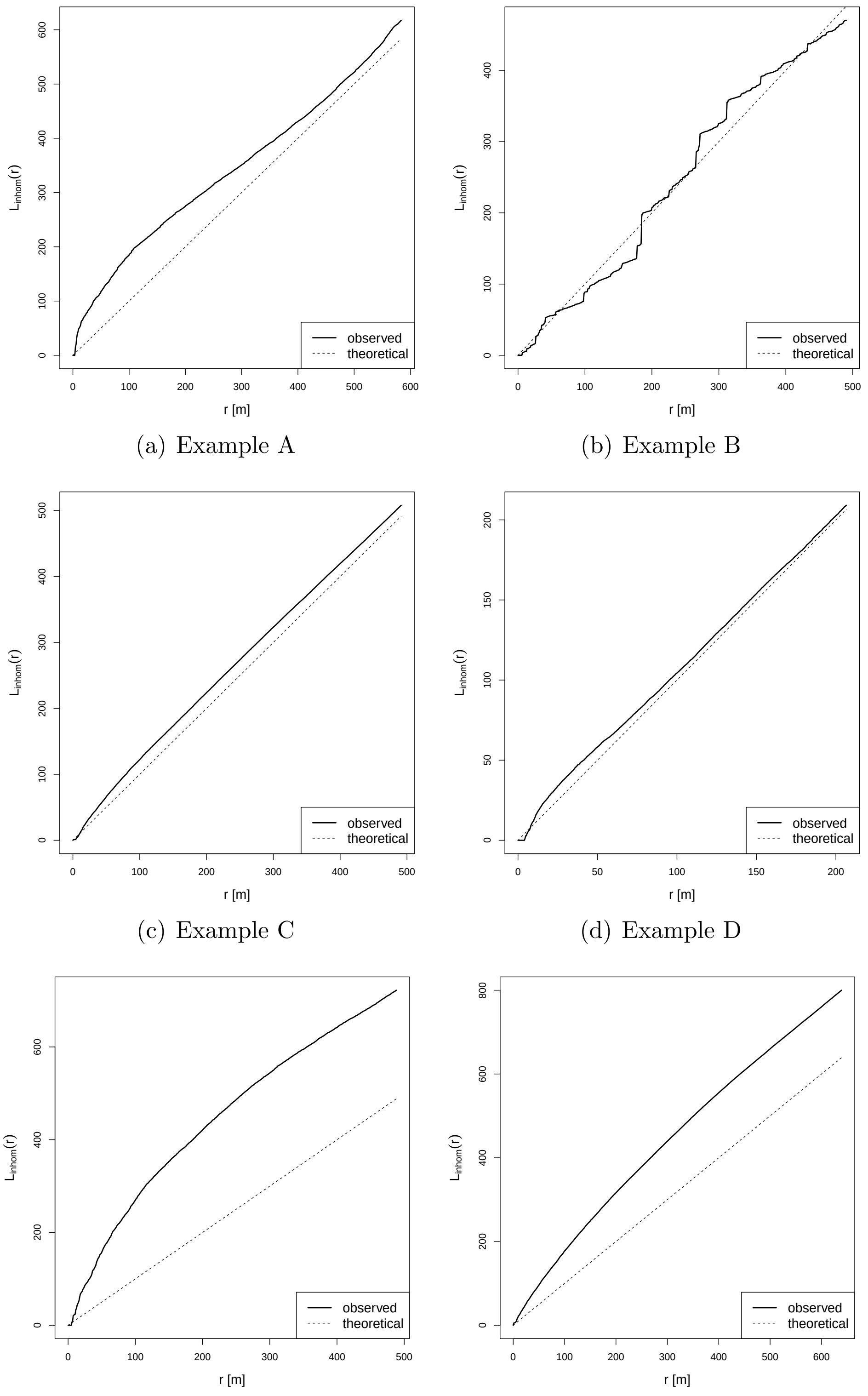
(c) clustered process (Fig. 2.2(c))

(d) inhomogeneous Poisson process (Fig. 2.3(c))

(e) inhomogeneous Poisson process (Fig. 2.4(a))

(f) inhomongenous Poisson process (Fig. 2.4(d))

Figure 3.15.: Inhomogeneous L-function: The solid lines represent the estimated inhomogeneous L-functions for the simulated patterns, the dashed lines correspond to the theoretical inhomogeneous L-functions of an (inhomogeneous) Poisson process.



(a) Example A

(b) Example B

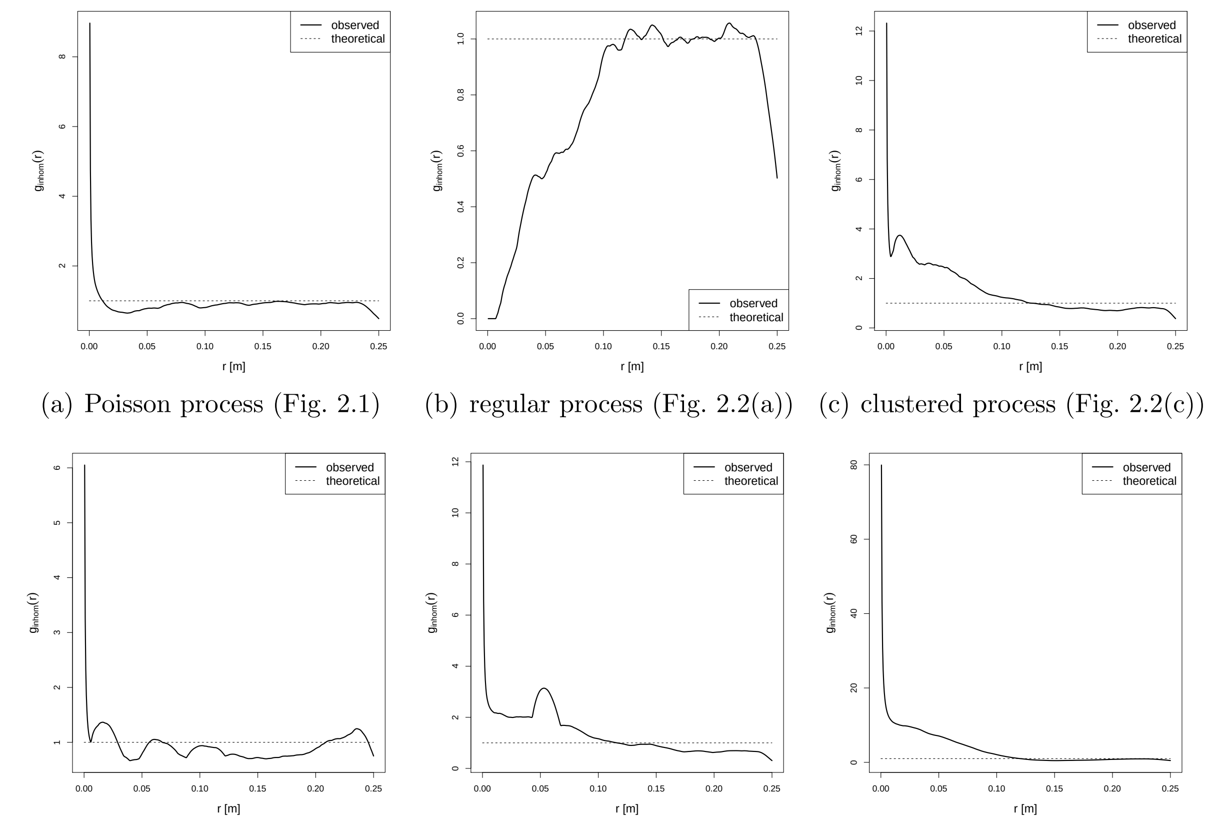
(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 3.16.: Inhomogeneous L-function: The solid lines represent the estimated inhomogeneous L-functions for the observed patterns, the dashed lines correspond to the theoretical inhomogeneous L-functions of an (inhomogeneous) Poisson process.



(a) Poisson process (Fig. 2.1)

(b) regular process (Fig. 2.2(a))

(c) clustered process (Fig. 2.2(c))

(d) inhomogeneous Poisson process (Fig. 2.3(c))

(e) inhomogeneous Poisson process (Fig. 2.4(a))

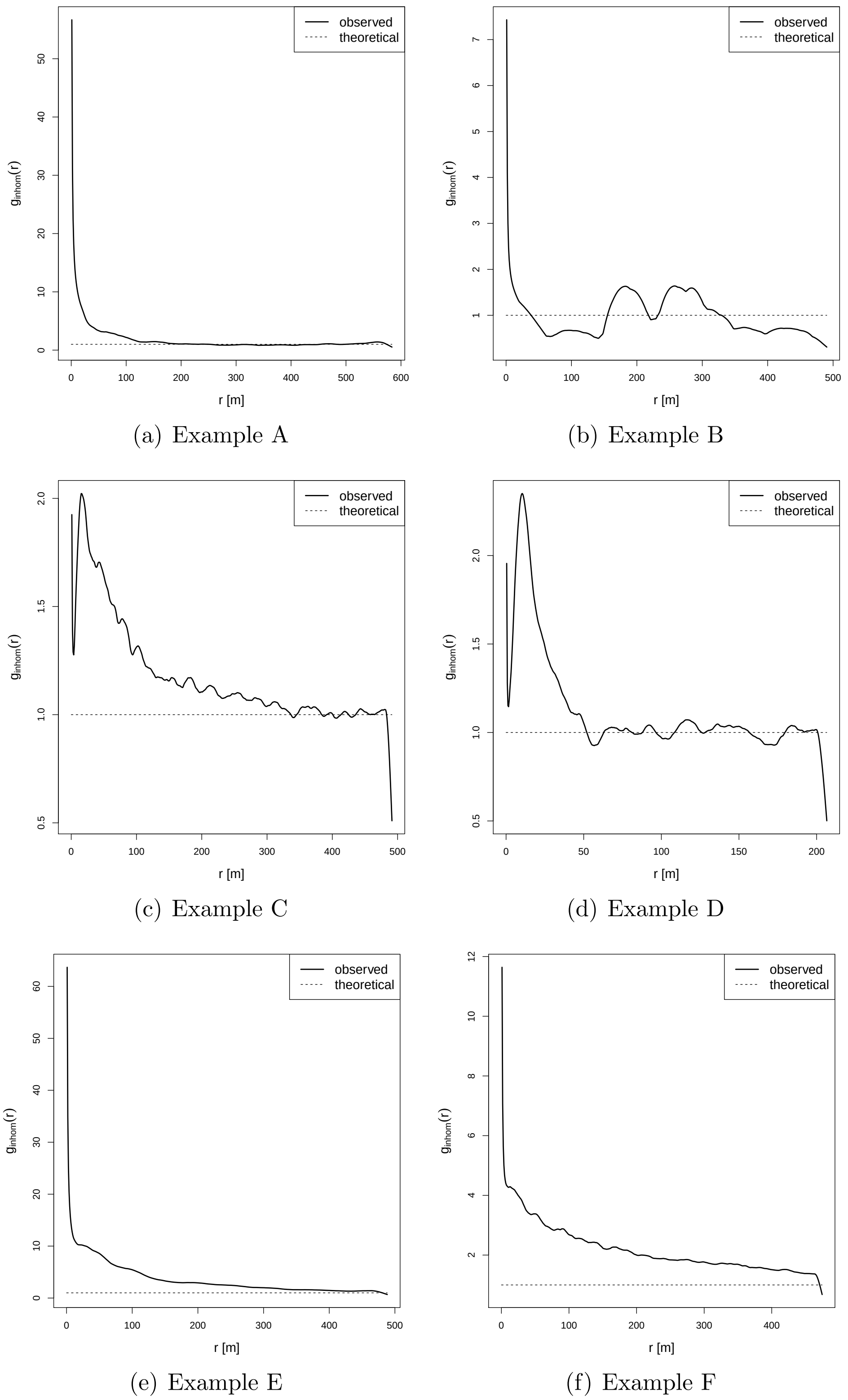
(f) inhomongenous Poisson process (Fig. 2.4(d))

Figure 3.17.: Inhomogeneous pair correlation function: The solid lines represent the estimated inhomogeneous pair correlation functions for the simulated patterns, the dashed lines correspond to the theoretical inhomogeneous pair correlation functions of an (inhomogeneous) Poisson process.

Second-order intensity-reweighted stationary processes have a pair correlation function which depends only on $r = ||\mathbf{x} - \mathbf{y}||$. If the constant intensity estimator is replaced by a variable intensity function estimator, $g(r) \equiv 1$ is obtained for inhomogeneous Poisson processes. The results for the simulated patterns (Figure 3.17) and for the observed patterns (Figure 3.18) are rough, although a larger bandwidth was chosen. The estimated inhomogeneous pair correlation functions for Examples A to F are slightly closer to 1 than the estimated pair correlation functions.

Inhomogeneous analogues can also be defined for the empty-space and the nearest-neighbour distance distribution functions, but are not considered here.

It is not possible to make any judgements about whether the high values of Ripley's K-function and the pair correlation function are due to clustering or inhomogeneity (see also Section 2.5.3 for the problems concerning the distinction of these two phenomena). But the inhomogeneous K-functions indicate that the observed patterns can be described sufficiently well by an inhomogeneous Poisson process. Moreover, according to OFD Niedersachsen, the inhomogeneous Poisson model fits the subject-matter theory better than cluster models. The assumption of several targets in a property is not justified for most examples. In some cases, like for Example B, the target of the attack even seems to be situated outside the property to be cleared.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 3.18.: Inhomogeneous pair correlation function: The solid lines represent the estimated inhomogeneous pair correlation functions for the observed patterns, the dashed lines correspond to the theoretical inhomogeneous pair correlation functions of an (inhomogeneous) Poisson process.

4. Methods for constructing high-risk zones

Recall that W denotes the observation window of the spatial point process X, i.e. the property of interest and the area for which data are available. The process X represents the locations of all events, observed as well as unobserved. Only a thinned version Y of the full process X (e.g. the exploded bombs recognized by the bomb craters) has been observed. It consists of $N_Y(W) = n_Y$ observations. The process of unobserved events is $Z = X \backslash Y$. The probability of non-observation q for every event is assumed to be homogeneous in W, which means that every $\mathbf{x} \in X$ is element of Z with probability q, regardless of its location $\mathbf{s} \in W$ and independently of the behaviour of the other elements of X.

Three methods for constructing high-risk zones are presented in this chapter: The traditional method, the quantile-based method (which is not a new approach, but has not been evaluated so far) and the intensity-based method, which is a novel approach based on spatial point process theory.

The methods have been presented in less detail in Mahling et al. (2013).

4.1. Traditional method

The method currently used for constructing high-risk zones for unexploded bombs consists in discs of a fixed radius r centered at each observed event, whose union gives the high-risk zone R_r:

$$R_r = \{\mathbf{s} \in W : \min_j ||\mathbf{s} - \mathbf{y}_j|| \leq r\}. \tag{4.1}$$

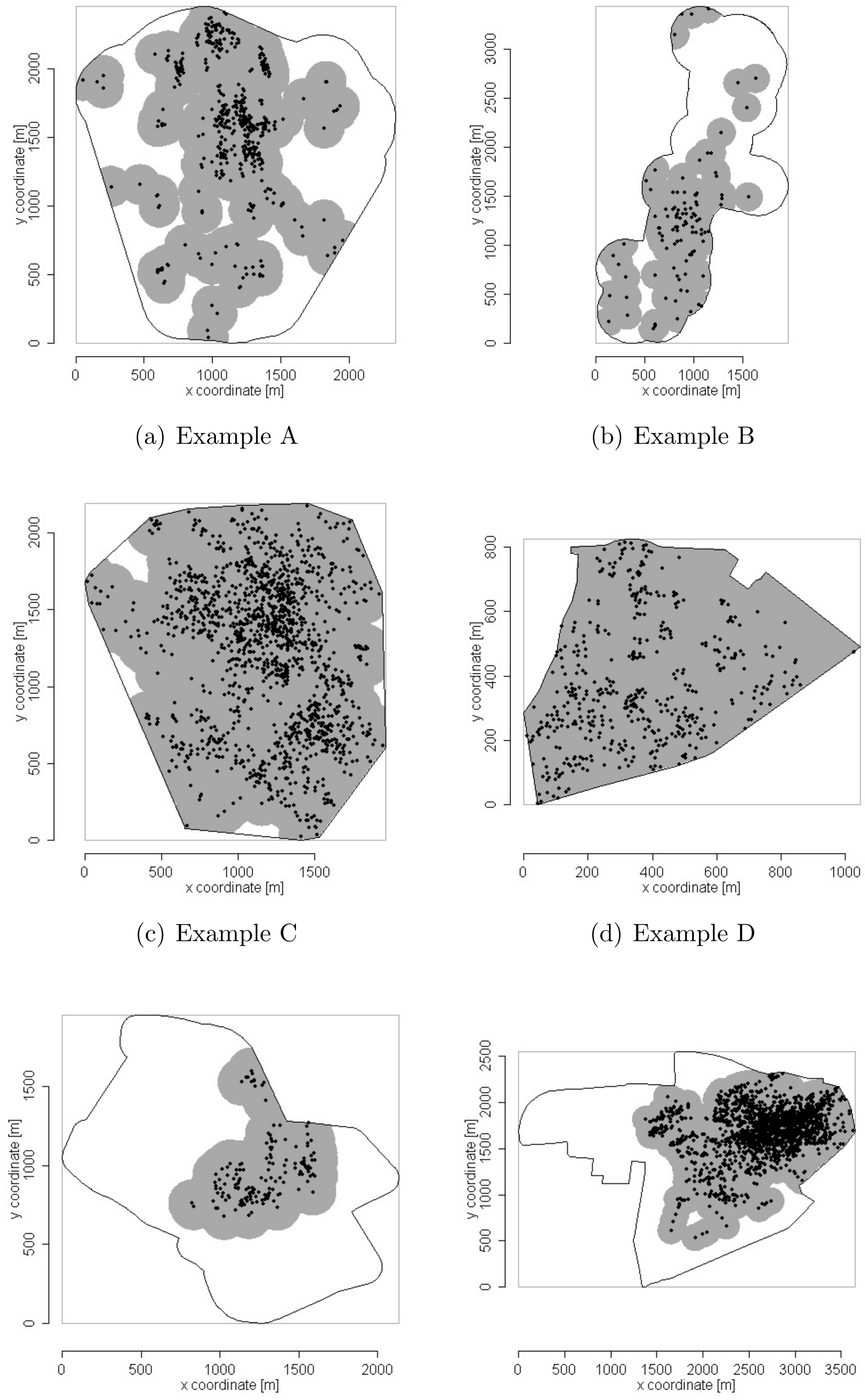
The radius r is chosen by an expert in advance. So besides expert knowledge, this approach uses only the coordinates of the observed events. General characteristics of the patterns are ignored completely.

In Figure 4.1, a radius of 150 m was used, which is rather a large value. As a consequence, the high-risk zone for Example D comprises the entire observation window. For Example C, most of the window is filled by the high-risk zone. The high-risk zones for Examples A and B seem ragged, whereas the radius seems to be too large for Example E and possibly also for Example F.

These high-risk zones can be interpreted as random sets, more specifically as *germ-grain models* (Illian et al. (2008), p. 43):

$$R_r = \bigcup_{\mathbf{y} \in Y} b(\mathbf{y}, r) = Y \oplus b(o, r), \tag{4.2}$$

where $b(o, r)$ denotes a disc of radius r centered on o and $\oplus$ denotes Minkowski addition, i.e. $A \oplus B = \{\mathbf{a} + \mathbf{b} : \mathbf{a} \in A \wedge \mathbf{b} \in B\}$ for $A, B \subseteq \mathbb{R}^d$.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 4.1.: High-risk zones (shaded grey areas) obtained for a fixed radius of 150 m.

4.2. Quantile-based method

A more sophisticated approach called *quantile-based construction method* represents a heuristic development of the traditional method. Instead of the arbitrary choice of the radius by an expert, the radius of the discs is determined as the p-quantile $Q(p)$ of the distribution of the nearest-neighbour distance:

For every bomb $\mathbf{y}_i$ in Y, the distance to the nearest other bomb in Y, the nearest-neighbour distance

$$t_i = \min_{\mathbf{y}_j \in Y; j \neq i} ||\mathbf{y}_i - \mathbf{y}_j||, \tag{4.3}$$

is computed and the empirical distribution function

$$G(r) = \frac{1}{n_Y} \sum_i \mathbb{1}\{t_i \leq r\} \tag{4.4}$$

of the nearest-neighbour distances of the point pattern is determined. No edge correction is performed, but the empirical distribution function is computed directly from the raw nearest-neighbour distances. The radius of the discs is then given by the p-quantile $Q(p)$ of the distribution of this nearest-neighbour distance, where $0 \leq p \leq 1$ is specified by the user.

Hyndman and Fan (1996) recommend the following estimator for the p-quantile $Q(p)$:

$$\hat{Q}(p) = (1-\gamma)t_{(j)} + \gamma t_{(j+1)}, \tag{4.5}$$

where $j = \lfloor pn + m \rfloor$, $\gamma = np + m - j$, $m = (p+1)/3$, n is the sample size and $t_{(j)}$ denotes the jth order statistic.

For a given value p, the high-risk zone consists of all locations $\mathbf{s}$ whose distance to the nearest observation does not exceed $\hat{Q}(p)$, i.e.

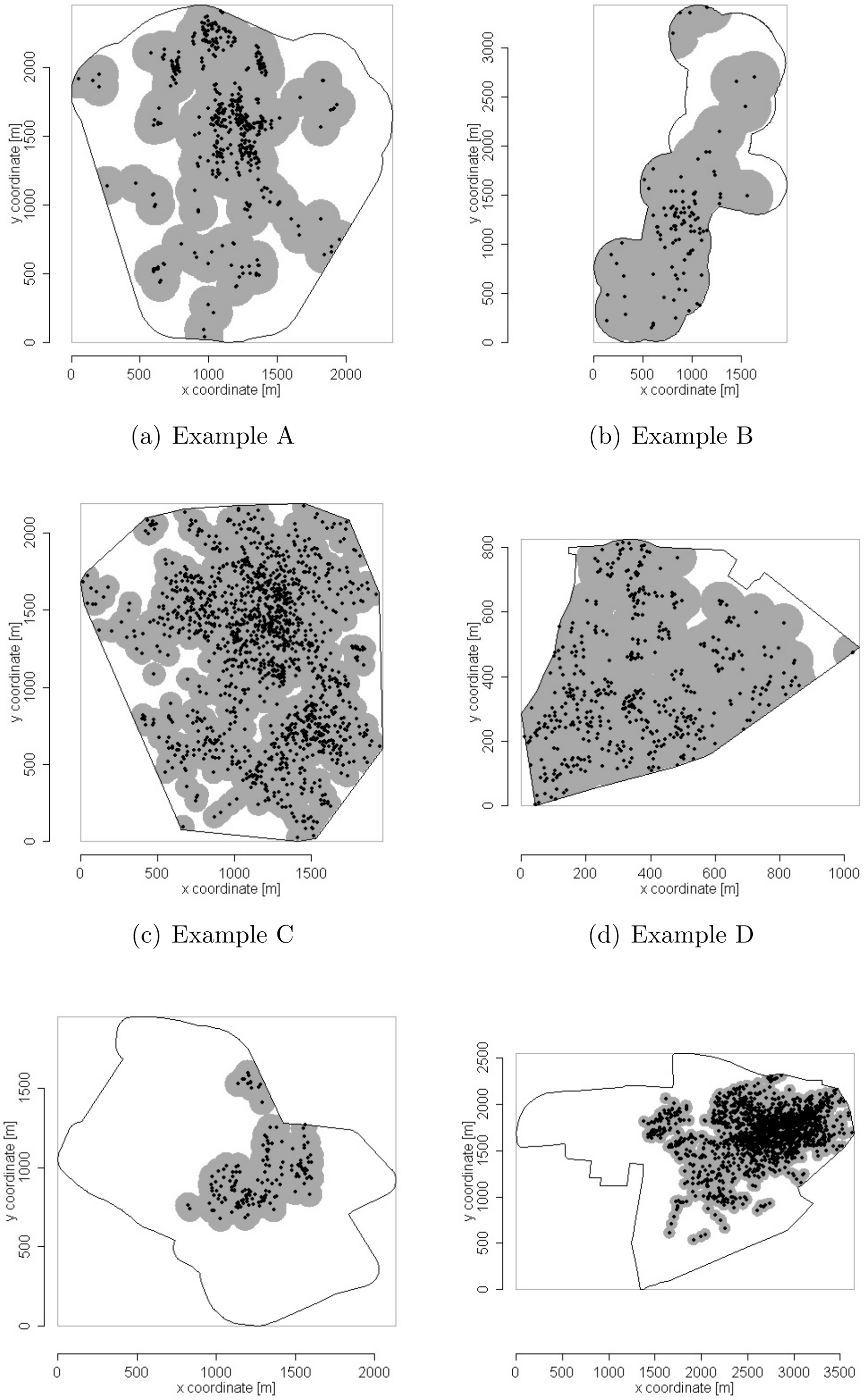
$$R_p = \{\mathbf{s} \in W : \min_j ||\mathbf{s} - \mathbf{y}_j|| \leq \hat{Q}(p)\}. \tag{4.6}$$

As both the traditional method and the quantile-based method rely on the distance to the nearest event, these two methods can be subsumed under the term *distance-based methods.*

Quantile-based high-risk zones for Examples A to F are shown in Figure 4.2. The high-risk zone for Example B is less ragged than before, the high-risk zone for Example C more. For Examples E and F, the high-risk zones become smaller. The high-risk zone for Example D does not comprise the entire observation window.

Note that this approach does not fix the global risk of leaving unobserved events outside the high-risk zone. Instead, the probability that a single unobserved event is covered by the high-risk zone should be close to p, so the individual risk for each event is specified.

As a modification, the radius could be determined from the estimated nearest-neighbour distance distribution function $D(r)$. This approach would yield very similar high-risk zones. Figure 4.3 compares the functions $D(r)$, which is corrected for edge effect, and $G(r)$. For Examples A and F, there is almost no difference at all. For Examples C, D and E, the values of $D(r)$ are slightly larger than those of $G(r)$, so the resulting radius would be a bit smaller. The resulting radius for Example B could be slightly smaller or slightly larger.



(a) Example A

(b) Example B

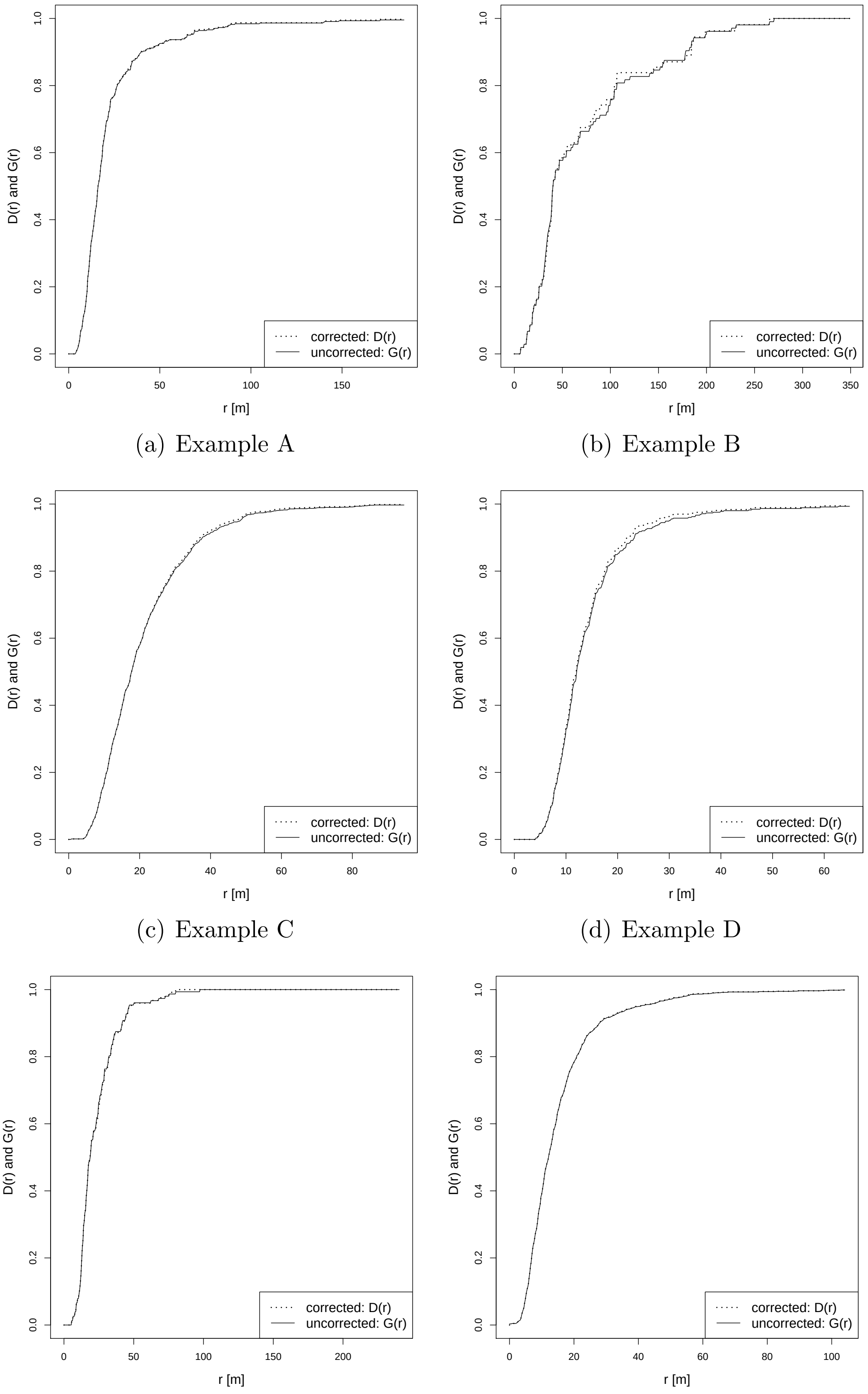
(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 4.2.: High-risk zones (shaded grey areas) obtained for the quantile-based method by using the 99 % quantile.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 4.3.: Empirical distribution functions of the nearest-neighbour distance with and without edge correction.

4.3. Intensity-based method

4.3.1. Basic idea

The aim is to construct a high-risk zone which comprises a large fraction of the unobserved events while covering a small area. This is achieved via the *intensity-based construction method*: The high-risk zone consists of those locations in the observation window W for which the intensity is largest.

The intensity function of the complete point process X is denoted by $\lambda_X(\mathbf{s})$. As the probability of non-explosion for every bomb q is assumed to be homogeneous in W, the intensity functions of the point process Y and Z are $\lambda_Y(\mathbf{s}) = (1-q) \cdot \lambda_X(\mathbf{s})$ and $\lambda_Z(\mathbf{s}) = q \cdot \lambda_X(\mathbf{s})$, respectively. For the application where Y is the process of bomb craters and Z the process of unexploded bombs it is indeed realistic to assume that the two intensities are proportional (see Tavakkoli et al. (2012)).

For a given threshold $c > 0$, a high-risk zone R_c is defined as

$$R_c = \{\mathbf{s} \in W : \hat{\lambda}_Z(\mathbf{s}) \geq c\}. \tag{4.7}$$

The general idea is closely related to the principle of highest posterior density intervals known from Bayesian inference. Highest posterior density intervals are credible intervals with minimal length (see Held, 2008, pages 159/160). Analogously, highest posterior density regions are credible regions with minimal area. As the point density of a spatial point process is proportional to the intensity function, R_c is the high-risk zone with minimal area comprising a certain fraction of the unobserved events.

Determining an intensity-based high-risk zone consists of the following steps: First of all, the intensity function $\lambda_Y(\mathbf{s})$ is estimated. As the probability of non-observation q for every event is assumed to be homogeneous, we can use $\hat{\lambda}_Y(\mathbf{s})$ to estimate the intensity function of the process Z: $\hat{\lambda}_Z(\mathbf{s}) = q/(1-q) \cdot \hat{\lambda}_Y(\mathbf{s})$. The region within the contours defined by $\hat{\lambda}_Z(\mathbf{s}) = c$ forms the high-risk zone. The estimation of $\lambda_Y(\mathbf{s})$ and the determination of the threshold $c > 0$ will now be explained in detail.

4.3.2. Estimation of the intensity function by kernel methods

The intensity $\lambda_Y(\mathbf{s})$ of the pattern of observed events can be estimated by using a kernel method (Diggle, 1985; Baddeley, 2008):

$$\hat{\lambda}_Y(\mathbf{s}) = e(\mathbf{s}) \cdot \sum_{i=1}^{n_Y} K_H(\mathbf{s} - \mathbf{y}_i), \tag{4.8}$$

where $K_H(\cdot)$ is an anisotropic Gaussian kernel and $e(\mathbf{s})$ is an edge effect bias correction. The variance-covariance matrix H of the Gaussian kernel determines the smoothing bandwidth. It is a symmetric positive definite $d \times d$ matrix.

Edge correction

Edge correction is necessary because observations lying outside the window, possibly near the boundary, are not taken into account, which would result in a negative bias around the boundary of the observation window. Diggle (1985) proposes an edge effect bias correction $e(\mathbf{s})$ of the form

$$e(\mathbf{s})^{-1} = \int_W K_H(\mathbf{s}-\mathbf{v})\; d\mathbf{v}, \tag{4.9}$$

which will be applied in this thesis. Note that Berman and Diggle (1989) introduced a modified edge correction with the same correction term.

Determination of the optimal variance-covariance matrix

Diggle (1985) determines the optimal smoothing bandwidth for one-dimensional point processes minimising the mean squared error (MSE) for stationary Cox processes and finds that the MSE is intractable for the two-dimensional analogue because of a "double integral [which] cannot be reduced to an explicit formula". Although this problem is solved in Diggle (2003), this approach is not applied here as anisotropy cannot be taken into account.

Instead, we make use of the fact that intensity estimation and density estimation are closely related. Of course, a kernel density estimator

$$\hat{f}(\mathbf{s}, H) = \frac{1}{n_Y} \cdot \sum_{i=1}^{n_Y} K_H(\mathbf{s}-\mathbf{y}_i) \tag{4.10}$$

does not comprise edge correction as the observations from a bivariate density are not restricted to some kind of observation window. The normalizing constant $1/n_Y$, however, does not affect the optimal choice of the bandwidth because n_Y is fixed. Two further differences of intensity estimation compared to density estimation are that "for a fixed region the number of observations does not increase to infinity" when asymptotic properties are considered and that "events in nearby regions can be correlated" (Guan, 2008b). For the special case of a stationary Cox process on the line, however, the bandwidth minimising MSE for intensity estimation is identical to the bandwidth determined via leave-one-out least squares cross-validation for density estimation (Diggle and Marron, 1988). So it seems justified to apply criterions which have initially been developed for density estimation.

A popular criterion for determining the optimal smoothing parameter for radially symmetric kernels with bandwidth h is least-squares cross-validation (LSCV) (Silverman, 1992), where the integrated square error

$$\int_W \left\{ \hat{f}_h(\mathbf{s}) - f(\mathbf{s}) \right\}^2 d\mathbf{s} = \int_W \hat{f}_h^2(\mathbf{s}) d\mathbf{s} - 2\int_W \hat{f}_h(\mathbf{s}) f(\mathbf{s}) d\mathbf{s} + \int_W f^2(\mathbf{s}) d\mathbf{s} \tag{4.11}$$

is minimized over h. However, the high-risk zones which resulted for these kernels turned out to be too small and anisotropy cannot be taken into account if a radially symmetric kernel is employed.

Wand and Jones (1993) recommend the use of unconstrained (i.e. not necessarily diagonal) symmetric matrices H (with diagonal elements h_1^2 and h_2^2 and off-diagonal element h_{12}), thus allowing the Gaussian kernel to have arbitrary orientation.

Such matrices can be selected using smooth cross-validation (Duong, 2007), which is based on two common criteria for bandwidth selectors, the Mean Integrated Squared Error (MISE) (see Jones et al., 1991) and the Asymptotic Mean Integrated Squared Error (AMISE) (see Scott, 1992, Chapter 6).

MISE The Mean Integrated Squared Error is

$$\mathrm{MISE}(H) = E\left[\int_{\mathbb{R}^2} \{\hat{f}(\mathbf{s}, H) - f(\mathbf{s})\}^2 d\mathbf{s}\right]. \tag{4.12}$$

The optimal bandwidth H_{MISE} minimizing MISE(H) does not have a closed form (cf. Wand and Jones (1995), p. 99). For this reason, the Asymptotic Mean Integrated Squared Error (AMISE) is commonly used to select H.

AMISE The Asymptotic Mean Integrated Squared Error is obtained by using a multivariate version of Taylor's theorem (Wand and Jones, 1995, Chapter 4.3) and can be written in the following form:

$$\mathrm{AMISE}(H) = \frac{1}{n_Y} \cdot \frac{1}{4\pi} \cdot |H|^{-\frac{1}{2}} + \frac{1}{4} \cdot (h_1^2 \; h_{12} \; h_2^2) \mathbf{\Psi}_4 \begin{pmatrix} h_1^2 \\ h_{12} \\ h_2^2 \end{pmatrix}, \tag{4.13}$$

where $\mathbf{\Psi}_4$ is a 3×3 matrix containing integrated density derivative functionals of f. Details about $\mathbf{\Psi}_4$ can be found in Wand and Jones (1995, pages 98/99) and in Duong and Hazelton (2003). Estimating $\mathbf{\Psi}_4$ yields a plug-in estimate PI(H) of the AMISE. A pilot bandwidth is required for the estimation of $\mathbf{\Psi}_4$. For this purpose, it is sufficient to employ a bandwidth matrix of the form $G = g^2 I$, where g is selected to minimise the Sum of AMSE (SAMSE) criterion (Duong and Hazelton, 2003). The plug-in bandwidth matrix $\hat{H}_{\mathrm{PI}}$ which is needed for estimating the intensity of the observed point pattern is then obtained by minimising the plug-in estimate of AMISE, PI(H). The first term on the right-hand side represents the asymptotic integrated squared bias, the second term represents the asymptotic integrated variance.

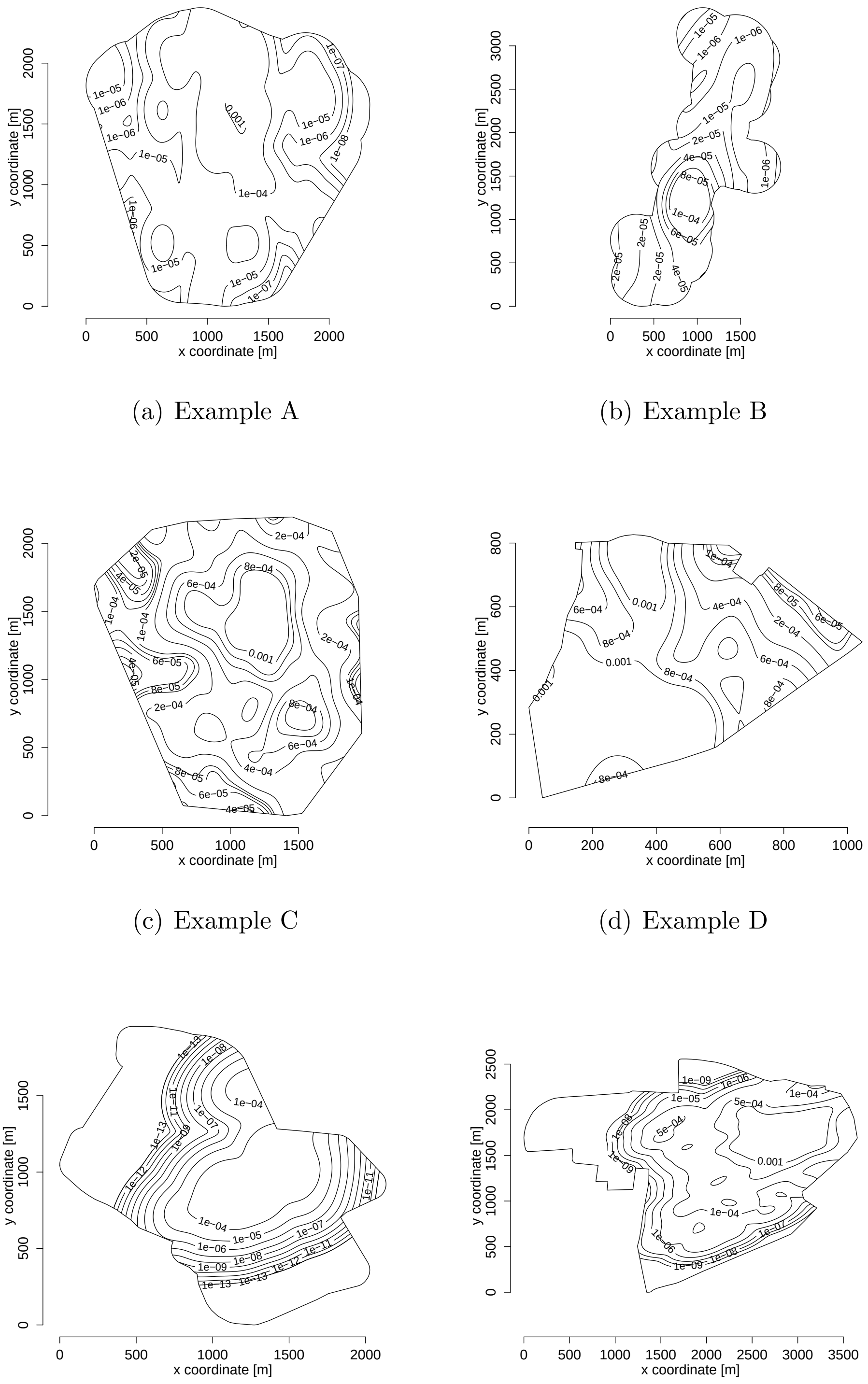
SCV Smooth cross-validation combines MISE and AMISE: The sum of the estimated exact integrated squared bias and the estimated asymptotic integrated variance is minimised. This can be interpreted as minimisation of the MISE for data which have been presmoothed with another kernel L_G with pilot bandwidth matrix G (Duong and Hazelton, 2005b). In case of Gaussian kernels K_H and L_G, the criterion is

$$\mathrm{SCV}(H) = \frac{1}{n_Y^2} \sum_{i=1}^{n_Y} \sum_{j=1}^{n_Y} (\phi_{2H+2G} - 2\phi_{H+2G} + \phi_{2G})(\mathbf{y}_i - \mathbf{y}_j) + \frac{1}{n_Y} \frac{1}{4\pi} |H|^{-\frac{1}{2}}, \tag{4.14}$$

where $\phi(\cdot)$ is the bivariate normal density with zero mean vector and identity covariance matrix and $\phi_H(\mathbf{y}) = |H|^{-\frac{1}{2}} \phi(H^{-\frac{1}{2}}\mathbf{y})$.

While the optimal bandwidth H is required to be unrestricted, it is sufficient to employ a bandwidth matrix of the form $G = g^2 I$ as pilot bandwidth for computing $\mathrm{SCV}(H)$. To obtain a reasonable result, the data are sphered before, i.e. the sample covariance matrix S is computed and transformed data $\tilde{\mathbf{y}} = S^{-\frac{1}{2}}\mathbf{y}$ are used to determine $\tilde{G}$, from which $G = S^{-\frac{1}{2}} \tilde{G} S^{-\frac{1}{2}}$ is derived. The choice of g is illustrated in detail in Duong and Hazelton (2005a).

Figure 4.4 shows the estimated intensity for Examples A to F. The variance-covariance matrix of the Gaussian kernel was selected minimising the smooth cross-validation criterion. Note that the contours are neither equidistant nor the same for all examples, but were chosen so that they can be compared with the borders of the high-risk zones later on.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 4.4.: Contour plots of the estimated intensity $\hat{\lambda}_Y(\mathbf{s})$ for the two properties of interest; the variance-covariance matrix of the Gaussian kernel was chosen automatically using smooth cross-validation.

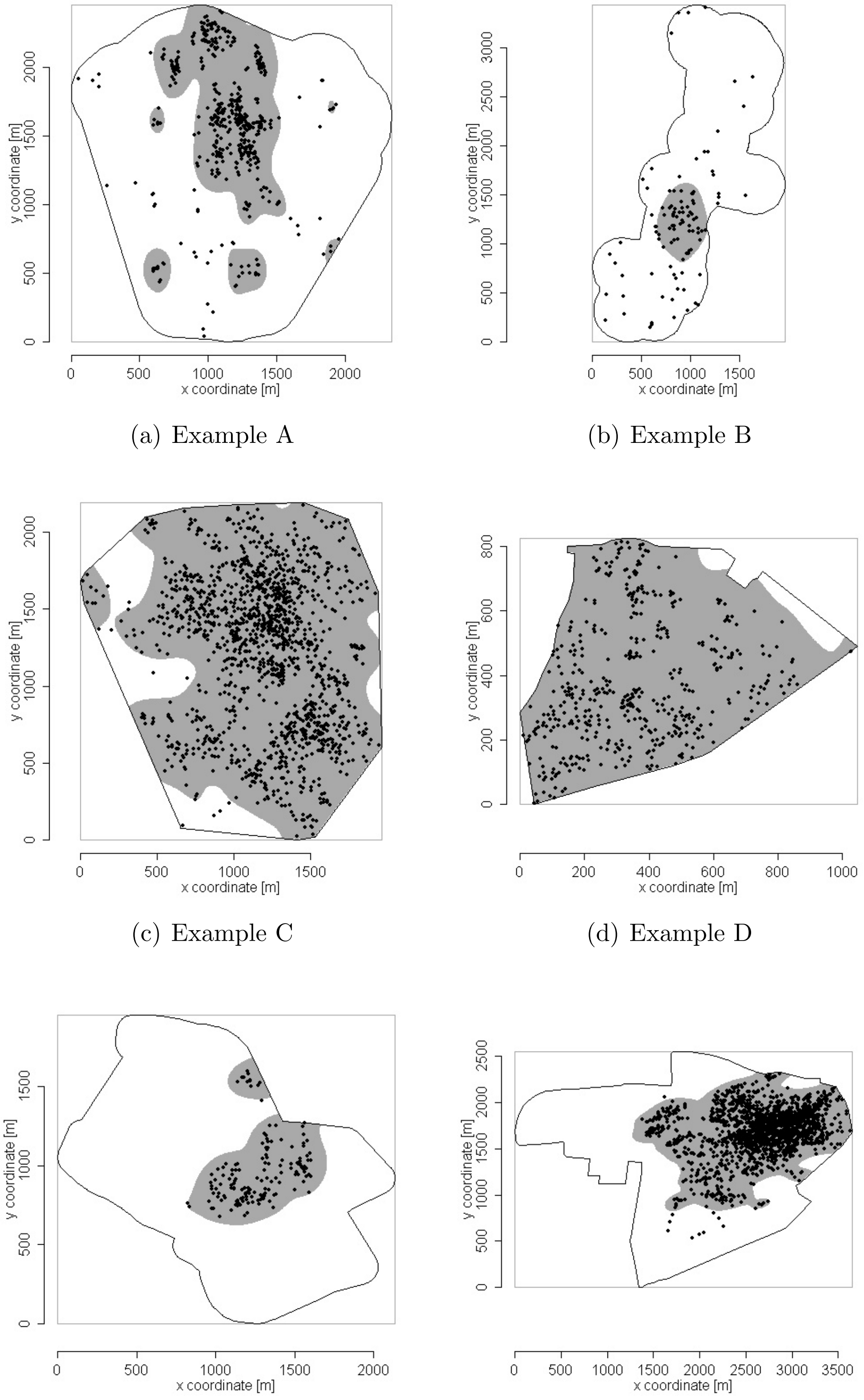
4.3.3. Direct specification of the threshold c

A simple approach for constructing high-risk zones based on the estimated intensity results if the cut-off value c is specified directly. The interpretation of a high-risk zone

$$R_c = \{\mathbf{s} \in W : \hat{\lambda}_Z(\mathbf{s}) \geq c\} \tag{4.15}$$

determined in this way is that the expected number of unexploded bombs per unit square does not exceed c in any location outside the high-risk zone.

High-risk zones obtained for $q = 0.10$ when cutting at $\hat{\lambda}_Z(\mathbf{s}) = 0.00001$ are shown in Fig. 4.5. This specific choice of c results in high-risk zones which are generally rather small, especially for Examples A and B. Only for Example D, all observations are contained in the high-risk zone. The differences between the examples make clear that one needs to take into account the average intensity: It is low for Examples A and B compared to Examples C, D and F. The average intensity for Example E, though, is larger than the average intensity in Example B. However, the events are concentrated on a small part of the observation window of Example E. It is interesting to consider the point density distribution function $G(t)$ (see Section 2.1.2) to gain more insight. Figure 4.6 depicts the estimated point density distribution function of Y and the value of $\hat{\lambda}_Y(\mathbf{s})$ corresponding to $\hat{\lambda}_Z(\mathbf{s}) = 0.00001$ for $q = 0.1$. The point density distribution function of Example D differs considerably from all others and is almost linear. For Examples C and D, the estimated intensity is below the threshold in less than 20 % of the observation window. For all other examples, this fraction is much higher (between 50 % and 90 %). The point density distribution functions for Examples E and F are extremely steep for values near 0, which reflects that the observations are concentrated on a part of the window only.



(a) Example A

(b) Example B

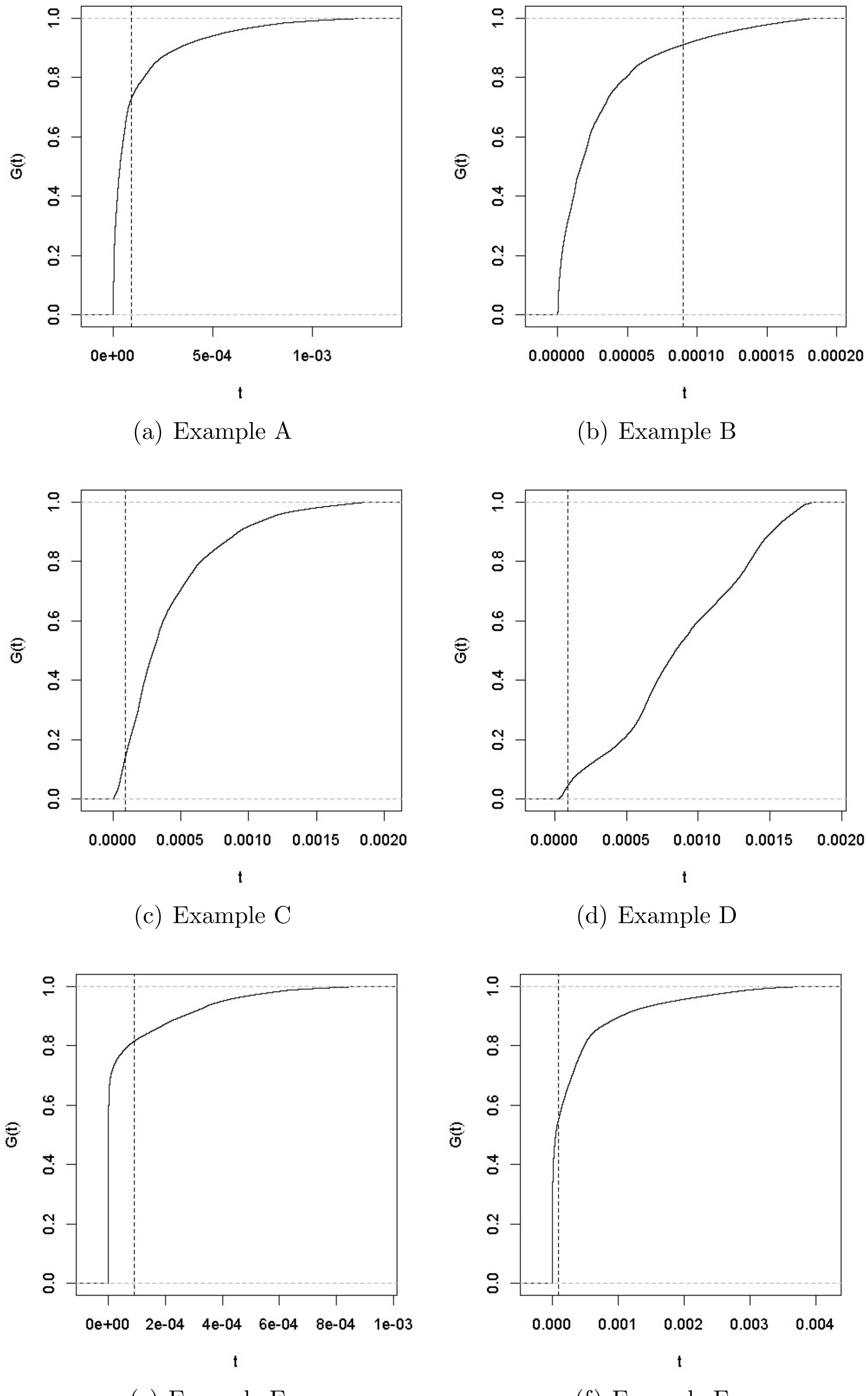
(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 4.5.: High-risk zones (shaded grey areas) obtained for the intensity-based method by cutting at $\hat{\lambda}_Z(\mathbf{s}) = 0.00001$ (probability of non-explosion $q = 0.10$).



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 4.6.: Estimated point density distribution function of Y and threshold determining the high-risk zones.

4.3.4. Specification of the threshold c via the global failure probability

To find an appropriate value for c, the *failure probability* of high-risk zone R_c,

$$P\{N_Z(W\backslash R_c) > 0\}, \tag{4.16}$$

i.e. the probability that not all unexploded bombs are covered by the high-risk zone, is considered. We want to find a high-risk zone R_c for which this probability equals a fixed value $0 \leq \alpha \leq 1$.

As pointed out, the number of unexploded bombs $N_Z(W)$ is unknown. If X is assumed to be an inhomogeneous Poisson point process, Y and Z are also Poisson point processes (Illian et al., 2008, page 367) and hence

$$N_Z(\mathcal{B}) \sim \text{Po}\{\Lambda_Z(\mathcal{B})\} \text{ with } \Lambda_Z(\mathcal{B}) = q\Lambda_X(\mathcal{B}) = \frac{q}{1-q}\Lambda_Y(\mathcal{B}). \tag{4.17}$$

Given the probability of non-explosion q and an estimate of the intensity function, the estimated failure probability can be computed as

$$\begin{aligned} \hat{P}\{N_Z(W\backslash R_c) > 0\} &= 1 - \exp\{-\hat{\Lambda}_Z(W\backslash R_c)\} \cdot \{\hat{\Lambda}_Z(W\backslash R_c)\}^0 \cdot \frac{1}{0!} \\ &= 1 - \exp\left[-\left\{\frac{q}{1-q}\left(\int_{(W\backslash R_c)} \hat{\lambda}_Y(\mathbf{y})d\mathbf{y}\right)\right\}\right]. \end{aligned} \tag{4.18}$$

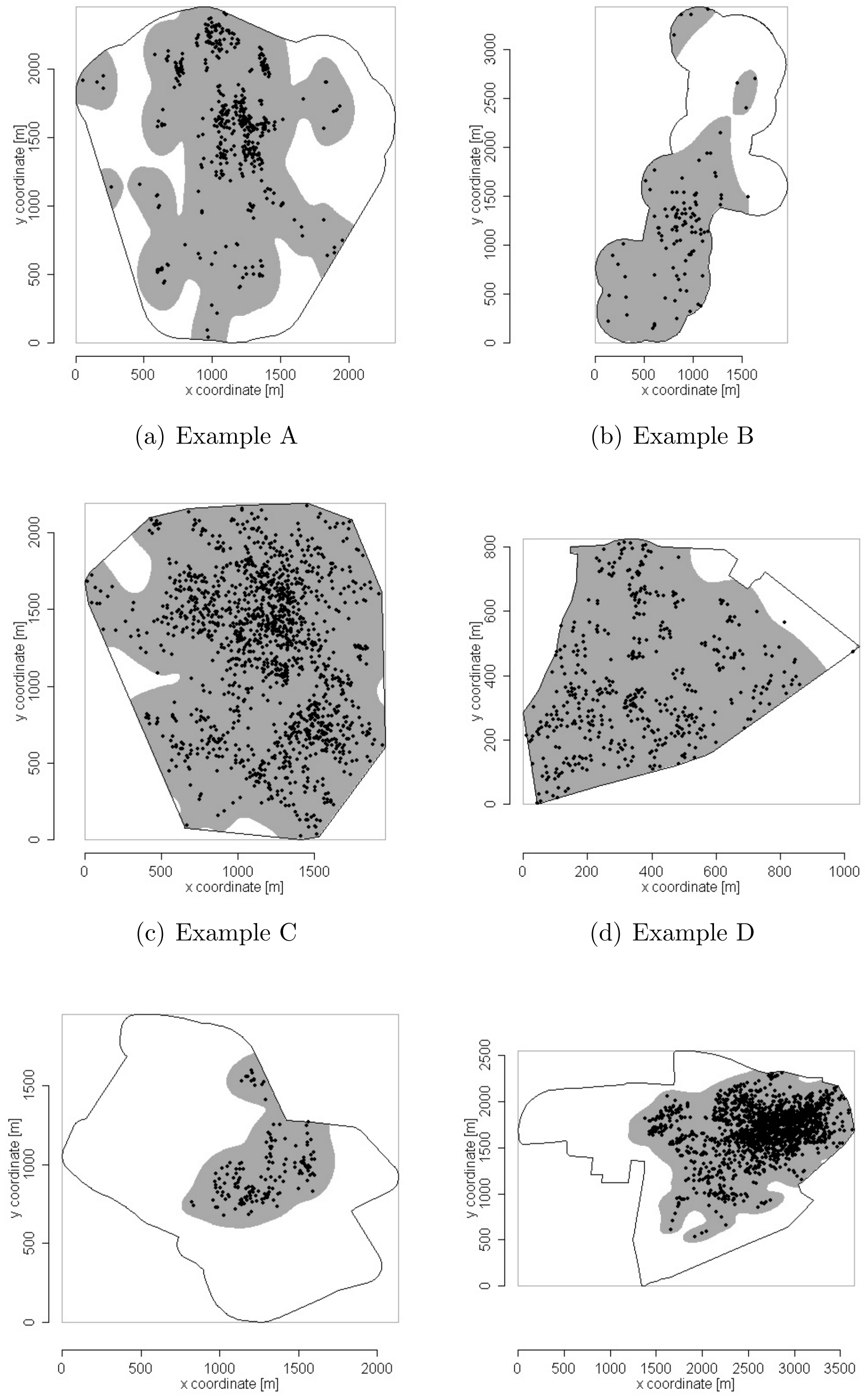
The threshold c for which $\hat{P}\{N_Z(W\backslash R_c) > 0\} = \alpha$ holds is determined by a numeric root finding procedure. Figure 4.7 illustrates the high-risk zones which are obtained if q is assumed to be 0.1 and $\alpha = 0.4$ is chosen. Most of the observed events are covered by the high-risk zones. The high-risk zone does not comprise the entire observation window for any of the examples. Compared to high-risk zones obtained by using a distance-based method, the intensity-based high-risk zones are less ragged and have smoother borders.

As the parameter α defines the global risk (i.e. the probability that not all unobserved events are covered) of a high-risk zone, it may be problematic to use the same α for properties of different size.

This does not play any role if the threshold c is specified directly. In this case, it is possible to estimate the corresponding failure probability

$$\hat{P}\{N_Z(W\backslash R_c) > 0\} = 1 - \exp\left[-\left\{\frac{q}{1-q}\hat{\Lambda}_Y(W\backslash R_c)\right\}\right] \tag{4.19}$$

afterwards.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 4.7.: High-risk zones (shaded grey areas) obtained for the intensity-based method with failure probability $\alpha = 0.4$ (probability of non-explosion $q = 0.10$).

5. Application and evaluation of the construction methods for high-risk zones

In this chapter, the construction methods introduced in Chapter 4 are applied to the bomb crater data. First, summary functions are used to investigate if the chosen inhomogeneous Poisson point process model with intensity function $\hat{\lambda}_Y(\mathbf{s})$ is appropriate. For this purpose, Monte Carlo tests based on the K-function and the pair correlation function are performed. Then, an evaluation procedure is presented and results obtained for the three construction methods are discussed. Finally, the properties of the intensity-based and the quantile-based construction methods are compared and a recommendation is given.

Some of the results for Examples A and B have been presented in Mahling et al. (2013), where the methodology has been introduced.

5.1. Model check

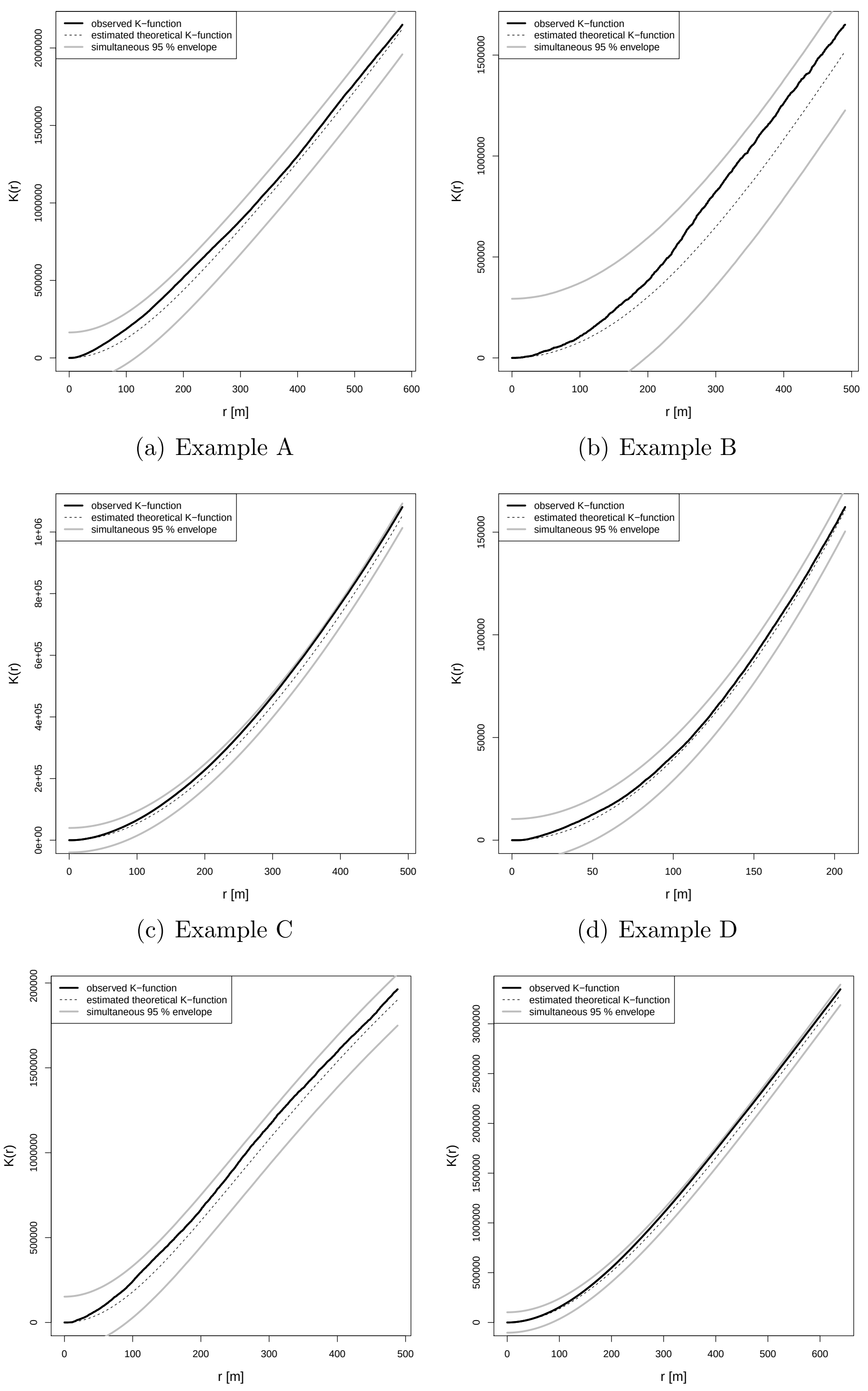
For the intensity-based method, we assume that the observed patterns are realisations of an inhomogeneous Poisson point process with intensity function $\hat{\lambda}_Y(\mathbf{s})$. As the inhomogeneous K-functions indicate that the observed patterns can be described sufficiently well by an inhomogeneous Poisson process, it generally seems justified to use the intensity-based construction method for high-risk zones. However, it is advisable to check the inhomogeneous Poisson model with intensity function $\hat{\lambda}_Y(\mathbf{s})$ estimated as described in Section 4.3.2 in a more rigorous way. Since the intensity of the Poisson process is estimated nonparametrically, it is not possible to consider residuals as proposed in Baddeley et al. (2005) and Baddeley et al. (2008) or Guan (2008a), as they are only defined for fully parametric models. Similar problems are encountered concerning the application of other approaches such as the tests proposed by Brix et al. (2001), where the observation window is split into cells and it is not possible to take into account the estimated intensity function $\hat{\lambda}_Y(\mathbf{s})$.

Therefore, a traditional approach was chosen, a Monte Carlo test (see Ripley, 1981). The initial idea to these tests is attributed to Barnard (1963). They were discussed in more detail by Ripley (1977). Diggle (1979) recommends "the use of several tests to investigate different aspects of the same data-set" and states that the significance level should not be interpreted too strictly. If the model which is tested was fitted via a summary characteristic (see Section 7.2), a summary characteristic should be chosen which is of different nature than the characteristic used for parameter estimation (Illian et al., 2008, page 456).

A Monte Carlo test based on the K-function with significance level 0.05 was performed, whose critical points are represented by a simultaneous envelope. The theoretical mean value of Ripley's K-function for patterns generated by our model was calculated as the average of 99 inhomogeneous Poisson point processes simulated from the estimated intensity function $\hat{\lambda}_Y(\mathbf{s})$. The symmetric envelopes are the result of adding and subtracting the fifth largest absolute difference between this average and the K-functions of 99 further simulated inhomogeneous Poisson point processes with intensity function $\hat{\lambda}_Y(\mathbf{s})$. Details can be found in Baddeley (2008). The same procedure was applied for the pair correlation function.

As seen from Figure 5.1, all K-functions are entirely situated inside the envelopes, so the K-functions of the observed patterns do not differ significantly from the K-functions of an inhomogeneous Poisson point process with the estimated intensity function. As a consequence, there is no evidence against using such modelling to describe the observed bomb patterns. In addition, recall that the inhomogeneous Poisson model fits the subject-matter theory better than cluster models (see Section 3.6).

Furthermore, we performed a Monte Carlo test based on the pair correlation function. Figure 5.2 shows the estimated pair correlation function which we obtain when we follow the recommendation of Stoyan and Stoyan (1992). As we can see, the estimated pair correlation function for Example A exceeds the values of the envelope for small r. The assumption of an inhomogeneous Poisson process is violated, possibly because of clustering in small r ($r < 50\ m$) in the case of Example A (note, however, that the estimation is difficult for $r < h$, cf. Section 3.5). This finding is contrary to the result of the Monte Carlo tests based on the K-function. We discuss the consequences of such clustering in Chapter 7. The estimated pair correlation functions for Examples B and E are entirely situated inside the envelope, so there is no evidence against the inhomogeneous Poisson process model. For Examples C, D and F, the estimated pair correlation function is close to the envelope for small r, which seems to be an estimation issue and not an indication that the model is inappropriate.



(a) Example A

(b) Example B

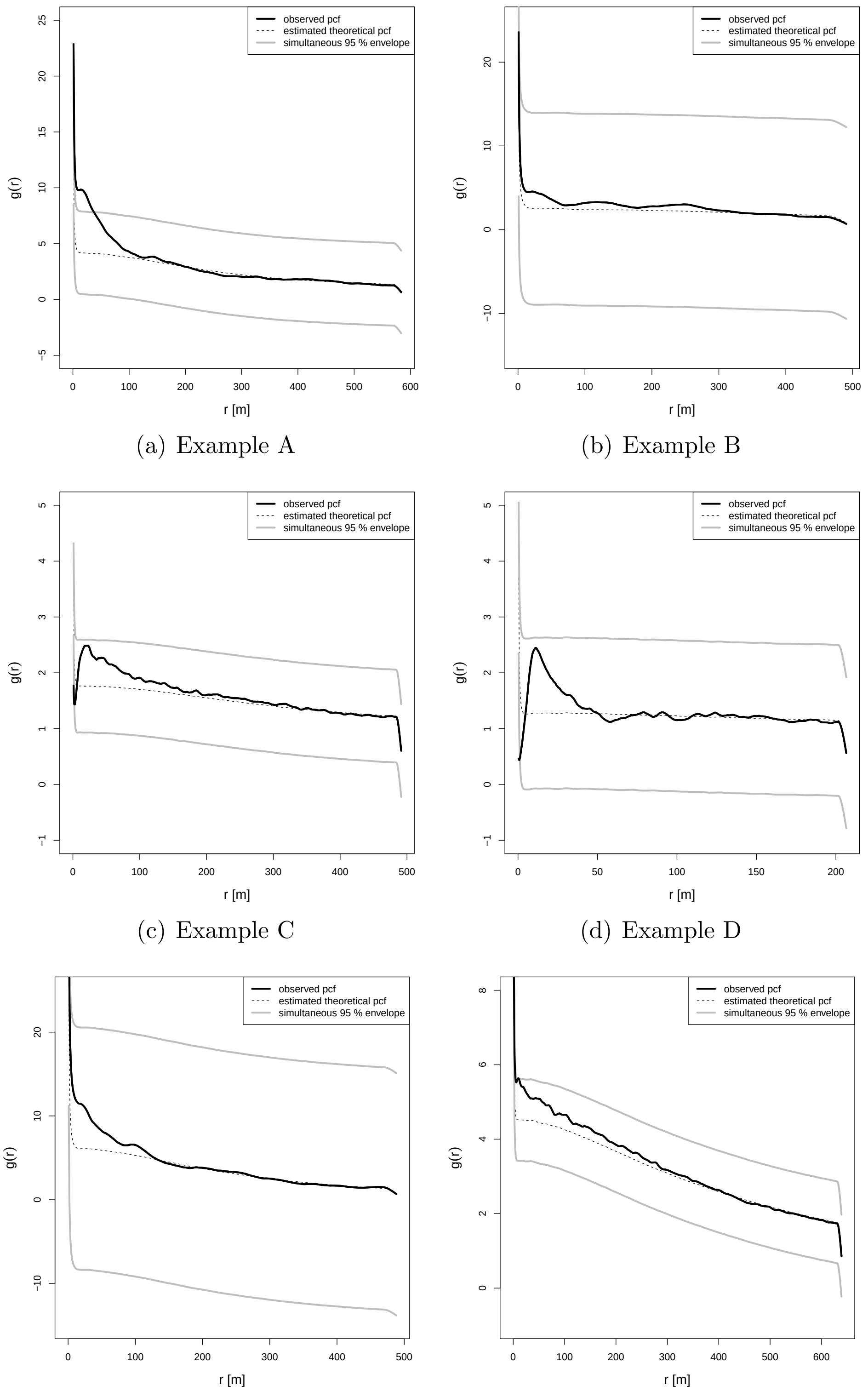
(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 5.1.: Ripley's K-function and simultaneous envelopes generated by inhomogeneous Poisson point processes representing the critical points of a Monte Carlo test with significance level 0.05.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 5.2.: Pair correlation function and simultaneous envelopes generated by inhomogeneous Poisson point processes representing the critical points of a Monte Carlo test with significance level 0.05.

5.2. Simulation study

The behaviour of the proposed construction methods is now examined in order to evaluate the properties of the methods and, finally, to find out which method yields the best high-risk zones. As no data about the location of unexploded bombs is available, a simulation-based procedure is used.

5.2.1. Setting

To obtain patterns that are as realistic as possible, the following approach was chosen: The observed patterns Y were taken as 'full patterns', which means that one pretends to know the full process X. Denote by $\tilde{X} = Y$ this artificially defined full process. In the next step, each of the points of $\tilde{X}$ was thinned with probability q (i.e. the retention probability was $1 - q$) by drawing independent Bernoulli distributed random variables, resulting in the process $\tilde{Y}$ of observed bombs and the process $\tilde{Z}$ of unobserved bombs. Of course, the intensities of $\tilde{X}$, $\tilde{Y}$ and $\tilde{Z}$ are smaller than those of the real processes X, Y and Z, but this procedure allows to perform a simulation without making an assumption about the underlying point process model: It is not necessary to assume an inhomogeneous Poisson point process or any other point process model to obtain $\tilde{X}$, $\tilde{Y}$ and $\tilde{Z}$. Moreover, the parameter q never exceeds 0.15, so only a small fraction of the original observations is thinned, which means that only little information is lost and the scenario reflects the real problem in an appropriate way. The high-risk zone was then computed based only on those observations assigned to $\tilde{Y}$, whereas the observations in $\tilde{Z}$ were used to evaluate the high-risk zone afterwards.

According to OFD Niedersachsen, a probability of non-explosion between 0.10 and 0.15 can be regarded as a well-established value. As global risk has never before been quantified in this field, no standard values for the failure probability α and the quantile p exist, so they were chosen based on expert knowledge of OFD Niedersachsen and α was set to 0.4, 0.2 or 0.1 for both examples. These values may seem large. However, one needs to take into account that an unexploded bomb outside the high-risk zone does not necessarily mean that somebody will die or be hurt. The bombs may be located in depths of several metres where they might not be affected by construction work. Even if they are found, this does not necessarily mean that they will cause damage. To ensure comparability of the quantile-based and the intensity-based construction method, different quantiles were used for Examples A to F: The 99 %, the 99.5 % and the 99.9 % quantile were used for Example A, C, D, E and F and the 95 %, the 97.5 % and the 99 % quantile for Example B. Three radii for the discs used in the traditional method were considered: According to OFD Niedersachsen, a radius of 50 m was used in the past, whereas 100 m and 150 m are common values nowadays. In each of those settings, 1000 iterations–in which $\tilde{X}$ was thinned and thus different processes $\tilde{Y}$ and $\tilde{Z}$ were obtained–were performed.

The aim was to find the method which yields the smallest zones covering as many unexploded bombs as possible. Moreover, it was investigated whether the parameters α and p of the intensity-based and the quantile-based method are adhered to: The probability that at least one unexploded bomb lies outside the high-risk zone should be α in case of the intensity-based construction method. Therefore, the computed fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside was compared with α. For the quantile-based construction method, the fraction p_{miss} of unexploded bombs outside the high-risk zone should take a value near $1-p$. To check this, the fraction p_{miss} was computed in every iteration, as well as the area of the zone.

5.2.2. Results for the bomb crater data

General results

Example A The results for Example A are shown in Table 5.1 for the quantile-based and the intensity-based method, as well as for the traditional method. The table contains the mean of p_{miss} and the area, as well as the fraction p_{out}. For the quantile-based method, the mean of p_{miss} of 1000 iterations is close to $1-p$ for all six combinations of parameters, whereas for the intensity-based construction method, the fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside exceeds α in most cases. The relative bias $(p_{\text{out}}-\alpha)/\alpha$ of the intensity-based method is between -0.063 and 1.840, the mean of the relative bias $(p_{\text{miss}}-(1-p))/(1-p)$ of the quantile-based method between -0.021 and 1.448.

Table 5.1.: Results of the simulation: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside and mean area of the zone, Example A, intensity-based method (INT), quantile-based method (QUANT) and traditional method (TRAD)

A	q	0.1	0.1	0.1	0.15	0.15	0.15
INT	α	0.4	0.2	0.1	0.4	0.2	0.1
	mean p_{miss}	0.011	0.006	0.005	0.010	0.005	0.005
	p_{out}	0.375	0.249	0.192	0.498	0.318	0.284
	mean area in m^2	2711785	2987652	3186754	2870287	3115226	3295427
QUANT	p	0.99	0.995	0.999	0.99	0.995	0.999
	mean p_{miss}	0.010	0.005	0.002	0.011	0.006	0.002
	p_{out}	0.324	0.196	0.086	0.465	0.285	0.139
	mean area in m^2	2663233	3097025	3345761	2701678	3108936	3367984
TRAD	radius	50 m	100 m	150 m	50 m	100 m	150 m
	mean p_{miss}	0.086	0.019	0.010	0.092	0.022	0.011
	p_{out}	0.979	0.575	0.377	1.000	0.751	0.545
	mean area in m^2	927549	1873954	2646576	907818	1847601	2622831

For the quantile-based method, the mean of p_{miss} is higher for $q = 0.15$ than for $q = 0.10$. The mean area of the high-risk zone and the fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside increased with the probability of non-explosion, as well. The situation is similar for the traditional method. The mean area, however, decreased for a higher probability of non-explosion. For the intensity-based method, p_{out} takes a higher value for $q = 0.15$ than for $q = 0.10$. Again, the mean area increased with the probability of non-explosion, but the mean of p_{miss} did not.

Example B The results for Example B are shown in Table 5.2. Like for Example A, the mean of p_{miss} corresponds to the quantile that was used for the quantile-based construction method. For the intensity-based construction method, the fraction p_{out} clearly exceeds the given α. The relative bias $(p_{\text{out}} - \alpha)/\alpha$ of the intensity-based method is between 0.233 and 1.185, whereas the mean of the relative bias $(p_{\text{miss}} - (1 - p))/(1 - p)$ of the quantile-based method is between -0.036 and 0.517. Again, the fractions p_{out} and p_{miss} often increased with the probability of non-explosion, but they decreased for $p = 0.99$ and $\alpha = 0.2$, respectively. The mean area increased for the intensity-based and the quantile-based method, whereas it decreased for the traditional method.

Table 5.2.: Results of the simulation: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside and mean area of the zone, Example B, intensity-based method (INT), quantile-based method (QUANT) and traditional method (TRAD)

B	q	0.1	0.1	0.1	0.15	0.15	0.15
INT	α	0.4	0.2	0.1	0.4	0.2	0.1
	mean p_{miss}	0.066	0.052	0.015	0.054	0.025	0.016
	p_{out}	0.493	0.437	0.140	0.583	0.281	0.215
	mean area in m^2	2102079	2438842	2680396	2281263	2580386	2794571
QUANT	p	0.95	0.975	0.99	0.95	0.975	0.99
	mean p_{miss}	0.048	0.026	0.015	0.051	0.028	0.015
	p_{out}	0.375	0.223	0.141	0.489	0.317	0.175
	mean area in m^2	2456074	2667156	2792131	2467116	2673899	2825624
TRAD	radius	50 m	100 m	150 m	50 m	100 m	150 m
	mean p_{miss}	0.444	0.268	0.161	0.458	0.276	0.166
	p_{out}	0.990	0.937	0.811	0.997	0.980	0.932
	mean area in m^2	514930	1330002	1980762	493881	1288033	1935110

Example C The results for Example C are shown in Table 5.3. For the intensity-based method, p_{out} is too small in four of six cases. The relative bias $(p_{\text{out}} - \alpha)/\alpha$ of the intensity-based method is between -0.333 and 0.300, the mean of the relative bias $(p_{\text{miss}} - (1-p))/(1-p)$ of the quantile-based method between 0.041 and 0.271. The mean area of the high-risk zone increased with q for the intensity-based and the quantile-based method in most cases. It decreased for the traditional method. The fraction p_{out} increased for all three methods, whereas the mean of p_{miss} decreased for the intensity-based method.

Table 5.3.: Results of the simulation: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside and mean area of the zone, Example C, intensity-based method (INT), quantile-based method (QUANT) and traditional method (TRAD)

C	q	0.1	0.1	0.1	0.15	0.15	0.15
INT	α	0.4	0.2	0.1	0.4	0.2	0.1
	mean p_{miss}	0.002	0.001	0.001	0.002	0.001	0.001
	p_{out}	0.267	0.161	0.102	0.339	0.166	0.130
	mean area in m^2	3160352	3228483	3268014	3199953	3253800	3285440
QUANT	p	0.99	0.995	0.999	0.99	0.995	0.999
	mean p_{miss}	0.010	0.005	0.001	0.010	0.005	0.001
	p_{out}	0.719	0.475	0.159	0.839	0.606	0.199
	mean area in m^2	2821615	2924161	3267627	2822275	2942857	3265169
TRAD	radius	50 m	100 m	150 m	50 m	100 m	150 m
	mean p_{miss}	0.043	0.002	0.001	0.048	0.003	0.002
	p_{out}	0.267	0.161	0.102	0.339	0.166	0.130
	mean area in m^2	2273650	3030836	3242156	2241303	3015633	3237421

Example D The results for Example D are shown in Table 5.4. The fraction p_{out} exceeded α in four of six cases. For the quantile-based method, the mean of p_{miss} was too large for $p = 0.999$. The relative bias $(p_{\text{out}} - \alpha)/\alpha$ of the intensity-based method is between -0.358 and 1.600, the mean of the relative bias $(p_{\text{miss}} - (1-p))/(1-p)$ of the quantile-based method between 0.229 and 3.520. For the quantile-based and the traditional method, the mean area decreased for larger q, the fraction p_{out} and the mean of p_{miss} increased. For the intensity-based method, the mean area increased, the mean of p_{miss} decrased in two of three cases and the fraction p_{out} increased.

Example E The results for Example E are shown in Table 5.5. The fraction p_{out} was clearly smaller than α. The mean of p_{miss}, in contrast, was clearly too large for $p = 0.999$. The relative bias $(p_{\text{out}} - \alpha)/\alpha$ of the intensity-based method is between -1.000 and -0.778, the mean of the relative bias $(p_{\text{miss}} - (1-p))/(1-p)$ of the quantile-based method between -0.137 and 6.049. For larger q, the area of the high-risk zones determined by using the intensity-based method increased, both the fraction p_{out} and the mean of p_{miss} decrased. For the quantile-based method, mean area and both fractions increased, whereas for the traditional method, the area and p_{out} decreased.

Example F The results for Example F are shown in Table 5.6. The fraction p_{out} was too small compared to α. The relative bias $(p_{\text{out}} - \alpha)/\alpha$ of the intensity-based method is between -0.800 and -0.025, the mean of the relative bias $(p_{\text{miss}} - (1-p))/(1-p)$ of the quantile-based method between -0.014 and 0.172. For $q = 0.15$, the mean area increased for the intensity-based and the quantile-based method, whereas it decreased for the traditional

Table 5.4.: Results of the simulation: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside and mean area of the zone, Example D, intensity-based method (INT), quantile-based method (QUANT) and traditional method (TRAD)

D	q	0.1	0.1	0.1	0.15	0.15	0.15
INT	α	0.4	0.2	0.1	0.4	0.2	0.1
	mean p_{miss}	0.006	0.006	0.004	0.006	0.005	0.004
	p_{out}	0.257	0.250	0.173	0.368	0.285	0.260
	mean area in m^2	469300	488269	500143	480154	495795	505061
QUANT	p	0.99	0.995	0.999	0.99	0.995	0.999
	mean p_{miss}	0.012	0.007	0.004	0.013	0.007	0.005
	p_{out}	0.390	0.260	0.172	0.547	0.371	0.259
	mean area in m^2	472870	491366	515957	471551	490914	514553
TRAD	radius	50 m	100 m	150 m	50 m	100 m	150 m
	mean p_{miss}	0.016	0.004	0.004	0.017	0.004	0.004
	p_{out}	0.257	0.250	0.173	0.368	0.285	0.260
	mean area in m^2	451480	507911	518090	448556	506760	517830

method. Both fractions decreased for the intensity-based method. For the quantile-based method, p_{out} increased, whereas p_{miss} increased for the traditional method.

Relation between area and fraction of unexploded bombs outside

Scatterplots illustrate the relation between the area of the high-risk zones and the fraction of unexploded bombs located outside the zones. Note that they are scaled individually for every example and for the traditional method compared to the intensity-based and quantile-based methods. In some cases, the high-risk zones obtained by the traditional method are much smaller.

Example A Figure 5.3 illustrates the relation between the area of the high-risk zone and the fraction of unexploded bombs outside the zone (p_{miss}) for each of the 6000 iterations in total for all three construction methods. As one would expect, a negative correlation between area and fraction p_{miss} is observed. In 4505 of the 6000 iterations for the quantile-based method, no unexploded bomb was located outside the zone at all. This was the case in 4084 iterations for the intensity-based method and 1773 iterations for the traditional method. As the number of unexploded bombs varies between the iterations, many distinct values for the fraction of unexploded bombs outside the zone have been obtained in those cases where at least one unexploded bomb was located outside the zone.

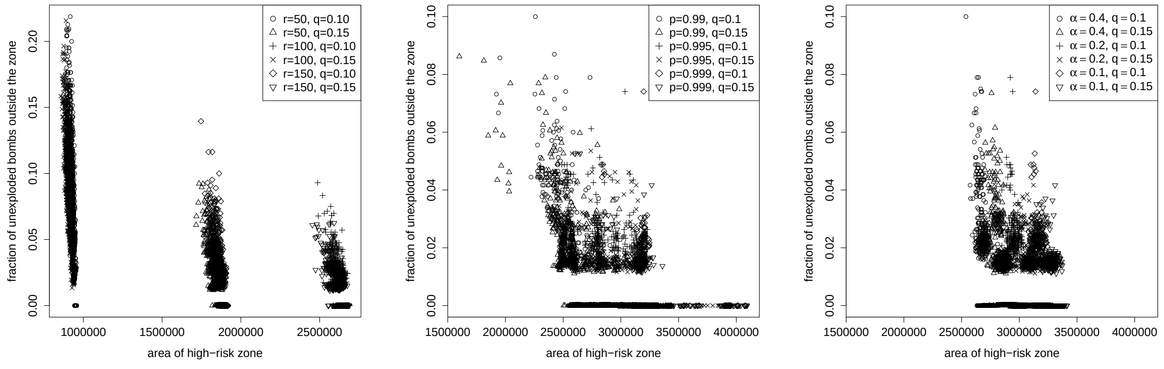
The area of high-risk zones determined by using the traditional method varies little for given radius. The fraction p_{miss} is generally large. The area of the quantile-based high-risk zones varies most. The performance of the quantile-based and the intensity-based

Table 5.5.: Results of the simulation: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside and mean area of the zone, Example E, intensity-based method (INT), quantile-based method (QUANT) and traditional method (TRAD)

E	q	0.1	0.1	0.1	0.15	0.15	0.15
INT	α	0.4	0.2	0.1	0.4	0.2	0.1
	mean p_{miss}	0.007	0.001	0	0.003	0	0
	p_{out}	0.089	0.008	0.002	0.049	0.004	0
	mean area in m^2	532551	615330	683260	580401	660279	726555
QUANT	p	0.99	0.995	0.999	0.99	0.995	0.999
	mean p_{miss}	0.009	0.007	0.007	0.009	0.007	0.007
	p_{out}	0.103	0.083	0.083	0.140	0.118	0.118
	mean area in m^2	508166	551519	552729	517562	559894	560015
TRAD	radius	50 m	100 m	150 m	50 m	100 m	150 m
	mean p_{miss}	0.058	0.002	0	0.068	0.003	0
	p_{out}	0.089	0.008	0.002	0.049	0.004	0
	mean area in m^2	339066	542426	707879	333000	537834	703681

construction method is similar. It is not possible to decide if one of the two methods yields smaller high-risk zones and a smaller fraction p_{miss} at the same time.

Example B Compared to Example A, higher fractions of unexploded bombs outside the zone were obtained for Example B (Figure 5.4), especially for the traditional method. The performance of the quantile-based and the intensity-based method was similar. The quantile-based high-risk zones had a very large area in several cases. This was not the case for the intensity-based high-risk zones.

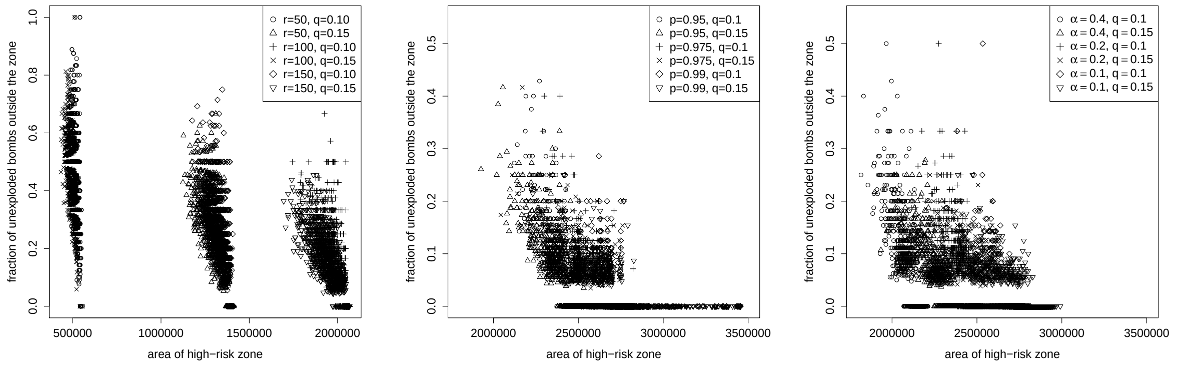


(a) traditional method (b) quantile-based method (c) intensity-based method

Figure 5.3.: Area of the high-risk zone and fraction of simulated unexploded bombs outside the zone for the traditional method, the quantile-based and the intensity-based method, Example A.

Table 5.6.: Results of the simulation: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside and mean area of the zone, Example F, intensity-based method (INT), quantile-based method (QUANT) and traditional method (TRAD)

F	q	0.1	0.1	0.1	0.15	0.15	0.15
INT	α	0.4	0.2	0.1	0.4	0.2	0.1
	mean p_{miss}	0.002	0.001	0	0.002	0.001	0
	p_{out}	0.348	0.195	0.025	0.332	0.177	0.020
	mean area in m^2	2804624	2977297	3101874	2901264	3054850	3168713
QUANT	p	0.99	0.995	0.999	0.99	0.995	0.999
	mean p_{miss}	0.010	0.005	0.001	0.010	0.005	0.001
	p_{out}	0.805	0.571	0.165	0.883	0.699	0.201
	mean area in m^2	2241518	2491832	2685726	2272046	2499031	2736677
TRAD	radius	50 m	100 m	150 m	50 m	100 m	150 m
	mean p_{miss}	0.032	0.003	0	0.035	0.003	0
	p_{out}	0.348	0.195	0.025	0.332	0.177	0.020
	mean area in m^2	1926788	2582710	2976585	1902359	2567389	2964595

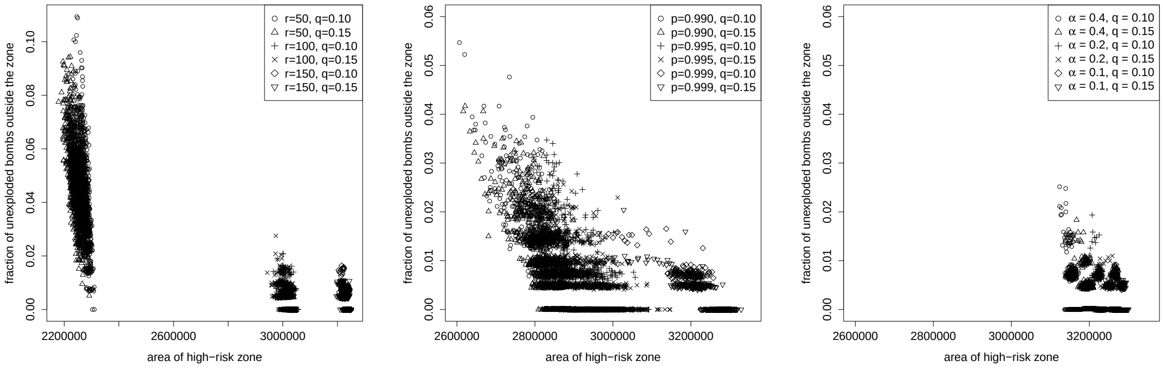


(a) traditional method (b) quantile-based method (c) intensity-based method

Figure 5.4.: Area of the high-risk zone and fraction of simulated unexploded bombs outside the zone for the traditional method, the quantile-based and the intensity-based method, Example B.

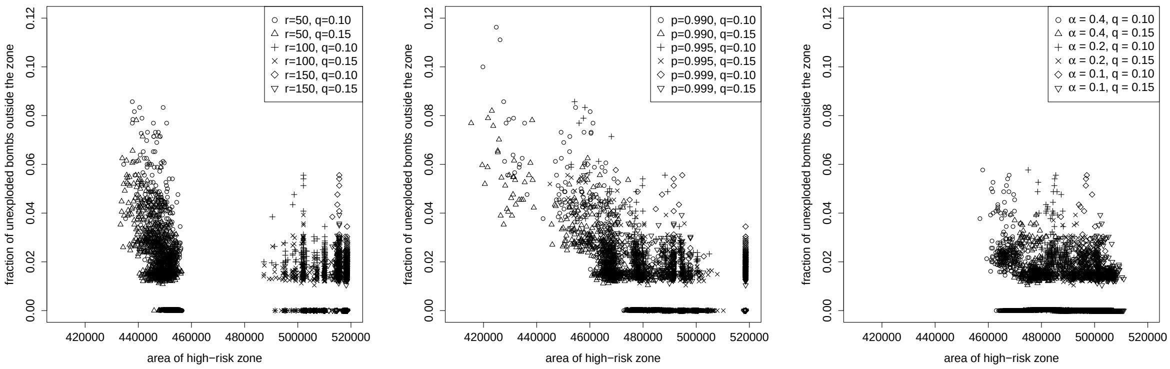
Example C The fraction p_{miss} was generally low for all three methods, especially for the intensity-based high-risk zones, whose area was large (Figure 5.5).

Example D For Example D, the fraction p_{miss} was generally low for all three methods (Figure 5.6). The results for the traditional method were close to those for the quantile-based high-risk zones. The intensity-based high-risk zones had the largest area and the smallest fraction p_{miss}.



(a) traditional method (b) quantile-based method (c) intensity-based method

Figure 5.5.: Area of the high-risk zone and fraction of simulated unexploded bombs outside the zone for the traditional method, the quantile-based and the intensity-based method, Example C.

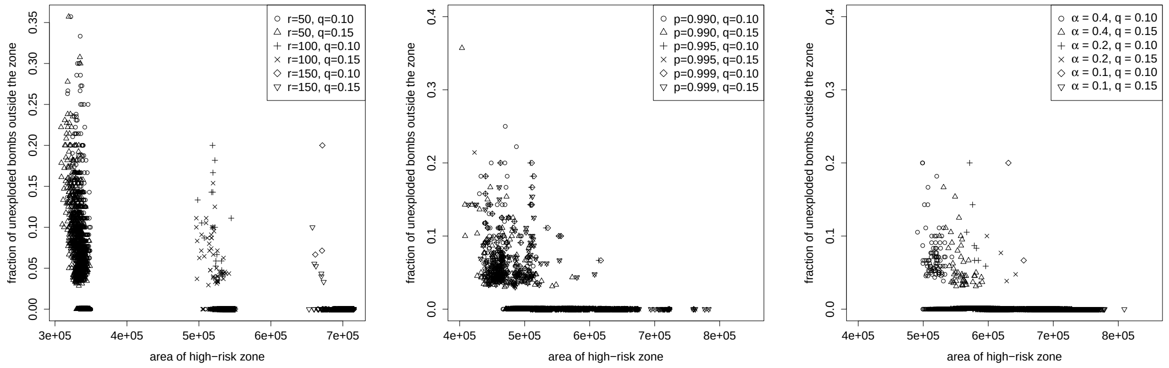


(a) traditional method (b) quantile-based method (c) intensity-based method

Figure 5.6.: Area of the high-risk zone and fraction of simulated unexploded bombs outside the zone for the traditional method, the quantile-based and the intensity-based method, Example D.

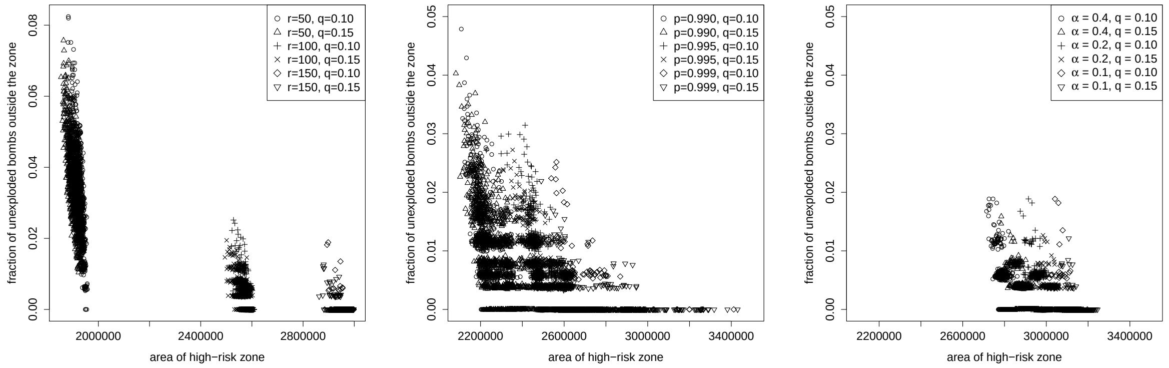
Example E For Example E, the relation between the area of the high-risk zone and the fraction of unexploded bombs outside the zone is depicted in Figure 5.7. For the intensity-based method and–less often–for the quantile-based method, $p_{\text{miss}} = 0$ was achieved in many cases. The area of some of these high-risk zones was very large.

Example F The fraction p_{miss} was generally low for all three methods for Example F (Figure 5.8). The intensity-based high-risk zones yielded the smallest fraction p_{miss}, but were rather large. Some of the quantile-based high-risk zones, however, were even larger.



(a) traditional method (b) quantile-based method (c) intensity-based method

Figure 5.7.: Area of the high-risk zone and fraction of simulated unexploded bombs outside the zone for the traditional method, the quantile-based and the intensity-based method, Example E.



(a) traditional method (b) quantile-based method (c) intensity-based method

Figure 5.8.: Area of the high-risk zone and fraction of simulated unexploded bombs outside the zone for the traditional method, the quantile-based and the intensity-based method, Example F.

Relation between specified parameters and resulting threshold

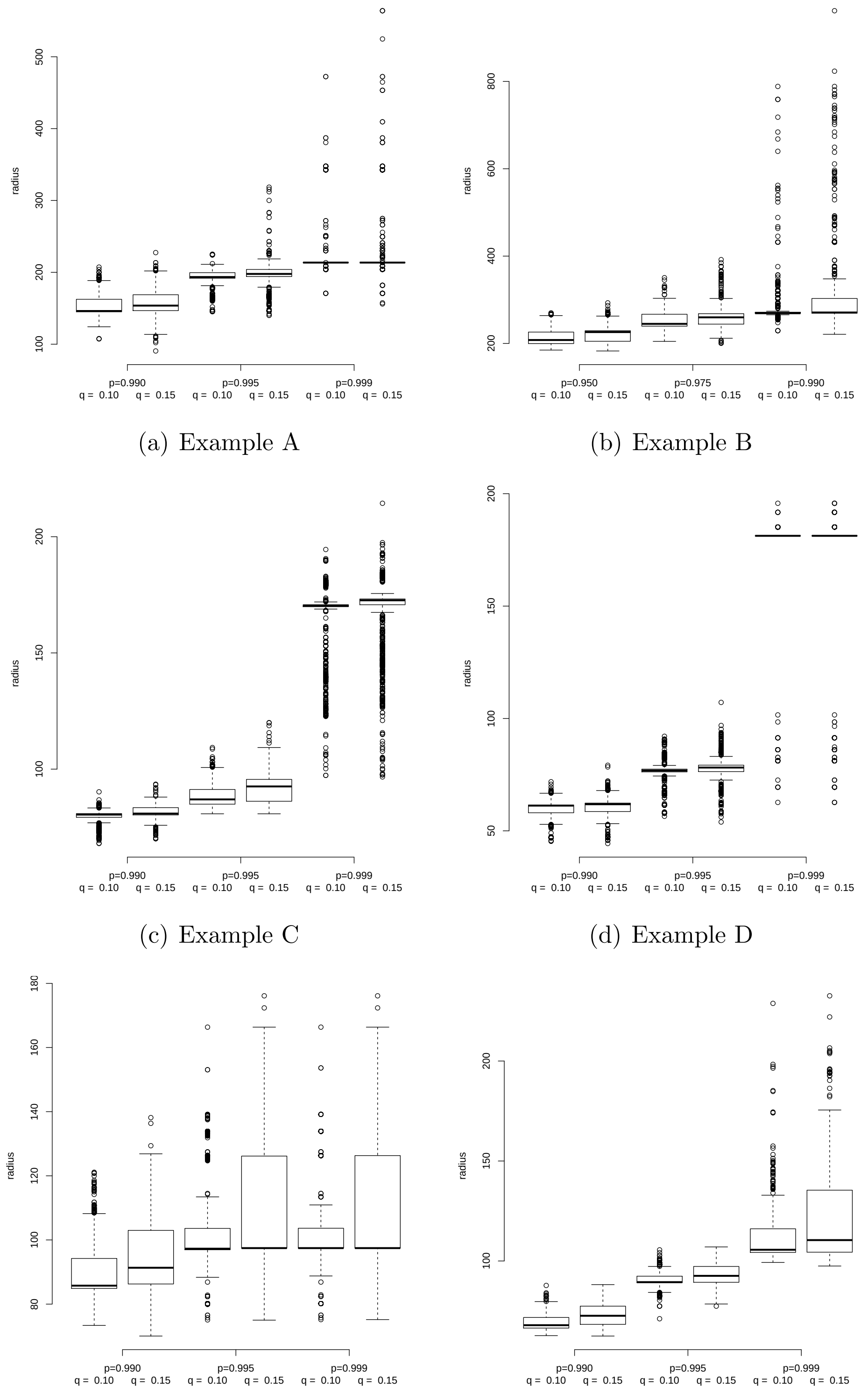
In Figure 5.9, the relation between the given quantile and the resulting radius for the quantile-based method is shown for $q = 0.10$ and $q = 0.15$. For Examples A and B, the radius increased only little for larger p. The radius is typically larger for a larger probability of non-explosion q. For Examples C and D, the radius obtained for $p = 0.999$ is extremely large in many cases.

Figure 5.10 shows the relation between the specified parameter α and the resulting threshold c for the intensity-based method. The resulting threshold c is larger for larger q and also for larger α. There are by far less outliers than for the radius of the quantile-based method. We observe a positive correlation between the average intensity (Chapter 3) of the six examples and the threshold c. If we consider the point density distribution functions of the six examples (Section 4.3.3), we see that the largest difference in c for different values of α is obtained for the Examples A, E and F, whose point density distribution function is steep for small values.

Summary

The resulting high-risk zones constructed using the traditional method are very small for some of the examples. The mean fraction of unexploded bombs outside the high-risk zone and the fraction of generated high-risk zones for which at least one unexploded bomb was located outside are large, especially for Example B. This would of course change if larger radii were chosen, but the choice of the radius will always remain arbitrary unless the quantile-based method is used. So the use of the traditional method cannot be recommended. A decision between the quantile-based and the intensity-based method is not possible at this point. Furthermore, the simulation revealed that the specified parameters α and p of the intensity-based and quantile-based method are not exactly adhered to.

An important aspect for the comparison of the construction methods is the way in which the probability of non-explosion q is taken into account: In the traditional method, q does not influence the shape of the high-risk zone at all. As the process $\tilde{Y}$ is obtained by independent thinning of $\tilde{X}$, it contains a smaller number of points for $q = 0.15$ than for $q = 0.10$. So the high-risk zones constructed with the traditional method are even smaller for a higher probability of non-explosion in our simulation setting. The quantile-based method does not explicitly take into account the probability of non-explosion q, either. However, the nearest-neighbour distances were typically larger for high values of q and so was the radius defined by the p-quantile. In this setting, the mean area of the high-risk zone increased in most cases if $q = 0.15$ was considered, but so did the mean of p_{miss} and the fraction p_{out}. In other words, the probability of non-explosion is not taken into account sufficiently: Indeed, the high-risk zones become larger with increasing q, but the failure probability rises nonetheless. The only approach which uses q as a parameter is the intensity-based method. The results show that the high-risk zones have a larger area for $q = 0.15$ than for $q = 0.10$. In some cases, p_{out} and the mean of p_{miss} decreased.



(a) Example A

(b) Example B

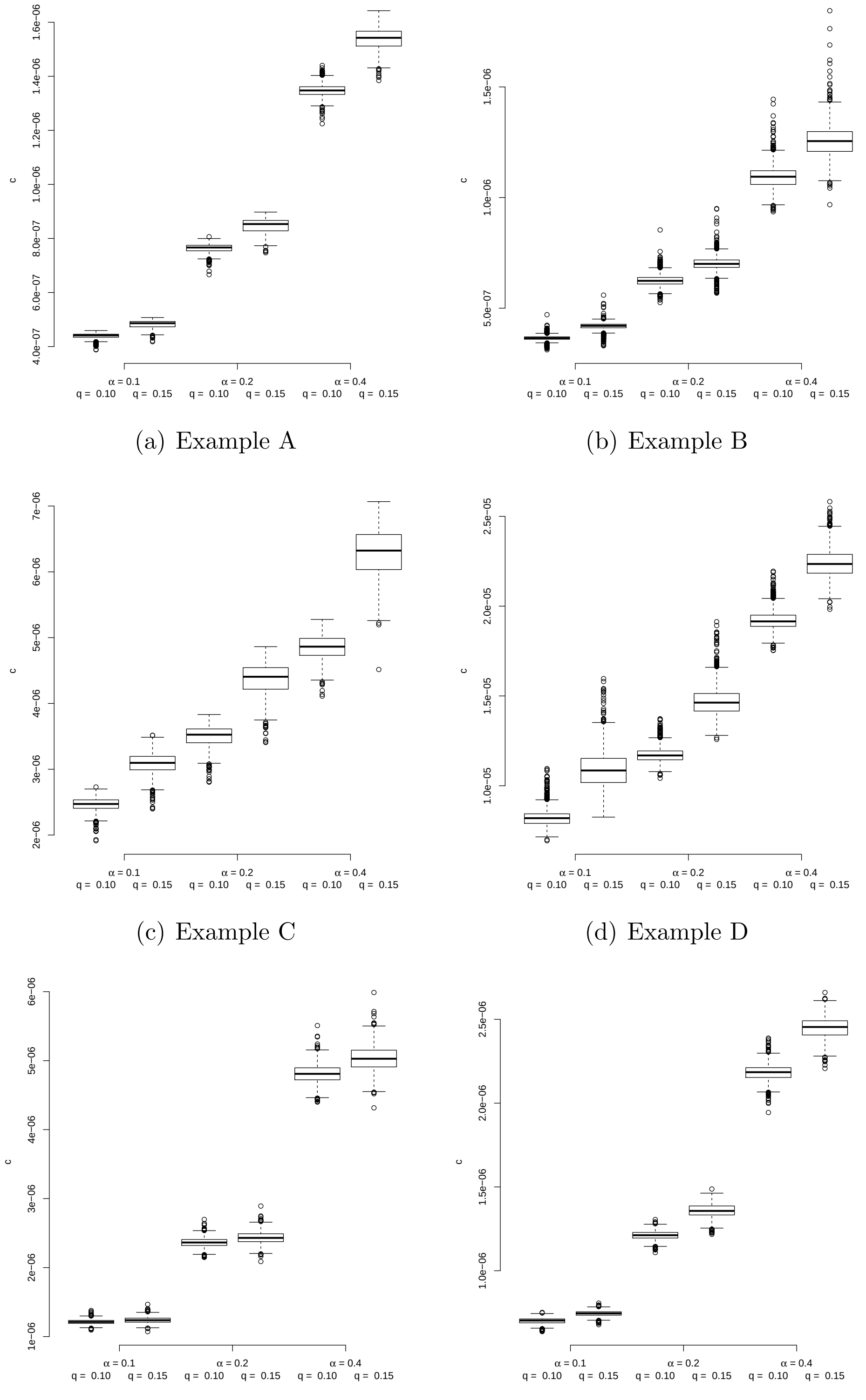
(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 5.9.: Relation between given quantile and resulting radius for the quantile-based method.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 5.10.: Relation between given α and resulting threshold c for the intensity-based method.

5.3. High-risk zones with fixed area

For both the intensity-based and the quantile-based construction method, it is possible to specify the desired area of the resulting high-risk zone. This approach allows a better comparison of the performance: If the area is fixed, it is easier to compare the intensity-based and the quantile-based method, because it is sufficient to consider p_{out} and p_{miss} to decide which method yields the better high-risk zones. The traditional method does not need to be considered in this context, since it can be interpreted as a simplified version of the quantile-based method.

5.3.1. Setting

In case of the quantile-based method, fixing the area of the high-risk zone means that the optimal radius must be determined: The union of all respective discs, centered at the observations, must result in the desired area. If this radius is not smaller than the minimum of the nearest-neighbour distances and not larger than their maximum, we can determine which quantile of the nearest-neighbour distances corresponds to this radius.

For the intensity-based method, the cut-off value c needs to be determined: The zone which results from all locations with an intensity of at least c must have the desired area. We can then determine to which value of

$$\alpha = \hat{P}\{N_Z(W \backslash R_c) > 0\} = 1 - \exp\left[- \left\{ \frac{q}{1-q} \hat{\Lambda}_Y(W \backslash R_c) \right\} \right] \tag{5.1}$$

this optimal c corresponds.

High-risk zones were determined for $q = 0.10$ and $q = 0.15$. The desired area of the high-risk zone was chosen individually for every example. For each example, three values were considered. In some cases where the desired area of the high-risk zone was large, no quantile matching the necessary radius could be determined.

5.3.2. Results for the bomb crater data

Comparison of the methods

Example A The results for Example A are depicted in Table 5.7. The quantile-based method gave better high-risk zones, i.e. its values of p_{miss} and p_{out} were smaller than for the intensity-based method. The comparison of the two methods is also illustrated in Figure 5.11(a), which shows boxplots of p_{miss} for the simulations. The figure underlines the better performance of the quantile-based method for Example A. The mean of the retrospectively determined values for α exceeds the fraction p_{out}, whereas the mean of the retrospectively determined p corresponds to the mean of p_{miss}.

Example B Table 5.8 contains the results for Example B. In five of six cases, the high-risk zones determined by using the intensity-based method gave better results. This relation is

Table 5.7.: Results of the simulation for given area: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside, mean area of the zone and mean of the retrospectively determined values for the parameters α and p, Example A, intensity-based method (INT) and quantile-based method (QUANT)

A	q	0.1	0.15	0.1	0.15	0.1	0.15
	area	1500000	1500000	2000000	2000000	2500000	2500000
INT	mean p_{miss}	0.076	0.078	0.028	0.032	0.014	0.016
	p_{out}	0.976	0.993	0.712	0.875	0.466	0.668
	mean area in m^2	1500001	1500001	2000001	2000000	2500001	2500001
	mean α	0.988	0.999	0.885	0.963	0.573	0.732
QUANT	mean p_{miss}	0.036	0.038	0.016	0.019	0.012	0.013
	p_{out}	0.809	0.938	0.528	0.721	0.446	0.608
	mean area in m^2	1500002	1500002	2000001	2000000	2500002	2500002
	mean p	0.961	0.960	0.982	0.980	0.986	0.986

also visible in Figure 5.11(b). For the quantile-based method, the mean fraction p_{miss} was generally slightly smaller than the mean of $1-p$, whereas the values of p_{out} were larger than the mean of α for the intensity-based method. The fraction p_{out} exceeded the mean of the retrospectively determined α, whereas $1 - p_{\text{miss}}$ is close to the retrospectively determined p. For the largest area of 2400000 m^2 and $q = 0.15$, no quantile matching the radius could be determined in five iterations.

Table 5.8.: Results of the simulation for given area: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside, mean area of the zone and mean of the retrospectively determined values for the parameters α and p, Example B, intensity-based method (INT) and quantile-based method (QUANT)

B	q	0.1	0.15	0.1	0.15	0.1	0.15
	area	2000000	2000000	2200000	2200000	2400000	2400000
INT	mean p_{miss}	0.079	0.076	0.057	0.056	0.050	0.046
	p_{out}	0.594	0.703	0.473	0.605	0.446	0.566
	mean area in m^2	1999998	1999998	2200000	2199999	2400000	2399999
	mean α	0.466	0.610	0.337	0.461	0.221	0.315
QUANT	mean p_{miss}	0.151	0.148	0.117	0.108	0.050	0.053
	p_{out}	0.830	0.915	0.775	0.884	0.444	0.605
	mean area in m^2	2000000	2000000	2199999	2199999	2399998	2399998
	mean p	0.843	0.844	0.877	0.882	0.939	0.938

Example C As Table 5.9 shows, the intensity-based method yielded the better high-risk zones with an area of 2000000 m^2, the quantile-based method for larger area (see also

Table 5.9.: Results of the simulation for given area: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside, mean area of the zone and mean of the retrospectively determined values for the parameters α and p, Example C, intensity-based method (INT) and quantile-based method (QUANT)

C	q area	0.1 2000000	0.15 2000000	0.1 2500000	0.15 2500000	0.1 3000000	0.15 3000000
INT	mean p_{miss}	0.089	0.091	0.034	0.034	0.008	0.008
	p_{out}	1	1	0.989	0.998	0.664	0.831
	mean area in m^2	1999999	1999999	2500000	2500000	3000000	3000000
	mean α	1	1	0.999	1	0.771	0.894
QUANT	mean p_{miss}	0.096	0.096	0.023	0.024	0.003	0.004
	p_{out}	1	1	0.975	0.995	0.401	0.542
	mean area in m^2	2000000	2000000	2499999	2499999	3000000	3000000
	mean p	0.904	0.904	0.977	0.976	0.996	0.996

Figure 5.11(c)). The mean of the retrospectively determined values for α is close to the fraction p_{out} in most cases. For an area of 3000000 m^2, no quantile could be determined in seven cases, six of them for $q = 0.15$.

Example D The intensity-based method yielded the better high-risk zones for Example D (see Table 5.10 and Figure 5.11(d)). The mean of the retrospectively determined values for α exceeds the fraction p_{out}. In many iterations, no quantile matching the necessary radius could be determined. For the smallest area of 460000 m^2, this happened three times (one time for $q = 0.10$/two for $q = 0.15$), for an area of 480000 m^2 23 times (7/16), and for the largest area of 500000 m^2 228 times (92/136).

Example E For an area of 280000 m^2 and 300000 m^2, the intensity-based method yielded the better high-risk zones, for an area of 320000 m^2, p_{out} and p_{miss} were smaller for the quantile-based method (see Table 5.11 and Figure 5.11(e)). Again, the mean of the retrospectively determined values for α exceeds the fraction p_{out}.

Example F As Table 5.12 shows, the intensity-based method yielded the better high-risk zones with an area of 1500000 m^2 and 2000000 m^2, whereas the quantile-based method was better for an area of 2500000 m^2 (see also Figure 5.11(f)).

Table 5.10.: Results of the simulation for given area: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside, mean area of the zone and mean of the retrospectively determined values for the parameters α and p, Example D, intensity-based method (INT) and quantile-based method (QUANT)

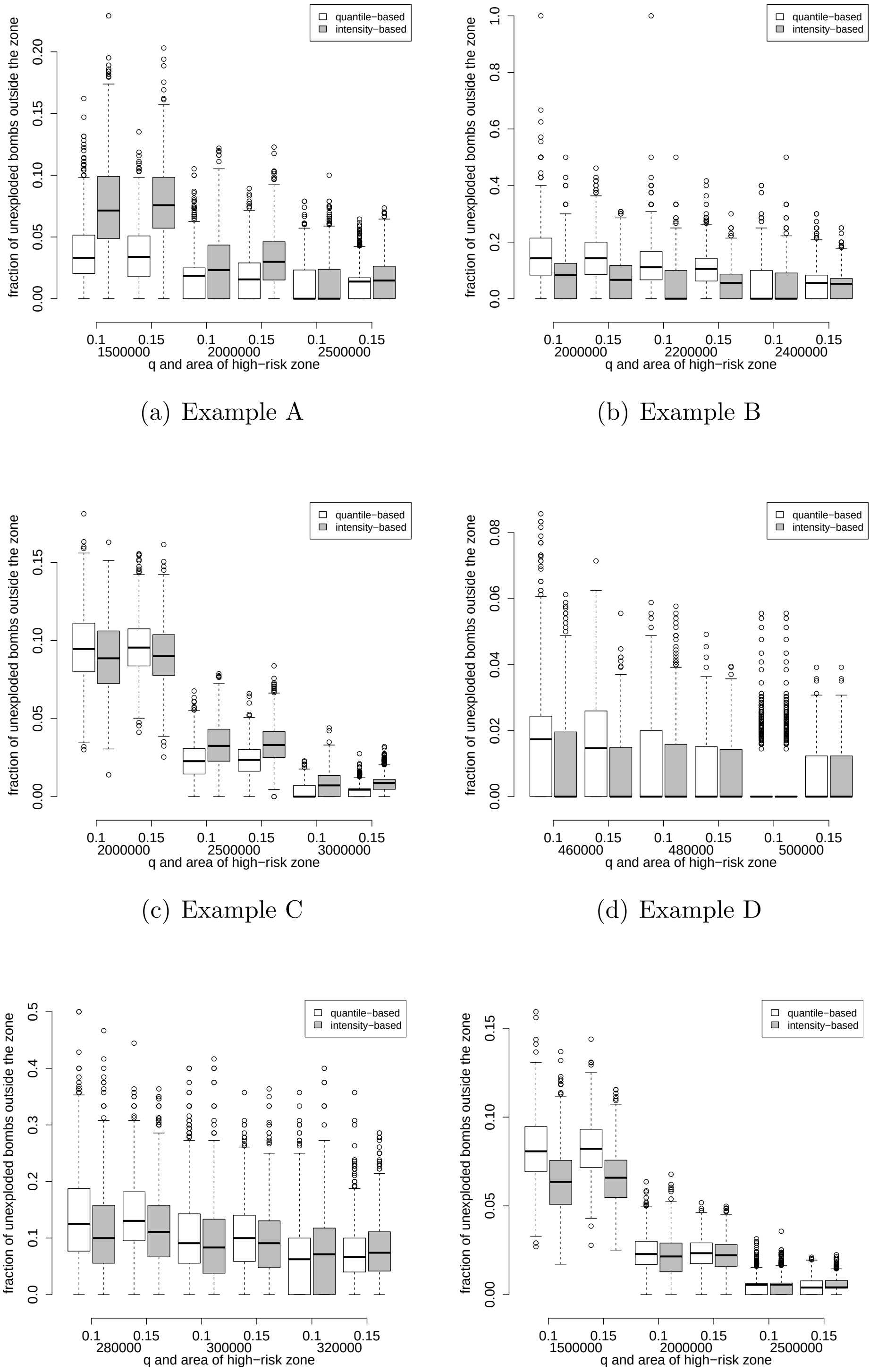
D	q area	0.1 460000	0.15 460000	0.1 480000	0.15 480000	0.1 500000	0.15 500000
INT	mean p_{miss}	0.008	0.008	0.006	0.006	0.004	0.004
	p_{out}	0.303	0.436	0.253	0.368	0.173	0.259
	mean area in m^2	460000	460000	480000	480000	500000	500000
	mean α	0.509	0.665	0.283	0.403	0.103	0.156
QUANT	mean p_{miss}	0.015	0.015	0.008	0.008	0.004	0.004
	p_{out}	0.513	0.682	0.320	0.472	0.174	0.260
	mean area in m^2	460000	460000	480000	480000	500000	500000
	mean p	0.986	0.986	0.992	0.992	0.996	0.996

Table 5.11.: Results of the simulation for given area: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside, mean area of the zone and mean of the retrospectively determined values for the parameters α and p, Example E, intensity-based method (INT) and quantile-based method (QUANT)

E	q area	0.1 280000	0.15 280000	0.1 300000	0.15 300000	0.1 320000	0.15 320000
INT	mean p_{miss}	0.108	0.117	0.090	0.095	0.076	0.078
	p_{out}	0.801	0.932	0.751	0.898	0.693	0.848
	mean area in m^2	280001	280000	300001	300000	319999	319999
	mean α	0.972	0.996	0.958	0.992	0.938	0.985
QUANT	mean p_{miss}	0.135	0.139	0.102	0.104	0.065	0.072
	p_{out}	0.880	0.969	0.819	0.944	0.648	0.818
	mean area in m^2	279999	280000	300002	300002	320000	320000
	mean p	0.858	0.856	0.891	0.890	0.924	0.923

Table 5.12.: Results of the simulation for given area: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside, mean area of the zone and mean of the retrospectively determined values for the parameters α and p, Example F, intensity-based method (INT) and quantile-based method (QUANT)

F	q area	0.1 1500000	0.15 1500000	0.1 2000000	0.15 2000000	0.1 2500000	0.15 2500000
INT	mean p_{miss}	0.064	0.066	0.022	0.022	0.005	0.006
	p_{out}	1	1	0.976	0.996	0.627	0.780
	mean area in m^2	1500000	1500001	1999998	1999998	2500001	2500001
	mean α	1	1	0.998	1	0.788	0.905
QUANT	mean p_{miss}	0.082	0.083	0.024	0.023	0.005	0.005
	p_{out}	1	1	0.991	0.999	0.564	0.741
	mean area in m^2	1499999	1499999	2000000	2000000	2500003	2500003
	mean p	0.919	0.917	0.975	0.976	0.995	0.995



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 5.11.: Fraction p_{miss} of simulated unexploded bombs outside the high-risk zone for given area. The boxplots compare the quantile-based and intensity-based method.

Retrospectively determined values of α and p

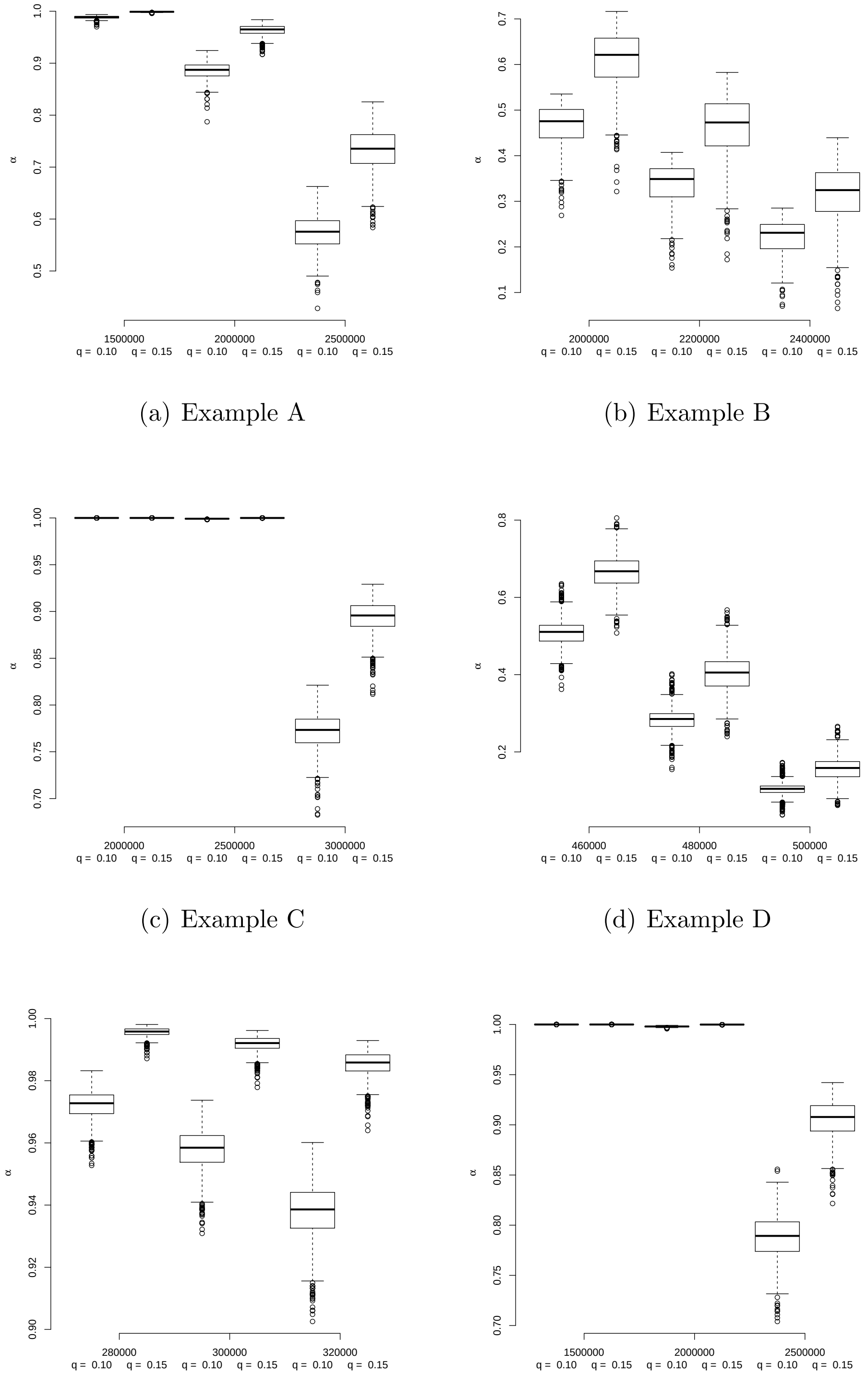
The retrospectively determined values of α for the considered values of q and given area are depicted in Figure 5.12. For Examples C, F and especially E, the values are high, often very close to 1. If the area is augmented, α decreases. It increases for $q = 0.15$.

The retrospectively determined quantiles for the considered values of q and given area are depicted in Figure 5.13. The quantile increases for a larger area. If q is increased, the quantile decreases slightly; in many cases, the variance of the quantile increases.

Shape of the high-risk zones

Figure 5.14 shows the frequency that every pixel in the observation window was part of the quantile-based high-risk zone in the 6000 total iterations. The same is depicted for the intensity-based method in Figure 5.15. We can see the general shape of the quantile-based and intensity-based high-risk zones and we get an impression of how much the high-risk zones change for the considered combinations of area and probability of non-explosion.

The high-risk zones obtained with the quantile-based method are often more ragged, single discs are visible in many cases. Intensity-based high-risk zones, in contrast, are often coherent.



(a) Example A

(b) Example B

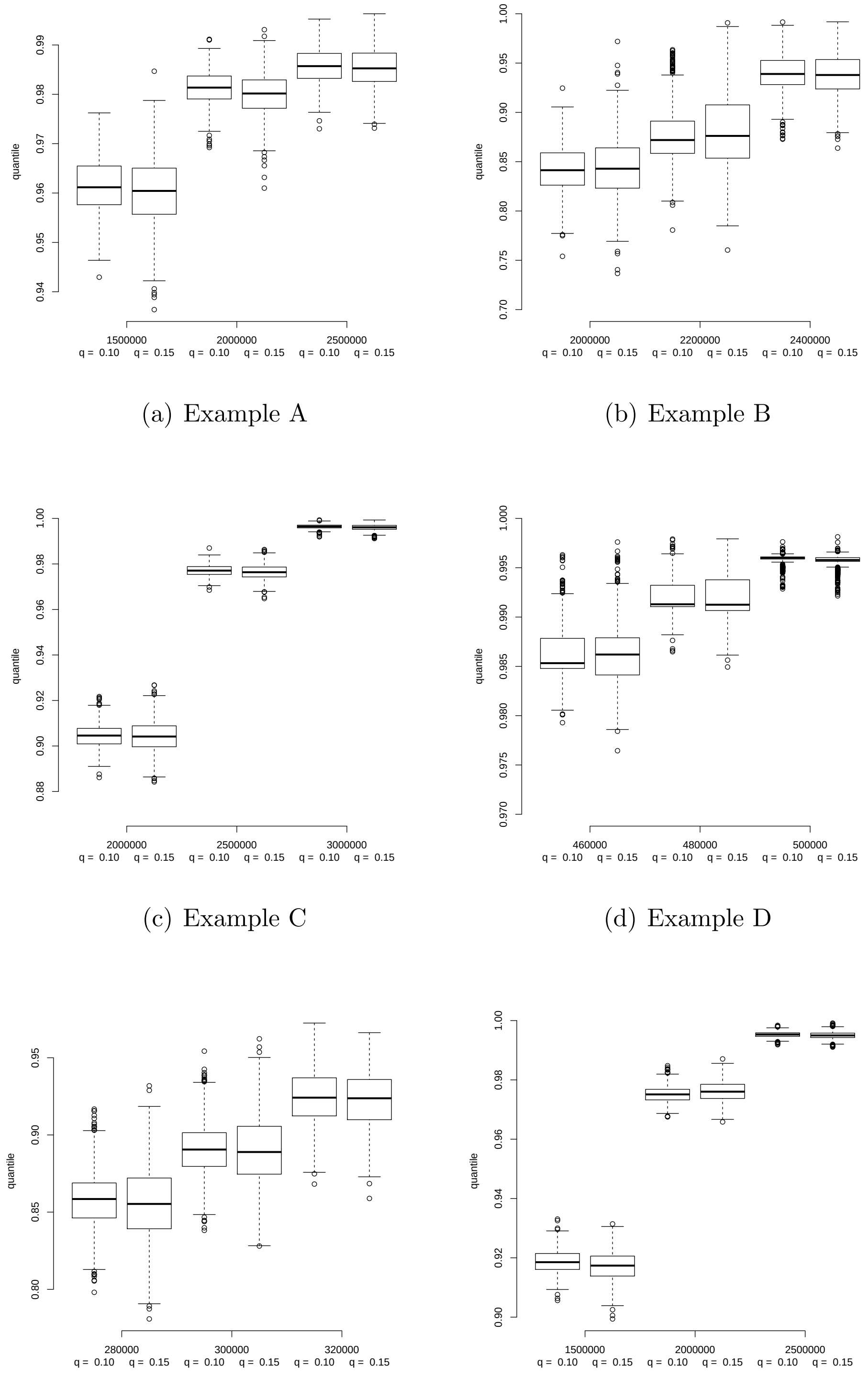
(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 5.12.: Retrospectively determined values of α for the considered values of q and area.



(a) Example A

(b) Example B

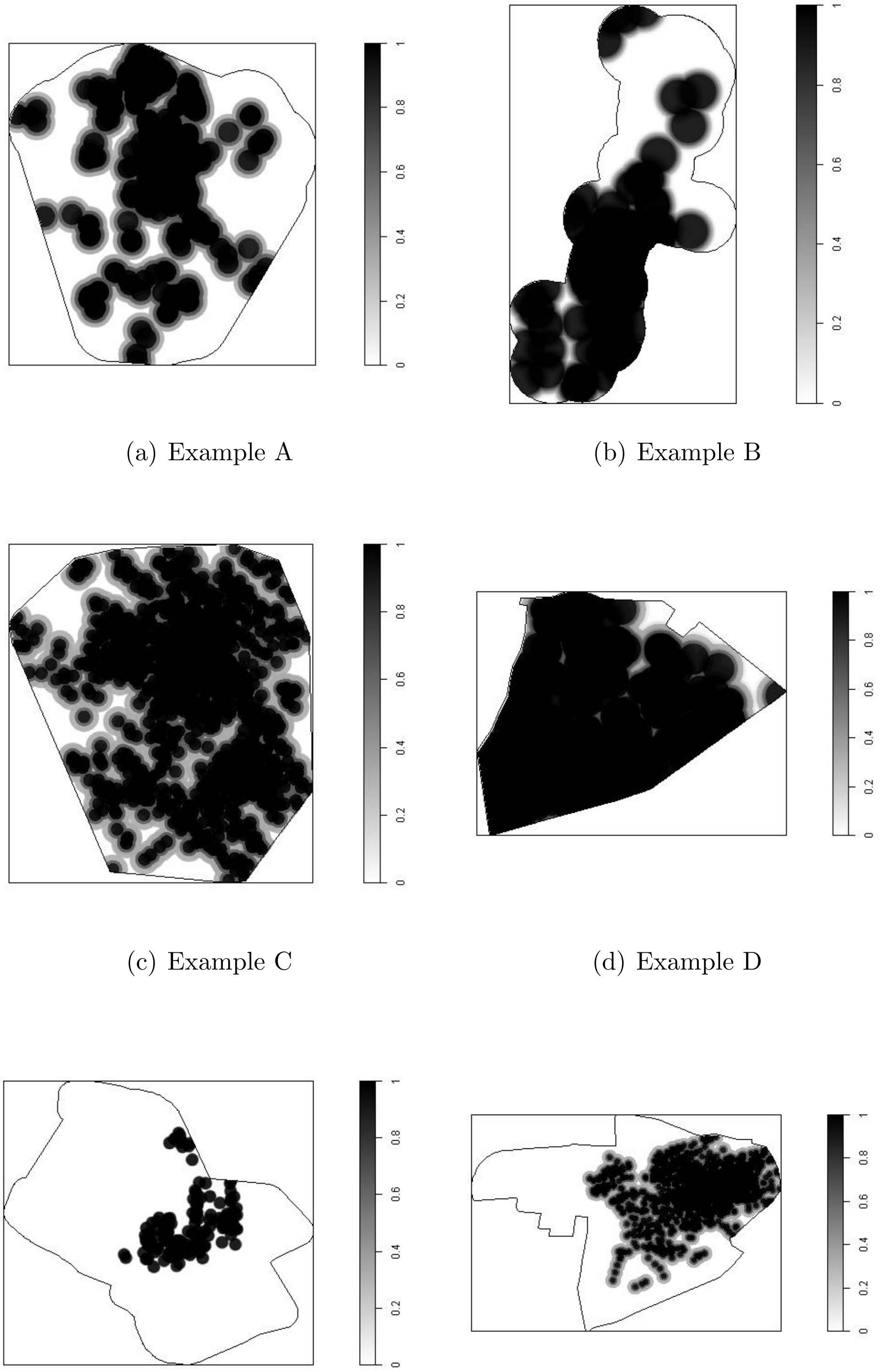
(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 5.13.: Retrospectively determined quantile for the considered values of q and area.



(a) Example A

(b) Example B

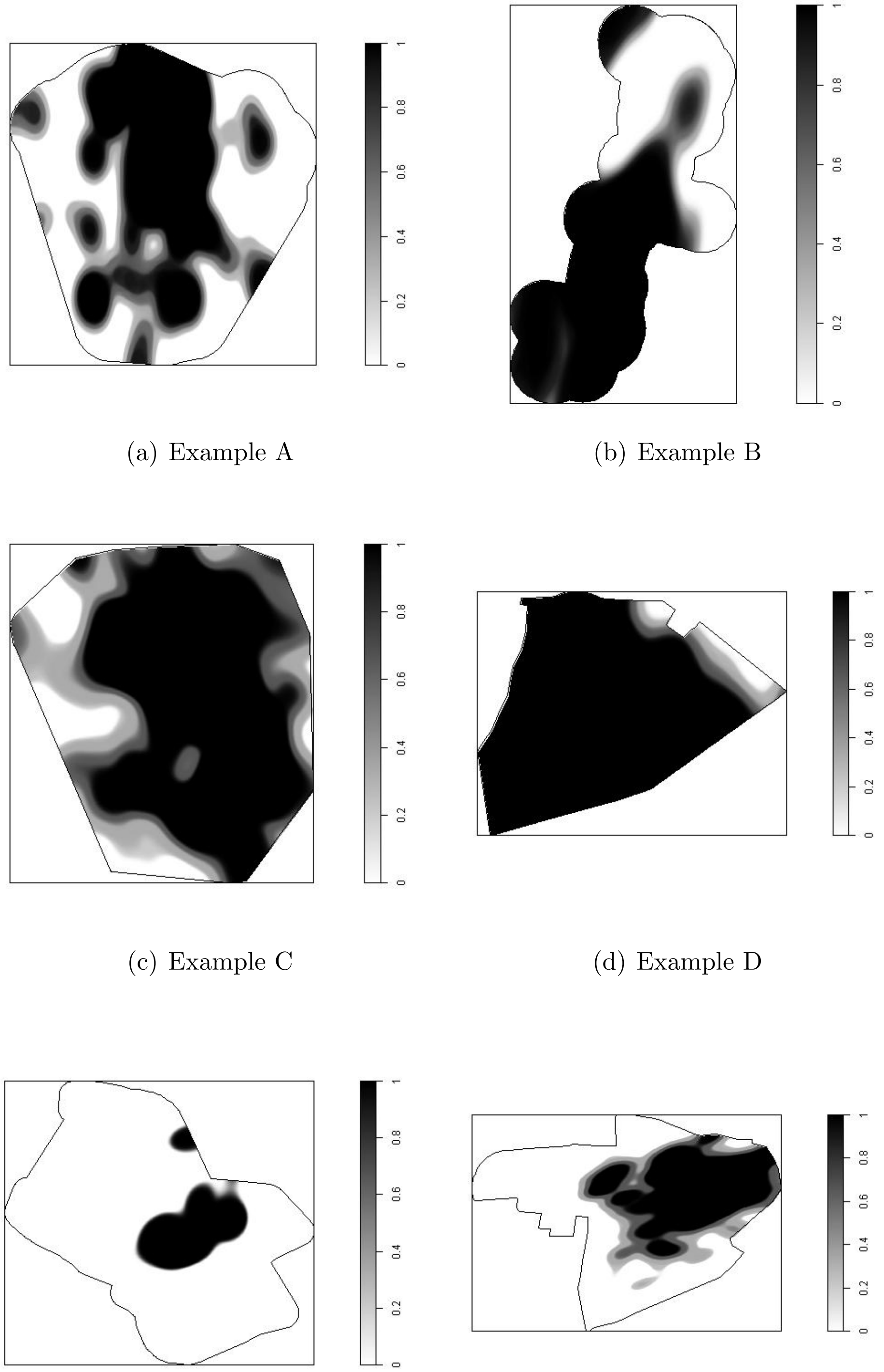
(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 5.14.: Frequency that every pixel in the observation window was part of the quantile-based high-risk zone in the 6000 total iterations.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 5.15.: Frequency that every pixel in the observation window was part of the intensity-based high-risk zone in the 6000 total iterations.

5.4. Properties of the methods for constructing high-risk zones

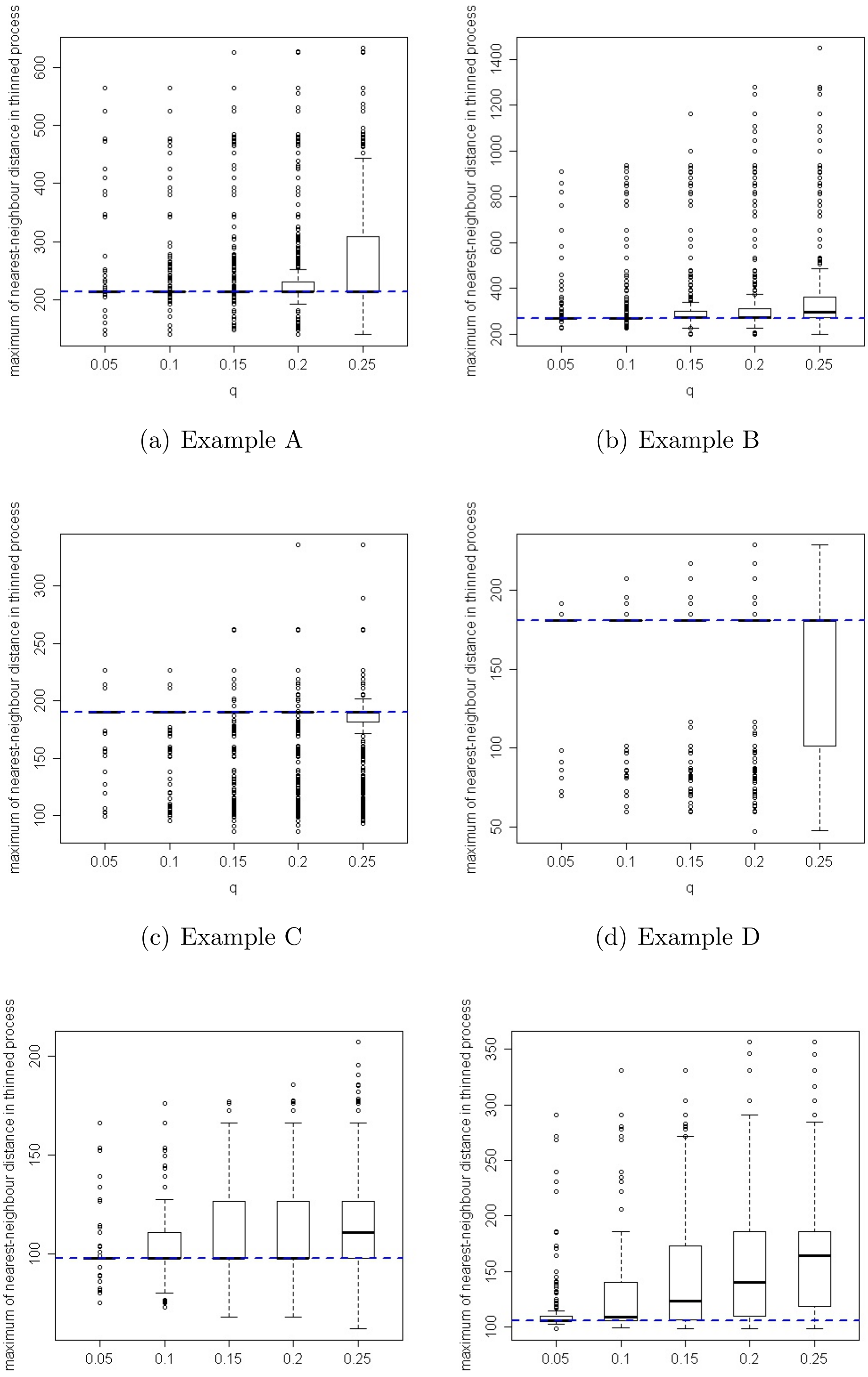
The quantile-based and the intensity-based construction methods yielded comparable results in the simulation. To decide which method should actually be applied, however, we should keep in mind that the more heuristic quantile-based construction method has a vague theoretical justification. As the p-quantile $Q(p)$ of the distribution of the nearest-neighbour distance is used, the fraction of unexploded bombs outside the high-risk zone is expected to be $1 - p$, i.e. risk is fixed for each unexploded bomb, but not globally.

The high-risk zones consist of the union of discs around the observations, so their shape is less flexible than the shape of intensity-based high-risk zones. In particular, possible anisotropy cannot be taken into account.

Moreover, the probability of non-explosion q is not a parameter of the quantile-based method and is therefore not accounted for sufficiently.

Another disadvantage is that it is not possible to construct a quantile-based high-risk zone with an arbitrarily small failure probability: The smallest failure probability is obtained when the maximum of the nearest-neighbour distances is used as radius of the discs. However, the probability that the maximum of the nearest-neighbour distances of the full process exceeds the maximum of those of the thinned process (which is the failure probability) depends on the characteristics of the point pattern and cannot be influenced in any way. For Examples A to F, the minimal failure probability of the quantile-based construction method was investigated in a simulation. Each observed pattern was thinned and the maximum nearest-neighbour distance of the thinned pattern was determined. This was repeated 10000 times for each example and five different values of q. Figure 5.16 compares the maximum nearest-neighbour distances of the thinned patterns and the maximum nearest-neighbour distance of the full patterns. For small values of q, the two values were identical for at least 50 % of the iterations for most examples. Deviations in both directions were observed. The effects were different for each example. The minimal failure probability is the probability that the maximum nearest-neighbour distance of the thinned pattern (i.e. the observed pattern of bomb craters) is smaller than the nearest-neighbour distance of the full pattern (the pattern consisting of bomb craters and unexploded bombs). The frequency that this happened in the simulation serves as an estimate for the minimal failure probability. The results are shown in Figure 5.17. For Examples A, C, D and E, the estimated minimal failure probability increases with q. For $q = 0.10$ and $q = 0.15$, the range is between 0.04 and 0.16. The values are small for Examples B and F and large for Examples A, C and D.

For these reasons, the intensity-based construction method is recommended. It is investigated in more detail in the following chapters.



(a) Example A

(b) Example B

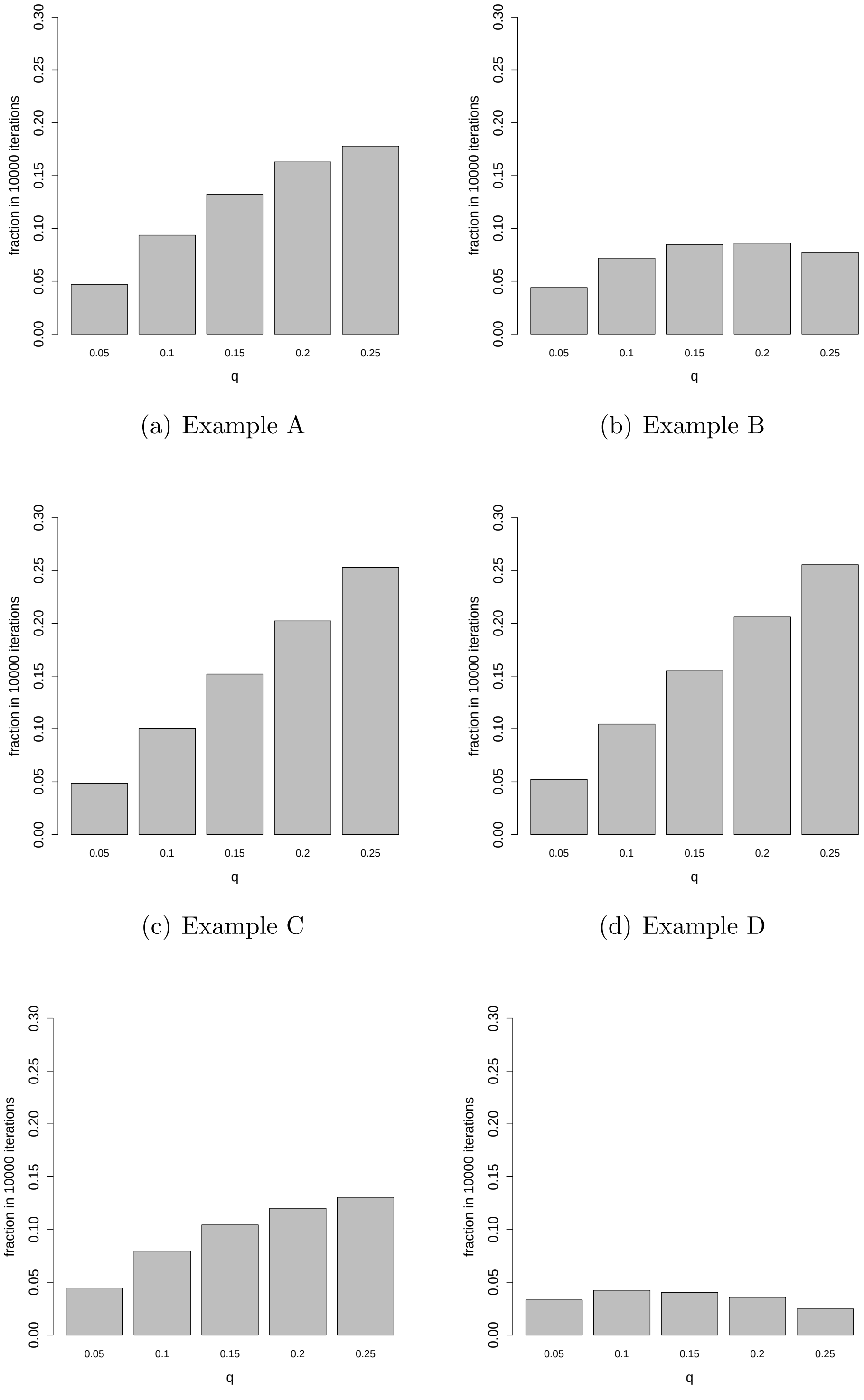
(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 5.16.: Comparison of the maximum of the nearest-neighbour distances in the full patterns (dashed blue lines) and in (1 - q)-thinned patterns (boxplots); 10000 iterations.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

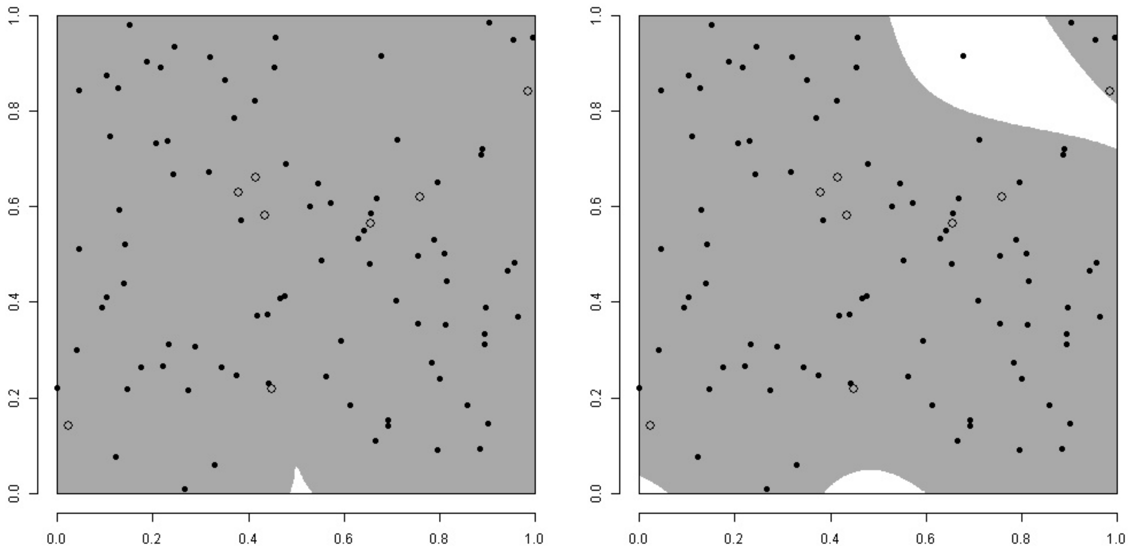
(e) Example E

(f) Example F

Figure 5.17.: Fraction of cases in which the maximum of the nearest-neighbour distances in the full pattern was smaller than in the $(1 - q)$-thinned pattern (10000 iterations).

5.5. Application to simulated patterns

To investigate the applicability of the construction methods for high-risk zones, the intensity-based method and the quantile-based method were tested in three cases which differ from an inhomogeneous Poisson point process: The patterns in Figure 2.2(a) (regular pattern), Figure 2.2(b) (homogeneous Poisson process) and Figure 2.2(c) (Thomas process) were considered. The 99 %, the 99.5 % and the 99.9 % quantile were used and α was set to 0.4, 0.2 or 0.1. The probability of non-explosion q was assumed to be 0.10 or 0.15.



(a) quantile-based method

(b) intensity-based method

Figure 5.18.: High-risk zones, observed events on which the high-risk zones are based (filled diamonds), unobserved events inside the zone (circles) and outside the zone (crosses; not present in this case) for the quantile-based method (99 % quantile) and the intensity-based method ($\alpha = 0.4$), homogeneous Poisson process.

Examples of resulting high-risk zones in a single simulation iteration are depicted in Figures 5.18, 5.19 and 5.20. In addition to the high-risk zones, the events on which the high-risk zones are based are depicted, as well as the points which were considered as unobserved events and used for evaluation in this iteration. For the homogeneous Poisson process and the regular process, the quantile-based high-risk zones cover (almost) the entire observation window.

The results from 1000 iterations are summarised in Tables 5.13, 5.14 and 5.15. For the homogeneous Poisson process, the high-risk zones are large. They were even larger for the regular process, especially if the quantile-based method was applied. The fraction p_{out} and the mean of p_{miss} were extremely large for intensity-based high-risk zones for the regular process. For the Thomas process, both fractions were small and the area of the high-risk zones was considerably smaller than for the homogeneous Poisson process and the regular process.

Table 5.13.: Results of the simulation: Mean fraction p_{miss} of unobserved events outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unobserved event was located outside and mean area of the zone, homogeneous Poisson process, intensity-based method (INT) and quantile-based method (QUANT)

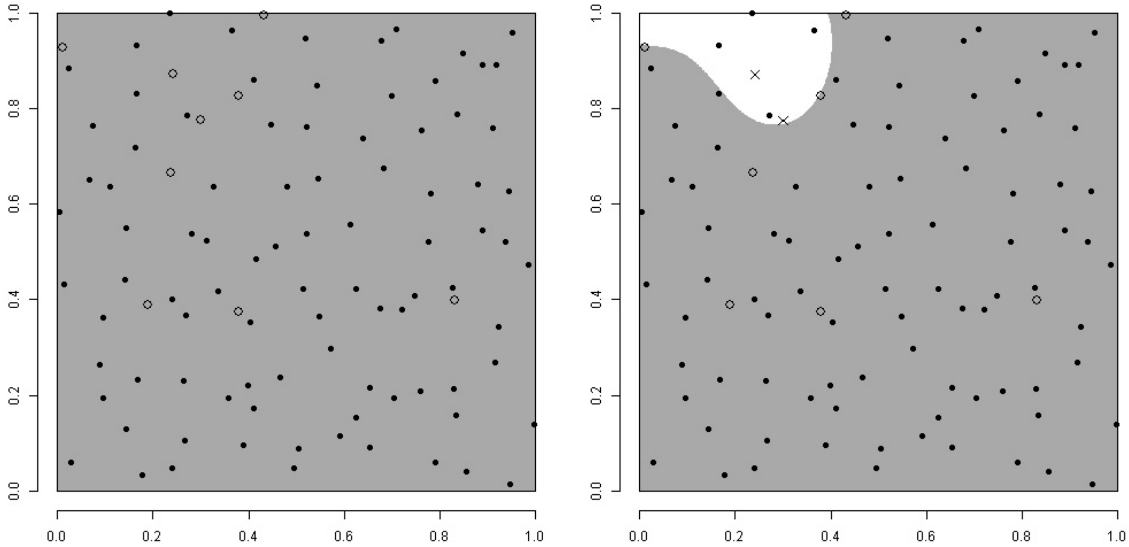
homogeneous	q	0.1	0.1	0.1	0.15	0.15	0.15
INT	α	0.4	0.2	0.1	0.4	0.2	0.1
	mean p_{miss}	0.060	0.018	0.012	0.034	0.016	0.012
	p_{out}	0.412	0.159	0.107	0.384	0.204	0.169
	mean area	0.9080	0.9553	0.9774	0.9346	0.9687	0.9844
QUANT	p	0.99	0.995	0.999	0.99	0.995	0.999
	mean p_{miss}	0.011	0.011	0.011	0.011	0.011	0.011
	p_{out}	0.091	0.091	0.091	0.135	0.135	0.135
	mean area	0.9932	0.9943	0.9943	0.9915	0.9924	0.9924

Table 5.14.: Results of the simulation: Mean fraction p_{miss} of unobserved events outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unobserved event was located outside and mean area of the zone, regular process, intensity-based method (INT) and quantile-based method (QUANT)

regular	q	0.1	0.1	0.1	0.15	0.15	0.15
INT	α	0.4	0.2	0.1	0.4	0.2	0.1
	mean p_{miss}	0.236	0.125	0.069	0.142	0.071	0.041
	p_{out}	0.966	0.792	0.572	0.960	0.771	0.533
	mean area	0.9404	0.9733	0.9872	0.9583	0.9813	0.9910
QUANT	p	0.99	0.995	0.999	0.99	0.995	0.999
	mean p_{miss}	0.016	0.011	0.011	0.016	0.013	0.013
	p_{out}	0.150	0.102	0.102	0.207	0.160	0.160
	mean area	0.9991	0.9992	0.9992	0.9985	0.9986	0.9986

Table 5.15.: Results of the simulation: Mean fraction p_{miss} of unobserved events outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unobserved event was located outside and mean area of the zone, Thomas process, intensity-based method (INT) and quantile-based method (QUANT)

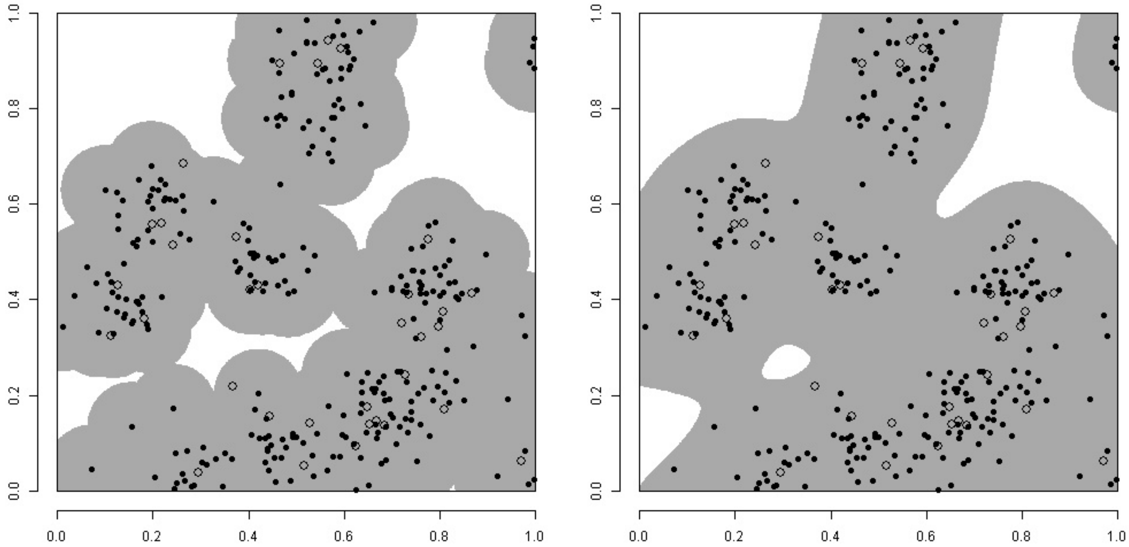
Thomas	q	0.1	0.1	0.1	0.15	0.15	0.15
INT	α	0.4	0.2	0.1	0.4	0.2	0.1
	mean p_{miss}	0.007	0.003	0.001	0.004	0.002	0
	p_{out}	0.193	0.096	0.021	0.154	0.104	0.014
	mean area	0.7784	0.8296	0.8645	0.8095	0.8534	0.8844
QUANT	p	0.99	0.995	0.999	0.99	0.995	0.999
	mean p_{miss}	0.010	0.004	0.003	0.009	0.004	0.003
	p_{out}	0.251	0.115	0.091	0.315	0.168	0.123
	mean area	0.7693	0.8125	0.8570	0.7675	0.8189	0.8555



(a) quantile-based method

(b) intensity-based method

Figure 5.19.: High-risk zones, observed events on which the high-risk zones are based (filled diamonds), unobserved events inside the zone (circles) and outside the zone (crosses) for the quantile-based method (99 % quantile) and the intensity-based method ($\alpha = 0.4$), regular process.

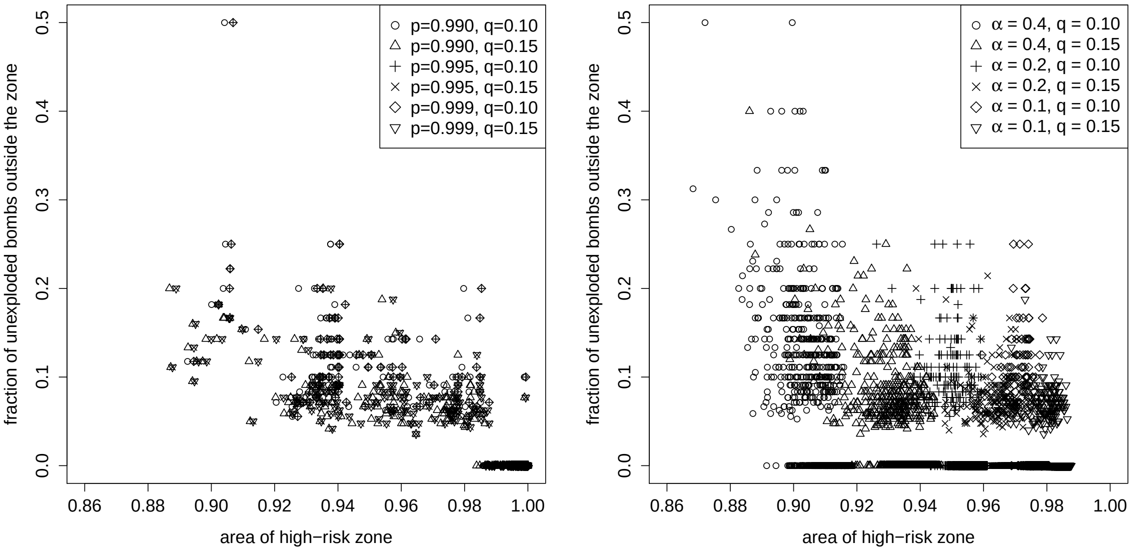


(a) quantile-based method

(b) intensity-based method

Figure 5.20.: High-risk zones, observed events on which the high-risk zones are based (filled diamonds), unobserved events inside the zone (circles) and outside the zone (crosses; not present in this case) for the quantile-based method (99 % quantile) and the intensity-based method ($\alpha = 0.4$), Thomas process.

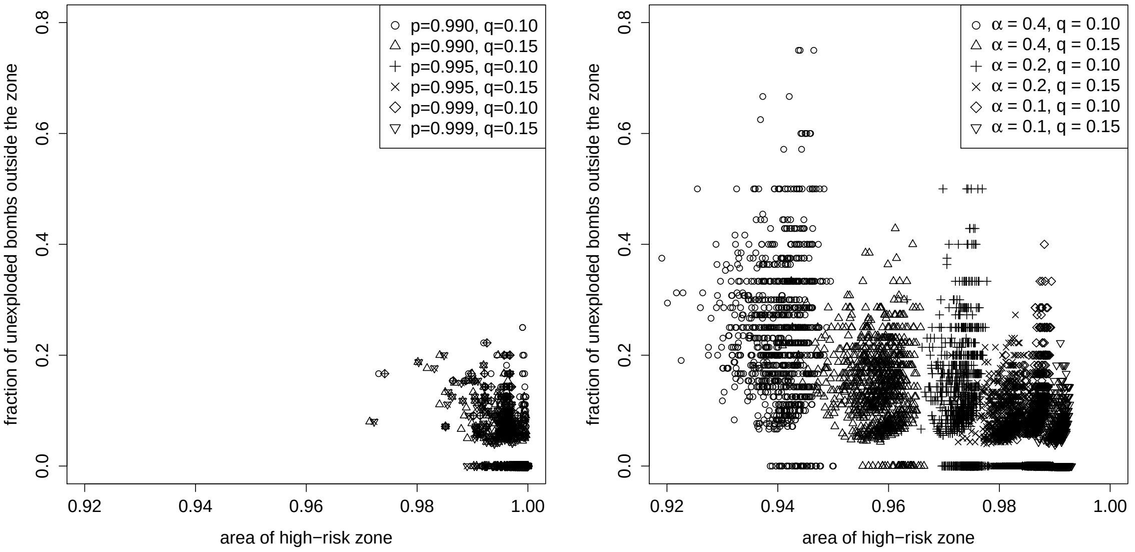
The scatterplots in Figures 5.21, 5.22 and 5.23 underline that both the quantile-based and the intensity-based method yield high-risk zones where the fractions p_{out} and p_{miss} are



(a) quantile-based method

(b) intensity-based method

Figure 5.21.: Area of the high-risk zone and fraction of simulated unobserved events outside the zone for the quantile-based and the intensity-based method, homogeneous Poisson process.

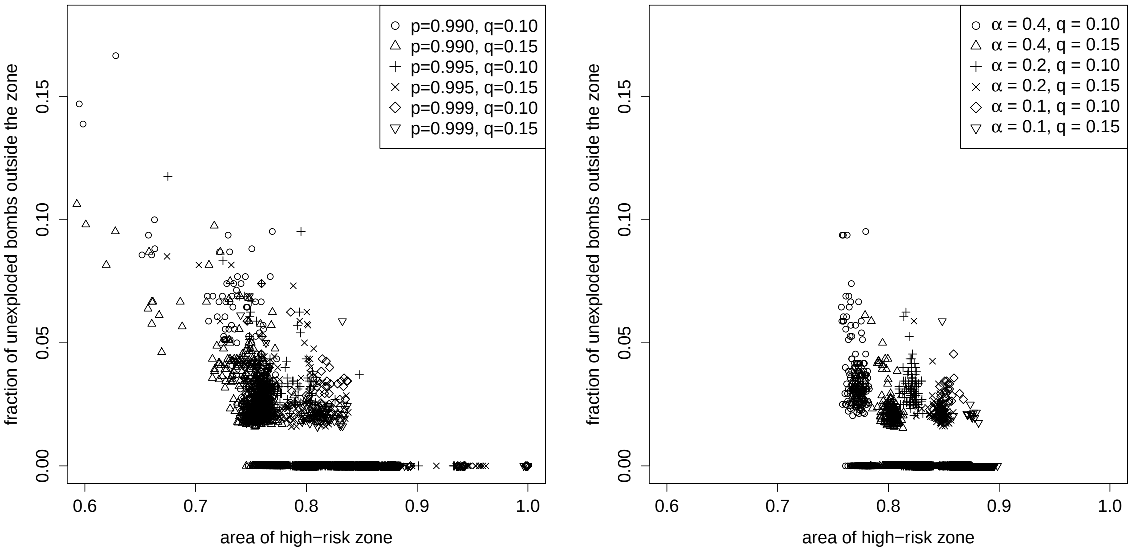


(a) quantile-based method

(b) intensity-based method

Figure 5.22.: Area of the high-risk zone and fraction of simulated unobserved events outside the zone for the quantile-based and the intensity-based method, regular process.

small for the Thomas process. High-risk zones for the homogeneous Poisson process are generally large. In practice, these large zones would suggest to search the whole property, which is exactly the correct proceeding if the intensity of unexploded bombs is more or



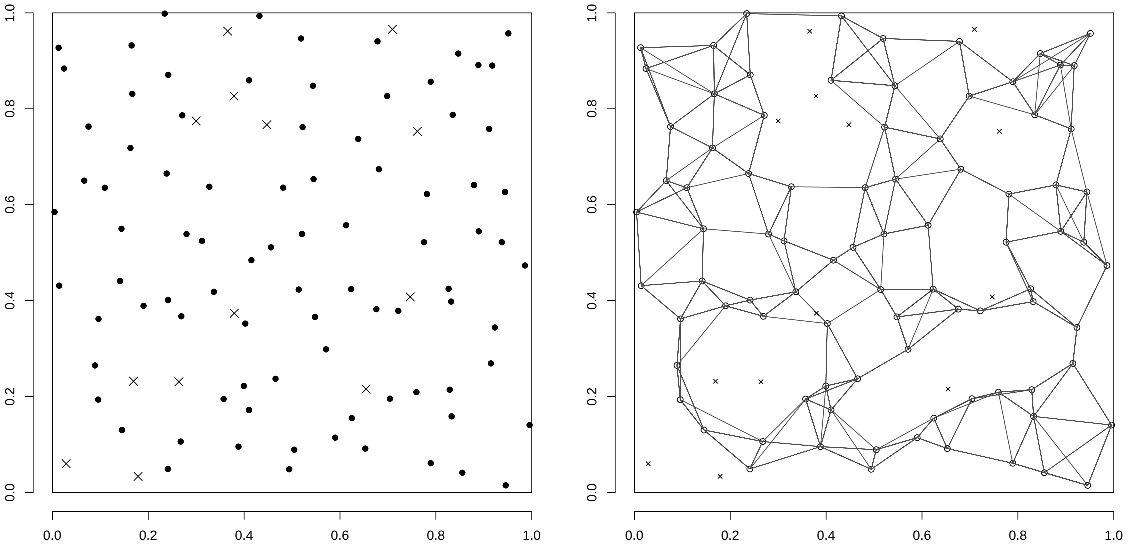
(a) quantile-based method

(b) intensity-based method

Figure 5.23.: Area of the high-risk zone and fraction of simulated unobserved events outside the zone for the quantile-based and the intensity-based method, Thomas process.

less constant. For the regular pattern, the fractions p_{out} and p_{miss} are large, even though the high-risk zones comprise a large fraction of the window.

In summary, the results for these three examples indicate that reasonable high-risk zones can be obtained for clustered patterns via both methods. Both methods can in principle also be applied for homogeneous patterns, where the high-risk zones will comprise most of the window and should be extended to the entire window. Furthermore, the results suggest that none of the methods can be applied successfully to regular patterns. Other approaches need to be developed for regular patterns. It may be useful to consider k-neighbour graphs (Illian et al., 2008, Section 1.8.5), which can be determined using the R package `spatgraphs` (Rajala, 2012). Large regions without edges suggest gaps (Illian et al., 2008, page 257) and can therefore be searched to find at least some of the unobserved events. An example of a thinned regular pattern and the corresponding 4-neighbour graph is depicted in Figure 5.24.



(a) thinned regular pattern (crosses indicate unobserved events)

(b) 4-neighbour graph

Figure 5.24.: Thinned regular pattern and corresponding 4-neighbour graph.

6. Risk assessment

The results for the intensity-based method in Section 5.2 showed that the specified parameter α is usually not exactly adhered to. A possible reason is that the estimation of the intensity function introduces uncertainty which is not accounted for in the construction method for high-risk zones. The problem is aggravated by thinning in the simulation scenario. All results shown in the previous section referred to high-risk zones based on the thinned version $\tilde{Y}$ of the observed process Y. A high-risk zone that is to be applied in reality, however, will be based on all observations available, i.e. the entire process Y. Therefore, it is not sufficient to choose a reasonable α expressing the intended failure probability, but it is indispensable to investigate the behaviour of the high-risk zone in a setting of realistically high intensity.

To do so, a simulation procedure is proposed where the full process X of bomb craters and unexploded bombs is assumed to be an inhomogeneous Poisson point process. Results are shown for Examples A to F. Finally, a correction procedure is introduced.

The simulation procedure has been introduced in Mahling et al. (2013), where results for Examples A and B have been shown, as well.

6.1. Simulation procedure based on the estimated intensity function

X is assumed to be an inhomogeneous Poisson point process with intensity $\lambda_X(\mathbf{s}) = 1/(1-q) \cdot \lambda_Y(\mathbf{s})$. In a first step, $\lambda_Y(\mathbf{s})$ is estimated using the kernel method described in Section 4.3.2. In each iteration performed, an inhomogeneous Poisson point process X^* with intensity $\lambda_{X^*}(\mathbf{s}) = 1/(1-q) \cdot \hat{\lambda}_Y(\mathbf{s})$ is simulated. As before, it is partitioned into the process Y^* of observed and the process Z^* of unobserved events by drawing independent Bernoulli distributed random numbers with probability q, i.e. $P(x^* \in Z^*) = q$. Finally, a high-risk zone based on Y^* is constructed via the intensity-based method. This procedure can be regarded as a parametric bootstrap (see Efron and Tibshirani, 1993; Gentle, 2005; Givens and Hoeting, 2005). The fraction p_{out} of high-risk zones for which at least one unobserved event from X^* is situated outside the zone expresses the risk. In addition, the fraction p_{miss} of unobserved events outside the high-risk zone and the area of the high-risk zone can be computed. Again, a probability of non-explosion q of 0.10 and 0.15 and $\alpha = 0.4$, $\alpha = 0.2$ and $\alpha = 0.1$ were considered.

To investigate the reason for the deviation of p_{out} from α, a modified version of the simulation procedure was applied additionally: Instead of estimating the underlying intensity of the simulated patterns in every iteration, the high-risk zone was determined based on the intensity which was used for simulation, so the high-risk zone was exactly the same in every iteration, but the patterns changed. This procedure yields an ‘oracle estimator’, which represents the most optimistic performance of the intensity-based method.

6.2. Results

For Examples A, B, D and E, 10000 iterations were performed. As Examples C and F comprise a large number of events, only 1000 iterations were performed for these two patterns.

The results for Example A are shown in Table 6.1. The fraction p_{out} was close to α for $q = 0.1$ and $\alpha = 0.4$, but clearly larger than α for all other combinations of parameters. The relative bias $(p_{\text{out}} - \alpha)/\alpha$ was between 0.0187 and 0.5890.

Table 6.1.: Bootstrap result: Fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside; Example A, intensity-based method (INT) and oracle estimator (ORACLE)

A	q	0.1	0.1	0.1	0.15	0.15	0.15
	α	0.4	0.2	0.1	0.4	0.2	0.1
INT	p_{out}	0.407	0.243	0.159	0.428	0.274	0.187
ORACLE	p_{out}	0.398	0.202	0.100	0.398	0.199	0.097

Even larger differences were observed for Example B (Table 6.2), where p_{out} exceeded α in all cases and the relative bias was between 0.1382 and 1.0180.

Table 6.2.: Bootstrap result: Fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside; Example B, intensity-based method (INT) and oracle estimator (ORACLE)

B	q	0.1	0.1	0.1	0.15	0.15	0.15
	α	0.4	0.2	0.1	0.4	0.2	0.1
INT	p_{out}	0.455	0.264	0.169	0.470	0.298	0.202
ORACLE	p_{out}	0.403	0.203	0.106	0.400	0.200	0.100

For Example C (Table 6.3), the fraction p_{out} was generally smaller than α. The relative bias was between -0.1800 and -0.0800.

Table 6.3.: Bootstrap result: Fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside; Example C, intensity-based method (INT) and oracle estimator (ORACLE)

C	q	0.1	0.1	0.1	0.15	0.15	0.15
	α	0.4	0.2	0.1	0.4	0.2	0.1
INT	p_{out}	0.357	0.174	0.092	0.344	0.164	0.085
ORACLE	p_{out}	0.395	0.190	0.085	0.389	0.205	0.098

Table 6.4.: Bootstrap result: Fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside; Example D, intensity-based method (INT) and oracle estimator (ORACLE)

D	q	0.1	0.1	0.1	0.15	0.15	0.15
	α	0.4	0.2	0.1	0.4	0.2	0.1
INT	p_{out}	0.260	0.129	0.073	0.261	0.135	0.080
ORACLE	p_{out}	0.408	0.202	0.101	0.400	0.204	0.100

Table 6.4 shows the results for Example D. The fraction p_{out} was considerably smaller than α for all six combinations of parameters. The relative bias was between -0.3560 and -0.2020. The results for Example E are similar, the fraction p_{out} was generally even smaller than for Example D, which results in a relative bias between -0.8170 and -0.6860.

Table 6.5.: Bootstrap result: Fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside; Example E, intensity-based method (INT) and oracle estimator (ORACLE)

E	q	0.1	0.1	0.1	0.15	0.15	0.15
	α	0.4	0.2	0.1	0.4	0.2	0.1
INT	p_{out}	0.126	0.045	0.018	0.107	0.038	0.016
ORACLE	p_{out}	0.405	0.201	0.103	0.403	0.198	0.099

For Example F, p_{out} was smaller than α in four of six cases. The relative bias was between -0.0700 and 0.2400 (Table 6.6).

Table 6.6.: Bootstrap result: Fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside; Example F, intensity-based method (INT) and oracle estimator (ORACLE)

F	q	0.1	0.1	0.1	0.15	0.15	0.15
	α	0.4	0.2	0.1	0.4	0.2	0.1
INT	p_{out}	0.379	0.205	0.137	0.372	0.207	0.124
ORACLE	p_{out}	0.413	0.189	0.105	0.399	0.182	0.080

In all cases, p_{out} was close to α for the oracle estimator: The relative bias $(p_{\text{out}} - \alpha)/\alpha$ was between -0.0260 and 0.0105 for Example A, between 0 and 0.0630 for Example B and between -0.1500 and -0.0125 for Example C. For Example D, a relative bias between -0.0010 and 0.0215 was obtained. The relative bias was between -0.0120 and 0.0330 for Example E and between -0.2000 and 0.0500 for Example F.

This indicates that the estimation of the intensity is a crucial issue for the procedure and that failure to meet the specification originates from this step.

A comparison of the fractions p_{out} from the bootstrap simulation and the fractions p_{out} which were obtained in the evaluation simulation in Section 5.2 reveals that Example E is the only example for which all values are larger than in Section 5.2. For all other examples, deviations were observed in both directions: The bootstrap fractions are smaller in five of six cases for Example A, in four cases for Example B and in five cases for Example D. For Example F, in contrast, they are larger in five of six cases. With regard to Example C, three of the bootstrap fractions are smaller and three are larger. A possible reason for these findings may be that for Examples E and F, the observed bomb craters are concentrated on a relatively small part of the observation window.

6.3. Bootstrap correction

To correct for the bias with respect to α and obtain a high-risk zone with the intended failure probability, the parametric bootstrap can be used in the following way:

1. For a given spatial point pattern, the intensity function is estimated and the simulation procedure based on the estimated intensity function as described in Section 6.1 is applied.

2. The fraction p_{out} represents an estimator for the failure probability, so the simulation procedure yields an updated failure probability, which is possibly different from the intended value.

This can be repeated with adapted values for the parameter α until we find the value which results in the intended failure probability. To find this value more quickly, the search can be realised in an adaptive procedure.

The algorithm for a possible adaptive procedure is given in Algorithm 1. The main idea is that a desired failure probability `failprob` can be specified and the procedure yields a value `cutoff`–which will usually differ from the desired failure probability–which is the value that should be used for the parameter α when a high-risk zone is determined.

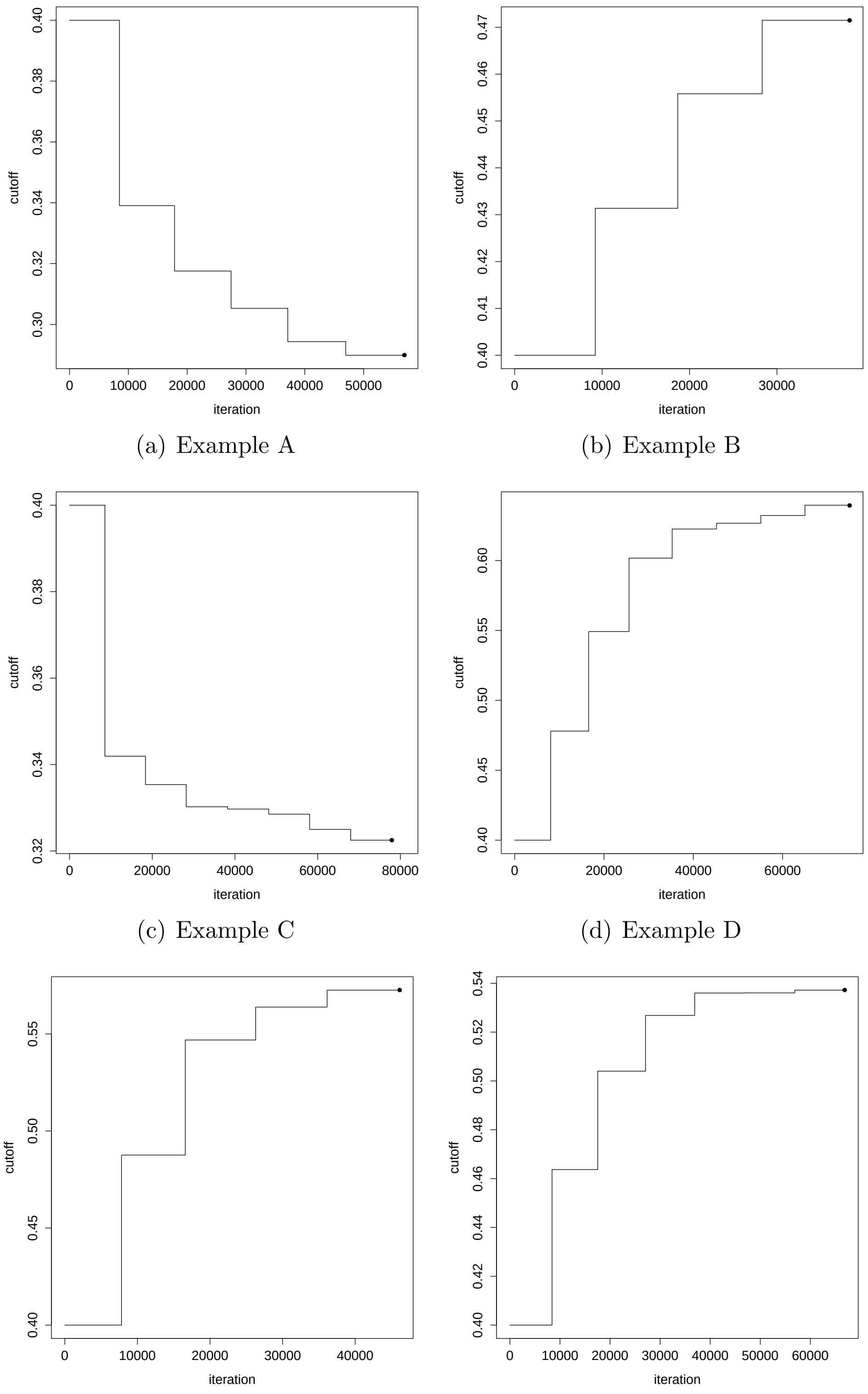
The precision of the routine depends on the number of iterations `numit` and the tolerance `tol`. In every iteration, the algorithm checks if `failprob - tol` $\leq p_{out} \leq$ `failprob` $+$ `tol` can still be achieved. If this is not the case, a new value for `cutoff` is proposed.

Figure 6.1 shows how the value of `cutoff` is adapted for Examples A to F, starting with $\alpha = 0.4$. The number of iterations `numit` was 10000 and `tol` was set to 0.01. For Examples A and C, the value which was obtained was smaller than 0.4 (0.28986 and 0.32250, respectively), whereas we obtained larger values for Examples B, D, E and F (0.47150, 0.63957, 0.57259 and 0.53720).

```
Input   : spatial point pattern pattern, desired failure probability failprob,
          number of iterations numit, tolerance tol, probability of
          non-observation q
Output: parameter cutoff which should be used to determine a high-risk zone
          with desired failure probability

numout $\leftarrow$ 0;
iter $\leftarrow$ 1;
cutoff $\leftarrow$ failprob;
while iter $\leq$ numit do
    simulate pattern from estimated intensity of observed pattern pattern;
    perform thinning;
    use thinned pattern to determine high-risk zone with parameter cutoff;
    evaluate high-risk zone and determine number of unobserved events outside
    high-risk zone (numbermiss);
    if numbermiss $>$ 0 then
        numout $\leftarrow$ numout + 1;
    poutmin $\leftarrow$ numout  / numit;
    poutmax $\leftarrow$ (numout + numit $-$ iter) / numit;
    if poutmin $>$ failprob + tol then
        cutoff $\leftarrow$ cutoff * iter/ (numit + 1);
        iter $\leftarrow$ 0;
        numout $\leftarrow$ 0;
    if poutmax $<$ failprob $-$ tol then
        cutoff $\leftarrow$ cutoff * (1 + (numit $-$ iter + 1) / numit);
        iter $\leftarrow$ 0;
        numout $\leftarrow$ 0;
    iter $\leftarrow$ iter + 1;
```

Algorithm 1: Bootstrap correction



(a) Example A

(b) Example B

(c) Example C

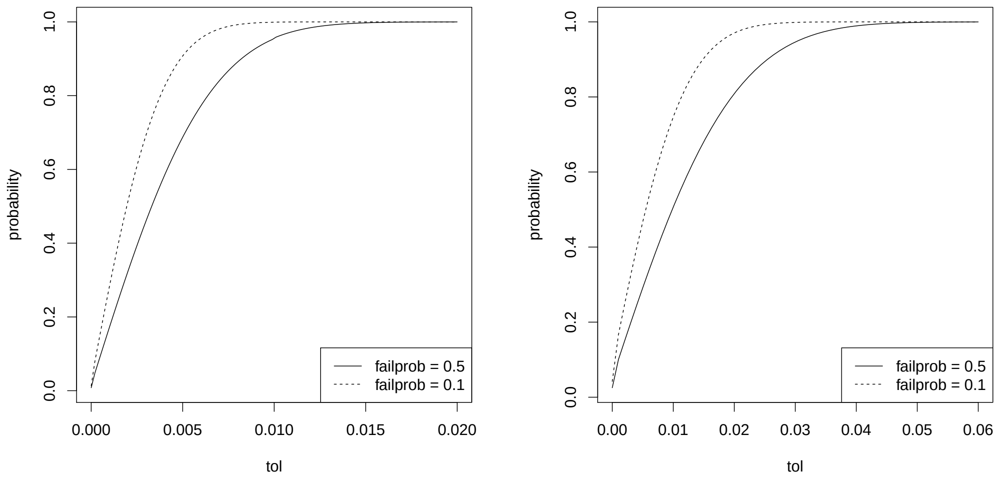
(d) Example D

(e) Example E

(f) Example F

Figure 6.1.: Bootstrap correction: Values tested for α in course of the procedure. The dot represents the final value.

The tolerance parameter `tol` must be chosen reasonably depending on `failprob` and `numit`. If it is chosen too small, the probability that the proposed `cutoff` is accepted after `numit` iterations is small even if it is correct, which results in unnecessary tries with further values for `cutoff`. If we assume that `cutoff` is correct and therefore `failprob` is adhered to (and we ignore the stopping criterion), the number `numout` of high-risk zones with at least one unobserved event outside follows a binomial distribution whose parameters are `failprob` and `numit`. Figure 6.2 shows the probability that `numout` takes a value between `failprob - tol` and `failprob` + `tol` as a function of `tol` for `numit` = 10000 and `numit` = 1000. The figure shows the worst case scenario `failprob` = 0.5, where the probability is smaller than for every other value of `failprob`, and the situation for `failprob` = 0.10. Figure 6.3 illustrates what happens if `tol` is chosen too small. Many different values for `cutoff` between 0.66 and 0.72 are tested, but none of them is accepted. Note that the bootstrap correction is based on a stochastic alogrithm, which explains why the final `cutoff` for Example E was smaller than the values which are tested in Figure 6.3.



(a) `numit` = 10000 (b) `numit` = 1000

Figure 6.2.: Probability that a random variable which follows a binomial distribution with parameters `numit` and `failprob` takes a value between `failprob - tol` and `failprob` + `tol`.

To recapitulate this chapter: A simulation procedure based on the estimated intensity function has been introduced. The risk assessment revealed that the obtained failure probability usually deviates from the parameter α. Therefore, a bootstrap correction procedure has been presented. It yields an updated value for α which leads to the intended failure probability.

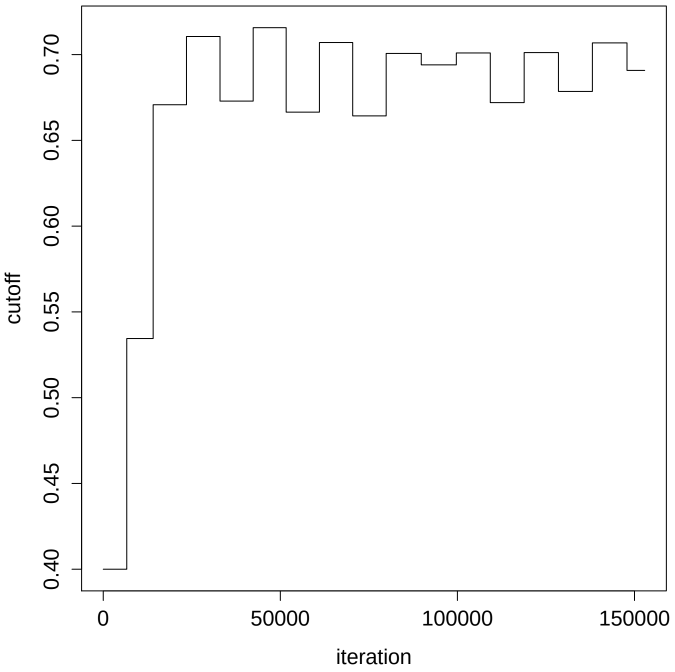


Figure 6.3.: Bootstrap correction: Values tested for α in course of the procedure for Example E with `numit` = 10000 and `tol` = 0.001.

7. Spatial clustering

The assumption of an inhomogeneous Poisson point process for the intensity-based method implies that beyond spatial variation in the intensity function, there is no stochastic dependence between observations. However, high values were obtained for Ripley's K-function and the pair correlation function in Chapter 3 and Section 5.1, which may indicate clustering. The inhomogeneous Poisson model fits the subject-matter theory better than cluster models, but as clustering cannot be ruled out, it appears advisable to perform a sensitivity analysis and investigate the behaviour of the intensity-based method in case that the bombing point pattern is generated by a cluster process instead of an inhomogeneous Poisson point process. A related question is if specific cluster models can be fitted to the data and how the intensity-based high-risk zones behave if these models are used in the simulation procedure from Section 6.1 instead of the inhomogeneous Poisson process.

Some of the results for Examples A and B and the procedure for the sensitivity analysis in Section 7.1 have been shown in Mahling et al. (2013). In most parts of this chapter, Examples C and F will not be considered because they comprise a large number of events, which makes simulations very time-consuming.

7.1. Sensitivity analysis

7.1.1. Simulation procedure based on the intensity

In order to obtain simulated processes resembling the observed patterns, the estimated intensity was used as a starting point and clustering was added. To achieve this, Neyman-Scott processes (Neyman and Scott, 1958) were used following the definition of Cressie (1993), which states that the cluster centres may form an inhomogeneous Poisson point process.

Using such a Neyman-Scott point process model results in a modified simulation procedure for X^* compared to Section 6.1: First, the cluster centres were simulated as an inhomogeneous Poisson point process with intensity function

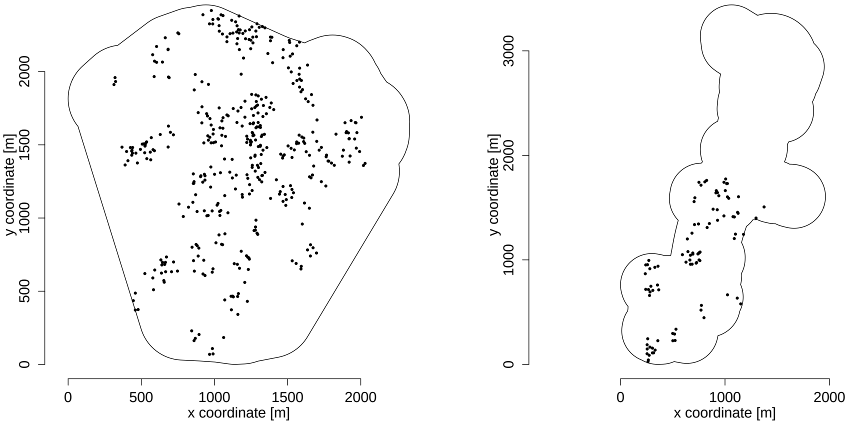
$$\lambda_{C^*}(\mathbf{s}) = \frac{1}{\tau \cdot (1-q)} \cdot \hat{\lambda}_Y(\mathbf{s}). \tag{7.1}$$

The number of points per cluster follows a Poisson(τ) distribution, the cluster points are placed independently and uniformly inside a disc of radius r centered at the cluster centres. The parameter $\tau > 0$ determines the extent of clustering. If τ is small, the process of cluster centres will contain almost as many points as the simulated Poisson processes in Section 6.1.

Each of these centres will be replaced by a small number of cluster points. Note that even for $\tau = 1$, the process is clustered: Even though the process of cluster centres is a Poisson process with the same intensity as in Section 6.1, the resulting process is not, since each cluster centre is replaced by a small number of cluster points and this number is not in all cases 1, but may be zero or larger than 1. If τ is large, the process of cluster centres will consist of a small number of points, whereas the clusters will comprise a large number of cluster points.

7.1.2. Behaviour of the high-risk zones

The radius r in the sensitivity analysis was chosen to be 80 m for Examples A and B, a value which is larger than most of the observed nearest-neighbour distances, but small enough to obtain clearly visible clusters for these two examples. Simulated Neyman-Scott processes for $r = 80\ m$ and $\tau = 5$ are depicted in Figure 7.1. The clusters are clearly visible. As the nearest-neighbour distances are generally smaller for Examples D and E, a smaller parameter value $r = 30\ m$ was applied for these two patterns (Figure 7.2). Six different values for τ were considered to study the consequences of different extents of clustering. 1000 iterations were performed for each combination of parameters. All other aspects of the simulation setting remained as in Section 6.1.

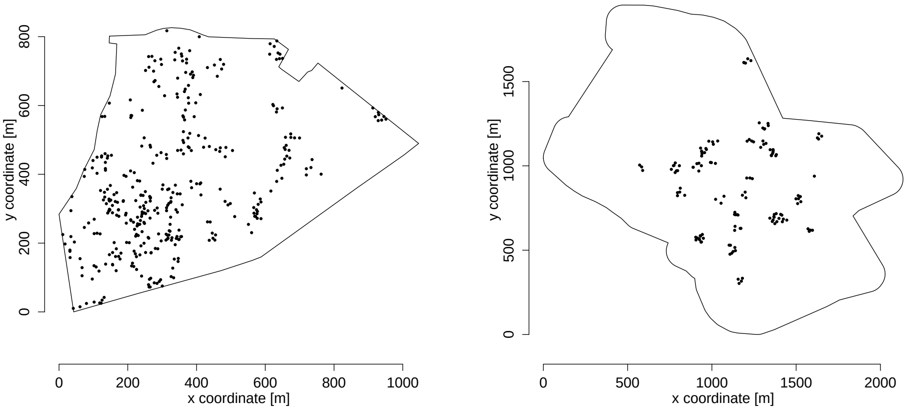


(a) Example A

(b) Example B

Figure 7.1.: Examples of simulated Neyman-Scott processes for $r = 80\ m$ and $\tau = 5$.

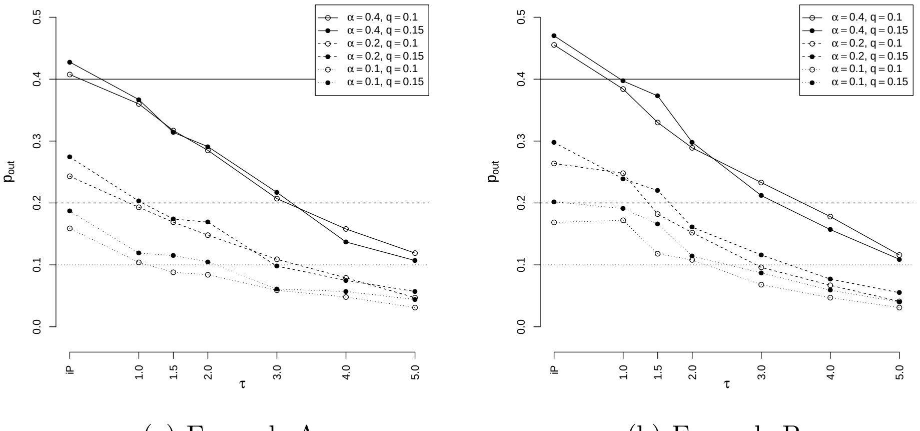
To facilitate the comparison between the cluster model and the inhomogeneous Poisson process model, the results obtained with the latter have been integrated into the figures although these results have already been shown in Section 6.2. As we can see in Figure 7.3, the intensity-based construction method for high-risk zones is conservative for Examples A and B if the pattern is a cluster process instead of an inhomogeneous Poisson process. In most cases, the mean fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside was smaller than the α values of 0.4, 0.2 and 0.1.



(a) Example D

(b) Example E

Figure 7.2.: Examples of simulated Neyman-Scott processes for $r = 30\ m$ and $\tau = 5$.

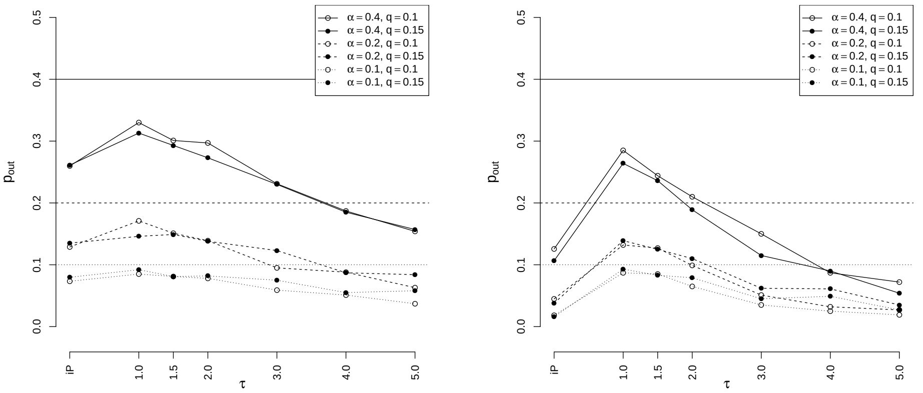


(a) Example A

(b) Example B

Figure 7.3.: Fraction of high-risk zones with at least one exploded bomb outside for different values of the cluster parameter τ and for the inhomogeneous Poisson process model ("iP"), $r = 80\ m$.

In particular, the numbers are smaller than for the inhomogeneous Poisson point process investigated in Section 6.2. The larger the value of τ, i.e. the more clustered the pattern, the smaller the value of p_{out}. The results for Examples D and E are depicted in Figure 7.4. Again, the mean fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside was smaller than the α values of 0.4, 0.2 and 0.1. However, p_{out} is larger than for the inhomogeneous Poisson point process investigated in Section 6.2 if τ is small.

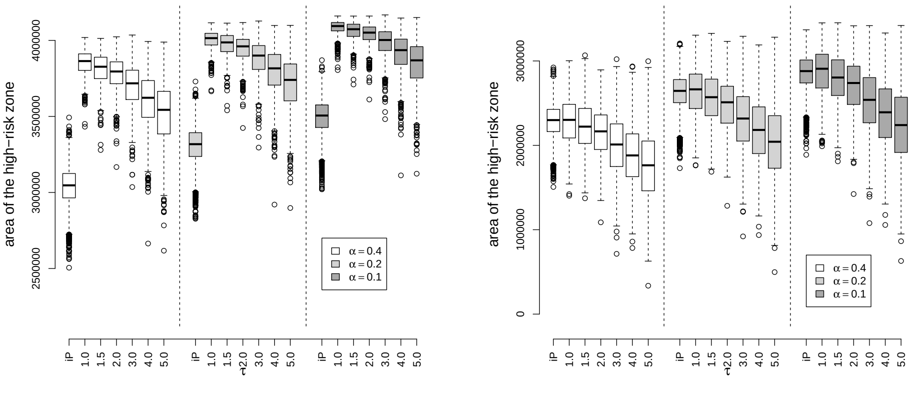


(a) Example D

(b) Example E

Figure 7.4.: Fraction of high-risk zones with at least one exploded bomb outside for different values of the cluster parameter τ and for the inhomogeneous Poisson process model (“iP”), $r = 30\ m$.

Figures 7.5 and 7.6 illustrate that smaller high-risk zones result as a consequence of larger τ. For Examples A and E, the high-risk zones were usually larger than in case of an inhomogeneous Poisson process, whereas they were often smaller for Examples B and D.

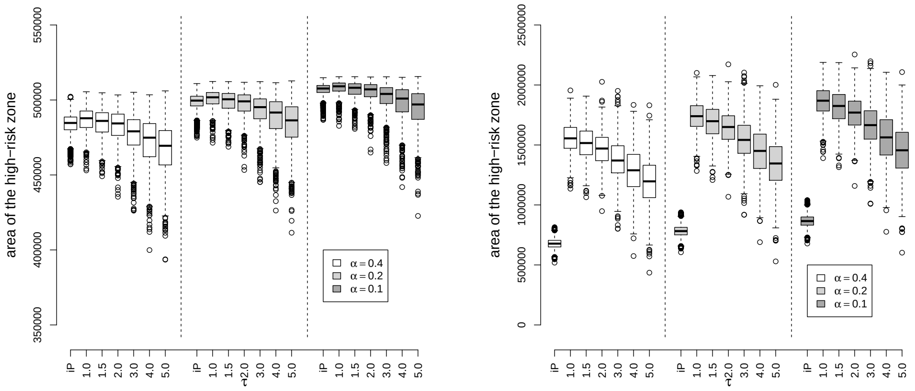


(a) Example A

(b) Example B

Figure 7.5.: Area of high-risk zones for $q = 0.10$ for different values of the cluster parameter τ and for the inhomogeneous Poisson process model (“iP”), $r = 80\ m$.

In summary, the sensitivity analysis indicates that the intensity-based method can be used if the bomb crater pattern is clustered. However, the resulting high-risk zones will usually be too large, so the approach is conservative.



(a) Example D

(b) Example E

Figure 7.6.: Area of high-risk zones for $q = 0.10$ for different values of the cluster parameter τ and for the inhomogeneous Poisson process model ("iP"), $r = 30\ m$.

7.2. Fitting cluster models to the bomb crater patterns

For the bootstrap correction introduced in Section 6.3, the inhomogeneous Poisson process served as the point process model from which patterns were simulated. If the assumption of an inhomogeneous Poisson process is not appropriate, other point process models can be used instead. For patterns which show a tendency to clustering, a classical cluster process such as the Thomas or the Matérn process, whose parameters can be estimated with the method of minimum contrast using one of the summary functions, is a natural choice.

In this section, these models are studied in more detail than in Section 2.5. Estimation is explained and the results for Examples A to F are discussed. Finally, the bootstrap simulation introduced in Section 6.1 is repeated, where the inhomogeneous Poisson point process is replaced by a Thomas process.

7.2.1. Summary functions for Thomas and Matérn processes

For Thomas and Matérn processes, explicit formulae for Ripley's K-function and the pair correlation function can be given. Let κ denote the intensity of the homogeneous Poisson point process of cluster centres (i.e. the intensity of the parent process). The number of cluster points per cluster (also called *offspring* or *daughter points*) follows a Poisson distribution with parameter μ. For the Thomas process, the positions of the cluster points relative to the cluster centres are given by a Gaussian distribution with mean $\mathbf{0}$ and standard deviation σ, whereas the cluster points of a Matérn process are uniformly distributed in a sphere of radius R centered at the cluster centre.

For a Thomas process, the pair correlation function is

$$g(r) = 1 + \frac{\exp\left\{-\frac{r^2}{4\sigma^2}\right\}}{4\pi\kappa\sigma^2}$$

and Ripley's K-function is

$$K(r) = \pi r^2 + \frac{1 - \left\{\frac{r^2}{4\sigma^2}\right\}}{\kappa}$$

(see Illian et al. (2008, page 377) for the pair correlation function and Møller and Waagepetersen (2003, page 62) for both functions).

Analogue expressions, which are however more complicated, can be given for Matérn processes, see Illian et al. (2008, page 376) and Stoyan (1992) for the pair correlation function. A formula for Ripley's K-function is given in the `spatstat` documentation.

7.2.2. Method of minimum contrast

The basic idea of the method of minimum contrast (also called 'minimum contrast method') is described in Illian et al. (2008, pages 450–452). One chooses a suitable summary characteristic S depending on the unknown parameters θ. The difference between the theoretical

S_θ for a specific choice of parameters and $\hat{S}$ estimated from the given data is minimised with respect to θ. For functional summary characteristics, a least-squares approach can be used:

$$\Delta(\theta) = \int_{s_1}^{s_2} |\hat{S}(r) - S_\theta(r)|^\beta dr$$

is minimised, where $\beta = 2$ is often used. In practice, the integral is approximated by a sum:

$$\Delta(\theta) \approx \sum_{i=0}^{k} |\hat{S}(\rho_i) - S_\theta(\rho_i)|^\beta \cdot \delta,$$

where $\rho_0 = s_1$, $\rho_k = s_2$, $\rho_i = s_1 + i\delta$ and $\delta = \frac{s_2 - s_1}{k}$ for an integer k. Illian et al. (2008) recommend using the pair correlation function and choosing a positive s_1 near the estimated mean nearest-neighbour distance.

If S_θ is unknown, it can be estimated using a simulation (Diggle, 1978; Diggle and Gratton, 1984). In `spatstat`, a modified version with

$$\Delta(\theta) = \int_{s_1}^{s_2} |(\hat{S}(r))^q - (S_\theta(r))^q|^\beta dr$$

is implemented, where $q = 0.25$ is recommended following Diggle (2003) and Waagepetersen (2007).

7.2.3. Application to Examples A to F

Thomas and Matérn cluster processes are fitted to Examples A to F, both with Ripley's K-function and the pair correlation function as summary characteristic S. As both functions are invariant under p-thinning, it is a sensible approach to fit the model based on the bomb crater patterns to obtain models which are useful for the whole process of bomb craters and unexploded bombs.

Table 7.1.: Parameter estimates for Thomas model

	estimated using $K(\cdot)$			estimated using $g(\cdot)$		
Example	$\hat{\kappa} \cdot \nu(W)$	$\hat{\sigma}^2$	$\hat{\mu}$	$\hat{\kappa} \cdot \nu(W)$	$\hat{\sigma}^2$	$\hat{\mu}$
A	4.5	15388	97	4	19912	111
B	3.2	34360	33	3.4	32343	30
C	10.3	24737	133	9.5	27346	145
D	22.4	3359	20	18.5	4653	24
E	2.2	12952	70	2.7	8727	57
F	2.3	47387	751	2.1	52530	817

The estimated parameter values are given in Tables 7.1 and 7.2. The estimates for κ are multiplied by the area of the observation window to see how many clusters can be expected. The number of expected clusters is generally low for Examples A, B, E and

Table 7.2.: Parameter estimates for Matérn cluster model

	estimated using $K(\cdot)$			estimated using $g(\cdot)$		
Example	$\hat{\kappa} \cdot \nu(W)$	$\hat{R}$	$\hat{\mu}$	$\hat{\kappa} \cdot \nu(W)$	$\hat{R}$	$\hat{\mu}$
A	4.6	235	96	7.1	166	62
B	3.3	341	32	3.8	310	27
C	73	41	19	10.1	300	136
D	22.4	110	20	19.5	123	23
E	2.2	213	69	2	241	77
F	12.5	24	137	2.3	414	758

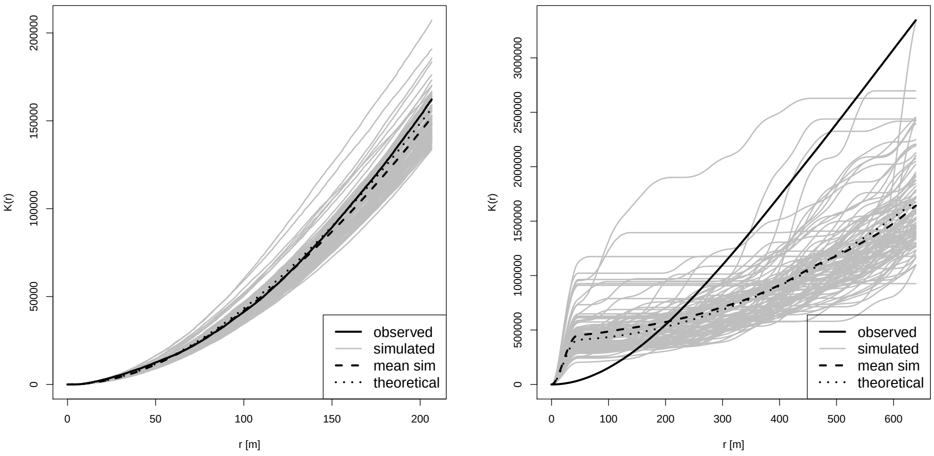
F. The parameter estimates for both Thomas models are similar, whereas the parameters of the Matérn models show some peculiarities: Large differences for $\hat{\mu}$ are observed for Example C, where $\hat{\mu}$ is small for the Matérn model fitted by using the K-function. For Examples C and F, the expected number of clusters was much larger for the Matérn model fitted by using the K-function. In what follows, the four types of models are discussed in more detail. Examples of patterns simulated from the fitted models are shown. The estimated K-functions, pair correlation functions and empty-space functions of the observed patterns are compared to those of 100 simulated patterns. In contrast to Section 5.1, the summary functions of the simulated patterns are not combined to an envelope, but shown individually. This way, it is easier to get an impression of differences between single simulations, which is more useful than a formal Monte Carlo test in this context.

Matérn processes fitted by using the K-function

Matérn processes fitted by using the K-function cannot describe the observed patterns in an appropriate way: Simulations based on the fitted models yielded patterns whose points were concentrated on only a small fraction of the original area. This problem is especially severe for Examples A, C and F (Figure 7.8). The only case in which the simulated pattern resembles the observed pattern is Example D.

The estimated K-function for the observed pattern is compared to the estimated K-functions of 100 patterns simulated from the model fitted to the data and the theoretical K-function for this model in Figure 7.7. Results are not shown for all six examples. Instead, one positive and one negative case is illustrated. For Example D, the observed K-function and the mean of the simulated K-function are close, whereas the observed K-functions and the simulated K-function differ very much for Example F. Examples C and E also represent negative cases, but are not shown here. A similar depiction is given in Figure 7.9 for the pair correlation function. Negative cases in addition to Example C are Examples D and F. Example A represents the case in which observed and simulated pair correlation function agree best. As the empty-space function is estimated at discrete points which are chosen individually for each pattern by the `spatstat` routine, the mean of the simulated empty-space functions cannot be computed (Figure 7.10). For all examples except Example

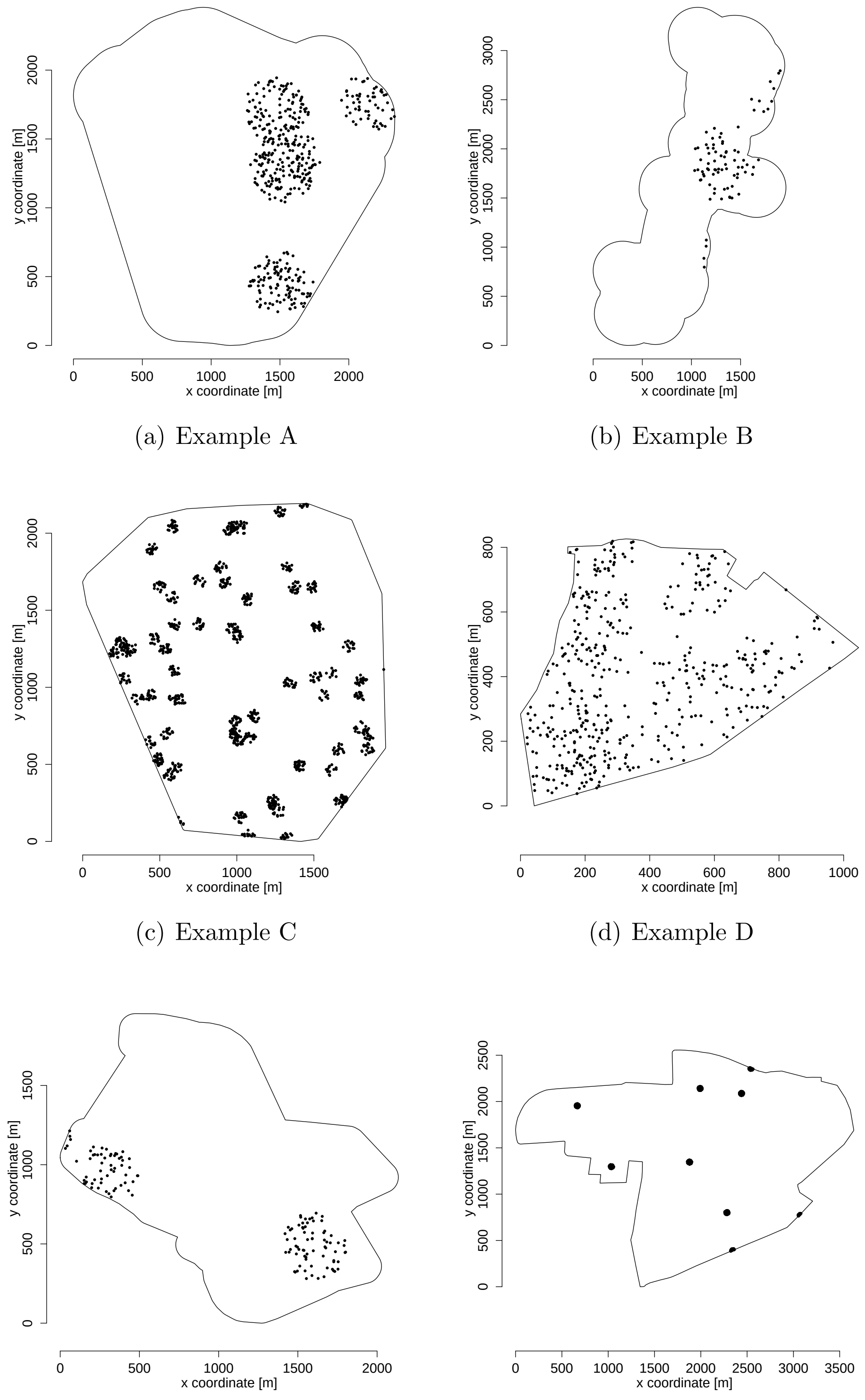
D, large deviations between the observed and the simulated empty-space functions were observed.



(a) Example D

(b) Example F

Figure 7.7.: Ripley's K-function: The solid black lines represent the estimated K-functions of the observed patterns, the grey lines correspond to the estimated K-functions for patterns which were simulated as a Matérn process (fitted by using the K-function), the dashed lines represent the mean of the estimated K-functions of the simulated patterns, the dotted lines give the theoretical K-functions for the fitted model.



(a) Example A

(b) Example B

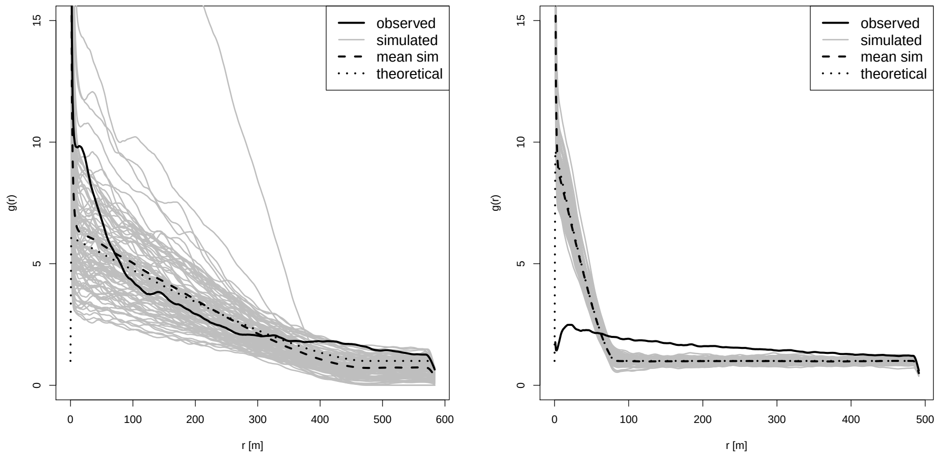
(c) Example C

(d) Example D

(e) Example E

(f) Example F

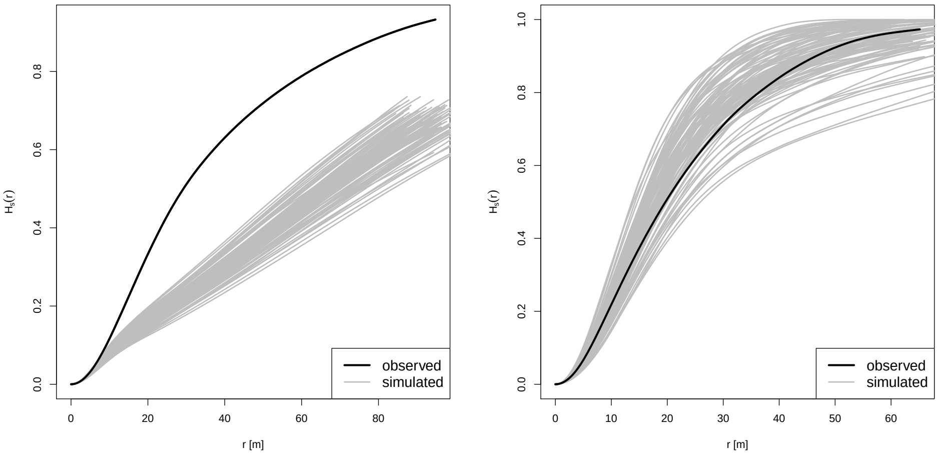
Figure 7.8.: Simulated examples of Matérn processes fitted by using the K-function.



(a) Example A

(b) Example C

Figure 7.9.: Pair correlation function: The solid black lines represent the estimated pair correlation functions of the observed patterns, the grey lines correspond to the estimated pair correlation functions for patterns which were simulated as a Matérn process (fitted by using the K-function), the dashed lines represent the mean of the estimated pair correlation functions of the simulated patterns, the dotted lines give the theoretical pair correlation functions for the fitted model.



(a) Example C

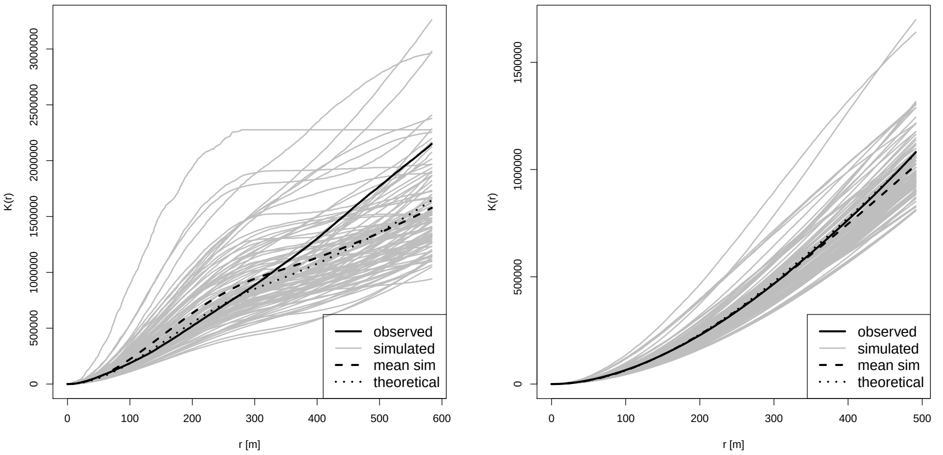
(b) Example D

Figure 7.10.: Empty-space function: The solid black lines represent the estimated empty-space functions of the observed patterns, the grey lines correspond to the estimated empty-space functions for patterns which were simulated as a Matérn process (fitted by using the K-function).

Matérn processes fitted by using the pair correlation function

The findings for Matérn processes fitted by using the pair correlation function are similar to those for Matérn processes fitted by using the K-function: Simulations based on the fitted models yielded patterns whose points were dispersed on a smaller area than the original patterns. Again, this problem is especially severe for Examples A, C and F (Figure 7.12) and the only case in which the simulated pattern resembles the observed pattern is Example D.

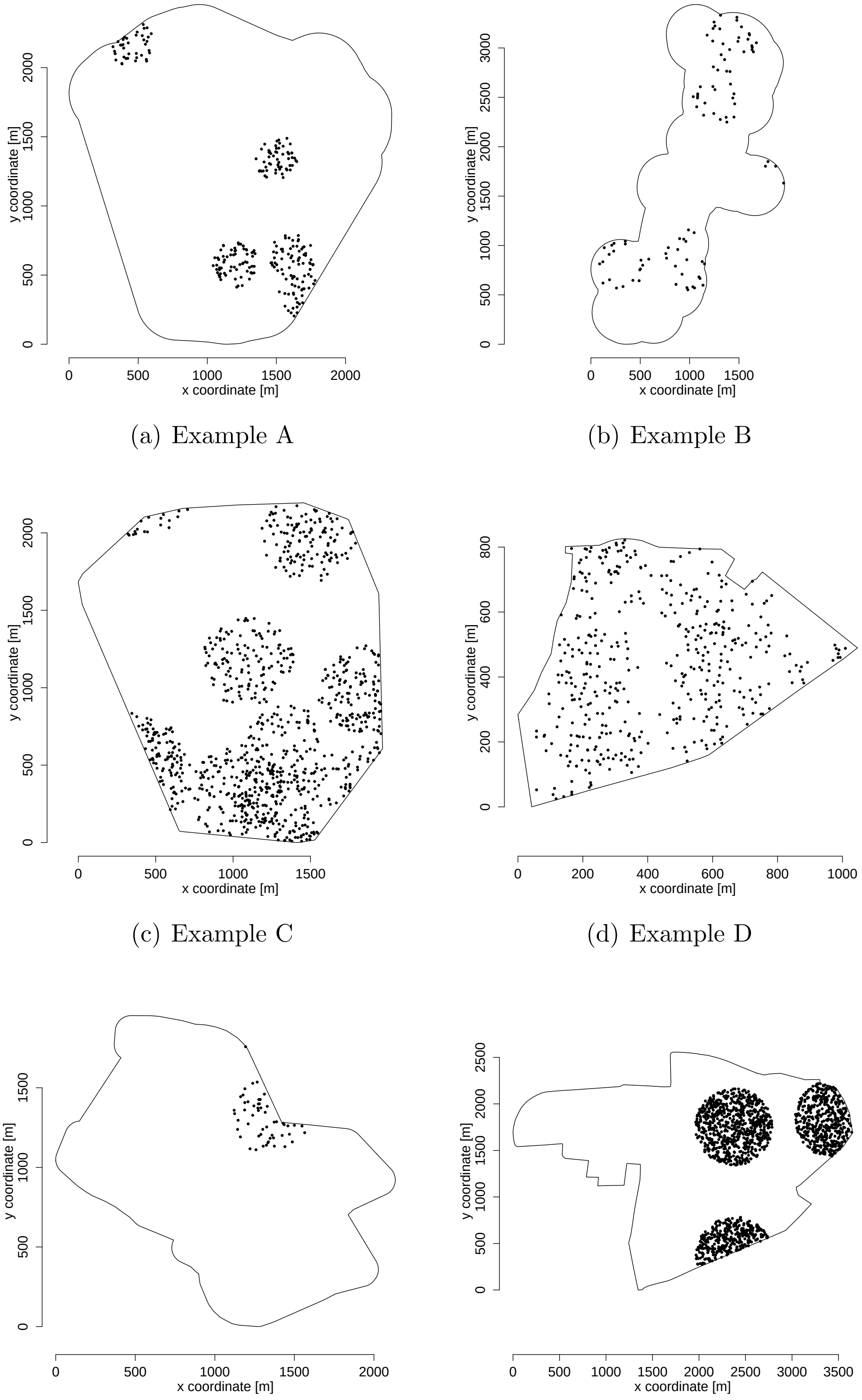
Regarding the K-function, a good fit was obtained for Examples C and D, whereas the deviations were large for Examples A, E and F (Figure 7.11). The pair correlation function could not be described very well for any of the examples. The best situation was observed for Example C, whose pair correlation function is depicted in Figure 7.13. A negative case (Example A) and the only positive case (Example D) in terms of the empty-space function are shown in Figure 7.14.



(a) Example A

(b) Example C

Figure 7.11.: Ripley's K-function: The solid black lines represent the estimated K-functions of the observed patterns, the grey lines correspond to the estimated K-functions for patterns which were simulated as a Matérn process (fitted by using the pair correlation function), the dashed lines represent the mean of the estimated K-functions of the simulated patterns, the dotted lines give the theoretical K-functions for the fitted model.



(a) Example A

(b) Example B

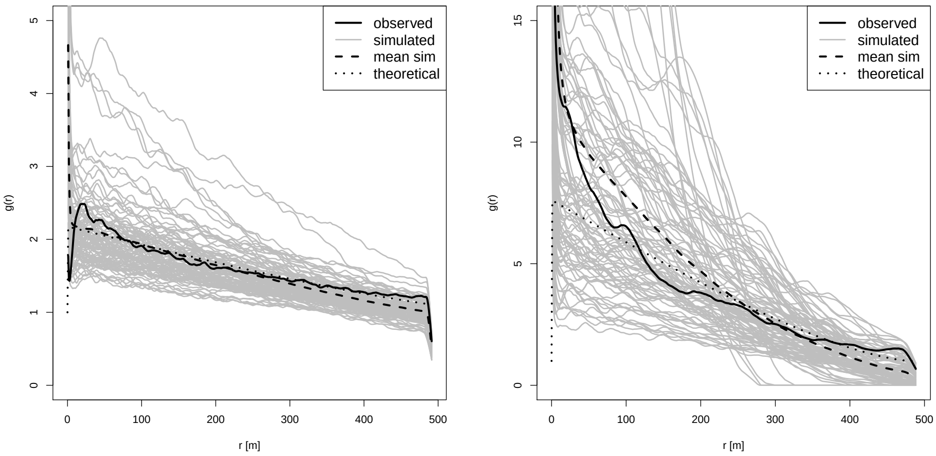
(c) Example C

(d) Example D

(e) Example E

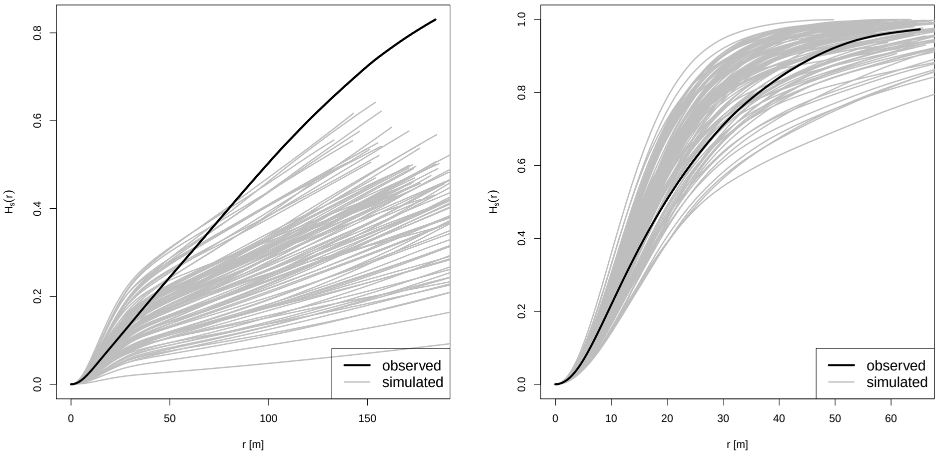
(f) Example F

Figure 7.12.: Simulated examples of Matérn processes fitted by using the pair correlation function.



(a) Example C (b) Example E

Figure 7.13.: Pair correlation function: The solid black lines represent the estimated pair correlation functions of the observed patterns, the grey lines correspond to the estimated pair correlation functions for patterns which were simulated as a Matérn process (fitted by using the pair correlation function), the dashed lines represent the mean of the estimated pair correlation functions of the simulated patterns, the dotted lines give the theoretical pair correlation functions for the fitted model.

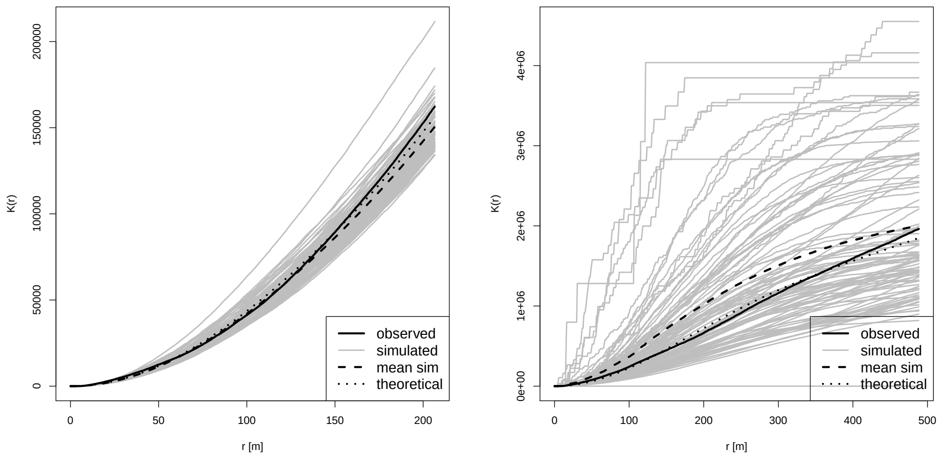


(a) Example A (b) Example D

Figure 7.14.: Empty-space function: The solid black lines represent the estimated empty-space functions of the observed patterns, the grey lines correspond to the estimated empty-space functions for patterns which were simulated as a Matérn process (fitted by using the pair correlation function).

Thomas processes fitted by using the K-function

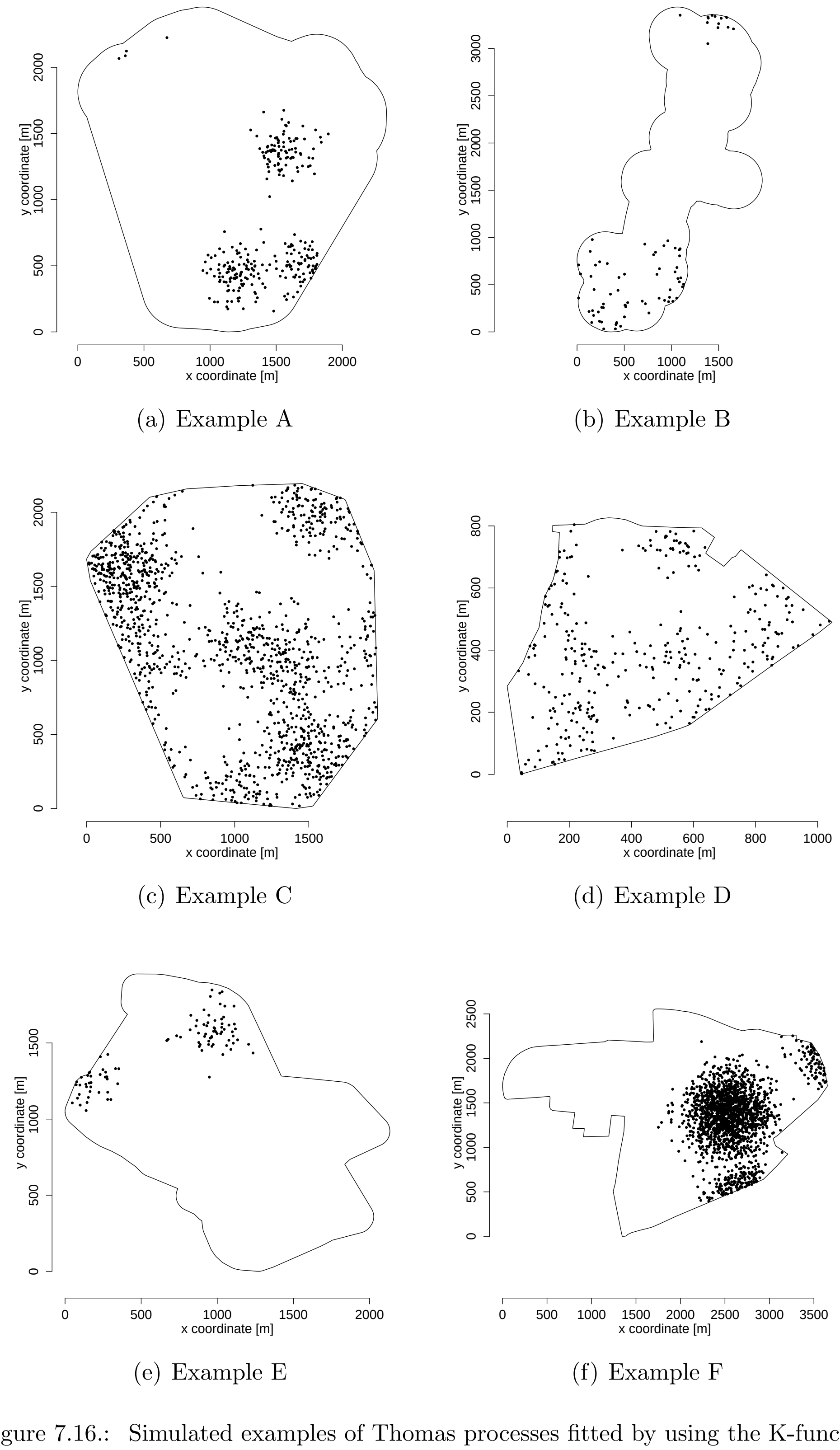
The fitted Thomas processes describe the observed patterns better than Matérn processes. As depicted in Figure 7.16, there are nonetheless examples for which simulated patterns look quite different from the observed patterns, in particular Examples A and C. The simulated K-functions are close to the observed K-functions for Examples B, C and D. Deviations were observed for Example A, E and F, the worst of them for Example E (see Figure 7.15). Regarding the pair correlation function, the largest deviations were observed for Example F, the best fit was achieved for Example B, which is depicted in Figure 7.17. Despite substantial deviations such as for Example A, the fit regarding the empty-space function was improved by far compared to the Matérn models (Figure 7.18).



(a) Example D

(b) Example E

Figure 7.15.: Ripley's K-function: The solid black lines represent the estimated K-functions of the observed patterns, the grey lines correspond to the estimated K-functions for patterns which were simulated as a Thomas process (fitted by using the K-function), the dashed lines represent the mean of the estimated K-functions of the simulated patterns, the dotted lines give the theoretical K-functions for the fitted model.



(a) Example A

(b) Example B

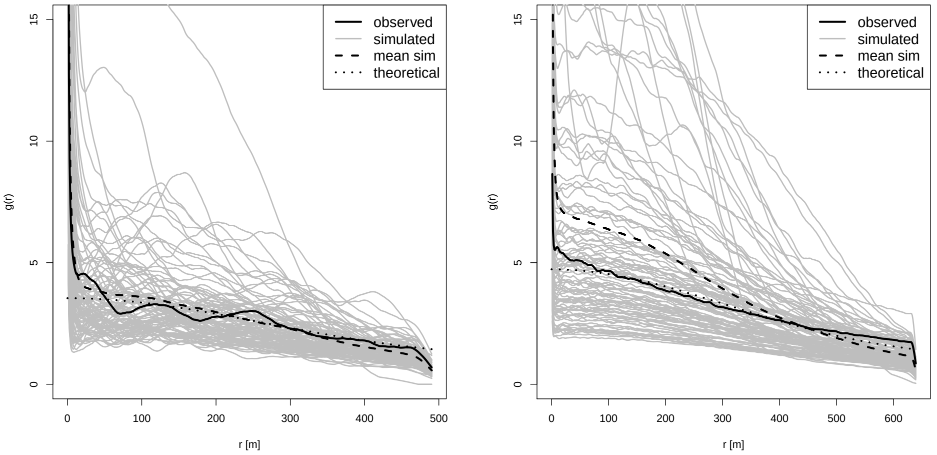
(c) Example C

(d) Example D

(e) Example E

(f) Example F

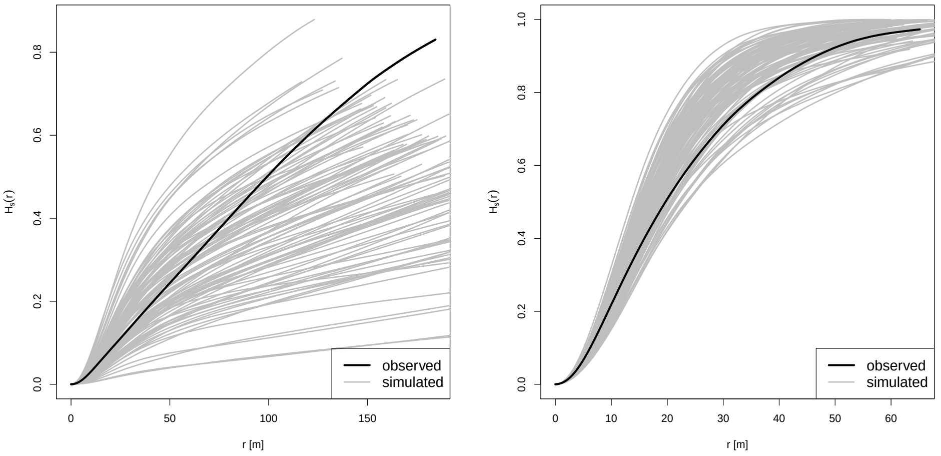
Figure 7.16.: Simulated examples of Thomas processes fitted by using the K-function.



(a) Example B

(b) Example F

Figure 7.17.: Pair correlation function: The solid black lines represent the estimated pair correlation functions of the observed patterns, the grey lines correspond to the estimated pair correlation functions for patterns which were simulated as a Thomas process (fitted by using the K-function), the dashed lines represent the mean of the estimated pair correlation functions of the simulated patterns, the dotted lines give the theoretical pair correlation functions for the fitted model.



(a) Example A

(b) Example D

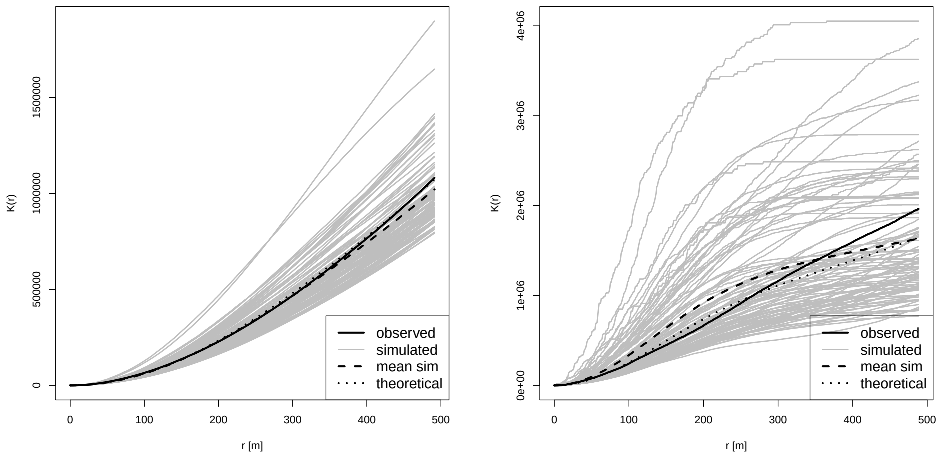
Figure 7.18.: Empty-space function: The solid black lines represent the estimated empty-space functions of the observed patterns, the grey lines correspond to the estimated empty-space functions for patterns which were simulated as a Thomas process (fitted by using the K-function).

Thomas processes fitted by using the pair correlation function

The results for Thomas processes fitted by using the pair correlation function are extremely similar to the results for Thomas processes fitted by using Ripley's K-function.

As depicted in Figure 7.20, patterns still look quite different from the observed patterns for Examples A and C. The simulated K-functions are close to the observed K-functions for Examples B, C and D (Figure 7.19). Regarding the pair correlation function, large deviations were observed for Examples A, E and F, the best fit was achieved for Example B, which is depicted in Figure 7.21, and for Example C. The simulated empty-space functions (Figure 7.22) were similar to those for the Thomas processes fitted by using Ripley's K-function.

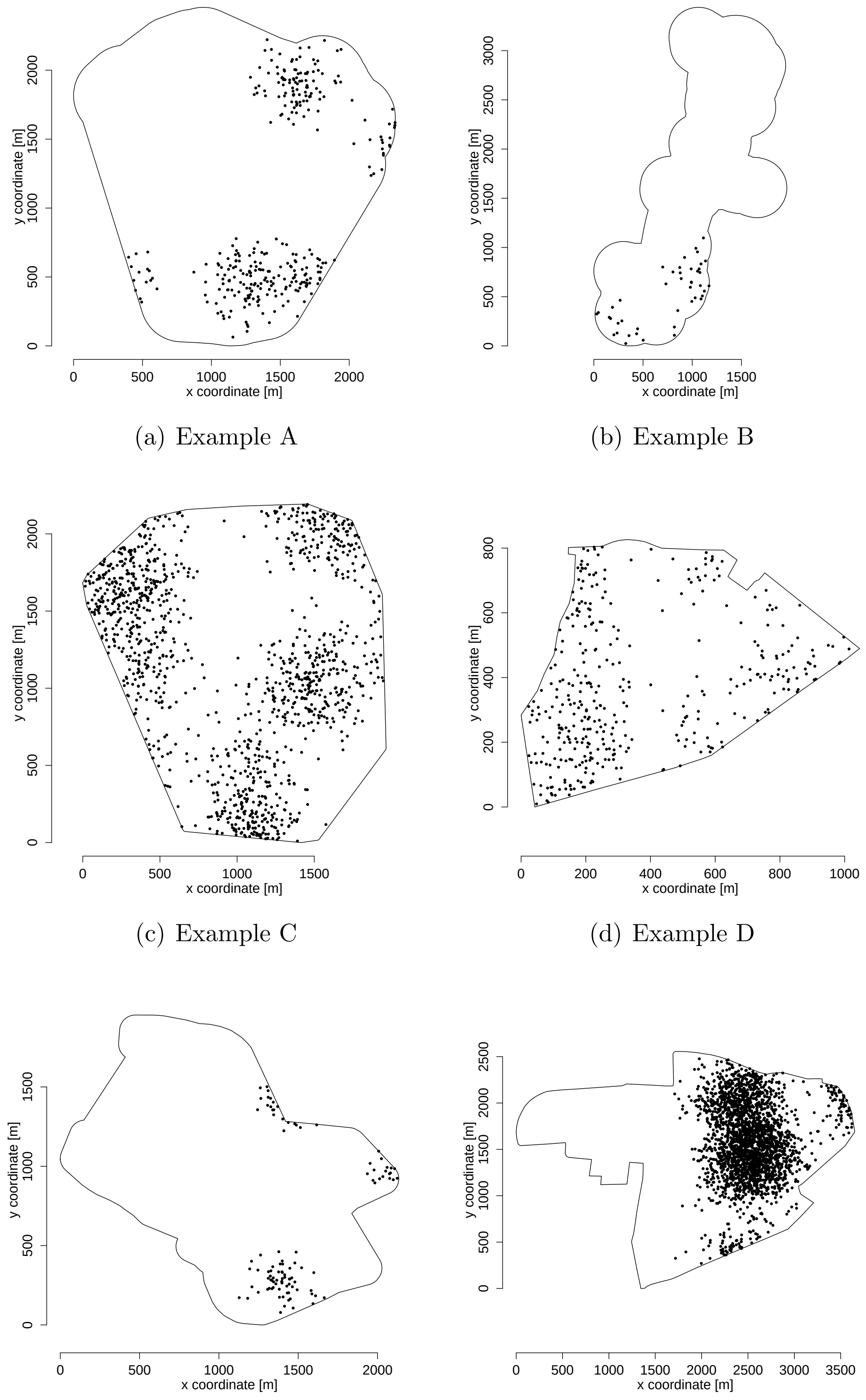
The fitted Thomas processes generally describe the observed patterns better than Matérn processes, so they are chosen for the sensitivity analysis. However, as the cluster centres are distributed randomly, single realisations may look very different compared to the observed patterns. From a user's point of view, this is unsatisfactory. A more natural approach should fix the cluster centres, e.g. by estimating the conditional intensity function given the cluster centres. This concept will be pursued in the following section.



(a) Example C

(b) Example E

Figure 7.19.: Ripley's K-function: The solid black lines represent the estimated K-functions of the observed patterns, the grey lines correspond to the estimated K-functions for patterns which were simulated as a Thomas process (fitted by using the pair correlation function), the dashed lines represent the mean of the estimated K-functions of the simulated patterns, the dotted lines give the theoretical K-functions for the fitted model.



(a) Example A

(b) Example B

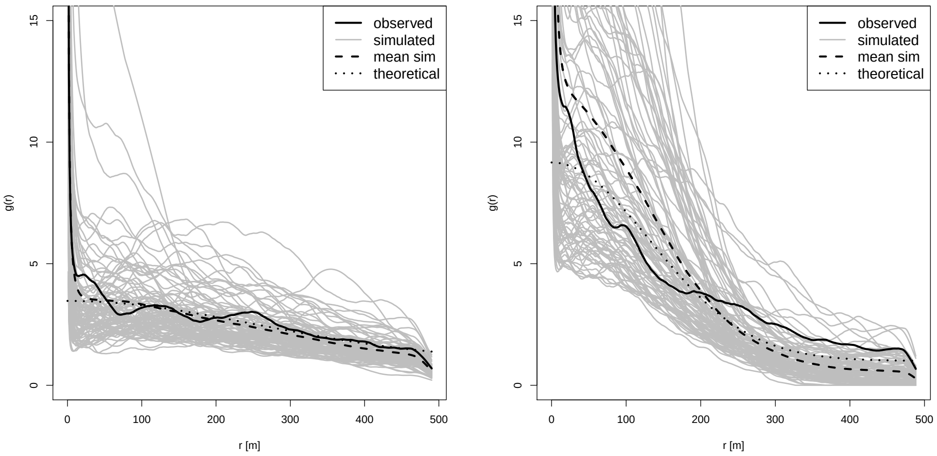
(c) Example C

(d) Example D

(e) Example E

(f) Example F

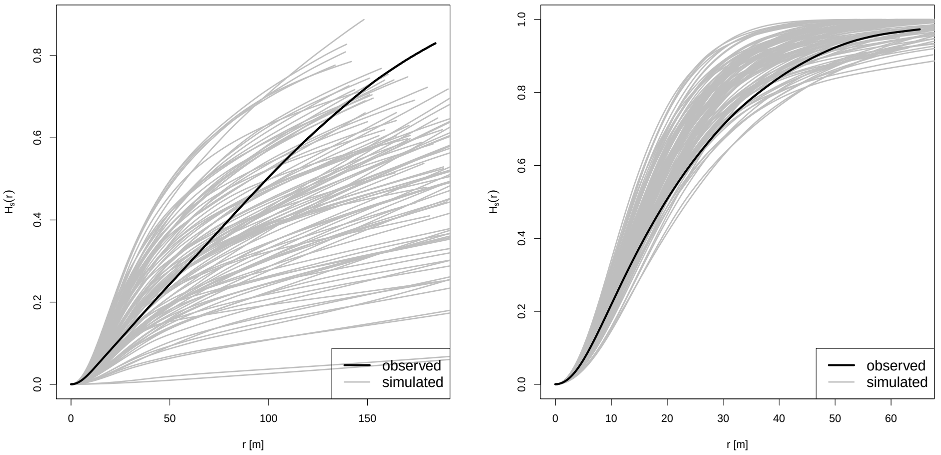
Figure 7.20.: Simulated examples of Thomas processes fitted by using the pair correlation function.



(a) Example B

(b) Example E

Figure 7.21.: Pair correlation function: The solid black lines represent the estimated pair correlation functions of the observed patterns, the grey lines correspond to the estimated pair correlation functions for patterns which were simulated as a Thomas process (fitted by using the pair correlation function), the dashed lines represent the mean of the estimated pair correlation functions of the simulated patterns, the dotted lines give the theoretical pair correlation functions for the fitted model.



(a) Example A

(b) Example D

Figure 7.22.: Empty-space function: The solid black lines represent the estimated empty-space functions of the observed patterns, the grey lines correspond to the estimated empty-space functions for patterns which were simulated as a Thomas process (fitted by using the pair correlation function).

7.2.4. Sensitivity analysis based on Thomas processes

As the fit for the Thomas model was better than for the Matérn model, the inhomogeneous Poisson point process in the bootstrap simulation introduced in Section 6.1 is replaced by a Thomas process. Following the recommendation of Illian et al. (2008), the pair correlation function was chosen as summary characteristic S.

In the bootstrap simulation, full processes are simulated in the first step. The Thomas model, however, was fitted to the thinned pattern, which must be taken into account in the simulation. It seems more natural to assume that a certain fraction of the cluster points of every cluster was thinned than to assume that a certain fraction of clusters was thinned entirely. Therefore, $\hat{\mu}$ was multiplied by $\frac{1}{1-q}$ for simulating the full pattern, whereas $\hat{\kappa}$ remained unchanged.

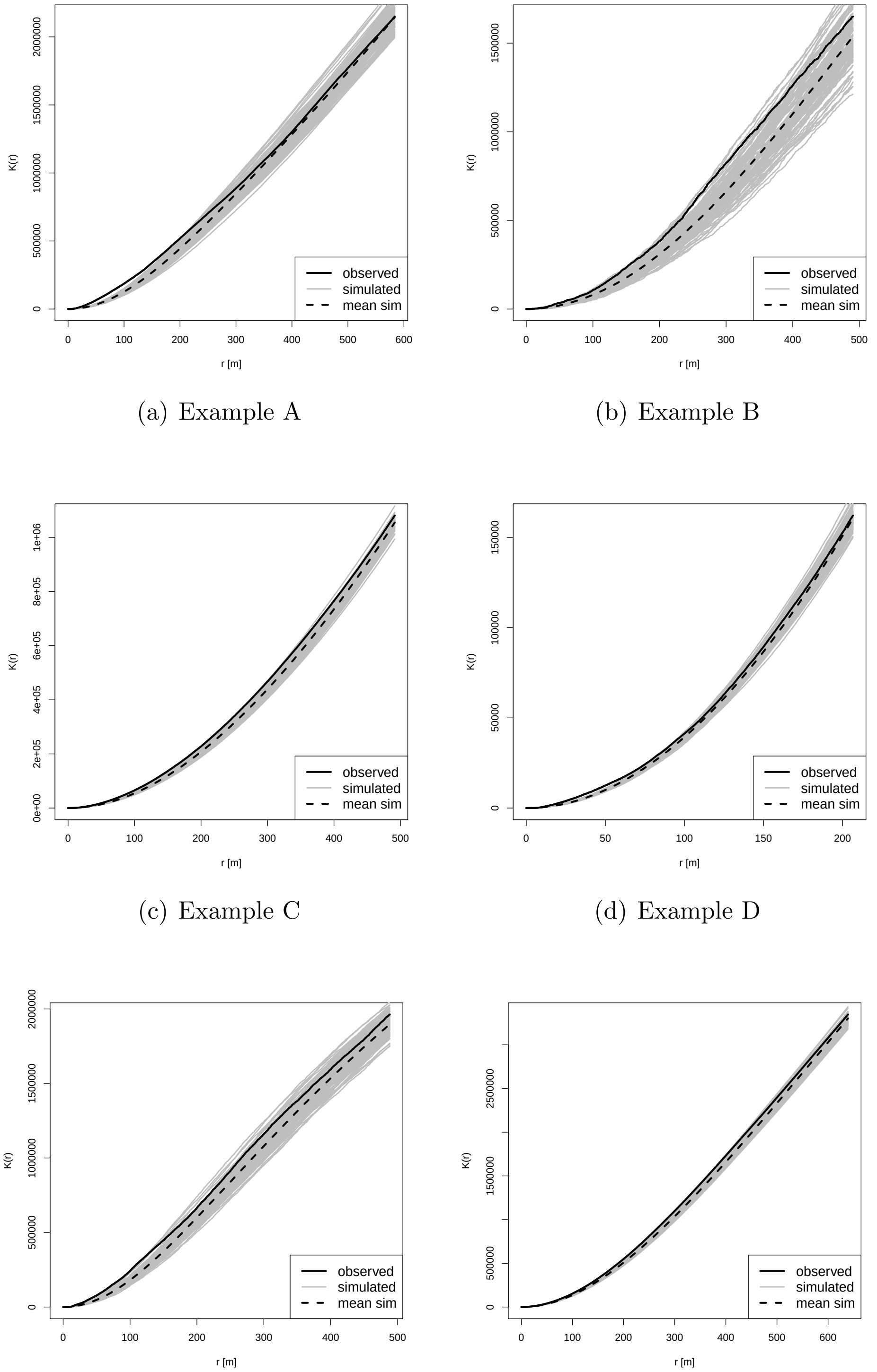
Table 7.3.: Result of the bootstrap simulation based on a Thomas process: Fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside in 1000 iterations; Examples A, B, D and E, intensity-based method (INT)

Example	q	0.1	0.1	0.1	0.15	0.15	0.15
	α	0.4	0.2	0.1	0.4	0.2	0.1
A	p_{out}	0.384	0.312	0.266	0.420	0.356	0.328
B	p_{out}	0.272	0.174	0.125	0.328	0.240	0.203
D	p_{out}	0.464	0.353	0.292	0.505	0.419	0.376
E	p_{out}	0.235	0.156	0.124	0.285	0.216	0.171

As Table 7.3 shows, the fraction p_{out} was smaller than in Section 6.2 for Example B. For Example A, however, this was only the case when $\alpha = 0.4$, but not for smaller values of α. All values of p_{out} for Examples D and E exceeded the values which had been obtained in Section 6.2. While the fraction p_{out} was nonetheless smaller than α for Example E, it was too large for Example D in all cases.

These findings indicate that larger high-risk zones would be obtained if the Thomas process was used instead of an inhomogeneous Poisson point process in the bootstrap correction procedure, i.e. smaller values for `cutoff` would result in most cases.

In order to facilitate the interpretation of this result, summary functions obtained for simulated inhomogeneous Poisson point processes with intensity function $\hat{\lambda}(\mathbf{s})$ are depicted in Figures 7.23, 7.24 and 7.25. We can see that the variation between the simulated patterns is considerably smaller than for Thomas and Matérn processes. The values of Ripley's K-function were often slightly too small. For small r, the pair correlation function of the simulated patterns took smaller values than for the observed patterns. The values of the empty-space function were usually too large. In general, the fit of the inhomogeneous Poisson process models seems better than for the Thomas process.



(a) Example A

(b) Example B

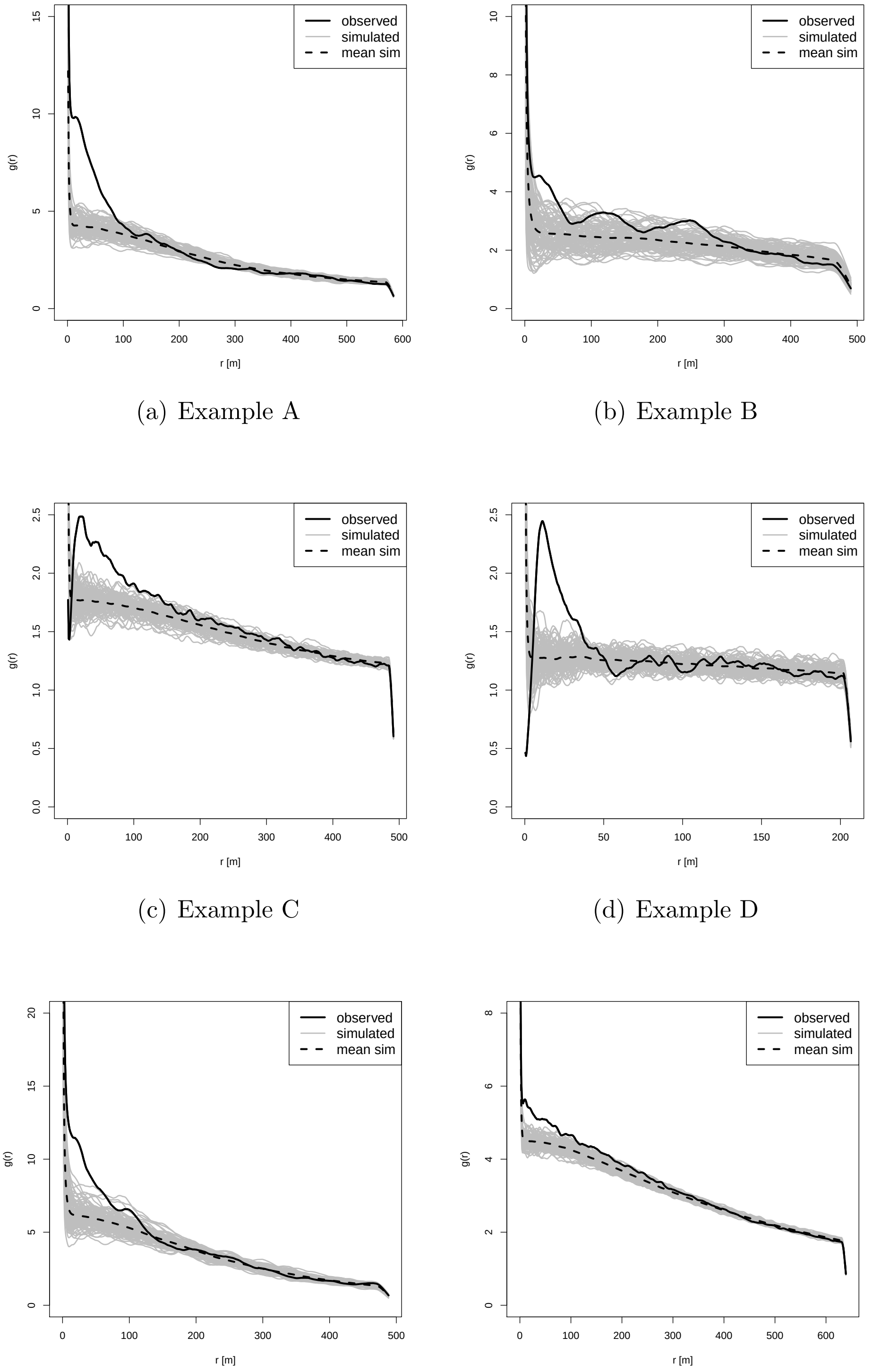
(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 7.23.: Ripley's K-function: The solid black lines represent the estimated K-functions of the observed patterns, the grey lines correspond to the estimated K-functions for patterns which were simulated as inhomogeneous Poisson processes, the dashed lines represent the mean of the estimated K-functions of the simulated patterns.



(a) Example A

(b) Example B

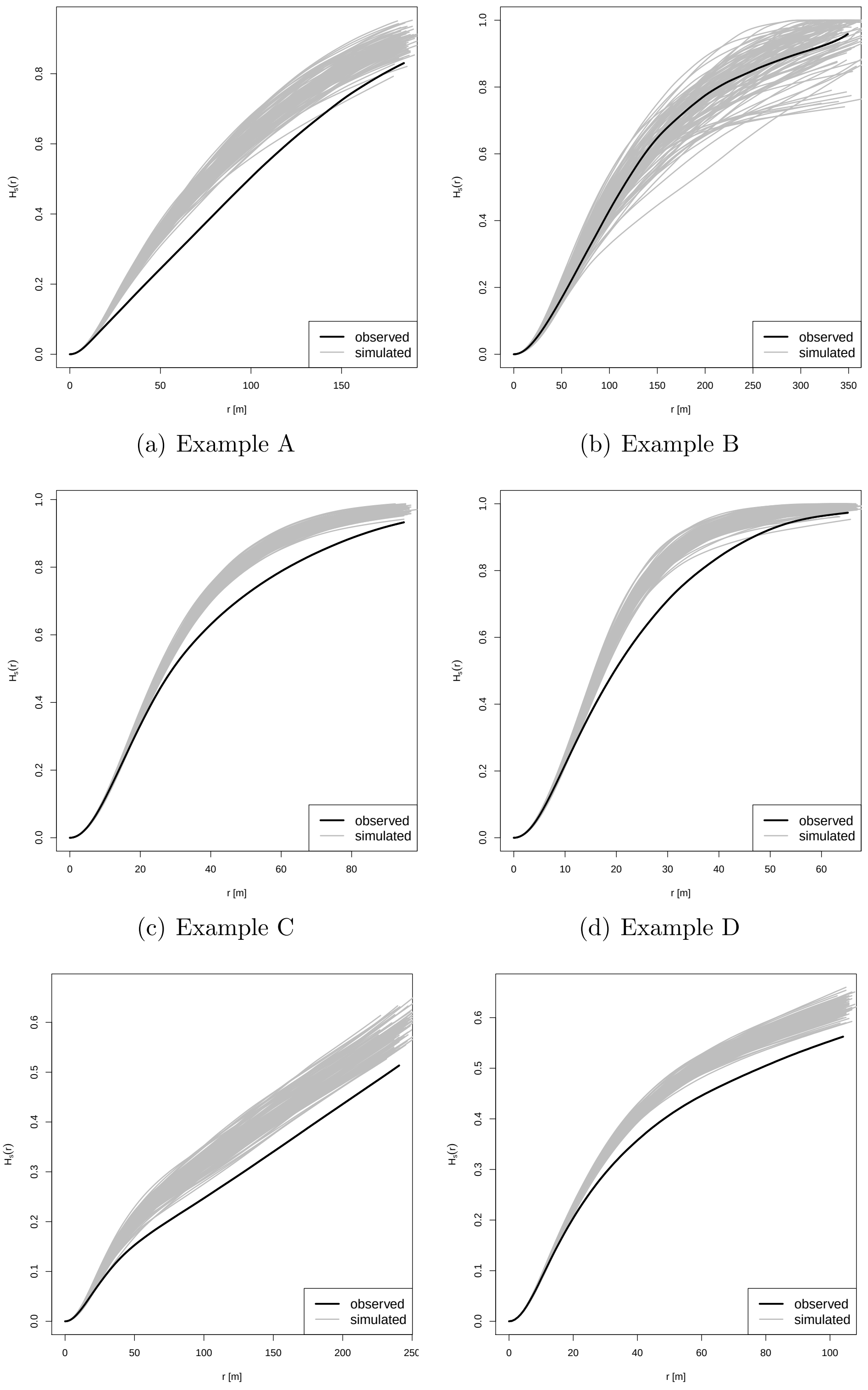
(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 7.24.: Pair correlation function: The solid black lines represent the estimated pair correlation functions of the observed patterns, the grey lines correspond to the estimated pair correlation functions for patterns which were simulated as inhomogeneous Poisson processes, the dashed lines represent the mean of the estimated pair correlation functions of the simulated patterns.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 7.25.: Empty-space function: The solid black lines represent the estimated empty-space functions of the observed patterns, the grey lines correspond to the estimated empty-space functions for patterns which were simulated as inhomogeneous Poisson processes.

7.3. Modelling the intensity of clustered patterns by using a mixture of bivariate normal distributions

7.3.1. Model and estimation of model parameters

An essential disadvantage of the Thomas process from a user's point of view (in particular, from the point of view of OFD Niedersachsen) is that the cluster centres are not fixed, so single realisations of a Thomas process may look very different from the pattern that was observed. Therefore, a simulation approach is introduced in which the cluster centres are fixed. The resulting model will have some similarities with the Thomas model.

As a starting point, a larger class of models–which comprises the Thomas process as a special case–is considered: *Shot noise Cox processes* are Cox processes whose intensity function is a realisation of a random field

$$Z(\mathbf{s}) = \sum_j \gamma_j k(c_j, \mathbf{s}), \tag{7.2}$$

where $k(c_j, \cdot)$ is a kernel (Møller, 2003). When we think of the shot noise Cox process as a cluster process, c_j is a cluster centre and γ_j is the intensity of the respective cluster, i.e. the parameter of the Poisson distribution of the number of cluster points per cluster. For a Thomas process in the classical definition, $k(c_j, \cdot)$ is a Gaussian kernel and all γ_j are equal.

Given the vector $\mathbf{c}$ of all cluster centres and the vector $\boldsymbol{\gamma}$ of all cluster intensity parameters, the conditional intensity of a shot noise Cox process is

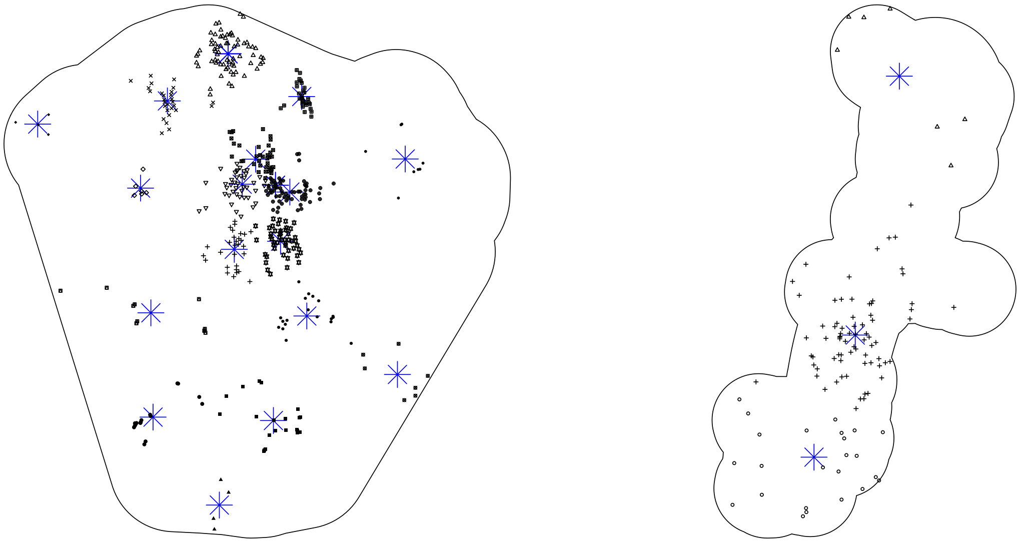
$$\lambda(\mathbf{s}|\mathbf{c}, \boldsymbol{\gamma}) = \sum_j \gamma_j k(c_j, \mathbf{s}). \tag{7.3}$$

This conditional intensity can be modelled by using a *finite mixture model.* A random variable X has a *finite mixture distribution* if its density function takes the form

$$p(x) = \sum_{i=1}^{k} \pi_i f_i(x), \tag{7.4}$$

where the *mixing weights* $\pi_j > 0 \quad \forall j$ and $\sum_{i=1}^{k} \pi_i = 1$ (Titterington et al., 1985). The *component densities* $f_j(\cdot)$ can take parametric forms with parameters θ_j.

Fraley and Raftery (2002) use finite mixture models for modelling clustering. Each component corresponds to a cluster. An implementation for $f_j(\cdot)$ multivariate normal is given in the R package `mclust` (Fraley et al., 2012). The parameters π_j and θ_j are estimated by using the EM algorithm (Dempster et al., 1977) for maximum likelihood estimation. The number of components k is determined by using the Bayesian information criterion BIC (Schwarz, 1978). BIC could also be used for choosing the complexity of the model in a more general way, e.g. to decide if a more flexible or component-specific covariance matrix should be used. However, we restrict the covariance matrix to $\sigma^2 I$ to obtain a model which is close to the Thomas process, i.e. all clusters are spherical and the variance does not vary between clusters.



(a) Example A

(b) Example B

Figure 7.26.: Classification as mixture of Normal distributions: The asterisks illustrate the cluster centres. Different plotting symbols were used for the cluster points of each component.

The resulting classification for Examples A to F is shown in Figures 7.26 and 7.27. For most examples, the number of clusters of the selected model is higher than the expected number of clusters in the Thomas and Matérn models: 18 clusters were obtained for Example A, 3 for Example B, 12 for Example C, 9 clusters for Example D, 8 for E and 6 clusters for F.

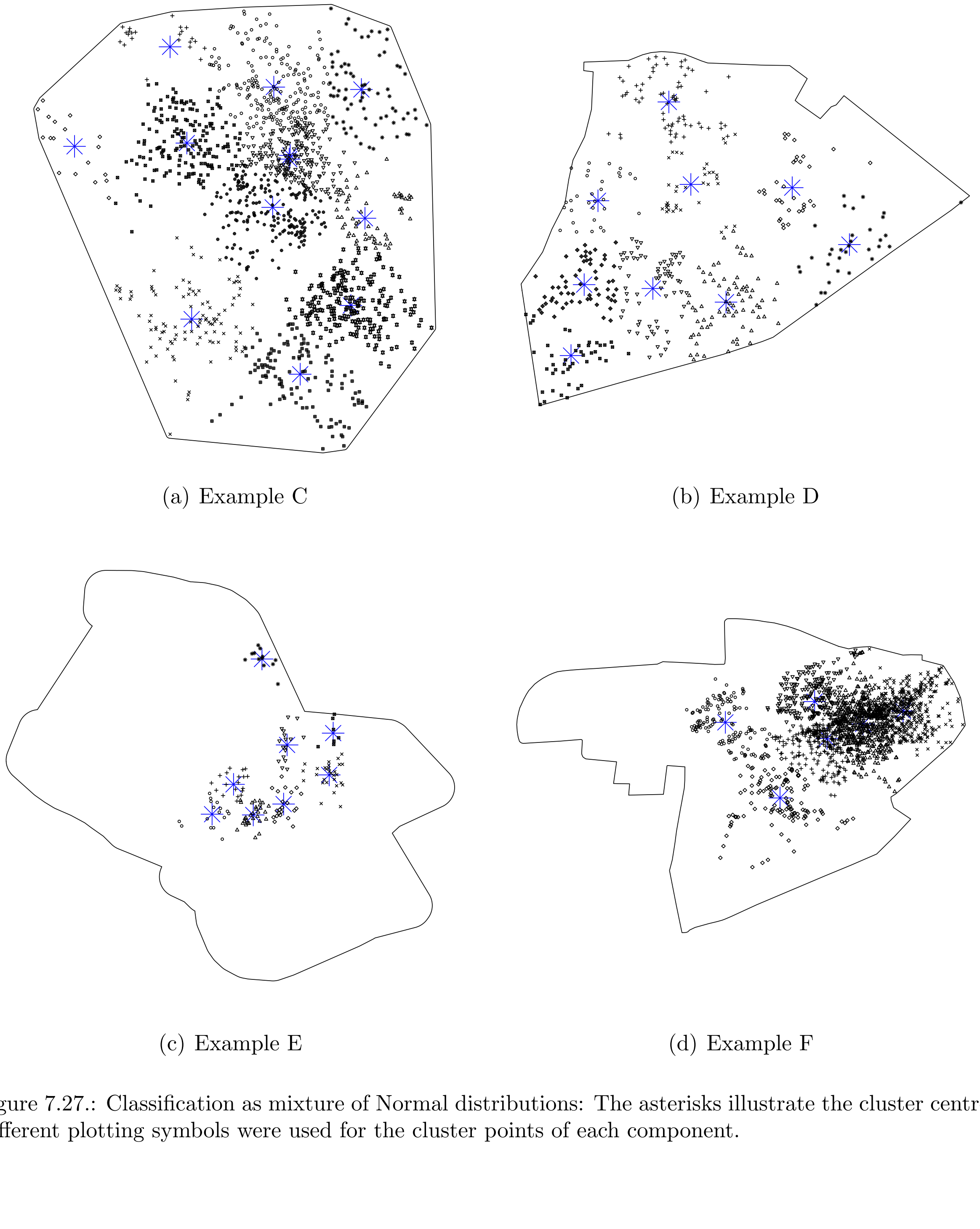
7.3.2. Properties

If we keep in mind that the intensity of a spatial point process and the density are proportional and use a bivariate Gaussian kernel for $k(c_j, \cdot)$ and the density of a bivariate normal distribution for $f_j(\cdot)$ (both of them with covariance $\sigma^2 I$), Equations 7.3 and 7.4 look very similar.

The main differences between the mixture approach and the Thomas model are as follows:

- The observation window is not taken into account in the mixture approach. The observations are assumed to be 'complete'. A Thomas process, on the other hand, is stationary, so the spatial censoring which results from the restriction of observations to the window is taken into account by using edge correction methods when the summary characteristics are estimated.

- In the mixture approach, the optimal model is determined via BIC. Summary characteristics are not considered.



(a) Example C

(b) Example D

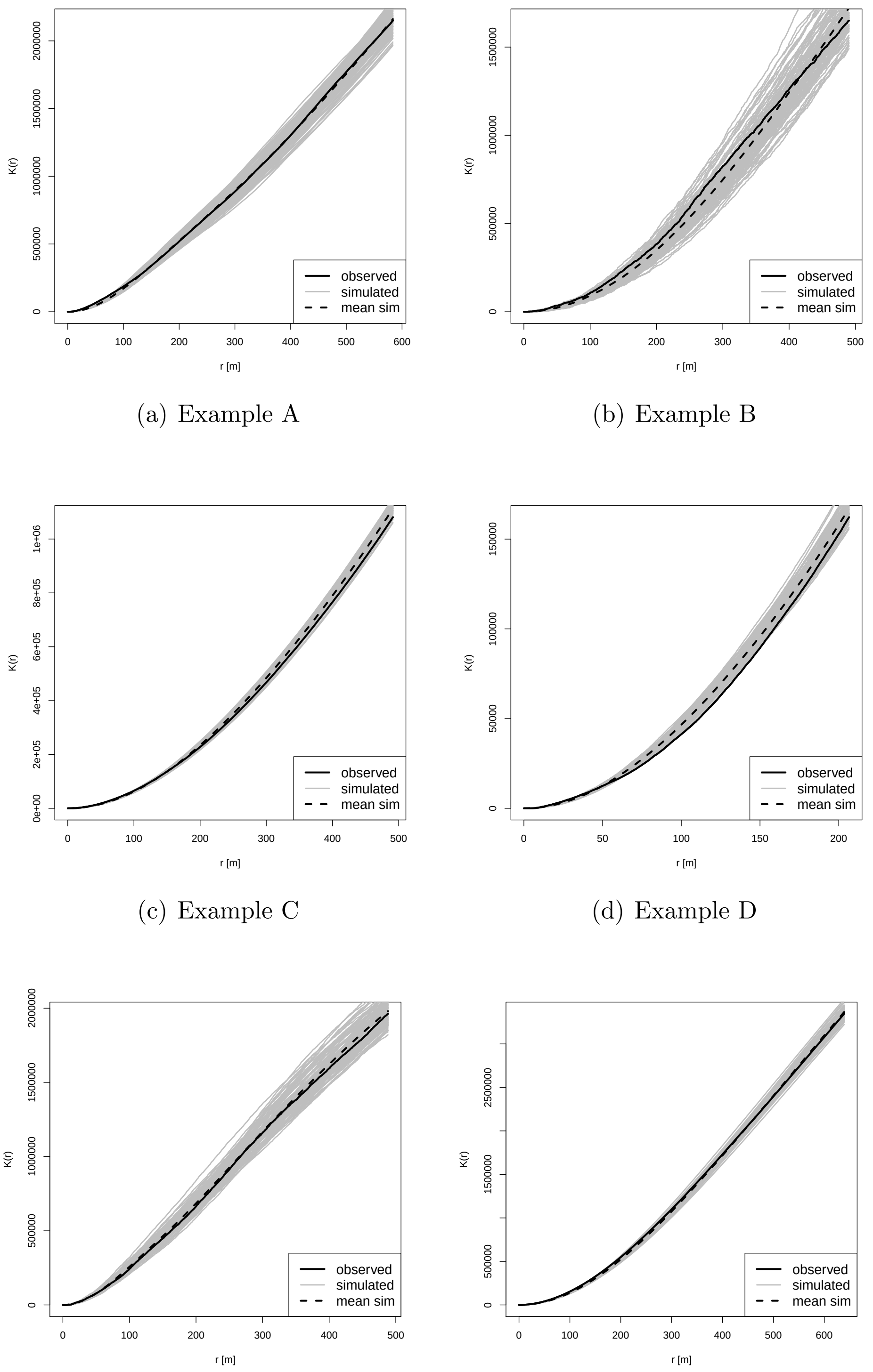
(c) Example E

(d) Example F

Figure 7.27.: Classification as mixture of Normal distributions: The asterisks illustrate the cluster centres. Different plotting symbols were used for the cluster points of each component.

- Compared with the ‘classical’ Thomas model, the mixture approach is more flexible: The distribution of cluster centres does not correspond to a homogeneous Poisson process, but the cluster centres are determined using maximum likelihood estimation. The number of cluster points does not necessarily follow a Poisson distribution. In addition to these aspects, it is possible to allow the shape of clusters to vary, e.g. different covariance matrices can be assigned to each component.

For a further investigation of the properties of the mixture model, 100 patterns were simulated as inhomogeneous Poisson processes whose intensity functions are the mixture intensity. We obtain a slightly better fit with regard to Ripley's K-function than with the Thomas model (Figure 7.28) for models A, E and F. For the pair correlation function, the fit was similar compared to the fit for the Thomas model (Figure 7.29). The empty-space function for the patterns simulated based on the mixture models were much closer to the observed empty-space function than for the Thomas models (Figure 7.30). In general, the variability of the functions was much smaller. In general, the fit was slightly better than for the inhomogeneous Poisson point processes with intensity function $\hat{\lambda}(\mathbf{s})$ estimated by using the kernel method.



(a) Example A

(b) Example B

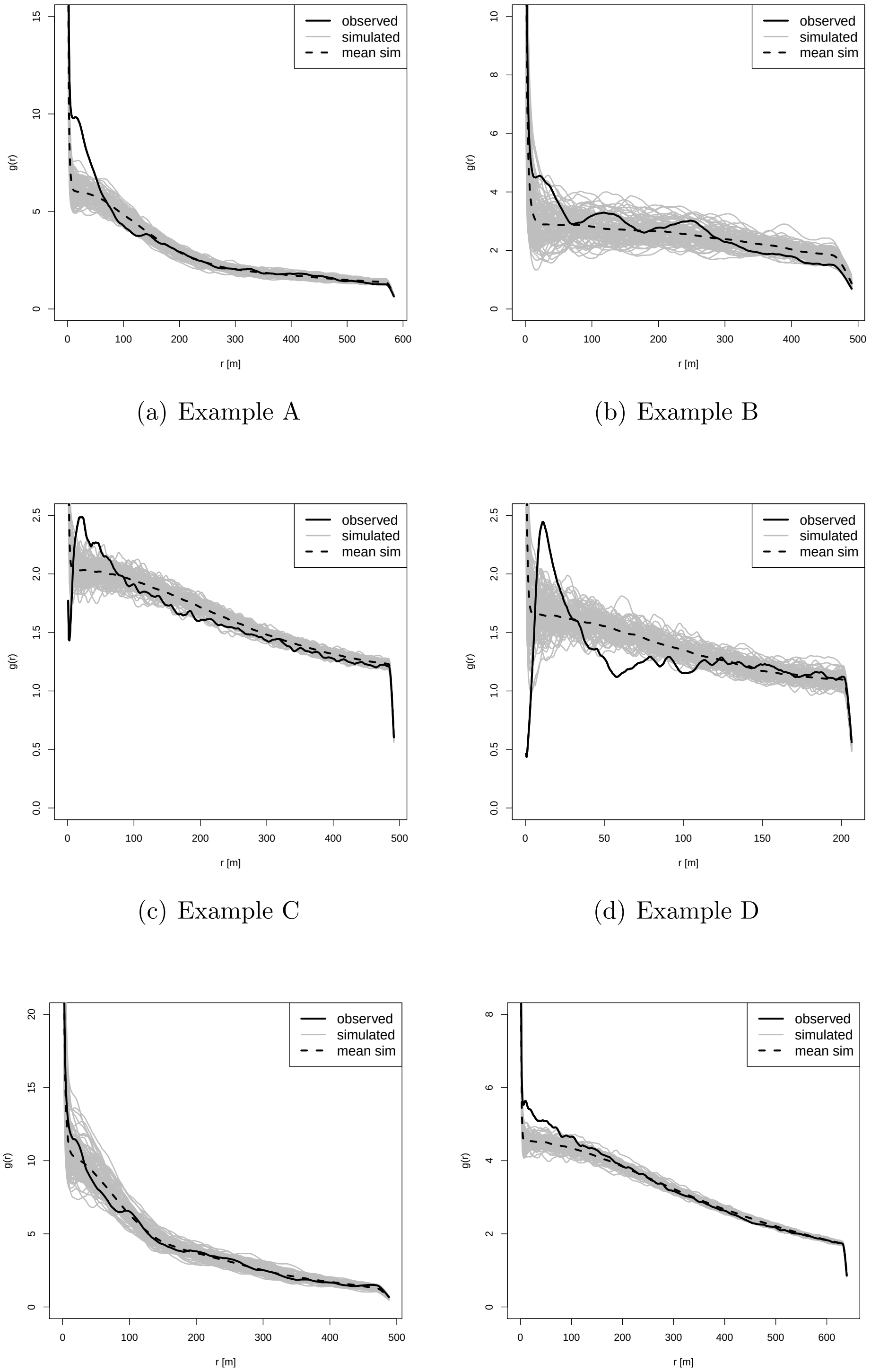
(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 7.28.: Ripley's K-function: The solid black lines represent the estimated K-functions of the observed patterns, the grey lines correspond to the estimated K-functions for patterns which were simulated using a mixture of Normal distributions, the dashed lines represent the mean of the estimated K-functions of the simulated patterns.



(a) Example A

(b) Example B

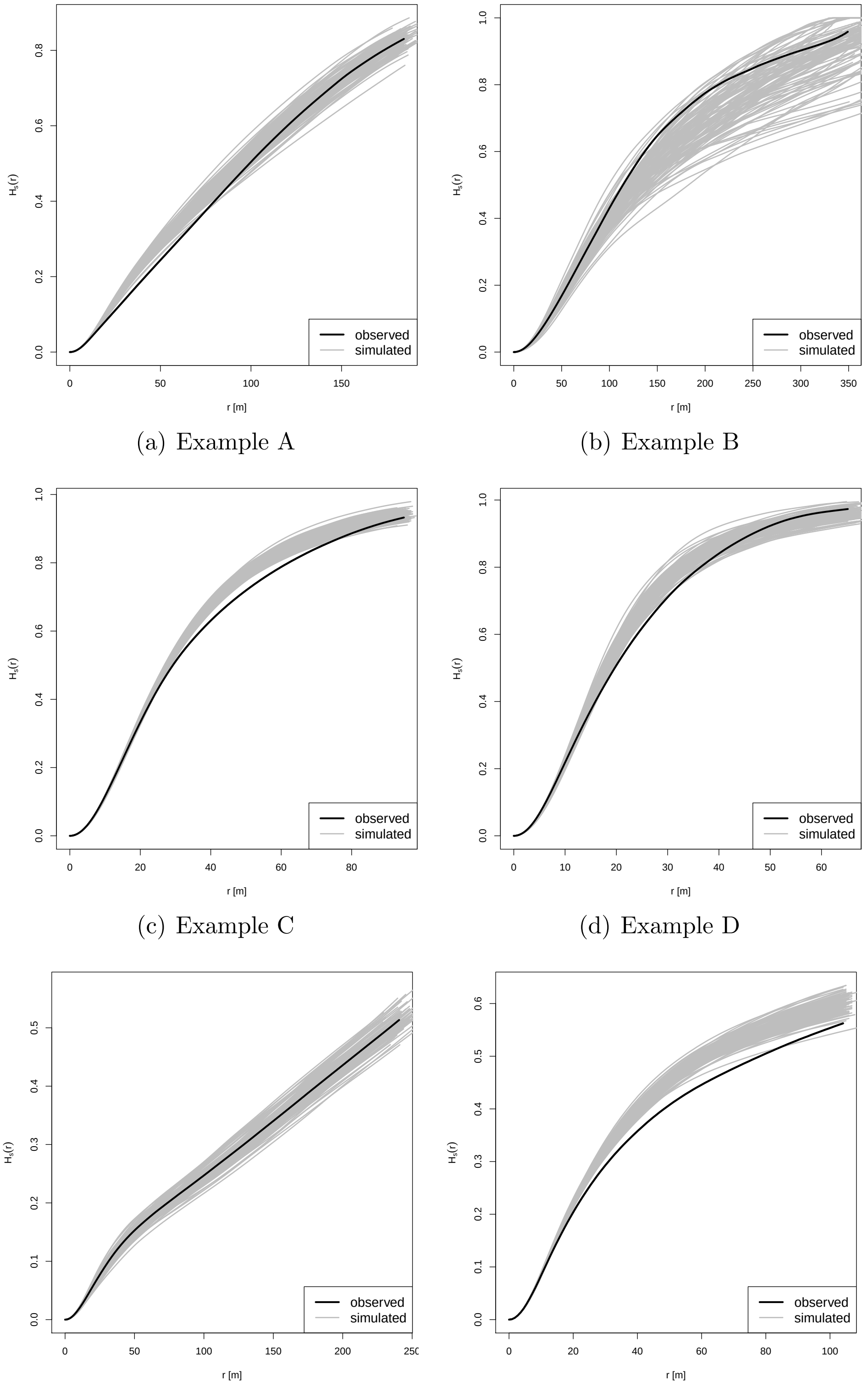
(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 7.29.: Pair correlation function: The solid black lines represent the estimated pair correlation functions of the observed patterns, the grey lines correspond to the estimated pair correlation functions for patterns which were simulated using a mixture of Normal distributions, the dashed lines represent the mean of the estimated pair correlation functions of the simulated patterns.



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure 7.30.: Empty-space function: The solid black lines represent the estimated empty-space functions of the observed patterns, the grey lines correspond to the estimated empty-space functions for patterns which were simulated using a mixture of Normal distributions.

7.3.3. Behaviour of the high-risk zones

The inhomogeneous Poisson point process in the bootstrap simulation introduced in Section 6.1 was replaced by a mixture model: The mixture intensity of the bomb crater patterns was multiplied by $\frac{1}{1-q}$ for simulating the full pattern as an inhomogeneous Poisson process in every iteration.

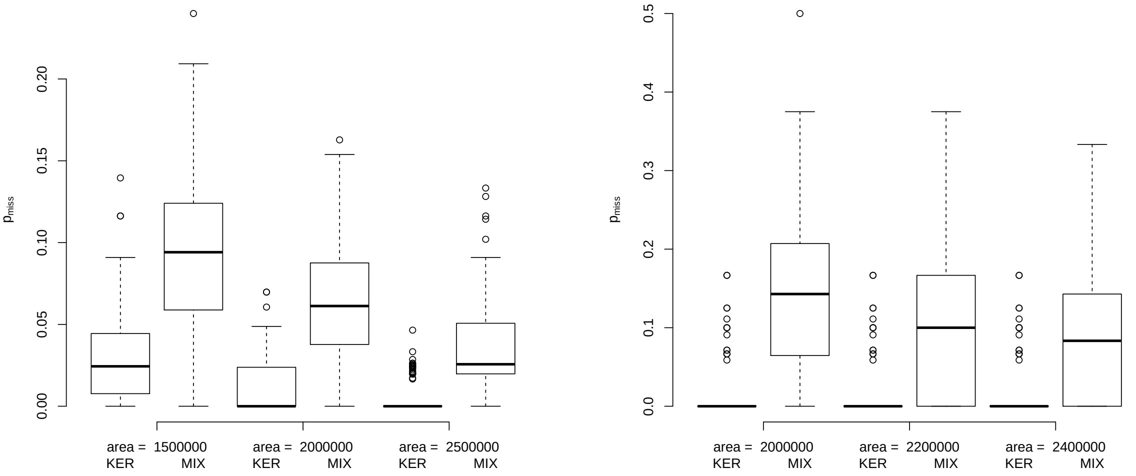
Table 7.4.: Result of the bootstrap simulation based on the mixture model: Fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside in 1000 iterations; Examples A, B, D and E, intensity-based method (INT)

Example	q	0.1	0.1	0.1	0.15	0.15	0.15
	α	0.4	0.2	0.1	0.4	0.2	0.1
A	p_{out}	0.248	0.129	0.077	0.288	0.156	0.102
B	p_{out}	0.393	0.256	0.174	0.444	0.303	0.214
D	p_{out}	0.140	0.057	0.023	0.129	0.049	0.022
E	p_{out}	0.056	0.011	0.004	0.071	0.020	0.010

Table 7.4 shows the behaviour of the high-risk zones. For Examples A, D and E, all fractions p_{out} were smaller than in Section 6.2. For Example B, this was only the case for $\alpha = 0.4$ and for $\alpha = 0.2$ with $q = 0.10$. This means that the high-risk zones would usually be smaller (and α larger) if the mixture model was used for the bootstrap correction.

7.3.4. Constructing high-risk zones based on the mixture intensity

The mixture model yields an estimate for the intensity function. In contrast to the kernel method, no bandwidth needs to be specified. This rises the question if the intensity-based method could be improved by using the mixture intensity instead of the kernel estimator. However, the mixture intensity turns out not to be flexible enough. High-risk zones with given area were determined and evaluated like in Section 5.3. The fractions p_{miss} which are obtained if intensity-based high-risk zones are constructed based on the estimated mixture intensity are much higher than for the kernel method (Figure 7.31). Moreover, it is computationally more expensive to use the mixture intensity. Therefore, this modification cannot be recommended.



(a) Example A

(b) Example B

Figure 7.31.: Fraction of unexploded bombs outside the high-risk zone: Comparison for intensity estimation by using the kernel method (KER) and by using a mixture (MIX) (100 iterations)

8. R package “highriskzone”

The R package highriskzone (Seibold and Mahling, 2012) contains an implementation of all methods for constructing and evaluating high-risk zones which were introduced in Chapters 4 and 5 and Sections 6.1 and 7.1. The package is described in detail in Seibold (2012), where many examples are given. The R package spatstat (Baddeley and Turner, 2005) for the analysis of spatial point patterns served as the main implementational toolbox, e.g. point patterns were represented in its ppp data format and package functionality was used to estimate the intensity and to simulate point patterns. Bandwidth selection as described in Section 4.3.2 was taken from the R package ks (Duong, 2007).

As soon as the package highriskzone has been installed, it can be loaded. It is usually advisable to increase the number of pixels which is used for pixel images in spatstat and therefore determines the precision of intensity estimation and all types of high-risk zones. For more details, please refer to the spatstat documentation.

```
> library("highriskzone")
> spatstat.options(npixel=1000)
```

The main functions of the package highriskzone are det_hrz() to determine a high-risk zone, eval_hrz() to evaluate a single high-risk zone and eval_method() to evaluate a construction method for high-risk zones. The main functionalities are presented in the following sections. For more details and further options, please refer to the documentation of highriskzone.

8.1. Data structure

The spatial point patterns for Examples A and B are included in the package as ppp objects craterA and craterB.

```
> data(craterB)
> craterB

 planar point pattern: 104 points
window: polygonal boundary
enclosing rectangle: [0, 1961.5682] x [0, 3440.013] units

> plot(craterB)
```

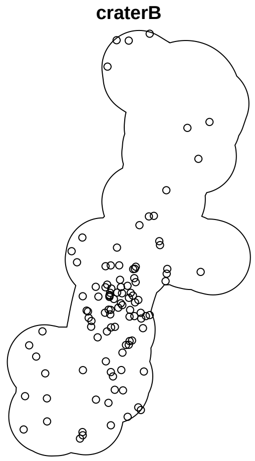


Figure 8.1.: Result of plot(craterB).

Additionally, the function read_pppdata() enables the user to determine high-risk zones for his own data. The coordinates of the observed events and of a polygon describing the border of the observation window can be given as vectors and are converted to a ppp object. Note that the coordinates of the polygon need to be sorted anti-clockwise.

```
> ppxcoord <- c(0, 1, -2, -2, 3, -1, 1, 3)
> ppycoord <- c(4, 0, 1, -3, 5, 2, 3, 1)
> winxcoord <- c(-3, -4.5, -3, 0, 4, 6, 5, 2)
> winycoord <- c(5.5, -2, -4.5, -4, -2, 2, 7, 5)
> simpleexample <- read_pppdata(xppp=ppxcoord, yppp=ppycoord,
+   xwin=winxcoord, ywin=winycoord)
> simpleexample
```

```
 planar point pattern: 8 points
window: polygonal boundary
enclosing rectangle: [-4.5, 6] x [-4.5, 7] units
```

```
> plot(simpleexample)
```

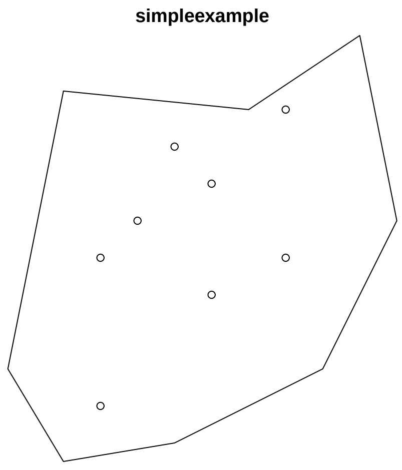


Figure 8.2.: Result of `plot(simpleexample)`.

Two classes of objects introduced by the package `highriskzone` are `highriskzone` and `hrzeval`, which result when the functions `det_hrz()` and `eval_hrz()` are employed. The package `highriskzone` comprises methods to print, plot and summarise such objects. They are presented in the following section.

8.2. Determining high-risk zones

Intensity-based and quantile-based high-risk zones as defined in Chapter 4, as well as high-risk zones based on the traditional method, can be determined by using the function `det_hrz(ppdata, type, criterion, cutoff, nxprob)`. It is also possible to construct a high-risk zone with given area as described in Section 5.3. The arguments `type` and `criterion` define which of these approaches is employed to determine the high-risk zone. Table 8.1 shows which approach corresponds to which combination of arguments. Remember that the traditional and the quantile-based method can be subsumed under 'distance-based methods'. The argument `type` can therefore take the values `"dist"` or `"intens"` and `criterion` can be `"direct"`, `"indirect"` or `"area"`.

The meaning of the argument `cutoff` depends on the values of `type` and `criterion` and is shown in Table 8.2. Note that for `type = "dist"` and `criterion = "direct"` the user does not specify c, the threshold with respect to $\hat{\lambda}_Z(\mathbf{s})$, but a transformation of it, the threshold with respect to $\hat{\lambda}_Y(\mathbf{s})$.

The argument `ppdata` defines the `ppp` object (a spatial point pattern including the observation window) which is used to determine the high-risk zone. Another important

Table 8.1.: Approach which is used to determine a high-risk zone for all possible combinations of arguments `type` and `criterion`

	`type = "dist"`	`type = "intens"`
`criterion = "direct"`	traditional method (Section 4.1)	intensity-based method where threshold c is specified directly (Section 4.3.3)
`criterion = "indirect"`	quantile-based method (Section 4.2)	intensity-based method where global failure probability α is specified (Section 4.3.4)
`criterion = "area"`	distance-based method where area is specified (Section 5.3)	intensity-based method where area is specified (Section 5.3)

Table 8.2.: Meaning of the argument `cutoff` for all possible combinations of arguments `type` and `criterion`

	`type = "dist"`	`type = "intens"`
`criterion = "direct"`	radius r	threshold $c_Y = \frac{1-q}{q} \cdot c$
`criterion = "indirect"`	p to define the quantile	global failure probability α
`criterion = "area"`	area of the high-risk zone	area of the high-risk zone

argument of the function `det_hrz()` is `nxprob`, the probability of non-explosion or, more generally, the probability of non-observation q.

The function `det_hrz()` returns an object of class `highriskzone`. It contains the high-risk zone as an object of class `owin`, the data format which is used for observation windows in `spatstat`, and–depending on the value of `criterion`–additional information such as the resulting threshold (for `criterion = "indirect"` and `criterion = "area"`) and the retrospectively determined values for p and α (for `criterion = "area"`).

A quantile-based high-risk zone with $p = 0.99$ can be determined as follows:

```
> hrz1 <- det_hrz(craterB, type = "dist", criterion = "indirect",
+   cutoff = 0.99)
> hrz1

high-risk zone of type dist
criterion: indirect
cutoff: 0.99

> summary(hrz1)
```

```
high-risk zone of type dist
criterion: indirect
cutoff: 0.99

threshold: 268.6246
area of the high-risk zone: 2801749
```

```
> plot(hrz1, main="High-risk zone", zonecol="lightgrey",
+   win=craterB$window, plotwindow=TRUE, pattern=craterB,
+   plotpattern=TRUE)
```

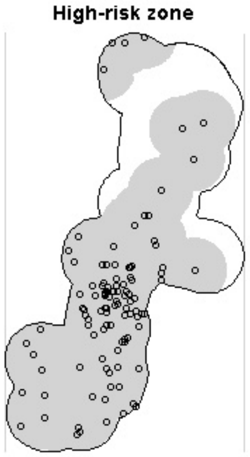


Figure 8.3.: `hrz1`.

An intensity-based high-risk zone with $\alpha = 0.2$ and $q = 0.10$ can be determined as follows:

```
> hrz2 <- det_hrz(craterB, type = "intens", criterion = "indirect",
+   cutoff = 0.2, nxprob = 0.1)
> hrz2

high-risk zone of type intens
criterion: indirect
cutoff: 0.2

> summary(hrz2)
```

```
high-risk zone of type intens
criterion: indirect
cutoff: 0.2

threshold: 5.798054e-06
estimated covariance matrix of Gaussian kernel: 12268.26 7610.874
                                                7610.874 49206.59

area of the high-risk zone: 2474041
```

```
> plot(hrz2, main="High-risk zone", zonecol="lightgrey",
+   win=craterB$window, plotwindow=TRUE, pattern=craterB,
+   plotpattern=TRUE)
```

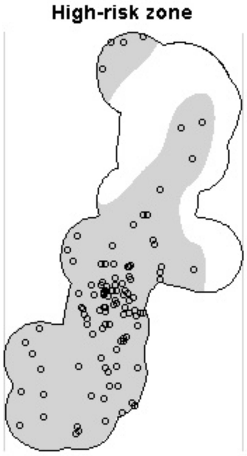


Figure 8.4.: `hrz2`.

A traditional high-risk zone with $r = 150$ can be determined as follows:

```
> hrz3 <- det_hrz(craterB, type = "dist", criterion = "direct",
+   cutoff = 150)
> hrz3

high-risk zone of type dist
criterion: direct
cutoff: 150
```

```
> summary(hrz3)

high-risk zone of type dist
criterion: direct
cutoff: 150

threshold: 150
area of the high-risk zone: 2068173

> plot(hrz3, main="High-risk zone", zonecol="lightgrey",
+   win=craterB$window, plotwindow=TRUE, pattern=craterB,
+   plotpattern=TRUE)
```

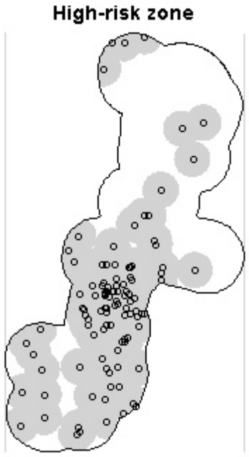


Figure 8.5.: `hrz3`.

8.3. Evaluating a single high-risk zone

The function `eval_hrz(hrz, unobspp, obspp)` is used to evaluate a single high-risk zone. It is invoked by the function `eval_method()`, but can also be used for illustrative purposes, e.g. to depict the situation in one iteration of `eval_hrz()`. Moreover, it can be used to evaluate a high-risk zone if data on the unobserved events are available.

The three arguments are `hrz` (an object of class `owin` representing a high-risk zone), `unobspp` (the point pattern of unobserved events) and `obspp` (the point pattern of observed events). Usually, `unobspp` and `obspp` are simulated from a complete pattern by thinning, for which the function `thin()` can be used.

The function `eval_hrz()` returns an object of class `hrzeval`, which contains (among other things) the number and fraction of unobserved events outside the high-risk zone and the area of the high-risk zone, as well as the subpatterns of unobserved events inside and outside the high-risk zone. The package comprises a method to plot these subpatterns together with the high-risk zone and the pattern of observed events.

```
> thdata <- thin(craterB, nxprob=0.1)
> hrz4 <- det_hrz(thdata$observed, type = "intens",
+   criterion = "indirect", cutoff = 0.6, nxprob = 0.1)
> evaluation <- eval_hrz(hrz = hrz4$zone,
+   unobspp = thdata$unobserved, obspp = thdata$observed)
> evaluation

evaluation of a high-risk zone based on 90 observed events
number of unobserved events: 14
number of unobserved events located outside the high-risk zone: 3

> summary(evaluation)

evaluation of a high-risk zone based on 90 observed events
number of unobserved events: 14
number of unobserved events located outside the high-risk zone: 3

fraction of unobserved events located outside the high-risk zone: 0.2142857
area of the high-risk zone: 1671827
```

```
> plot(evaluation, hrz = hrz4, obspp = thdata$observed,
+   plothrz = TRUE, plotobs = TRUE)
```

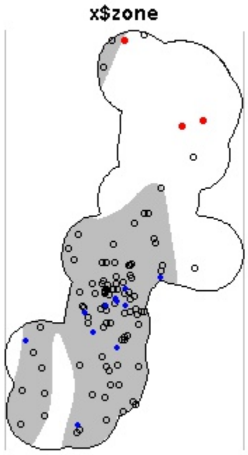


Figure 8.6.: Evaluation of high-risk zone `hrz4`.

The function which plots the evaluation results is flexible, e.g. plotting symbols can be specified by the user.

```
> plot(evaluation, hrz = hrz4, obspp = thdata$observed, plothrz = TRUE,
+   plotobs = TRUE, insidecol = "red", outsidecol = "red",
+   obscol = "blue", insidepch = 19, outsidepch = 4, main = "Evaluation")
> legend(2400, 2456.4061, c("observed", "unobs inside", "unobs outside"),
+   col = c("blue", "red", "red"), yjust=1, pch=c(1, 19, 4), cex=0.8)
```

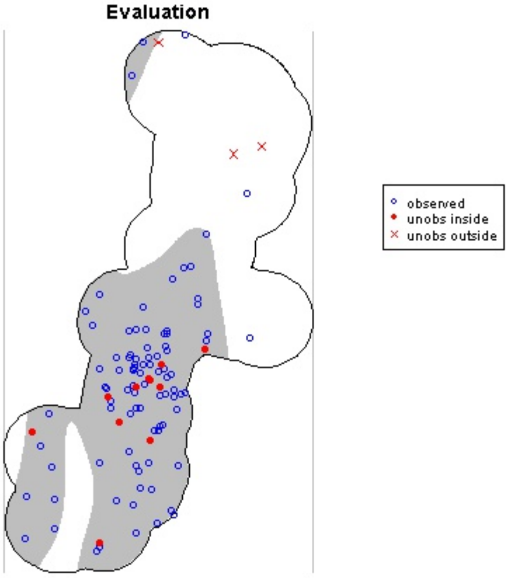


Figure 8.7.: Evaluation of high-risk zone `hrz4`.

8.4. Evaluating a construction method for high-risk zones

The function `eval_method(ppdata, type, criterion, cutoff, nxprob, numit, simulate)` allows to evaluate the behaviour of high-risk zones for one or more construction methods. It invokes the functions `det_hrz()` and `eval_hrz()`. Therefore, the arguments `ppdata`, `type`, `criterion`, `cutoff` and `nxprob` have exactly the same meaning as for `det_hrz()`. However, as high-risk zones constructed by applying different approaches and referring to different values for `cutoff` (but based on the same data and the same probability of non-observation) can be evaluated simultaneously, `type`, `criterion` and `cutoff` can be given as vectors. The argument `numit` defines the number of iterations to be performed.

If the argument `simulate` takes the value `"thinning"`, the evaluation is performed by thinning the observed pattern as described in Section 5.2. For `simulate = "intens"`, the simulation procedure based on the estimated intensity (as described in Section 6.1) is applied. A sensitivity analysis with regard to clustering (Section 7.1) can be performed by setting `simulate = "clintens"`. In this case, the additional parameter τ is given by the argument `clustering` and `radiusClust` is the radius of the simulated clusters.

The function `eval_method()` returns a data frame which contains the number and fraction of unobserved events outside the high-risk zone and the area of the high-risk zone for every iteration, as well as additional information such as the resulting threshold (for `criterion = "indirect"` and `criterion = "area"`) and the retrospectively determined values for p and α (for `criterion = "area"`).

```
> set.seed(321)
> evalm <- eval_method(craterB, type = c("dist", "intens"),
+   criterion = c("area", "area"), cutoff = c(2500000, 2500000),
+   nxprob = 0.1, numit = 1000, simulate = "thinning", pbar = FALSE)

> head(evalm)

  Iteration   Type Criterion  Cutoff nxprob    threshold
1         1   dist      area 2500000    0.1 2.417165e+02
2         1 intens      area 2500000    0.1 3.727658e-06
3         2   dist      area 2500000    0.1 2.230912e+02
4         2 intens      area 2500000    0.1 4.627532e-06
5         3   dist      area 2500000    0.1 2.103224e+02
6         3 intens      area 2500000    0.1 5.873129e-06
  calccutoff covmatrix11 covmatrix12 covmatrix21
1  0.9650292    13816.89    8583.149    8583.149
2  0.1104404    13816.89    8583.149    8583.149
3  0.9400810    12543.24    8108.112    8108.112
4  0.1488846    12543.24    8108.112    8108.112
5  0.9528919    14112.04   11188.562   11188.562
6  0.2182055    14112.04   11188.562   11188.562
  covmatrix22 numbermiss numberunobserved missingfrac
1    53190.34          1               13  0.07692308
2    53190.34          1               13  0.07692308
3    53945.32          0                8  0.00000000
4    53945.32          1                8  0.12500000
5    64350.06          0               13  0.00000000
6    64350.06          0               13  0.00000000
  arearegion numberobserved
1    2500000             91
2    2499987             91
3    2500000             96
4    2500007             96
5    2500000             91
6    2499987             91
```

9. Discussion

9.1. Summary

This thesis is concerned with the construction of high-risk zones for incompletely observed spatial point processes. The starting point is the application of unexploded bombs. To construct such high-risk zones, bomb crater patterns derived from aerial pictures are considered as realisations of spatial point processes. High-risk zones comprising unexploded bombs with high probability should be as small as possible. In most cases, the probability of non-explosion is assumed to be constant (Chapter 1).

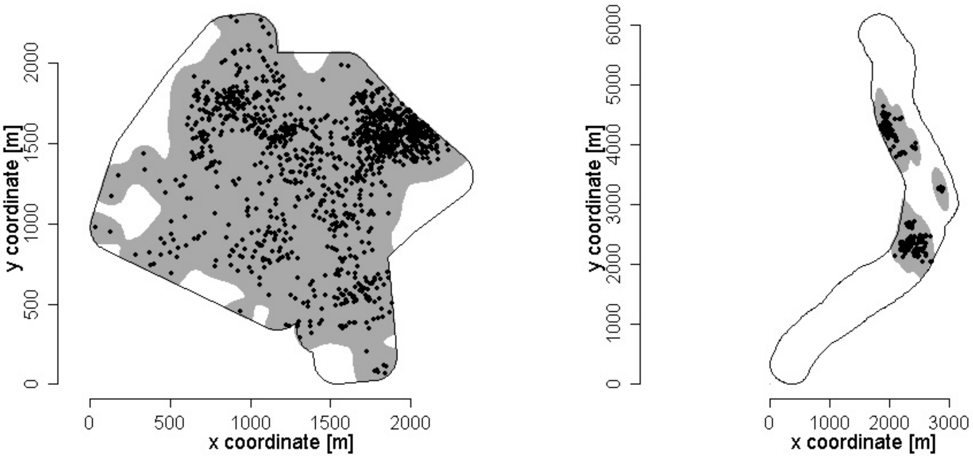
Six examples of bomb crater point pattern data were investigated using functional summary characteristics. Edge correction methods were applied. All patterns tend to be clustered, which may possibly be due to inhomogeneity (Chapter 3).

In addition to the traditional method, which is based on discs of a fixed radius centered at the events, a quantile-based construction method for high-risk zones, which is a development of the traditional method, is presented. A quantile of the nearest-neighbour distance is used instead of a fixed radius. For the six examples, it makes hardly any difference if the observed nearest-neighbour distances are considered directly or if edge correction is applied (Chapter 4).

The intensity-based method is introduced as a new construction method for high-risk zones. An intensity-based high-risk zone consists of those locations for which the intensity attains or exceeds a threshold. The threshold can be specified directly or with respect to the failure probability, which expresses the global risk that not all unobserved events are covered by the high-risk zone. In the latter case, it is necessary to assume that the underlying point process model is the inhomogeneous Poisson process. The intensity function is estimated by using a bivariate Gaussian kernel with unconstrained covariance matrix which is determined via smooth cross-validation. Edge correction needs to be applied (Chapter 4).

To investigate the applicability of the intensity-based construction method, high-risk zones were determined for a variety of bomb crater patterns which partly represent very complex or particular situations. Examples of such high-risk zones have been shown in Section 4.3, further high-risk zones can be found in Chapters A and D in the Appendix. High-risk zones for two of the properties which have not been considered in this thesis so far are depicted in Figure 9.1 to illustrate two further possible constellations.

The intensity-based high-risk zones clearly differ from the high-risk zones which have been determined up to now by using the traditional method. According to Oberfinanzdirektion Niedersachsen, they have a plausible shape. As mentioned in Mahling et al. (2013), single observations which are not covered by the high-risk zone (as for Examples B and D



(a) Example I

(b) Example J

Figure 9.1.: High-risk zone obtained with the intensity-based construction method for $q = 0.10$ and $\alpha = 0.4$.

in Figure 4.7) are not regarded as problematic. In general, intensity-based high-risk zones turned out to be less ragged than quantile-based high-risk zones. This is an advantage of the method, as the on-site determination of the border of the high-risk zones is laborious. Insular areas inside a high-risk zone would only be excluded from the high-risk zone if they are big enough to justify this additional work.

In Chapter 5, a Monte Carlo test based on the K-function showed no evidence against the assumption of an inhomogeneous Poisson process. However, the results based on the pair correlation function were different, as the estimated pair correlation function of Example A exceeds the envelope for small arguments. For Examples C, D and F, the estimated pair correlation functions are close to the envelope for small arguments of the pair correlation function. In summary, it seems justified to apply the intensity-based method, but clustering cannot be ruled out. A simulation study where the observed patterns were taken as full patterns revealed that the failure probability is usually not kept to very well. The traditional method cannot be recommended as almost no information contained in the data is used and the choice of the radius by an expert remains slightly arbitrary. The performance of intensity-based and quantile-based high-risk zones was compared in another simulation study where the area of the high-risk zones was fixed. It was comparably good for both methods. However, the quantile-based high-risk zones tend to be more ragged. Moreover, the theoretical properties are not convincing: Risk is fixed for each unobserved event separately, but not globally. The global risk can only be influenced indirectly, but cannot be chosen arbitrarily small. The shape of the high-risk zones is predetermined by the use of discs, which means that quantile-based high-risk zones are less flexible and, in particular, cannot take into account anisotropy. Moreover, the probability of non-observation is not taken into account when a quantile-based high-risk zone is determined.

A bootstrap simulation procedure based on the estimated intensity function was proposed to assess the risk which is associated with a high-risk zone (Chapter 6). It is necessary to assume that the patterns are realisations of inhomogeneous Poisson processes. As the global failure probability is not kept to, a bootstrap correction was introduced to find the necessary parameter values to obtain a high-risk zone which is associated with the desired risk.

As spatial clustering cannot be ruled out, its consequences on the bootstrap simulation procedure are investigated in a sensitivity analysis in Chapter 7. In most cases, the fraction of high-risk zones which did not cover all unexploded bombs was smaller than for the inhomogeneous Poisson model. Fitting classical cluster models to the observed patterns via the minimum contrast method did not yield convincing results, especially when Matérn processes were used as point process model. The results for Thomas processes were better, even if the empty-space functions showed considerable differences between the observed and simulated patterns. In a sensitivity analysis based on Thomas processes, the fraction of high-risk zones which did not cover all unobserved events was larger than for the original bootstrap simulation in most cases. Compared to the Thomas processes, the fit to the observed patterns was slightly better if mixtures of bivariate normal distributions were used to model the intensity. In most cases, the fraction of high-risk zones which did not cover all unobserved events was smaller than for the original bootstrap simulation. However, using the mixture intensity to construct high-risk zones cannot be recommended. The advantage compared to the kernel method is that no bandwidth or covariance matrix needs to be determined, but the mixture intensity is not flexible enough to be the basis of the high-risk zone and its determination is computationally more expensive than kernel methods.

The main methods presented in this thesis are implemented in an R package called `highriskzone`. It contains functionalities to determine and evaluate high-risk zones. A brief introduction is given in Chapter 8.

9.2. Outlook

As demonstrated in Section 6.2, the estimation of the intensity function is a crucial issue for the success of the intensity-based method. Other approaches to estimate the intensity, such as the adaptive estimator in the R package `sparr` (Davies et al., 2011), which also corrects for edge effect bias, or the estimator proposed by Bernardeau and van de Weygaert (1996), which is based on the Voronoi tessellation, might help to improve the intensity-based method in general.

To account for the clustered structure of the bomb crater data, more complicated models could be used. A major restriction in the choice of models was that even a user with little statistical knowledge should be able to fit them in an automated procedure and that no covariates are available. A solution to the latter aspect could be to use constructed covariates as proposed by Illian and Rue (2010) and Illian et al. (2012). They fit log-Gaussian Cox processes and perform a Bayesian analysis by using the integrated nested Laplace approximation (Rue et al., 2009).

As mentioned in Section 4.1, high-risk zones can be represented as random sets (for the theory of random sets, see Molchanov, 2005). An investigation of the properties of high-risk zones as random sets could inspire further developments of the construction methods.

The methods for constructing high-risk zones which were presented in this thesis cannot be used for regular patterns. Suitable methodology needs to be developed. However, examples in the literature–as an example on fallen trees in Illian et al. (2008, page 257), where the k-neighbour graph (see Section 5.5) is applied–indicate that it might be possible to find out that and where observations of a regular process are missing without determining high-risk zones as such.

A related problem is the determination of high-risk zones for point processes on linear networks (Ang et al., 2012). An application example is the assessment of deer-vehicle collision risk, where the aim is not to find out where unknown collisions have taken place in the past, but where they are likely to happen in the future. This could help to take measures to prevent collisions at the places where these measures are most useful. Hothorn et al. (2012) analysed data on more than 74000 deer-vehicle collisions in Bavaria by aggregating the data for municipalities. However, as exact coordinates of the collisions are available, point process models can be applied (Hornung, 2011). As the collisions are in fact located on roads, these could be regarded as a linear network, which would allow an analysis which does not refer to entire municipalities or the geographical position in general, but to the roads. For measures such as fencing and green bridges, the roads are the entity which is of importance, so the linear network approach might yield useful new insights.

With regard to high-risk zones for unexploded bombs, the main achievements up to now are that existing methods and the new intensity-based method have been evaluated and that it is finally possible to perform a risk assessment. Neither the quantile-based nor the intensity-based construction method have been applied in practice so far, but OFD Niedersachsen and Mull und Partner are currently working on their adoption. The next step which needs to be taken is the implementation of a plug-in for a Geographical Information System (GIS) such as Quantum GIS (Quantum GIS Development Team, 2013) to make all functionalities implemented in the R package `highriskzone` available for analysts of aerial pictures.

While the use of spatial point process methodology is a novel development in the field of high-risk zones for unexploded bombs, it has a long tradition for other fields of application, such as forestry, epidemiology and ecology. The methods presented in this thesis may be of use in these fields, as well. The R package `highriskzone` facilitates application and further development of the methods for constructing and evaluating high-risk zones.

Appendix

Some further issues which arose in practical application are discussed. This comprises extensions such as the development of high-risk zones for incompletely observed bomb crater patterns or for patterns with a spatially varying probability of non-explosion. Moreover, the definition of guard regions, the construction of separate high-risk zones for subpatterns and the consequences of outliers are considered.

A. Incomplete bomb crater patterns

In some of the data examples provided by OFD Niedersachsen, the observation window comprised water areas (such as rivers or lakes) or forest areas. These areas are referred to as "restriction areas", as–like in cities–not all bomb craters in these properties can be derived from the aerial pictures, so the bomb crater pattern is incomplete. This should be taken into account when a high-risk zone is determined. In this chapter, possible scenarios are introduced. The consequences of incompletely observed patterns Y are investigated for two simulated examples. Finally, high-risk zones accounting for the restriction areas are determined for two real-data examples.

To keep appellations as clear and simple as possible, the process Y will be identified with bomb craters and Z with unexploded bombs, although the methods are not restricted to this special case. For applications in other fields, however, the following considerations are only relevant if the processes Y and Z are in some way of a different type. Otherwise (i.e. if the only difference between Y and Z is that Y is observed and Z is unobserved) it does not make sense to say that the observations of Y are incomplete. A more useful approach in this situation might be to assume a spatially varying probability of observation as in Section E.

A.1. Scenarios

If no bomb crater within the restriction area can be derived from the aerial pictures, this can be represented by an observation probability function which takes the value 0 inside the restriction area and 1 outside. In this situation, the restriction areas can be taken into account by modifying the observation window. These modified observation windows contain holes where no observations of Y could be made.

Another possible scenario, which is mainly relevant for forest areas, is a constant positive observation probability (e.g. 30 %) inside the restriction area. In this situation, a modified observation window is a good option for correction if the observation probability is low. All observations within the restriction area are ignored. For a high observation probability, in contrast, it is desirable to use the observations from the restriction area. This can be achieved by weighting all observations with their reciprocal observation probability when the intensity is estimated. Of course, the observation probability does not have to be constant inside the restriction area. It may vary continuously, for example one might divide it by 2 for every unit of distance from the border of the restriction area. Such an observation probability might be realistic for water areas where the water becomes continuously deeper. Another example of a continuously varying observation probability is a forest area in which the density of the trees varies continuously.

Alternatively, the observation probability can vary discretely. All these scenarios can be combined and a very flexible function for the observation probability can be assumed. In practice, however, there is no data concerning the observation probability and one has to resort to expert knowledge and experience, which is both rough. For this reason, only the two simplest scenarios (observation probability of 0 % inside the restriction area and constant positive observation probability inside the restriction area) will be discussed further.

A.2. Consequences of incompletely observed bomb crater patterns for simulated examples

To investigate the consequences of incompletely observed bomb crater patterns, the estimated intensity functions and the resulting high-risk zones were considered on the basis of two simulated examples for complete patterns as well as for observation probabilities of 75 %, 50 %, 25 % and 0 %. Note that besides *observation probability*, the term *extent of restriction* is used, which denotes 1− observation probability.

The intensity function was estimated without correction, weighted (where the weights are obtained as reciprocal observation probability) and using a modified observation window, where the restriction areas are integrated as holes. The estimated intensity functions were then used to determine high-risk zones. For each setting, 1000 iterations with $\alpha = 0.4$ and $q = 0.10$ were performed. The full patterns of bomb craters and unexploded bombs were simulated as inhomogeneous Poisson point processes with the specified intensity function in every iteration. Weighted intensity estimation was used for observation probabilities of 75 %, 50 % and 25 %. For complete patterns, the results would be exactly like without correction, whereas for an observation probability of 0 %, weighting is not possible. The only correction which is possible in this situation is the modification of the observation window. Note that this approach was only used for an observation probability of 0 %, as the results for positive observation probability would be exactly the same. Holes were always fully integrated into the high-risk zones.

The intensity function which was used for the first simulated example is depicted in Figure A.1. The intensity is very high in the centre and decreases quickly towards the east and west. As Table A.1 shows, $\alpha = 0.4$ is kept to quite well if the full bomb crater pattern is used. The fraction p_{out} and the mean of p_{miss} increase considerably with the extent of restriction if no correction is applied. The mean area of the high-risk zones decreases.

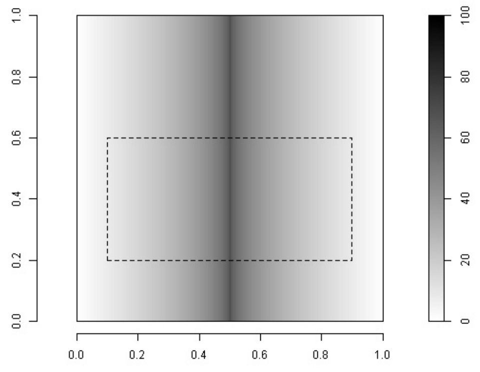


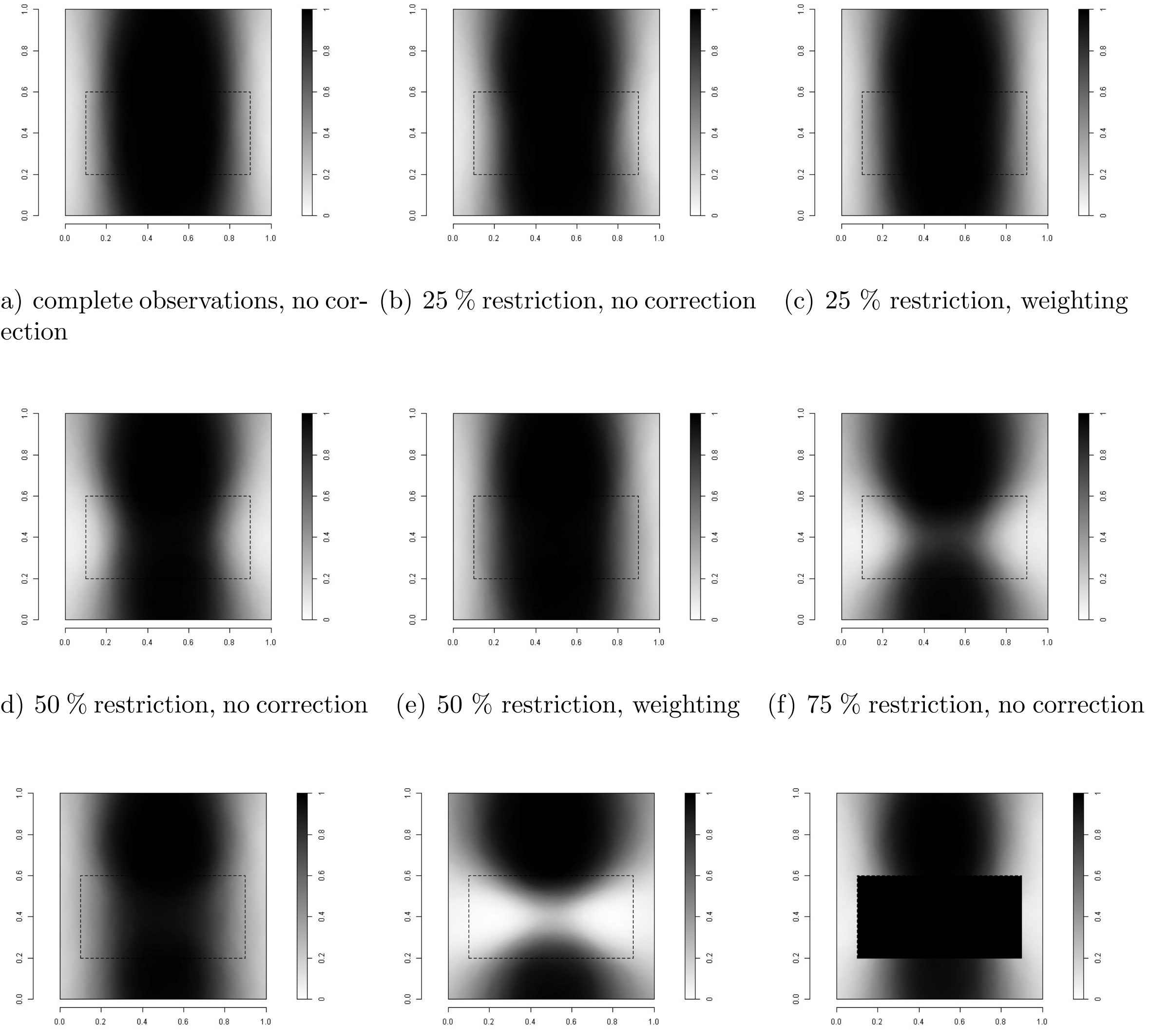
Figure A.1.: Intensity function used for the first simulated example and border of the restriction area (dashed line).

Table A.1.: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside and mean area of the zone; first simulated example, intensity-based method

WITHOUT CORRECTION	extent of restriction	0 %	25 %	50 %	75 %	100 %
	mean p_{miss}	0.1060	0.1218	0.1473	0.1964	0.3243
	p_{out}	0.384	0.423	0.491	0.582	0.765
	mean area in m^2	0.7040	0.6875	0.6640	0.6261	0.5516
WITH CORRECTION	extent of restriction	0 %	25 %	50 %	75 %	100 %
	mean p_{miss}		0.1085	0.1153	0.1403	0.1142
	p_{out}		0.390	0.400	0.463	0.411
	mean area in m^2		0.7061	0.7055	0.6914	0.7115

If we correct for the restrictions, the mean area of the high-risk zones is relatively constant. The fraction p_{out} and the mean of p_{miss} increase only slightly. For a restriction of 75 %, it would be better to use a modified window instead of performing a weighted intensity estimation.

Figures A.2 and A.3 show the shape of the high-risk zones (more precisely, the relative frequency that every pixel was part of the high-risk zones in 1000 iterations) and the mean estimated intensity functions from 1000 iterations. Without correction, the high-risk zones become slimmer in the restriction area. This can be avoided by weighting, at least for 25 % and 50 % restriction. If the modified window is used, we can see that the high-risk zones become slimmer near the border of the hole. This means that the edge correction for the hole cannot fully compensate the lack of observations in the restrictions area. However, this phenomenon is not visible for the mean estimated intensity itself in Figure A.3(i). Apart from this, the results for the estimated intensity functions are very similar to those for the high-risk zones.

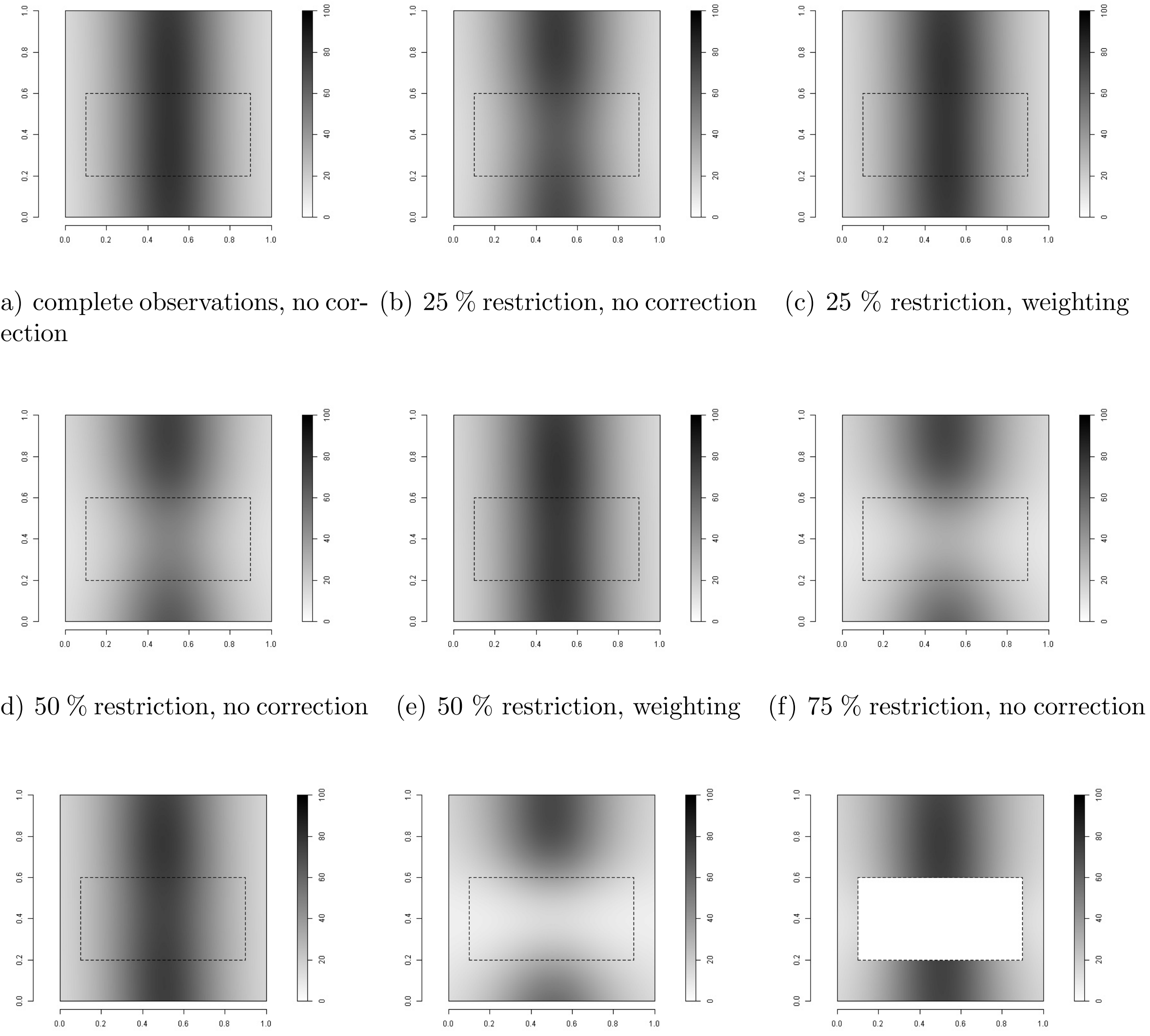


(a) complete observations, no correction (b) 25 % restriction, no correction (c) 25 % restriction, weighting

(d) 50 % restriction, no correction (e) 50 % restriction, weighting (f) 75 % restriction, no correction

(g) 75 % restriction, weighting (h) 100 % restriction, no correction (i) 100 % restriction, modified window

Figure A.2.: Resulting high-risk zones (first simulated example): Relative frequency that every pixel was part of the high-risk zones in 1000 iterations.



(a) complete observations, no correction

(b) 25 % restriction, no correction

(c) 25 % restriction, weighting

(d) 50 % restriction, no correction

(e) 50 % restriction, weighting

(f) 75 % restriction, no correction

(g) 75 % restriction, weighting

(h) 100 % restriction, no correction

(i) 100 % restriction, modified window

Figure A.3.: Mean estimated intensity functions from 1000 iterations (first simulated example).

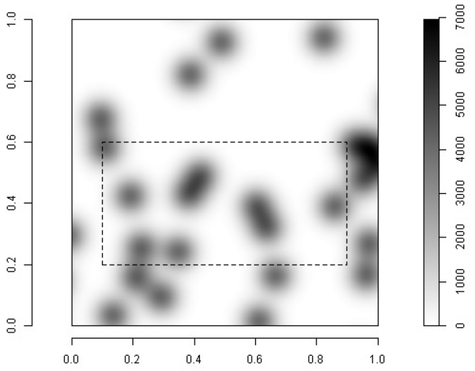


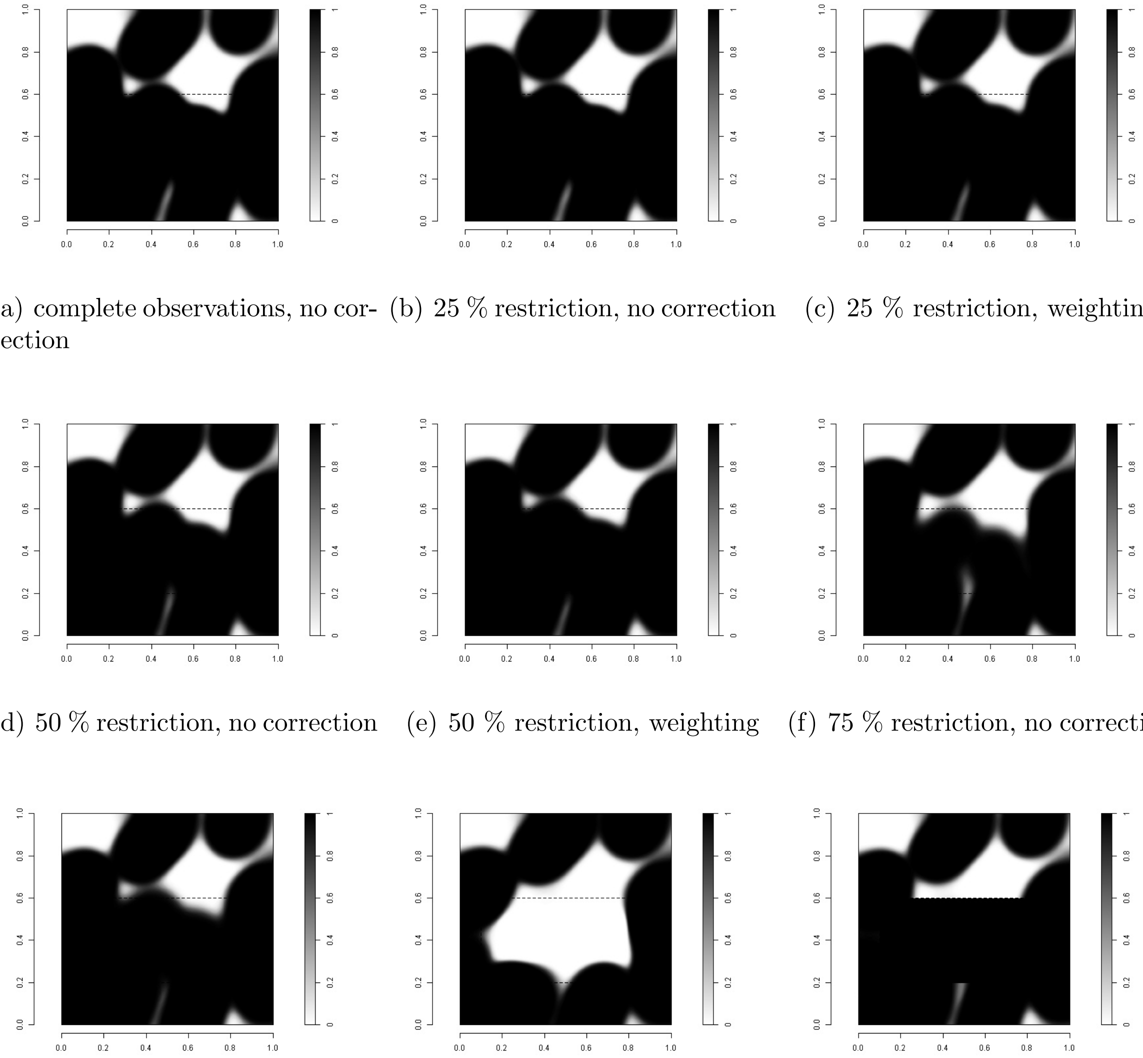
Figure A.4.: Intensity function used for the second simulated example and border of the restriction area (dashed line).

The intensity function which was used for the second simulated example is depicted in Figure A.4. As Table A.2 shows, p_{out} is much smaller than $\alpha = 0.4$ if the extent of restriction is smaller than 100 %. The fraction p_{out} and the mean of p_{miss} even decrease for a restriction of 25 % and 50 % if no correction is applied. The mean area of the high-risk zones changes little. Only for 100 % restriction, both fractions rise considerably and the mean area of the high-risk zones is smaller than when the complete bomb crater patterns are used. Figure A.5 underlines that the shape of the high-risk zones hardly changes for restrictions between 25 % and 75 %. The reason is that–although the estimated intensity decreases with increasing extent of restriction, as we can see in Figure A.6–the intensity in the relevant area is so high that the threshold which defines the high-risk zone is even exceeded for a thinned pattern.

Table A.2.: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside and mean area of the zone; second simulated example, intensity-based method

WITHOUT CORRECTION	extent of restriction	0 %	25 %	50 %	75 %	100 %
	mean p_{miss}	0.00040	0.00033	0.00031	0.00117	0.22430
	p_{out}	0.037	0.030	0.029	0.092	1.000
	mean area in m^2	0.8466	0.8491	0.8475	0.8363	0.6689

WITH CORRECTION	extent of restriction	0 %	25 %	50 %	75 %	100 %
	mean Anteil p_{miss}		0.00034	0.00031	0.00035	0.00050
	p_{out}		0.031	0.028	0.031	0.046
	mean area in m^2		0.8525	0.8561	0.8558	0.8489

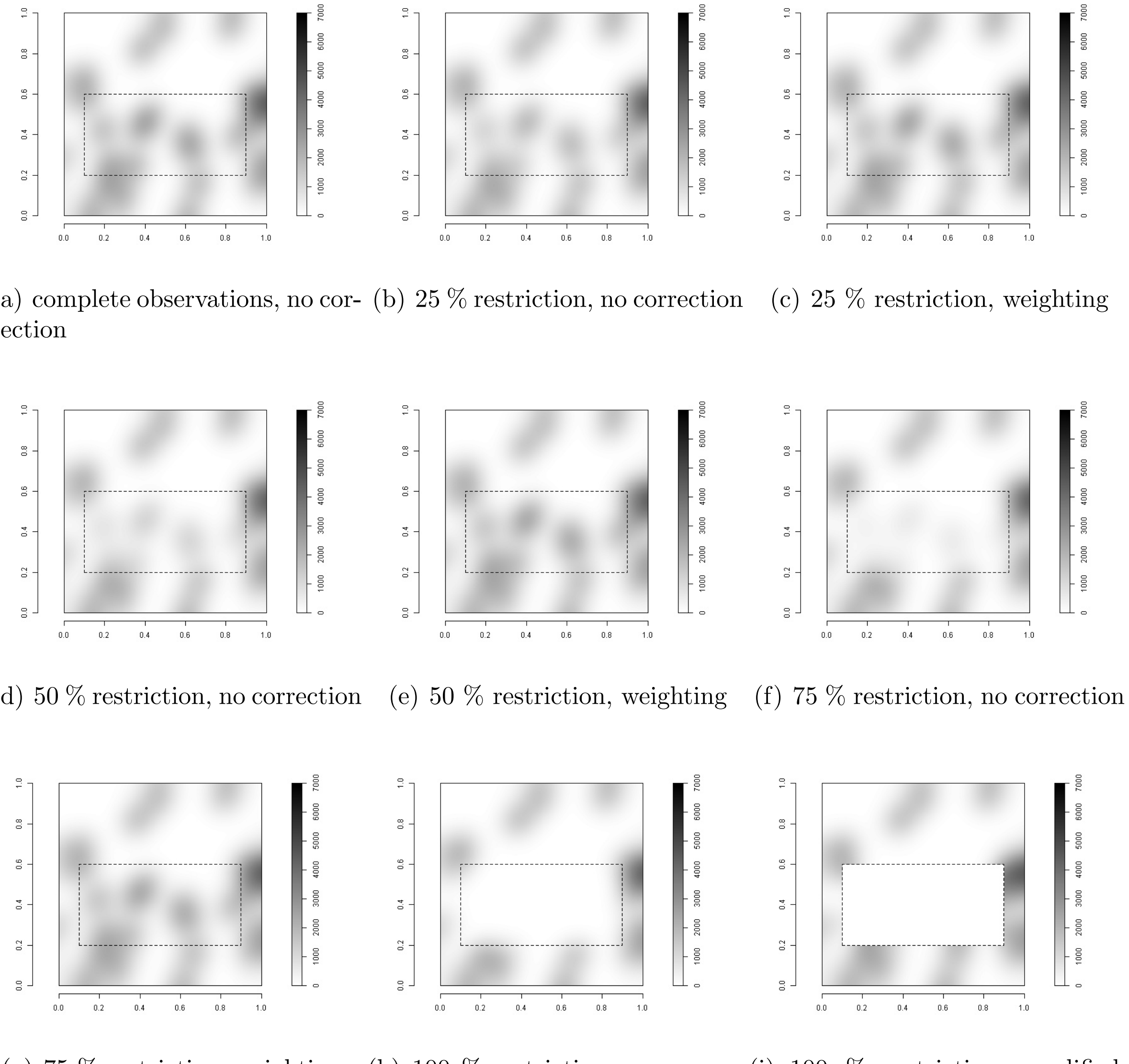


(a) complete observations, no correction (b) 25 % restriction, no correction (c) 25 % restriction, weighting

(d) 50 % restriction, no correction (e) 50 % restriction, weighting (f) 75 % restriction, no correction

(g) 75 % restriction, weighting (h) 100 % restriction, no correction (i) 100 % restriction, modified window

Figure A.5.: Resulting high-risk zones (second simulated example): Relative frequency that every pixel was part of the high-risk zones in 1000 iterations.



(a) complete observations, no correction

(b) 25 % restriction, no correction

(c) 25 % restriction, weighting

(d) 50 % restriction, no correction

(e) 50 % restriction, weighting

(f) 75 % restriction, no correction

(g) 75 % restriction, weighting

(h) 100 % restriction, no correction

(i) 100 % restriction, modified window

Figure A.6.: Mean estimated intensity functions from 1000 iterations (second simulated example).

A.3. Real-data examples

Both correction methods are now applied to Example F and to an additional example. The estimated intensity and the resulting high-risk zones for $\alpha = 0.4$ and $q = 0.10$ are shown.

Example F

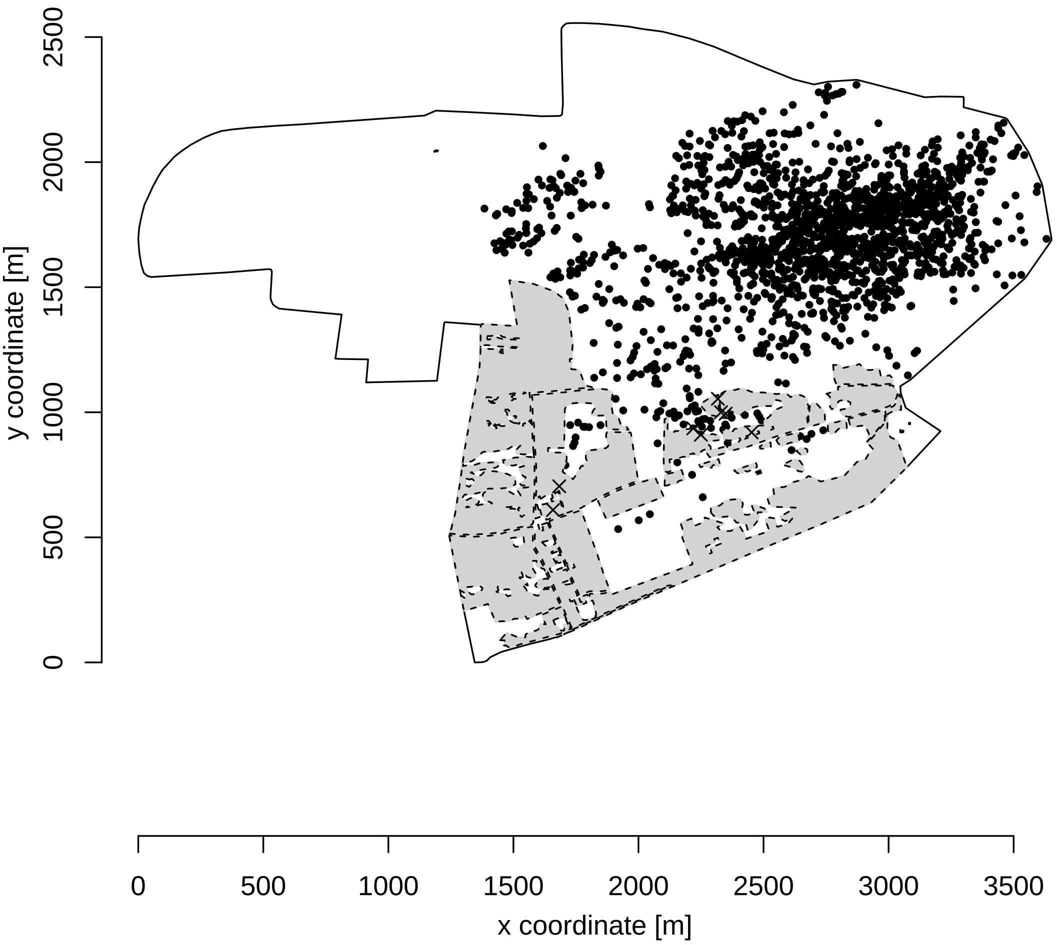
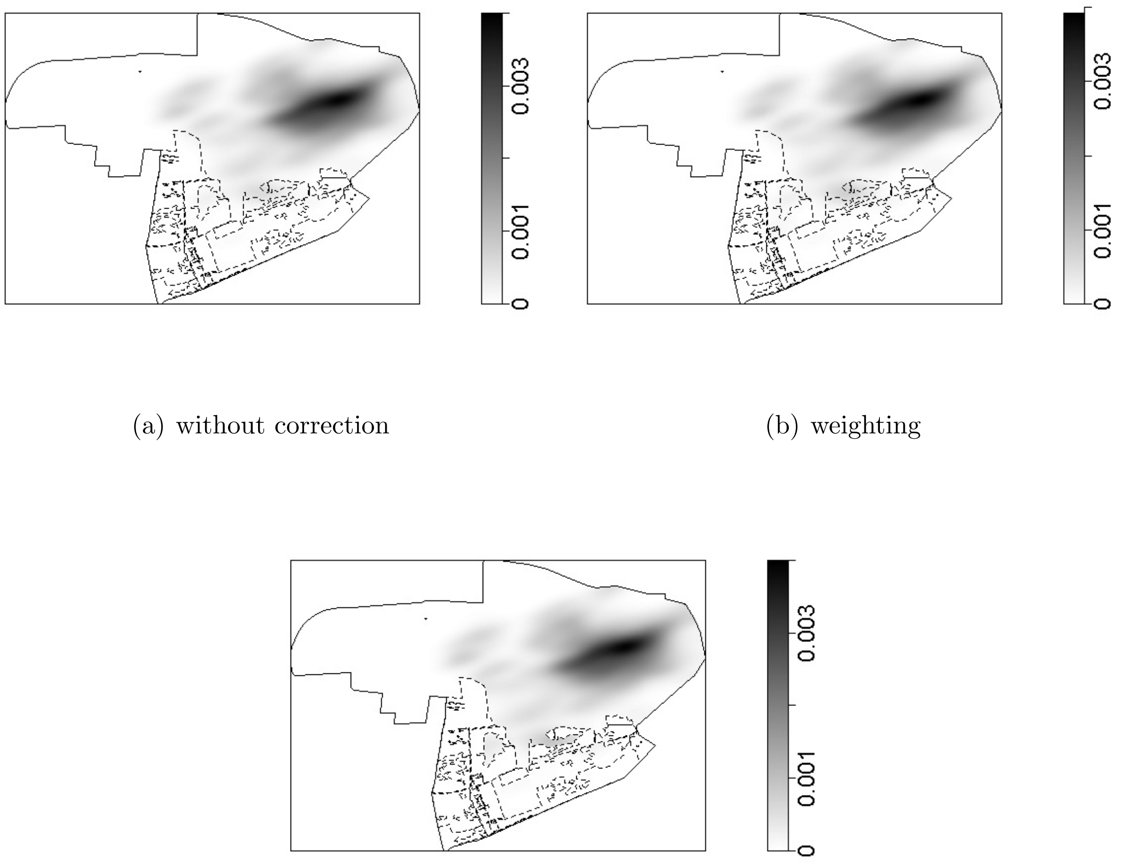


Figure A.7.: Example F: The dashed line represents the border of the restriction area. Because of its complex shape, it is additionally shaded. Observations within the restriction area are marked with a cross.

Example F contains some very small areas with an observation probability of 0 %. These are expanses of water. For the forest areas, which are far larger, OFD Niedersachsen suggested an observation probability of 30 %. Eight observations are located in the restriction area.



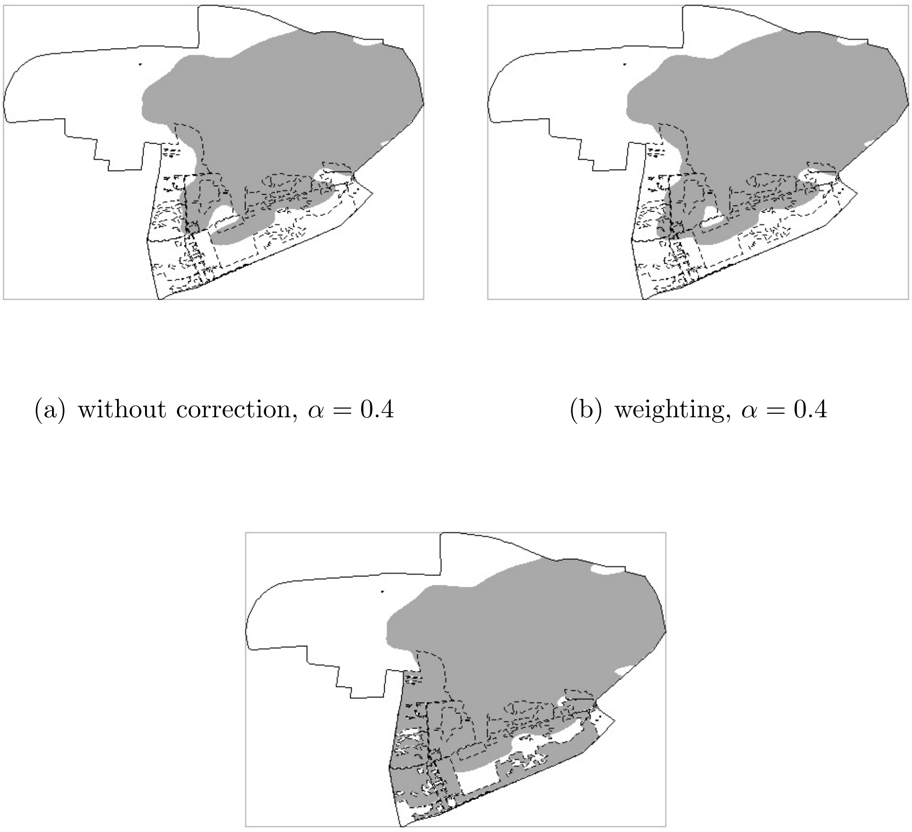
(a) without correction

(b) weighting

(c) modified window

Figure A.8.: Estimated intensity for Example F.

It is difficult to perceive any changes for the estimated intensities (Figure A.8). However, Figure A.9 shows that the high-risk zones become larger for both correction approaches.



(a) without correction, $\alpha = 0.4$

(b) weighting, $\alpha = 0.4$

(c) modified window, $\alpha = 0.4$

Figure A.9.: High-risk zones for Example F, $q = 0.10$.

Example G

Example G comprises 206 observations in an area of approximately 172 *ha*. As the observation probability inside the restriction area is 0 %, it was not possible to correct for this by performing a weighted estimation of the intensity.

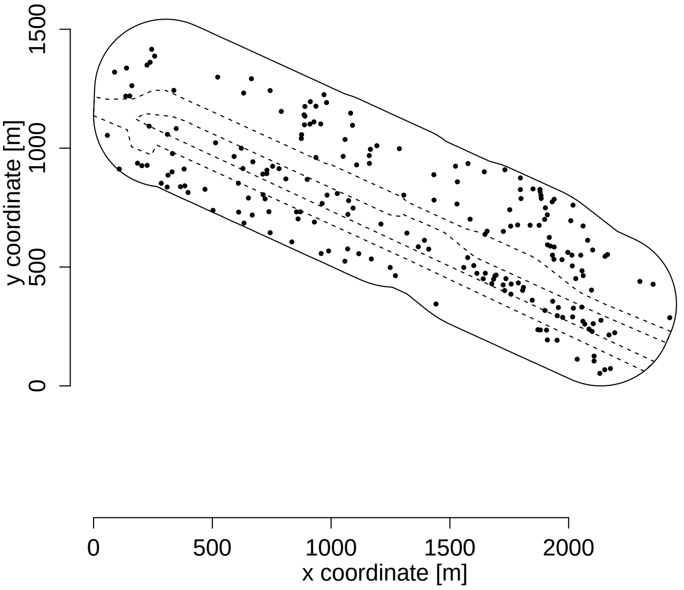
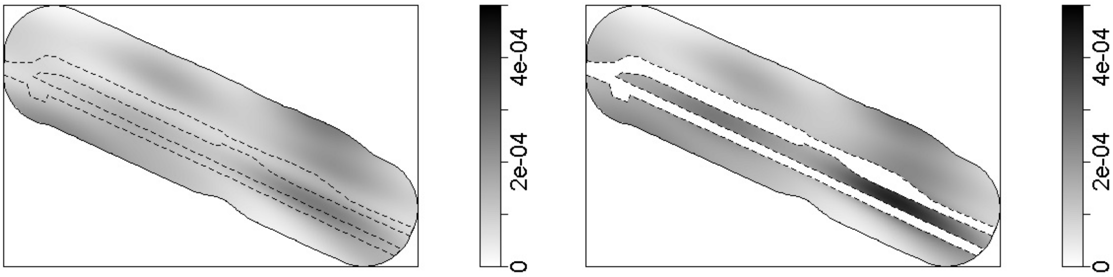


Figure A.10.: Example G: The dashed line represents the border of the restriction area.

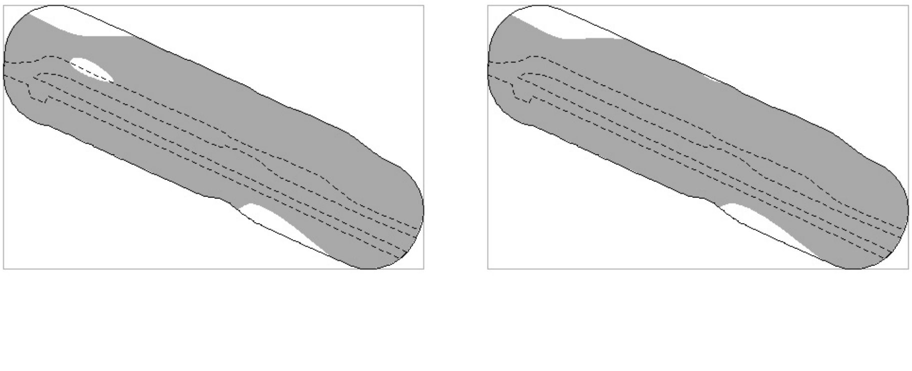
Both the estimated intensity and the high-risk zone change if the observation window is modified.



(a) without correction

(b) modified window

Figure A.11.: Estimated intensity for Example G.



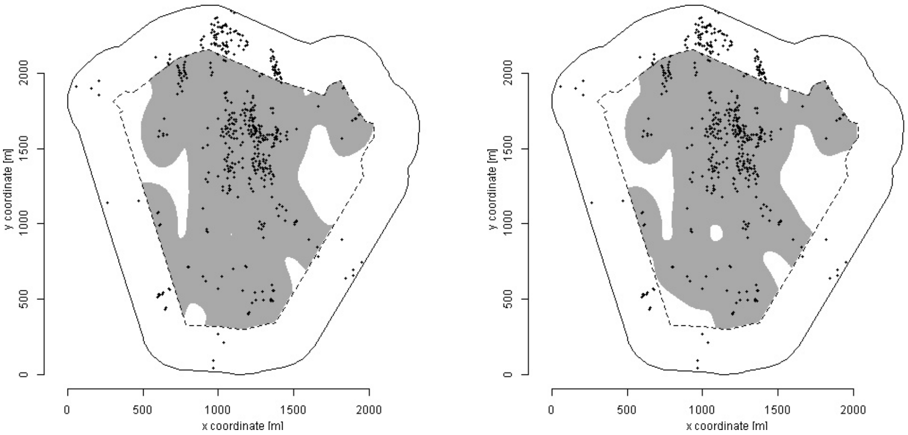
(a) without correction, $\alpha = 0.4$

(b) modified window, $\alpha = 0.4$

Figure A.12.: High-risk zones for Example G, $q = 0.10$.

B. Definition of guard regions

In practice, data is often not only available for the property itself, but also for a *guard region* (or *guard area*, see Baddeley (1999, page 40)) surrounding the property. The high-risk zones then only comprise locations within the inner window, which represents the property, but the complete data can be used for determining the high-risk zone. Although the intensity is estimated using an edge correction, high-risk zones based on the complete data differ from high-risk zones for whose determination only the data from the inner window was used. For Example A (Figure B.1), the high-risk zones for $q = 0.10$ and $\alpha = 0.2$ becomes generally smaller if only data from the inner window is used, as the edge correction does not sufficiently account for several observations located outside the property, near the border of the inner window. In the north-eastern corner, however, a small field is added to the high-risk zone in a region where no observations are located outside the property. A similar behaviour is observed for Example F (Figure B.2), where a small field is added in the south-west, whereas a larger part in the north is omitted.

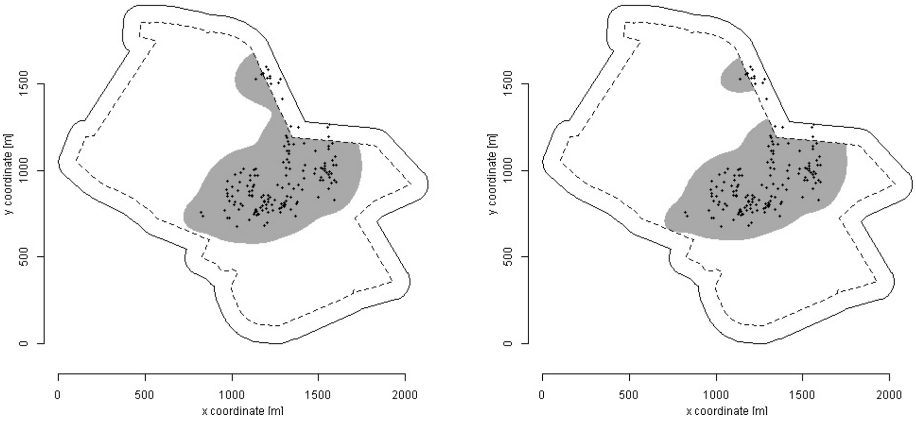


(a) complete data

(b) only inner window

Figure B.1.: Intensity-based high-risk zone for $q = 0.10$, $\alpha = 0.2$; Example A.

As expenses are incurred if additional aerial pictures need to be procured and analysed to gather information on bomb craters in the guard area, it is important to assess the necessary width of the guard region. Therefore, one needs to reflect which additional observations (or the lack of which observations) in the guard region would affect the shape of the high-risk zones.



(a) complete data

(b) only inner window

Figure B.2.: Intensity-based high-risk zone for $q = 0.10$, $\alpha = 0.2$; Example E.

If the quantile-based method is applied, all observations within the guard region whose distance from the inner window is not larger than the determined radius are relevant. Indeed, the radius itself may change with additional observations, so a possible strategy is to use the maximum nearest-neighbour distance of the observations in the inner window as width of the guard region.

The required width of the guard region for the intensity-based method is deduced from the covariance matrix which is used for intensity estimation, as the shape of the high-risk zone is determined by the estimated intensity. To keep the following deliberations simple, we will ignore that edge correction is used when the intensity is estimated. The width of the guard region will at least not become too small if we do so. An observation in the guard region is relevant if its contribution to the estimated intensity in any location inside the inner window is large, say at least η. The definition of η will be discussed later on. As the contribution of observations in the guard area is largest for those locations in the inner window which are close to the border, it is sufficient to consider the estimated intensity at the border.

The weight which is attributed to an arbitrary observation $\mathbf{x}$ when the intensity at a certain location $\mathbf{s}$ is estimated depends on the covariance matrix of the Gaussian kernel and is a function of $\mathbf{d_x} = \mathbf{x} - \mathbf{s}$. If an isotropic kernel is used, the situation can be described in an even simpler form as the weight only depends on the Euclidean distance of $\mathbf{x}$ and $\mathbf{s}$. As explained in Section 4.3.2, an anisotropic kernel is used for the intensity-based method.

OFD Niedersachsen wishes the guard region to have equal width in all directions, so the maximal distance an observation $\mathbf{x}$ with $\mathbf{d_x} = (d_1, d_2)$ can have from the border of the inner window W to be associated with a contribution of η is determined. The value of the density function of a bivariate normal distribution with mean $\mathbf{0}$, standard deviations σ_1 and σ_2 and correlation ρ

$$\varphi(\mathbf{d}_\mathbf{x}) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{d_1^2}{\sigma_1^2} + \frac{d_2^2}{\sigma_2^2} - \frac{2\rho d_1 d_2}{\sigma_1\sigma_2}\right)\right\}, \tag{B.1}$$

where σ_1, σ_2 and ρ are derived from the estimated covariance matrix of the Gaussian kernel, corresponds to the weight which is attributed to $\mathbf{x}$ when the intensity is estimated at the border of the inner window W, i.e. $\mathbf{s} \in \partial W$.

For a fixed value η and a given value of d_1, the positive solution (if existent) is

$$d_2 = d_1 \cdot \frac{\rho\sigma_2}{\sigma_1} + \sigma_2\sqrt{\frac{d_1^2(\rho^2-1)}{\sigma_1^2} - 2(1-\rho^2)\cdot\ln(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\cdot\eta)}, \tag{B.2}$$

so to maximize the distance of $\mathbf{x}$ from the border for fixed η, we can maximize the squared distance

$$\begin{aligned} d_1^2 + d_2^2 \quad = \quad & d_1^2 \cdot (1 + \frac{\sigma_2^2}{\sigma_1^2}\cdot(2\rho^2-1)) \\ & + d_1 \cdot 2\rho \cdot \frac{\sigma_2^2}{\sigma_1}\sqrt{d_1^2 \cdot \frac{\rho^2-1}{\sigma_1^2} - 2(1-\rho^2)\cdot\ln(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\cdot\eta)} \\ & - 2(1-\rho^2)\cdot\sigma_2^2\ln(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\cdot\eta). \end{aligned} \tag{B.3}$$

There are basically two strategies to determine η: The first possible strategy is to specify the probability mass of the Gaussian kernel which is to be taken into account, i.e. to demand that

$$\int_E \varphi(\mathbf{y})d\mathbf{y} = \gamma,$$

where $E = \{\mathbf{d}_\mathbf{x} : \varphi(\mathbf{d}_\mathbf{x}) > \eta\}$ is the region inside the contour for the value η and γ gives the desired probability mass, e.g. 95 %. The dashed contours in Figure B.3, which shows contour lines of the Gaussian kernels which are used for intensity estimation for Examples A to F, were determined in this way.

The second possible strategy is to derive η from the threshold c which is obtained for a high-risk zone with given parameters α and q based on the observations in the inner window only. For $\eta = c$, the interpretation is that one single observation inside the guard region would increase the estimated intensity at the inner border so that the threshold c is attained even if the estimated intensity in that region was zero before. If we ignore the edge correction and the fact that the threshold c might change for the pattern comprising observations from the guard region, this would mean that the high-risk zone would be enlarged in the region in question.

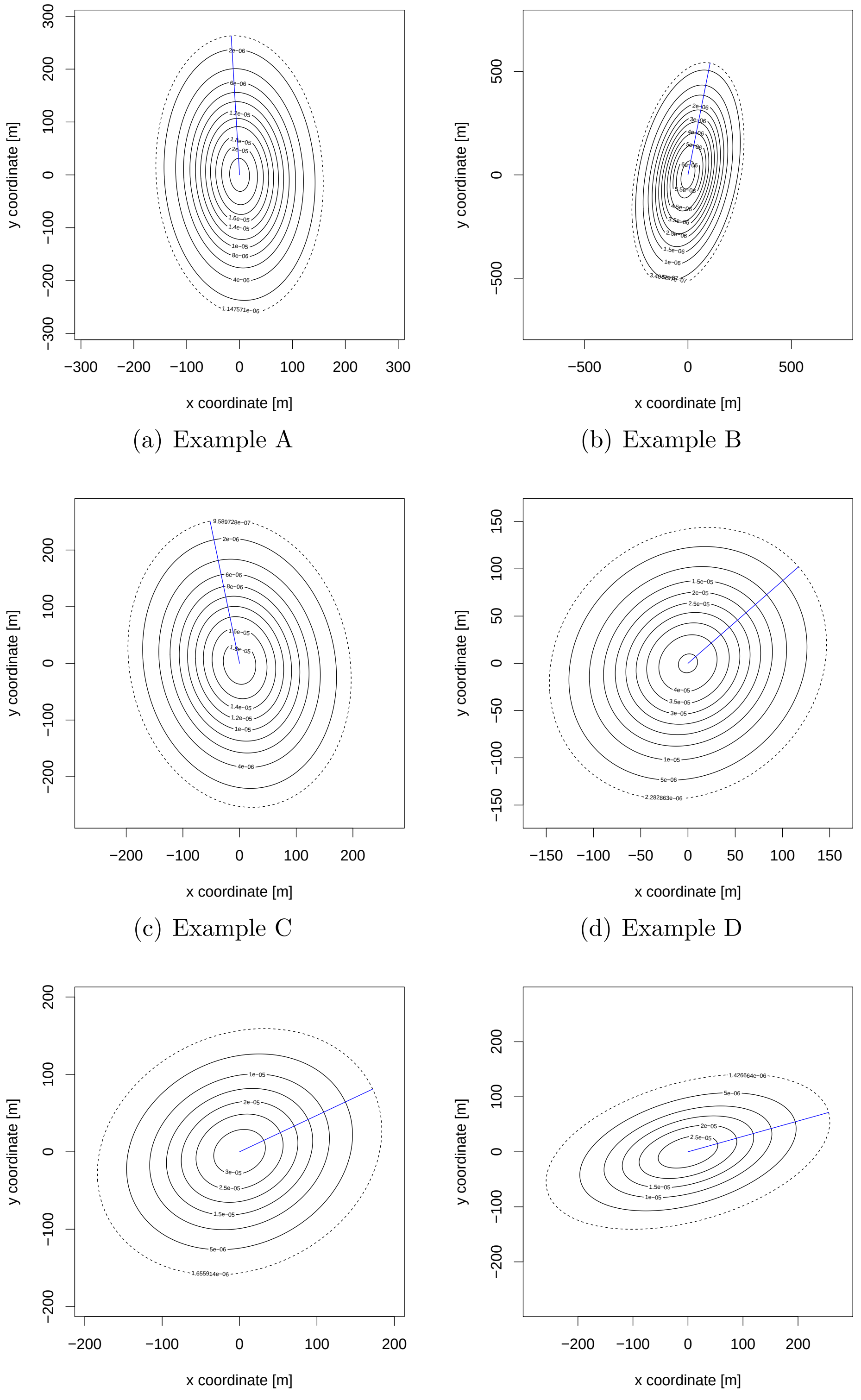
However, we need to take into account that the estimated intensity might have been larger than zero before or that there could be several observation close to each other in the guard region. For this reason, it is useful to consider $\eta = \frac{c}{2}$ and $\eta = \frac{c}{4}$ as well to gain a better understanding.

The resulting widths of the guard regions for both strategies are shown in Table B.1. Note that a value of 0 means that no guard region would be necessary at all.

Example	width of guard region for criterion				
	probability mass $\gamma = 0.95$	$\eta = c$ $\alpha = 0.2$, $q = 0.1$	$\alpha = 0.1$, $q = 0.1$	$\alpha = 0.2$, $q = 0.15$	$\alpha = 0.1$, $q = 0.15$
A	263.3	164.7	200.0	187.0	219.4
B	551.3	0	0	0	0
C	256.3	0	0	0	0
D	155.2	0	0	0	0
E	190.1	71.1	114.7	100.1	135.0
F	265.4	148.4	186.4	172.0	209.4
		$\eta = \frac{c}{2}$			
A		207.8	236.7	225.9	253.4
B		0	377.8	0	419.0
C		0	103.7	87.7	125.5
D		0	0	0	54.1
E		115.8	146.7	135.6	163.1
F		195.8	225.9	214.2	245.2
		$\eta = \frac{c}{4}$			
A		243.3	268.5	258.9	283.3
B		396.2	461.5	437.9	495.9
C		134.3	161.1	151.3	176.0
D		65.3	84.2	78.0	92.2
E		147.6	172.9	163.6	187.0
F		233.7	259.5	249.4	276.5

Table B.1.: Width of guard region

Figure B.4 shows the guard regions which result for $\gamma = 0.95$. The corresponding contour line of the Gaussian kernel was added, centered at some locations of the border of the window (where the original observation window of Examples A to F was interpreted as inner window). As anisotropic kernels are used, the width of the guard region could be chosen smaller in some directions, e.g. to the west and to the east for Example A or to the north and the south for Example F.



(a) Example A

(b) Example B

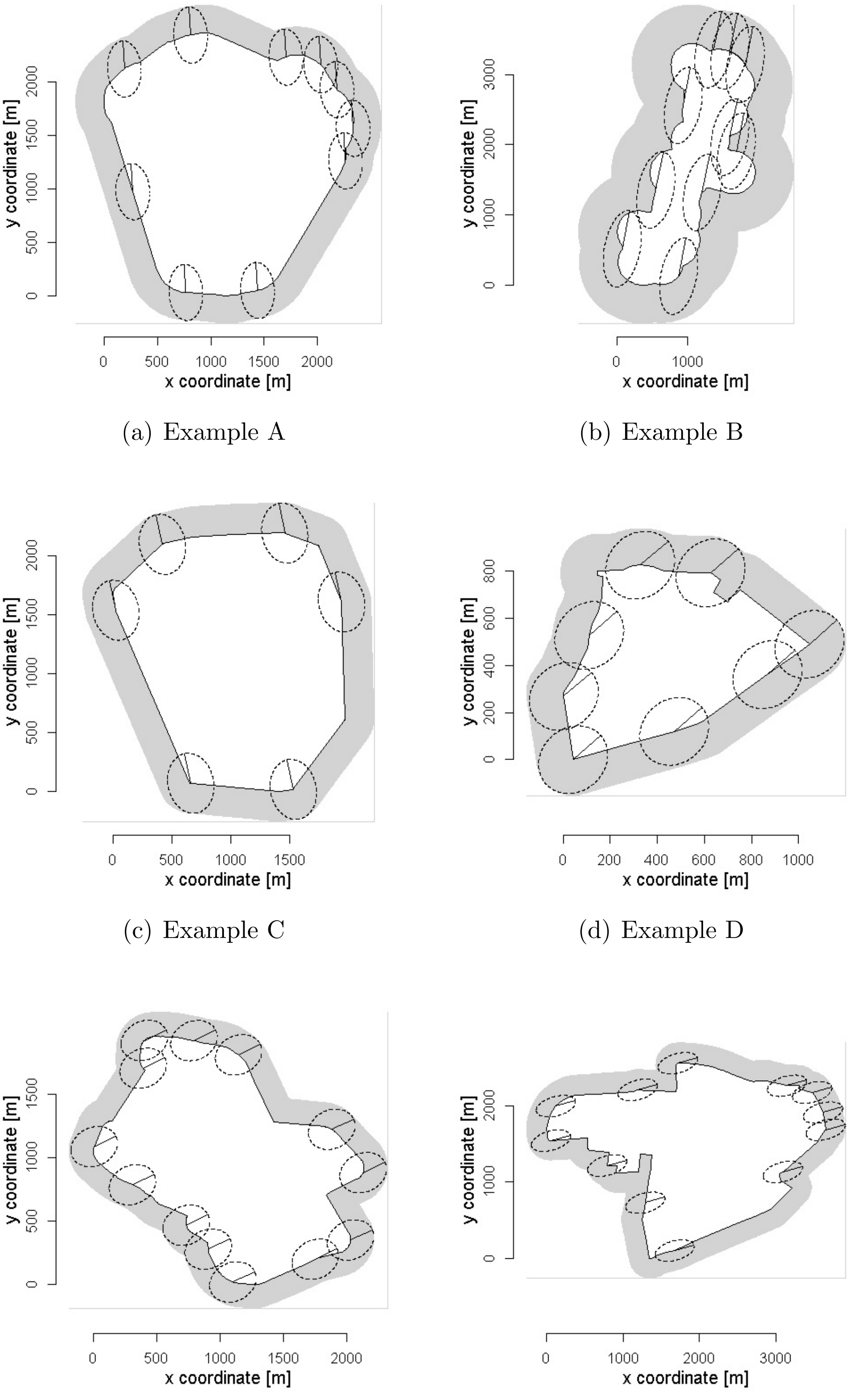
(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure B.3.: Gaussian kernel which is used for intensity estimation: contours (solid lines), contour for desired weight η with $\gamma = 0.95$ (dashed line), corresponding maximal distance (blue line).



(a) Example A

(b) Example B

(c) Example C

(d) Example D

(e) Example E

(f) Example F

Figure B.4.: Guard regions (grey) and corresponding contours of the Gaussian kernels.

C. Marked point processes

Some of the data examples which were provided by OFD Niedersachsen consist of several subpatterns which are attributed to different aerial attacks. This is possible for properties of which aerial pictures are available for different dates, which enables experts to find out which bomb craters belong to which attack. An example is shown in Figure C.1. Subpattern 1 comprises 381 events, subpattern 2 consists of 32 bomb craters and pattern 3 of 30 bomb craters. The exact dates are not given because this would facilitate the identification of the property. In some cases, additional information (such as historical records) may be available from which can be derived that different types of bombs were used. If these different types of bombs are associated with unequal potential for danger, it seems appropriate to use different failure probabilities for the subpatterns, which means that the high-risk zones have to be determined separately for each subpattern. As we will see, this approach is problematic as the combined high-risk zones become too large.

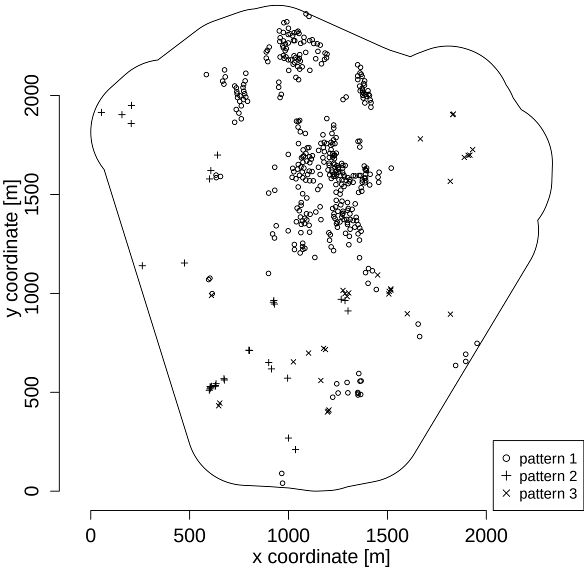
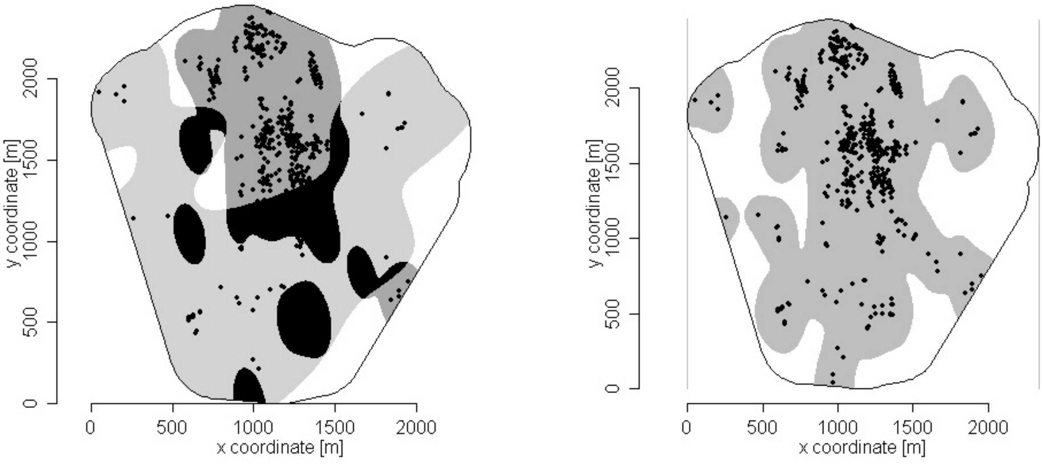


Figure C.1.: Example A: Distinction of three separate patterns.

As no data with the additional information on bomb types are available, the approach of separate high-risk zones for subpatterns is tested on Example A. Subpatterns 2 and 3 are combined to ‘pattern 2/3’ to keep the setting simple and obtain subpatterns which are not too small. In a first step, the high-risk zone obtained for a probability of non-explosion of

$q = 0.10$ and a failure probability of $\alpha = 0.4$ if the two patterns are considered separately was compared to the high-risk zone which results if the two patterns are considered jointly (Figure C.2). The darker part of the high-risk zone in Figure C.2(a) corresponds to pattern 1, the lighter part to pattern 2/3, the intersection is shaded in black. The combination is much larger than the high-risk zone depicted in Figure C.2(b).



(a) considering the two patterns separately

(b) considering the two patterns jointly

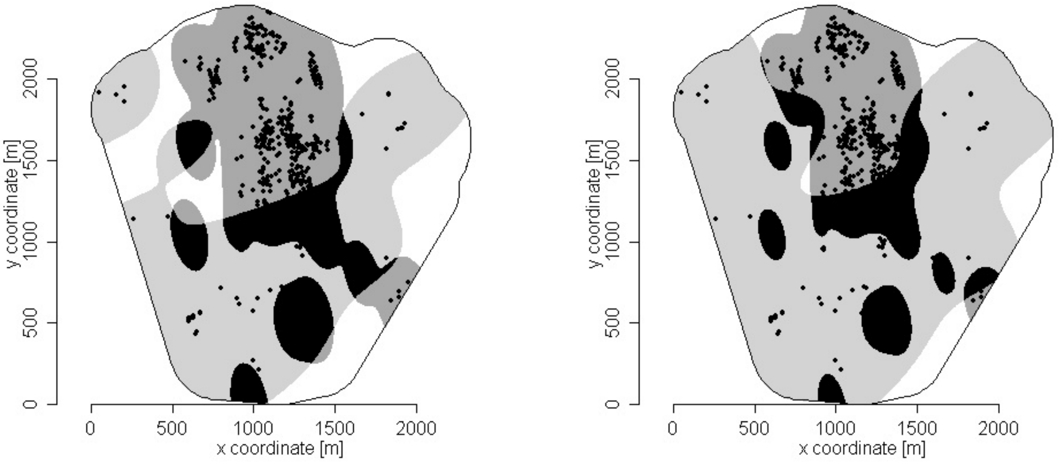
Figure C.2.: Example A: Intensity-based high-risk zone obtained for a probability of non-explosion of $q = 0.10$ and a failure probability of $\alpha = 0.4$.

For the second step, different values for α were assigned to the two patterns. The resulting high-risk zones are depicted in Figure C.3. The part of the high-risk zone which corresponds to the pattern with $\alpha = 0.2$ is enlarged compared to Figure C.2(a), the part with $\alpha = 0.6$ is demagnified. The combination of both parts is large.

Finally, the behaviour of the high-risk zones was evaluated in a simulation where the observed patterns were thinned. Four different values for the failure probability were considered. The fraction p_{out} was determined separately for the two patterns. It is given for all sixteen combinations of the failure probability α for the considered pattern and $\tilde{\alpha}$ for the respective other pattern. All values are by far too small, especially for pattern 1 (upper part of the table), where one observes also that the values depend heavily on the failure probability chosen for pattern 2/3, as the values in the lower rows ($\tilde{\alpha} = 0.2$ and $\tilde{\alpha}$ $= 0.1$) are clearly smaller than in the upper rows.

The values for pattern 2/3 are slightly closer to α and the effect of $\tilde{\alpha}$ is smaller. There is no influence of $\tilde{\alpha}$ if $\alpha = 0.1$.

In summary, determining separate high-risk zones for subpatterns cannot be recommended, at least not if the patterns are located close to each other, but overlap little. Instead, it would be better to apply the smallest of the relevant values for α to the entire pattern.



(a) High-risk zone with $\alpha = 0.2$ for pattern 1 und $\alpha = 0.6$ for pattern 2/3

(b) High-risk zone with $\alpha = 0.6$ for pattern 1 und $\alpha = 0.2$ for pattern 2/3

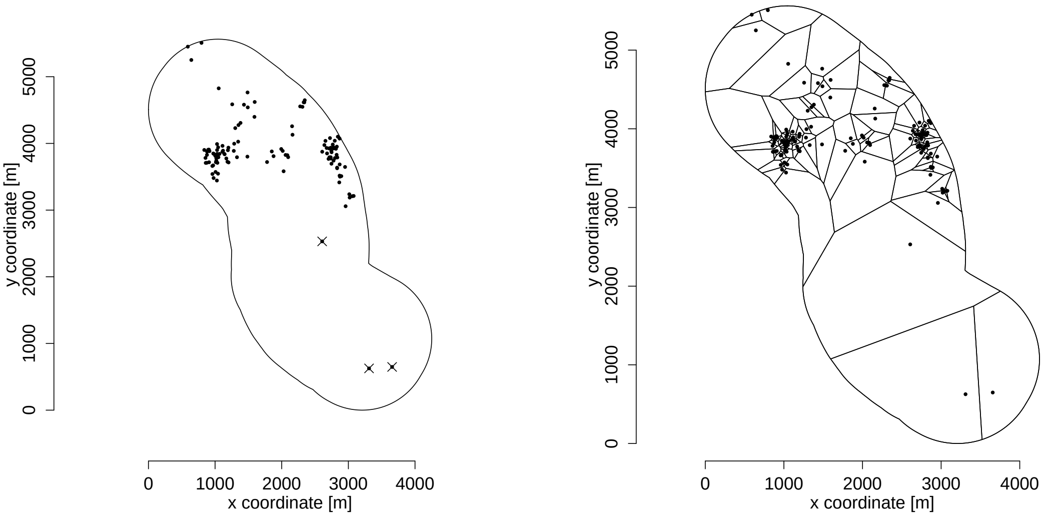
Figure C.3.: High-risk zones with different values for α for the two patterns.

Table C.1.: Evaluation results: Fraction p_{out} of high-risk zones with at least one unexploded bomb located outside (1000 iterations); Example A, intensity-based method, results for pattern 1 (P1) und combination of patterns 2 and 3 (P2) depending on parameters α for the pattern which is considered and $\tilde{\alpha}$ for the other pattern

A	q		0.1	0.1	0.1	0.1
	α		0.6	0.4	0.2	0.1
P1	$\tilde{\alpha} = 0.6$	p_{out}	0.066	0.014	0.004	0.003
	$\tilde{\alpha} = 0.4$	p_{out}	0.049	0.010	0.001	0.001
	$\tilde{\alpha} = 0.2$	p_{out}	0.035	0.005	0	0
	$\tilde{\alpha} = 0.1$	p_{out}	0.008	0.001	0	0
P2	$\tilde{\alpha} = 0.6$	p_{out}	0.333	0.261	0.015	0.008
	$\tilde{\alpha} = 0.4$	p_{out}	0.327	0.260	0.015	0.008
	$\tilde{\alpha} = 0.2$	p_{out}	0.324	0.259	0.015	0.008
	$\tilde{\alpha} = 0.1$	p_{out}	0.237	0.219	0.010	0.008

D. Consequences of outliers

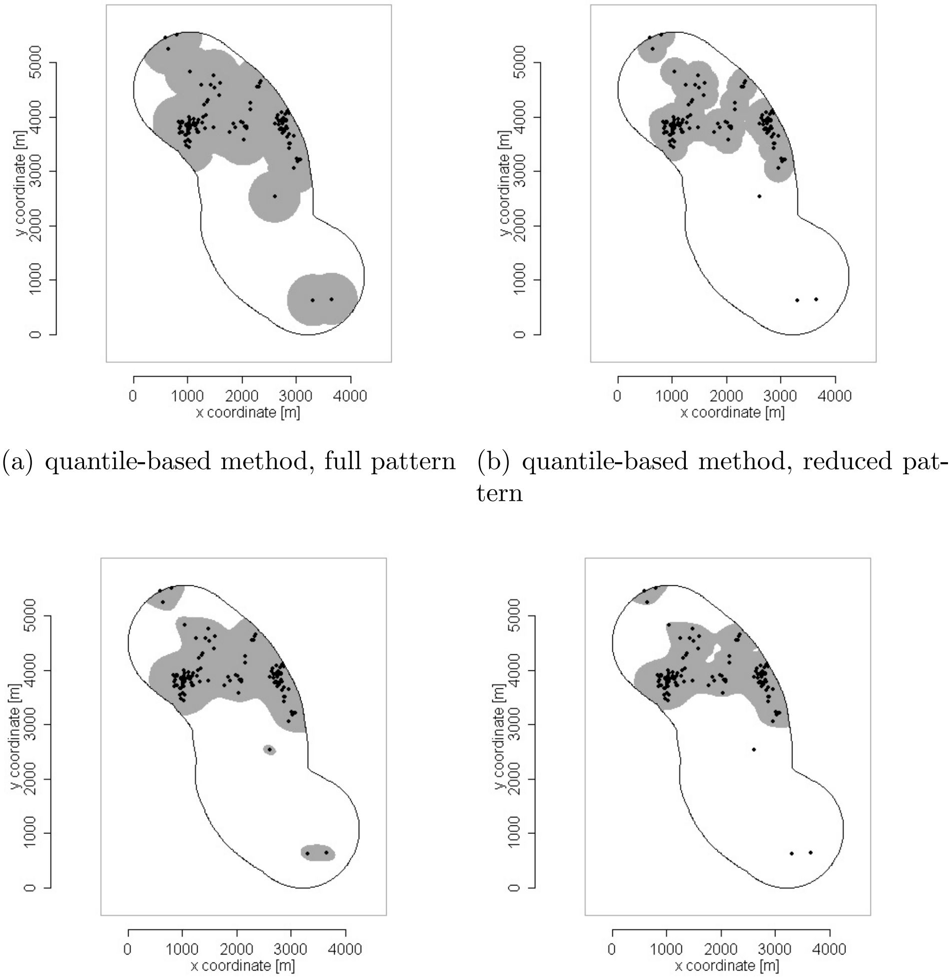
In one of the data examples provided by OFD Niedersachsen, only three of the 120 bomb craters on an area of approximately 12400000 m^2 were located in the southern half of the observation window, two of them very far away from all other bomb craters. A person evaluating aerial pictures would classify them as outliers. The pattern is shown in Figure D.1(a). In Figure D.1(b), the Voronoi tessellation (see Illian et al., 2008, Section 1.8) is depicted. It underlines that these three points would be classified as outliers from a statistical point of view, as well, as the corresponding Voronoi cells are very large. As depicted in Figure D.2, the shape of the high-risk zones is influenced considerably by these outliers. The consequences of outliers for intensity-based and quantile-based high-risk zones were investigated on the basis of this Example H. High-risk zones were determined and evaluated for the full pattern and for a reduced pattern without the three outliers. The failure probability α was set to 0.4, 0.2 and 0.1. To keep the results as comparable as possible, the 95 %, 97.5 % and 99 % quantile were considered. The probability of non-explosion was 0.10 or 0.15. For each combination of parameters, 1000 iterations were performed. As



(a) The solid line represents the border of the observation window, the three points in the southern part (marked with crosses) are perceived as outliers.

(b) Voronoi tessellation

Figure D.1.: Example H: Observed bomb craters and Voronoi tessellation.



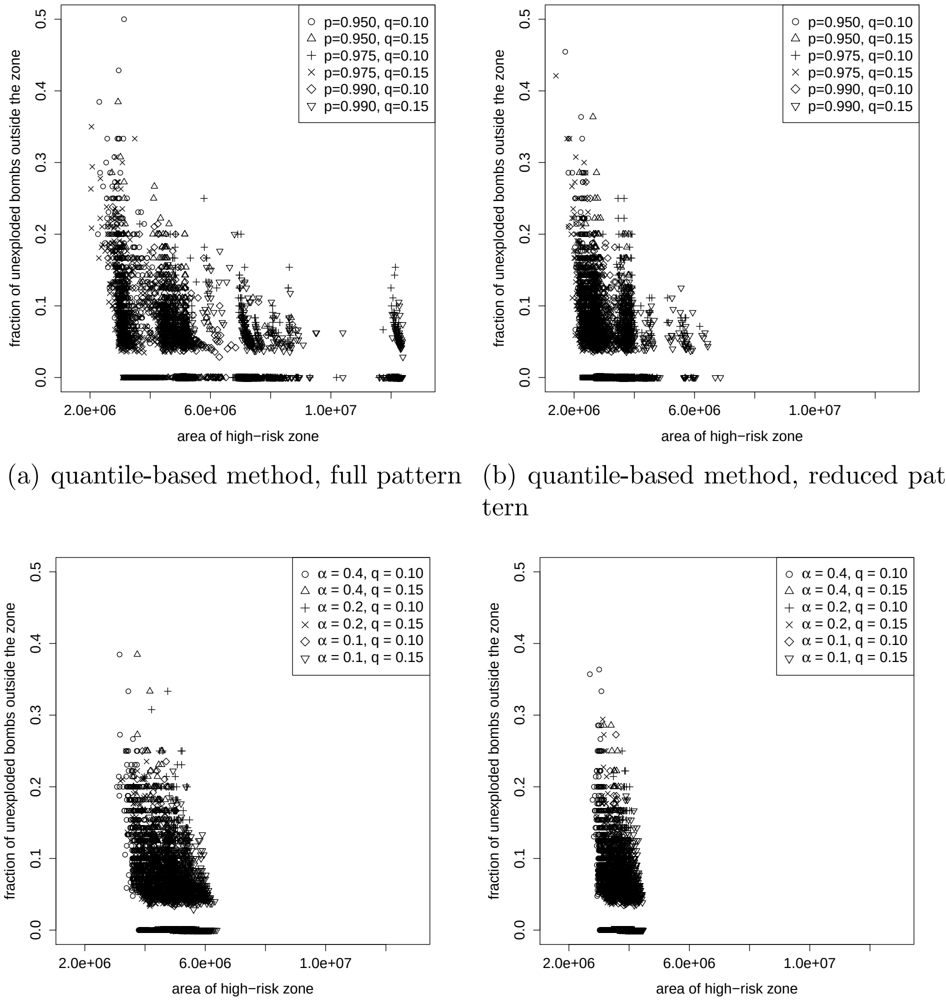
(a) quantile-based method, full pattern (b) quantile-based method, reduced pattern

(c) intensity-based method, full pattern (d) intensity-based method, reduced pattern

Figure D.2.: High-risk zones constructed by using the quantile-based method (99 % quantile) and intensity-based method ($\alpha = 0.4$) for $q = 0.10$, Example H.

depicted in Figure D.3, the area of the quantile-based high-risk zones generally varies more than the area of the intensity-based high-risk zones, which makes direct comparison of the two construction methods difficult. The scatterplots show that the area of the high-risk zones decreases if the reduced pattern is used instead of the full pattern, especially for the quantile-based method. The fraction of unexploded bombs outside the zone, which is generally higher for the quantile-based method, is slightly reduced, especially for the intensity-based method.

Figure D.4 shows how the variation in terms of the radius obtained for the quantile-based method is reduced. For high quantiles, the radius was extremely large for some cases if the full pattern was used. For the reduced pattern, there were still some cases with a large



(a) quantile-based method, full pattern

(b) quantile-based method, reduced pattern

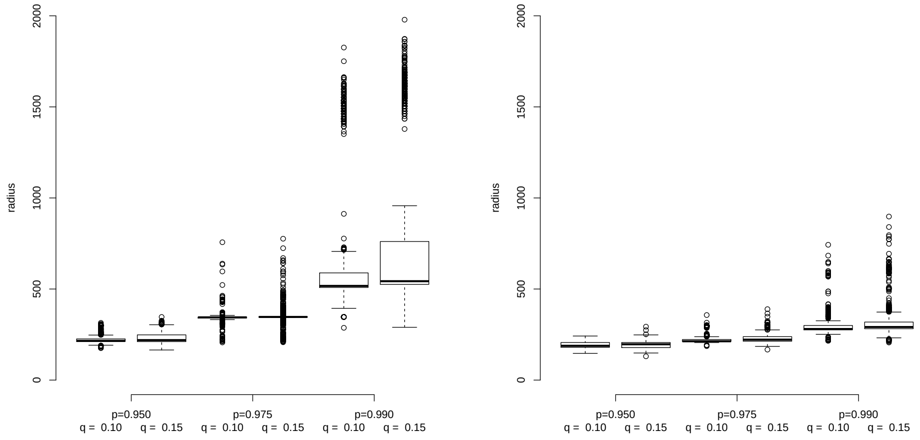
(c) intensity-based method, full pattern

(d) intensity-based method, reduced pattern

Figure D.3.: Area of the high-risk zone and fraction of simulated unexploded bombs outside the zone for the quantile-based and the intensity-based method, Example H.

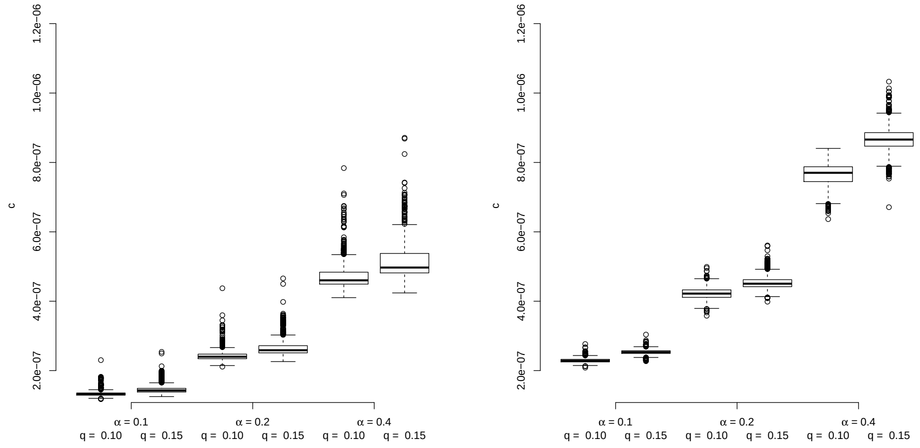
radius, but the radius was much smaller in general. The differences between the results for the full and the reduced pattern were smaller with regard to the threshold c for the intensity-based method, as the outliers for the full pattern were not as extreme as in case of the quantile-based method (Figure D.5). In general, the values of c were larger for the reduced pattern than for the full pattern.

In summary, we can say that both methods are affected by outliers. For the quantile-based method, high-risk zones can become extremely large.



(a) full pattern (b) reduced pattern

Figure D.4.: Relation between given quantile and resulting radius for the quantile-based method, Example H.



(a) full pattern (b) reduced pattern

Figure D.5.: Relation between given α and resulting threshold c for the intensity-based method, Example H.

E. Spatially varying probability of non-explosion

In general, the probability of non-explosion is not exactly known. Among other factors, it depends on characteristics of the subsoil. Therefore, it may be necessary to generalise the intensity-based method, which is the only construction method which takes the probability of non-explosion into account at all.

Instead of a homogeneous probability of non-observation q for every event, a location-dependent function $q(\mathbf{s})$ is assumed. We can still use $\hat{\lambda}_Y(\mathbf{s})$ to estimate the intensity function of the process Z:

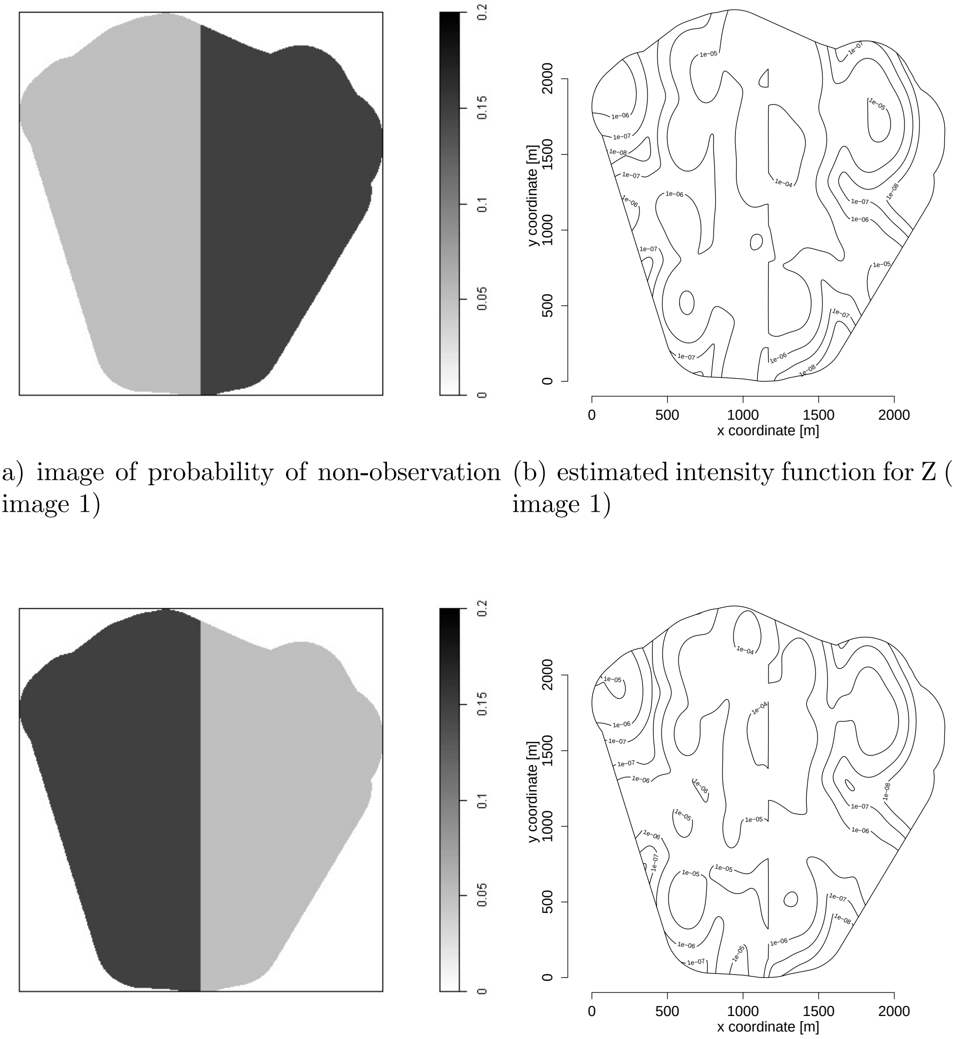
$$\hat{\lambda}_Z(\mathbf{s}) = \frac{q(\mathbf{s})}{1 - q(\mathbf{s})} \cdot \hat{\lambda}_Y(\mathbf{s}).$$

Y and Z are still independent inhomogeneous Poisson point processes (if X is assumed to be an inhomogeneous Poisson point process). As for homogeneous q, the region within the contours defined by $\hat{\lambda}_Z(\mathbf{s}) = c$ forms the high-risk zone.

No data with varying probability of non-observation could be provided. Therefore, Example A is considered and two simple fictitious functions for the probability of non-observation are used to illustrate this generalisation of the intensity-based method. Both functions $q(\mathbf{s})$ and the resulting estimated intensity function for Z are shown in Figure E.1. The functions $q(\mathbf{s})$ are depicted as images; for image 1, $q(\mathbf{s}) = 0.05$ in the western half of the window and $q(\mathbf{s}) = 0.15$ in the eastern half, whereas it is vice versa for image 2. The number of events is 218 in the western half of the observation window and 225 in its eastern half.

The resulting high-risk zones for $\alpha = 0.4$ are depicted in Figure E.2, together with high-risk zones which are obtained for a homogeneous intensity of non-explosion, where $q = 0.05$, $q = 0.10$ and $q = 0.15$. The high-risk zones based on the varying probability of non-explosion do not only differ from the high-risk zone with constant $q = 0.10$, both halves are also different from the corresponding halves of the high-risk zones for $q = 0.05$ and $q = 0.15$, respectively. More specifically, the western half of the high-risk zone in Figure E.2(d) (image 1) is even smaller than in Figure E.2(a), whereas the eastern part is even larger than in Figure E.2(c) and vice versa for the high-risk zone based on image 2 (Figure E.2(e)).

For a further investigation of the behaviour, a simulation based on thinning the observed patterns (as in Section 5.2) was performed. The results are shown in Table E.1. For both images, the fraction p_{out} is below α for $\alpha = 0.4$, whereas it exceeds α for smaller values.



(a) image of probability of non-observation (image 1)

(b) estimated intensity function for Z (for image 1)

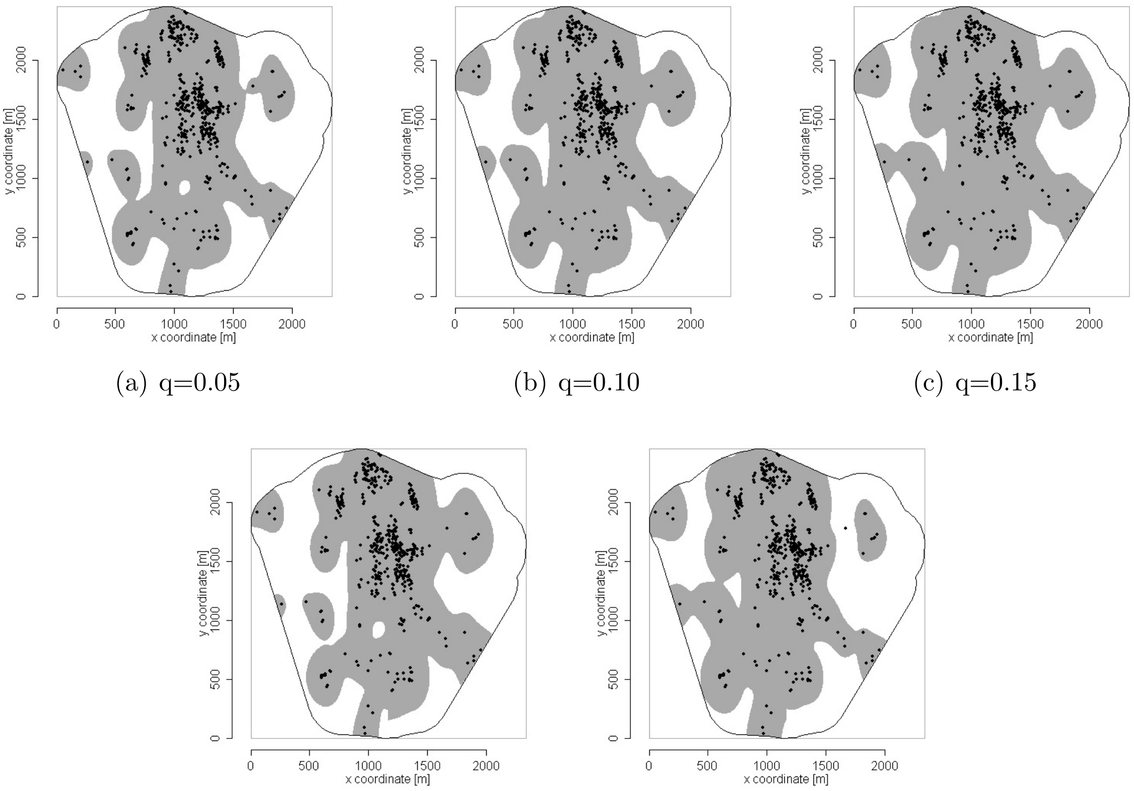
(c) image of probability of non-observation (image 2)

(d) estimated intensity function for Z (for image 2)

Figure E.1.: Fictitious examples of varying probability of non-observation and resulting estimated intensity functions for Z in Example A.

Both p_{out} and the mean fraction p_{miss} are larger for image 2 compared to image 1, while the mean area is smaller for small values of α.

A comparison with the results from Section 5.2 reveals that the high-risk zones for images 1 and 2 are smaller than for $q = 0.10$. The fraction p_{out} for high-risk zones based on image 1 is smaller than for $q = 0.10$. With regard to image 2, this is only the case for $\alpha = 0.4$, but all fractions p_{out} are smaller than for $q = 0.15$.



(a) q=0.05

(b) q=0.10

(c) q=0.15

(d) q=0.05 in the west, q=0.15 in the east (image 1)

(e) q=0.15 in the west, q=0.05 in the east (image 2)

Figure E.2.: High-risk zones (shaded grey areas) obtained for the intensity-based method with maximal failure probability of $\alpha = 0.4$, Example A.

Table E.1.: Results of the simulation: Mean fraction p_{miss} of unexploded bombs outside the high-risk zone from 1000 iterations, fraction p_{out} of generated high-risk zones for which at least one unexploded bomb was located outside and mean area of the zone, Example A, intensity-based method (INT) for varying probability of non-explosion

A	image	1	1	1	2	2	2
	α	0.4	0.2	0.1	0.4	0.2	0.1
INT	mean p_{miss}	0.009	0.005	0.004	0.011	0.007	0.006
	p_{out}	0.332	0.225	0.190	0.356	0.274	0.219
	mean area in m^2	2573198	2890718	3115477	2593287	2874754	3075703

Bibliography

Ang, Q. W., A. Baddeley, and G. Nair (2012). Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scandinavian Journal of Statistics 39*(4), 591–617.

Baddeley, A. (1999). Spatial sampling and censoring. In O. E. Barndorff-Nielsen, W. S. Kendall, and M. N. M. van Lieshout (Eds.), *Stochastic Geometry: Likelihood and Computation*, Chapter 2, pp. 37–78. Chapman and Hall.

Baddeley, A. (2008). Analysing spatial point patterns in R. Workshop notes. Technical report, CSIRO online technical publication. Available online at `www.csiro.au/resources/pf16h.html`.

Baddeley, A. and R. D. Gill (1997). Kaplan-Meier estimators of interpoint distance distributions for spatial point processes. *Annals of Statistics 25*, 263–292.

Baddeley, A., M. Kerscher, K. Schladitz, and B. T. Scott (2000). Estimating the J function without edge correction. *Statistica Neerlandica 54*(3), 315–328.

Baddeley, A., J. Møller, and A. Pakes (2008). Properties of residuals for spatial point processes. *Annals of the Institute of Statistical Mathematics 60*, 627–649.

Baddeley, A., J. Møller, and R. Waagepetersen (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica 54*(3), 329–350.

Baddeley, A. and R. Turner (2005). spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software 12*(6), 1–42. Available online at `http://www.jstatsoft.org`.

Baddeley, A. and R. Turner (2006). Modelling spatial point patterns in R. In A. Baddeley, P. Gregori, J. Mateu, R. Stoica, and D. Stoyan (Eds.), *Case Studies in Spatial Point Pattern Modelling*, Number 185 in Lecture Notes in Statistics, pp. 23–74. New York: Springer-Verlag.

Baddeley, A., R. Turner, J. Møller, and M. Hazelton (2005). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society, Series B 67*(5), 617–666.

Baddeley, A. J. and B. W. Silverman (1984). A cautionary example on the use of second-order methods for analyzing point patterns. *Biometrics 40*, 1089–1093.

Barnard, G. (1963). Contribution to the discussion of Professor Bartlett's paper. *Journal of the Royal Statistical Society, Series B 25*(2), 294.

Bartlett, M. S. (1964). The spectral analysis of two-dimensional point processes. *Biometrika 51*(3 and 4), 299–311.

Bedford, T. and J. van den Berg (1997). A remark on the Van Lieshout and Baddeley J-function for point processes. *Advances in Applied Probability 29*(1), 19–25.

Berman, M. and P. Diggle (1989). Estimating weighted integrals of the second-order intensity of a spatial point process. *Journal of the Royal Statistical Society, Series B 51*(1), 81–92.

Bernardeau, F. and R. van de Weygaert (1996). A new method for accurate estimation of velocity field statistics. *Monthly Notices of the Royal Astronomical Society 279*, 693–711.

Besag, J. (1977). Discussion of Dr Ripley's paper. *Journal of the Royal Statistical Society, Series B 39*(2), 193–195.

Borgefors, G. (1986). Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing 34*, 344–371.

Brix, A., R. Senoussi, P. Couteron, and J. Chadœuf (2001). Assessing goodness of fit of spatially inhomogeneous Poisson processes. *Biometrika 88*(2), 487–497.

Chiu, S. N. and D. Stoyan (1998). Estimators of distance distributions for spatial patterns. *Statistica Neerlandica 52*(2), 239–246.

Choi, E. and P. Hall (1999). Nonparametric approach to analysis of space-time data on earthquake occurrences. *Journal of Computational and Graphical Statistics 8*, 733–748.

Cox, D. R. (1955). Some statistical models related with series of events. *Journal of the Royal Statistical Society, Series B 17*, 129–164.

Cressie, N. A. C. (1993). *Statistics for spatial data* (Second ed.). New York: Wiley.

Daley, D. J. and D. Vere-Jones (1988). *An introduction to the theory of point processes.* New York: Springer.

Davies, T. M., M. L. Hazelton, and J. C. Marshall (2011). sparr: Analyzing spatial relative risk using fixed and adaptive kernel density estimation in R. *Journal of Statistical Software 39*(1), 1–14.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*(1), 1–38.

Diggle, P. J. (1978). On parameter estimation for spatial point processes. *Journal of the Royal Statistical Society, Series B 40*(2), 178–181.

Diggle, P. J. (1979). On parameter estimation and goodness-of-fit testing for spatial point processes. *Biometrics 35*, 87–101.

Diggle, P. J. (1983). *Statistical analysis of spatial point patterns.* London: Academic Press.

Diggle, P. J. (1985). A kernel method for smoothing point process data. *Journal of the Royal Statistical Society, Series C 34*, 138–147.

Diggle, P. J. (2003). *Statistical analysis of spatial point patterns* (Second ed.). London: Arnold.

Diggle, P. J., V. Gómez-Rubio, P. E. Brown, A. G. Chetwynd, and S. Gooding (2007). Second-order analysis of inhomogeneous spatial point processes using case-control data. *Biometrics 63*, 550–557.

Diggle, P. J. and R. J. Gratton (1984). Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society, Series B 46*, 193–212.

Diggle, P. J. and J. S. Marron (1988). Equivalence of smoothing parameter selectors in density and intensity estimation. *Journal of the American Statistical Association 83*(403), 793–800.

Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of Statistical Software 21*(7), 1–16. Available online at `http://www.jstatsoft.org/v21/i07`.

Duong, T. and M. L. Hazelton (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics 15*, 17–30.

Duong, T. and M. L. Hazelton (2005a). Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *Journal of Multivariate Analysis 93*, 417–433.

Duong, T. and M. L. Hazelton (2005b). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics 32*, 485–506.

Efron, B. and R. Tibshirani (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*(458), 611–631.

Fraley, C., A. E. Raftery, T. B. Murphy, and L. Scrucca (2012). mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report 597, Department of Statistics, University of Washington.

Gelfand, A. E., P. J. Diggle, M. Fuentes, and P. Guttorp (Eds.) (2010). *Handbook of Spatial Statistics.* Handbooks of Modern Statistical Methods. Boca Raton, FL: Chapman & Hall/CRC.

Gentle, J. E. (2005). *Elements of computational statistics.* New York: Springer.

Givens, G. H. and J. A. Hoeting (2005). *Computational statistics.* Hoboken, N.J.: Wiley-Interscience.

Guan, Y. (2007). A least-squares cross-validation bandwidth selection approach in pair correlation function estimations. *Statistics & Probability Letters 77*, 1722–1729.

Guan, Y. (2008a). A goodness-of-fit test for inhomogeneous spatial Poisson processes. *Biometrika 95*(4), 831–845.

Guan, Y. (2008b). On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. *Journal of the American Statistical Association 103*(483), 1238–1247.

Hanisch, K.-H. (1983). Reduction of n-th moment measure and the special case of the third moment measure of stationary and isotropic planar point processes. *Mathematische Operationsforschung und Statistik, Series Statistics 14*(3), 421–435.

Hanisch, K.-H. (1984). Some remarks on estimators of the distribution function of nearest neighbour distance in stationary spatial point processes. *Mathematische Operationsforschung und Statistik, Series Statistics 15*(3), 409–412.

Held, L. (2008). *Methoden der statistischen Inferenz: Likelihood und Bayes.* Heidelberg: Spektrum Akademischer Verlag.

Hornung, R. (2011). *Analyse von Wildunfalldaten mit Hilfe räumlicher Poissonprozesse.* Master thesis, Ludwig-Maximilians-Universität München.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685.

Hothorn, T., R. Brandl, and J. Müller (2012). Large-scale model-based assessment of deer-vehicle collision risk. *PLoS ONE 7*(2), 1–10.

Hyndman, R. T. and Y. Fan (1996). Sample quantiles in statistical packages. *The American Statistician 50*(4), 361–365.

Ickstadt, K. and R. L. Wolpert (1997). Multiresolution assessment of forest inhomogeneity. In C. Gatsonis, J. S. Hodges, R. E. Kass, R. McCulloch, P. Rossi, and N. D. Singpurwalla (Eds.), *Case Studies in Bayesian Statistics, Volume III*, Number 121 in Lecture Notes in Statistics, pp. 371–386. Chapman and Hall.

Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan (2008). *Statistical analysis and modelling of spatial point patterns.* Chichester, West Sussex: Wiley.

Illian, J. B. and H. Rue (2010). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). Technical Report No. 6/2010, University of Trondheim.

Illian, J. B., S. H. Sørbye, and H. Rue (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *Annals of Applied Statistics 6*(4), 1499–1530.

Jones, M. C., J. S. Marron, and B. U. Park (1991). A simple root n bandwidth selector. *Annals of Statistics 19*, 1919–1932.

Mahling, M., M. Höhle, and H. Küchenhoff (2013). Determining high-risk zones for unexploded World War II bombs by using point process methodology. *Journal of the Royal Statistical Society, Series C 62*(2), 181–199.

Matérn, B. (1960). Spatial variation: Stochastic models and their applications to problems in forest surveys and other sampling investigations. *Meddelanden från Statens Skogsforskningsinstitut 49*(5).

McDonald, J. A. and M. J. Small (2006). Assessing sites contaminated with unexploded ordnance: Statistical modeling of ordnance spatial distribution. *Environmental Science & Technology 40*(3), 931–938.

McDonald, J. A. and M. J. Small (2009). Statistical analysis of metallic anomaly patterns at former air force bombing ranges. *Stochastic Environmental Research and Risk Assessment 23*, 203–214.

Molchanov, I. (2005). *Theory of Random Sets.* London: Springer.

Møller, J. (2003). Shot noise Cox processes. *Advances in Applied Probability 35*(3), 614–640.

Møller, J., A. R. Syversveen, and R. P. Waagepetersen (1998). Log Gaussian Cox Processes. *Scandinavian Journal of Statistics 25*, 451–482.

Møller, J. and G. L. Torrisi (2005). Generalised shot noise Cox processes. *Advances in Applied Probability 37*, 48–74.

Møller, J. and R. P. Waagepetersen (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Boca Raton, FL: Chapman & Hall/CRC.

Møller, J. and R. P. Waagepetersen (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics 34*(4), 643–684.

Neyman, J. and E. L. Scott (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society, Series B 20*, 1–43.

Neyman, J. and E. L. Scott (1972). Processes of clustering and applications. In P. A. W. Lewis (Ed.), *Stochastic Point Processes*, pp. 646–681. New York: Wiley.

Ohser, J. (1983). On estimators for the reduced second moment measure of point processes. *Series Statistics 14*, 63–71.

Ohser, J. and D. Stoyan (1981). On the second-order and orientation analysis of planar stationary point processes. *Biometrical Journal 23*, 523–533.

Quantum GIS Development Team (2013). *GNU General Public License*. Available online at `http://qgis.osgeo.org`.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available online at `http://www.R-project.org`.

Rajala, T. (2012). spatgraphs. R package version 2.62. Available online at `http://cran.r-project.org/web/packages/spatgraphs/index.html`.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability 13*(2), 255–266.

Ripley, B. D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, Series B 39*(2), 172–212.

Ripley, B. D. (1981). *Spatial Statistics*. New York: John Wiley and Sons.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes*. Cambridge University Press.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion). *Journal of the Royal Statistical Society, Series B 71*, 319–392.

Schabenberger, O. and C. Gotway (2005). *Statistical Methods for Spatial Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice and visualization.* New York: Wiley.

Seibold, H. (2012). *Determining high-risk zones using point process methodology–Realization by building an* R *package.* Bachelor thesis, Ludwig-Maximilians-Universität München.

Seibold, H. and M. Mahling (2012). highriskzone. R package version 1.0. Available online at `http://cran.r-project.org/web/packages/highriskzone/index.html`.

Silverman, B. W. (1992). *Density estimation for statistics and data analysis.* London: Chapman & Hall.

Stoyan, D. (1992). Statistical estimation of model parameters of planar Neyman-Scott cluster processes. *Metrika 39*, 67–74.

Stoyan, D., W. S. Kendall, and J. Mecke (1987). *Stochastic geometry and its applications.* Wiley series in probability and mathematical statistics: Applied probability and statistics. Chichester: Wiley.

Stoyan, D., W. S. Kendall, and J. Mecke (1995). *Stochastic geometry and its applications* (Second ed.). Wiley series in probability and mathematical statistics: Applied probability and statistics. Chichester: Wiley.

Stoyan, D. and H. Stoyan (1992). *Fraktale, Formen, Punktfelder: Methoden der Geometrie-Statistik.* Berlin: Akademie-Verlag.

Stoyan, D. and H. Stoyan (1994). *Fractals, random shapes and point fields: methods of geometrical statistics.* New York: John Wiley and Sons.

Tavakkoli, M., D. Weth, and J. Agarius (2012). Bestimmung der Wahrscheinlichkeiten von Bombenblindgängern. *altlasten spektrum 3*, 124–127. Available online at `http://www.altlastendigital.de/AltS.03.2012.124`.

Thomas, M. (1949). A generalization of Poisson's binomial limit for use in ecology. *Biometrika 36*, 18–25.

Thönnes, E. and M. N. M. Van Lieshout (1999). A comparative study on the power of Van Lieshout and Baddeley's J function. *Biometrical Journal 41*, 721–734.

Titterington, D. M., A. F. Smith, and U. E. Makov (1985). *Statistical analysis of finite mixture distributions.* Chichester: Wiley.

Van Lieshout, M. N. M. and A. Baddeley (1996). A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica 50*, 344–361.

Vere-Jones, D. (1970). Stochastic models for earthquake occurrence. *Journal of the Royal Statistical Society, Series B 32*, 1–62.

Waagepetersen, R. P. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics 63*, 252–258.

Wand, M. P. and M. C. Jones (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association 88*, 520–528.

Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*. London: Chapman and Hall Ltd.

Wolpert, R. L. and K. Ickstadt (1998). Poisson/gamma random field models for spatial statistics. *Biometrika 85*(2), 251–267.

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 28.03.2013

Monia Mahling