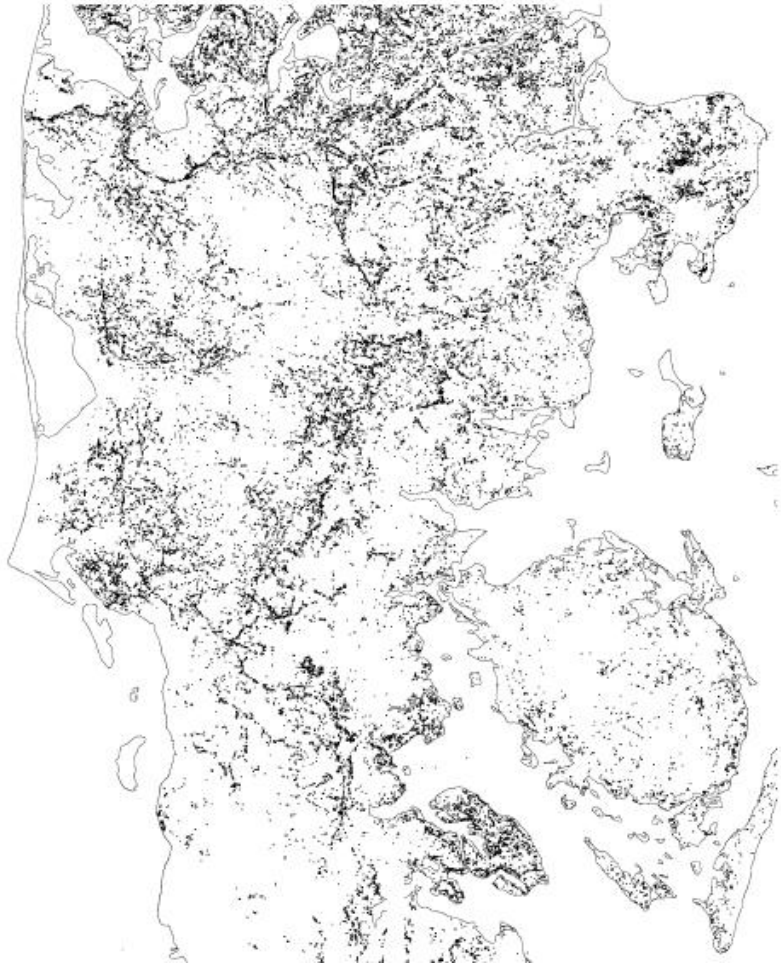
Rumlige punkt- og linjeprocesser

En introduktion med henblik på beskrivelse af punktmønstre med linjetendenser og en modelopstilling for datasæt af gravhøje

af Niels Peter Haals og Per Jensen



AALBORG UNIVERSITET

Institut for Matematiske Fag • Fredrik Bajers Vej 7G • 9220 Aalborg Ø •
Gruppe G3-111 • MAT6 - Speciale • Forår 2009 •

Aalborg Universitet
Institut for Matematiske Fag
Fredrik Bajers Vej 7G
9220 Aalborg Ø
http://www.math.aau.dk

Titel:

Rumlige punkt- og linjeprocesser
En introduktion med henblik på beskrivelse af punktmønstre med linjetendenser og en modelopstilling for datasæt af gravhøje

Specialeretning:

Statistik

Semester og gruppe:

MAT6-Speciale, G3-111

Projektperiode:

2. feb. - 5. jun. 2009

Gruppemedlemmer:

Niels Peter Haals

Per Jensen

Vejleder:

Jakob Gulddahl Rasmussen

Oplagstal: 7

Sidetal: 148

Appendiks start: Side 136

Forside: Plot af punktmønster for gravhøje i dele af Danmark. Scan fra [Johansen et al., 2004].

Synopsis:

Formålet med denne rapport er at opstille en model til beskrivelse af linjetendenser i et punktmønster bestående af koordinater for gravhøje. For at muliggøre dette foretages der en teoretisk analyse af rumlige punktprocesser, der er statistiske modeller til beskrivelse af punktmønstre.

Rapporten indledes med at introducere forskellige generelle begreber indenfor rumlige punktprocesser. Dernæst gennemgås Poisson-processen, og forskellige egenskaber herved introduceres og bevises. Poisson-processen er specielt interessant indenfor punktprocesteori, da den benyttes som referenceproces ved sammenligning af punktprocesser. En anden type punktproces, der beskrives i denne rapport er Cox-processen.

Summary statistics for punktprocesser beskrives ved første og anden-ordensegenskaber, og der foretages estimater heraf. Summary statistics muliggør tilpasning af modeller på data og modeltjek.

I forbindelse med modelopstillingen af datasættet for gravhøje er det nødvendigt at introducere linjeprocesser. Dette er blot en anden type af punktprocesser, der er defineret på et specielt repræsentationsrum.

Der opstilles en specifik model til beskrivelse af punktmønstret for gravhøje. Denne model opstilles på baggrund af gennemgået teori om linjeprocesser, således linjetendenserne i datasættet for gravhøje kan beskrives. Denne model er implementeret i `R`, og dertil er simulationsmetoder af punktprocesser benyttet, bl.a. en Birth-Death-Move-algoritme.

Afslutningsvis udføres et modeltjek på modellen og det vurderes hvorvidt den opstillede model er i stand til at beskrive datasættet af gravhøje.

© Gruppe G3-111 forår 2009

Summary

Spatial point and line processes are statistical models used to describe point and line patterns respectively. The common point processes can be used to describe repulsion, aggreation and clustering among points in a point pattern, but they fail to describe points located along lines. Points located along lines are called line tendencies in the following. Several point patterns contain line tendencies, for example locations of mountain tops, and it can be usefull to develop point process models which can describe and produce point patterns with line tendencies.

The main focus of this report is to establish a point process model to describe a point pattern consisting of locations of bronze age burial places. This point pattern contain line tendencies and it is therefore necessary to construct a point process model which can describe these line tendencies. In this report the line tendencies are modeled by combining theory from point and line processes.

Two different main types of point processes are described, the Poisson and Cox process. The Poisson process has an important role in the point process theory, because it describes complete spatial randomness. Cox processes on the other hand can be used when modelling aggregation and clustering in point patterns. Essential results and definitions regarding the Poisson and Cox processes are discussed. Results treated for Poisson processes include superposition, thinning and marked Poisson processes. Cox processes are mainly treated through special types of Cox processes like the Neyman-Scott processes: the Thomas process and the Matérn process.

Summary statistics for point patterns and point processes are also treated. The summary statistics for a point pattern give an idea about how the points are located, and by the summary statistics it is possible to describe a point pattern with a single graph. The summary statistics can indicate whether there is repulsion or aggregation among the points. This can be used to decide which point processes are suitable for describing the point pattern. To plot the summary statistic for a point pattern and compare it with the summary statistic calculated from the point process model is a useful tool to conclude if the model fits the data.

Often a point process is specified by a density with respect to the standard Poisson process. To simulate a point process from this density Markov chain Monte Carlo theory can be used. Therefore different types of Markov chain Monte Carlo algorithms are introduced including the Birth-Death-Move algorithm. The described algorithms are capable of simulating a point process when only an unnormalised density with respect to the standard Poisson process is known.

As an introduction to line processes different types of parametrization of lines is described, and it is explained how the line processes can be seen as point processes on special representation spaces. Only line processes in $\mathbb{R}^2$ are treated and a special kind of these line processes is the Poisson line processes, which describes complete spatial randomness among the lines. Other results about stationarity and intensity

of line processes is discussed and results about the number of lines from a line process intersecting a convex set are covered too.

Point and line processes and the results treated are illustrated through out the report by plots of simulations. Also examples of analysis, model fitting and model checking of point and line patterns is shown. This is performed to specify a procedure when working with point processes and as a starting point for the data analysis for barrows. Furthermore different Poisson and Cox point processes are fitted the barrows in an attempt to model the point pattern, but none of these seems to be able to describe the line tendencies.

A point process model to describe the line tendencies is constructed by combining a Poisson line process and Poisson point processes. To carry out the model construction different model assumptions are needed. These assumptions are based on the theory about the barrows and their tendency to be located along lines. Under the model assumption a likelihood function for the point process model is obtained with different unknown parameters and an unknown Poisson line process.

A Birth-Death-Move algorithm is used together with Bayesian statistics to simulate the lines and estimate the parameters from the point pattern of barrows. This algorithm is implemented by use of the statistical program `R`. With the estimated parameters and lines realisations of the model is obtained. To find out how well the point process model describes the barrows a model check i performed. This is done by visually comparing a realisation of the point process model and the point pattern. Also a correct model check is done by calculating a summary statistic and envelopes from the model and comparing it with the corresponding one from the point pattern. As a last check of the model an angle plot is made for realisations of the model and for barrows to see if the this indicates the line tendencies.

After the model check it is discussed how well the model fits the data. This is done by comparing results from the model with the barrows. Also the model's ability to predict prehistoric roads are discussed. In the end of the report different improvements of the model are discussed.

Forord

Denne rapport er resultatet af et MAT6-projekt(speciale) i statistik ved Institut for Matematiske Fag på Aalborg Universitet. Det overordnede emne for projektet er rumlige punkt- og linjeprocesser og anvendelse af disse.

Rapporten indeholder begreber og resultater inden for rumlige punkt- og linjeprocesser. En del af rapporten består derudover af en analyse af et datasæt bestående af koordinater for gravhøje, samt en opstilling af en model til beskrivelse af punkternes fordeling i dette datasæt. Til denne analyse er statistik-programmet `R` [R Development Core Team, 2008] benyttet med pakken `spatstat`, som er en bestemt pakke i `R` med funktioner til analysering af punktmønstre. Implementeringen af modellen i `R` er en tidskrævende proces, og derfor er `R`-koden hertil vedlagt rapporten i Appendiks B, som dokumentation og som hjælp til andre, der kan have nytte af det.

I rapporten benyttes to slags henvisninger. Når et henvisningsnummer er i parentes, som f.eks. (1.2) henvises der til en ligning, ulighed eller linje med dette nummer, mens henvisningen 'Sætning 1.2' henviser til en hel sætning med dette nummer. Sidstnævnte henvisning findes ligeledes til definition, algoritme, kapitel, afsnit og appendiks.

Litteraturhenvisninger anføres som regel i starten af afsnit og kapitler og noteres [Forfatter(e), udgivelsesår, evt. placering i kilden]. Litteraturhenvisningerne henviser til litteraturlisten, som findes bagerst i rapporten. Ved sætninger osv. henvises der specifikt til kilden umiddelbart efter sætningens nummer.

Visse af emnerne i denne rapport er behandlet med før i forbindelse med et MAT4-projekt. Indholdet i det tidligere projekt var dog rent teoretisk, hvorimod der i dette projekt især fokuseres på det anvendelsesmæssige. For at få rapporten til at være sammenhængende og selvindeholdende har det været nødvendigt at medtage dele af denne allerede behandlede teori. Denne teori indgår dog i en kraftig revideret udgave i form af en anvendelsesmæssig tilgangsvinkel og ved at have inkluderet yderligere resultater. Der er tale om kapitlerne 2,3,4 og 5.

Vi vil gerne takke vores vejleder, Lektor Jakob Gulddahl Rasmussen, for vejledning og for at stille datasættet for gravhøjene til rådighed.

Aalborg den 5. juni 2009

Niels Peter Haals

Per Jensen

Indhold

1 Indledning **1**

I Teori **11**

2 Punktprocesser **13**

2.1 Introduktion 13

2.2 Målteori 14

2.3 Punktprocesser 17

3 Poisson-processer **21**

3.1 Definitioner og resultater 21

3.1.1 Eksistens af Poisson-processer 22

3.1.2 Poisson-udvikling 24

3.1.3 Uafhængige Poisson-processer 26

3.1.4 Slivnyak-Meckes sætning 27

3.1.5 Superposition og uafhængig udtynding 27

3.1.6 Mærkede Poisson-processer 31

3.2 Tætheder for Poisson-processer 31

3.2.1 Likelihoodfunktion for Poisson-proces 33

4 Summary statistics **35**

4.1 Andenordens egenskaber 35

4.1.1 Parkorrelationsfunktionen og det andenordens reducerede momentmål 38

4.2 Andenordens summary statistics 40

4.2.1 Ikke-parametrisk estimation af intensitetsfunktionen 41

4.2.2 Ikke-parametrisk estimation af $\mathcal{K}$, K og L-funktionerne . . 42

4.2.3 Kantkorrektion 43

4.2.4 Ikke-parametrisk estimat af parkorrelationsfunktion 43

4.3 Afstandsbaserede summary statistics 44

4.3.1 Ikke-parametrisk estimation af F-, G- og J-funktionerne . . 45
4.4 Envelopes for summary staistics 46

5 Cox-processer 47
5.1 Definitioner og resultater 47
5.2 Generelt om clusterprocesser 49
5.3 Neyman-Scott processer 50
5.3.1 Thomas-processer 51
5.3.2 Matérn-processer 52
5.4 Inferens for Cox-processer 53
5.4.1 Minimumkontrastestimation 54

6 Simulering af punktprocesser 57
6.1 Metropolis-Hastings algoritmer 57
6.1.1 For punktprocesser med et fast antal punkter 57
6.1.2 For punktprocesser med skiftende antal punkter 58

7 Linjeprocesser 61
7.1 Definitioner og resultater 61
7.1.1 Stationaritet og isotropi 64
7.1.2 Typer af linjeprocesser 65
7.1.3 Intensitetsmål for linjeprocesser 66
7.1.4 Linjemål 67
7.1.5 Længden af linjer 71
7.2 Likelihood for Poisson-linjeproces 72

8 Eksempel på punkt og linjeprocesser 73
8.1 Eksempel på punktproces 73
8.2 Eksempel på linjeproces 77

II Databehandling 81

9 Databeskrivelse 83
9.1 Databeskrivelse 83

9.1.1 Summary statistics 84

10 Punktproces-modeller for gravhøje **87**

10.1 Tilpasning af Poisson-proces 87

10.2 Tilpasning af Cox-proces 88

11 Model med linjetendenser for gravhøje **91**

11.1 Opstilling af model 92

11.2 Implementering af model 96

11.2.1 Parametre 96

11.2.2 Linjer 97

11.3 Model med linjer fra BDM-algoritme 99

11.4 Model med passende fast linjemønster 107

12 Modeltjek for model med linjetendenser **113**

12.1 Beregning af summary statistic 113

12.1.1 Summary statistics ved vinkler 118

12.2 Modeltjek af model fra Afsnit 11.3 119

12.2.1 Modeltjek med linjer fra BDM-algoritmen 120

12.2.2 Modeltjek med linjer fra Poisson-linjeproces 122

12.2.3 Opsummering af modeltjek 125

III Epilog **127**

13 Afrunding **129**

13.1 Konklusion 129

13.2 Perspektivering 132

IV Appendiks **135**

A Matematiske definitioner **137**

B R-kode **138**

B.1 Model BDM 138

Litteratur **148**

Indledning

Et punktmønster er en angivelse af placeringen af bestemte hændelser indenfor et bestemt område, hvor hver hændelse kaldes et punkt. Der eksisterer statistiske modeller til beskrivelse af sådanne punktmønstre kaldet punktprocesser, og det er varianter af disse modeller dette projekt omhandler. Visse punktmønstre indeholder punkter, der har tendens til at ligge langs linjerstykker. Dette er netop tilfældet for et punktmønster, der beskriver placeringen af gravhøje i Vestjylland, som det følgende omhandler.

Der eksisterer flere teorier, om hvorfor gravhøjene har tendens til at ligge langs linjestykker, og disse teorier kan være nyttige i forbindelse med opstilling af passende modeller til beskrivelse af datasættet for gravhøje. Følgende afsnit omhandler derfor gravhøje generelt og teorier om gravhøje langs linjerstykker.

Gravhøje

Gravhøje blev i oldtiden bygget som gravsteder og var en typisk begravelsesmetode i Danmark fra den tidlige bondestenalder for omkring 5000 år siden frem til kristendommens indførsel i Danmark for omkring 1000 år siden. Udformningen og størrelsen af gravhøjene afspejler forskellige tidsperioder, og kun de mest markante er bevaret i dag, heriblandt hovedsageligt bronzealder-gravhøjene, der er ca. 3500 år gamle. I dag eksisterer der stadig en del af disse gravhøje, og en del er i årenes løb forsvundet, sløjfet eller ødelagte bl.a. som følge af dyrknig af jord og bebyggelse. [Online Leksikon, 2006] [Egeberg, 2005].

Et andet levn fra fortiden er spor fra gamle veje. Disse vejspor har deres oprindelse langt tilbage i tiden, og de veje der kendes i dag, menes blot at være en del af et større vejsystem. Der er fundet rester af vejspor på marker og heder, som menes, at være mindst lige så gamle som gravhøjene. Dette understøttes af, at der er fundet bronzealder-gravhøje placeret direkte oven på vejspor, hvilket vil sige, at visse vejspor sandsynligvis er 3500 år eller ældre. Dog er mange af de lokaliserede vejspor formodentlig yngre end gravhøjene. [Egeberg, 2005] [Egeberg, 2004].

Visse steder er der en tendens til, at gravhøjene ligger langs buede linjer gennem landskabet, og disse linjer følger ofte nutidige og/eller forhistoriske veje. Det er ikke tilfældigt, at gravhøje ligger i linjer langs veje, de har nemlig deres egen indbyrdes afhængighed. Blandt andet har arkæologen Sophus Müller allerede omkring år 1900 udtalt, at der ikke kan findes anden forklaring på linjerne af gravhøje end, at de må betegne gamle færdselsveje. Dette kunne skyldes, at færdselsvejene gik gennem de områder, hvor befolkningen boede, og hvor de også begravede de døde. Især er rækkerne tydelige og mange i det sydlige og vestlige Jylland. [Egeberg, 2005] [Egeberg, 2004].

På Sophus Müllers tid var antallet af kendte gravhøje lavere end i dag, da der dengang ikke var mulighed for lokalisering af gravhøje og gamle veje via luftfotografier og teknologiske metoder, og da der dengang ikke var gjort forsøg på lokalisering i samme grad, som der har været senere. Selvom teorien om, at linjer af gravhøje skulle ligge langs færdselsårer efterhånden blev anerkendt, var mulighederne for en nærmere undersøgelse begrænsede. Senere har teorien stået i skyggen af mere pompøse emner som rige grave og mosefund. I forbindelse med tilbageføringen af Skjern å omkring årene 1999-2002, blev der dog gjort opdagelser, som vagte interessen om sammenhængen mellem gamle veje og gravhøje blandt arkæologer. Der blev nemlig gjort fund af gamle overfartssteder over Skjern å, og det interessante var nu, om der var sammenhæng mellem linjer af gravhøje, altså de formodede veje, og overfartsstederne. [Egeberg, 2004].

Tilbageføringen af Skjern å førte til projektet 'Høj og Vad', som bestod i at indsamle koordinater for samtlige kendte gravhøje, direkte synlige eller lokaliserede via udgravninger eller luftfotos, i et område omkring Skjern å. Datasættet fra 'Høj og Vad'-projektet består således af koordinater for samtlige kendte gravhøje fra oldtiden, og udgør information, om hvor folk blev begravet i oldtiden, og dermed også information om, hvor de formodes at have boet og færdes. Derudover indsamledes information om gamle veje og specielt overfartssteder over Skjern å, således en nærmere undersøgelse af sammenhængene mellem veje og gravhøje blev muliggjort. [Egeberg, 2005] udtaler, hvorfor specielt Vestjylland er velegnet til undersøgelse af sammenhængen mellem gravhøje og veje: "Specielt Vestjylland rummer særlige muligheder. I ingen andre dele af landet lader gravhøje og hjulspor sig kortlægge i tilsvarende store mængder, samtidig med at fordelingerne af gravhøje og spor danner et finmasket net på kryds og tværs gennem landskabet. Det er dette netværk, der nu tolkes som forskellige tiders foretrukne kommunikationslinjer og dermed færdselskorridorer.". [Egeberg, 2005] [Egeberg, 2004] [Møller og Rasmussen, 2009].

Datasæt for gravhøje

Datasættet for gravhøje i området omkring Skjern å fra 'Høj og Vad'-projektet består af geografiske koordinater for samtlige kendte gravhøje i området. Som eksempel er fem forskellige gravhøje fra datasættet angivet ved koordinaterne:

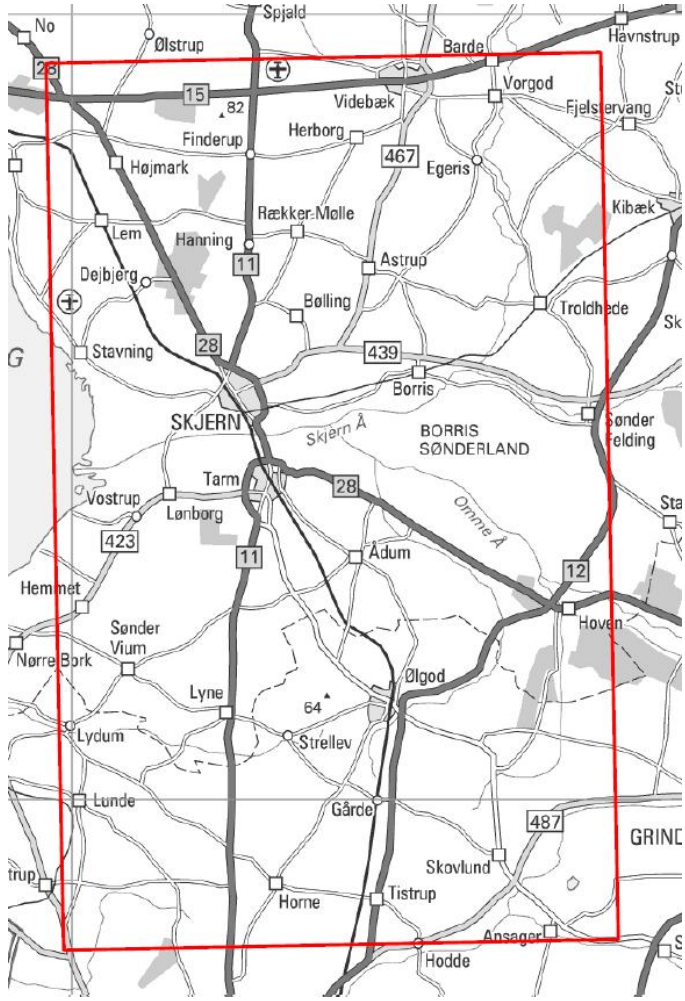
	V1	V2
1	470367	6205364
2	470360	6205336
3	470362	6205307
4	470365	6205275
5	470389	6205300

Koordinaterne er angivet i Universal Transverse Mercator(UTM)-koordinater, som er en anden måde at angive koordinater på end de klassiske længde- og breddegrader. Værdierne af V1 svarer til et udtryk for breddegraden og V2 er et udtryk

for længdegraden. Gravhøjene i datasættet er fordelt over et område givet ved minimum og maksimum for de to koordinatakser, nemlig:

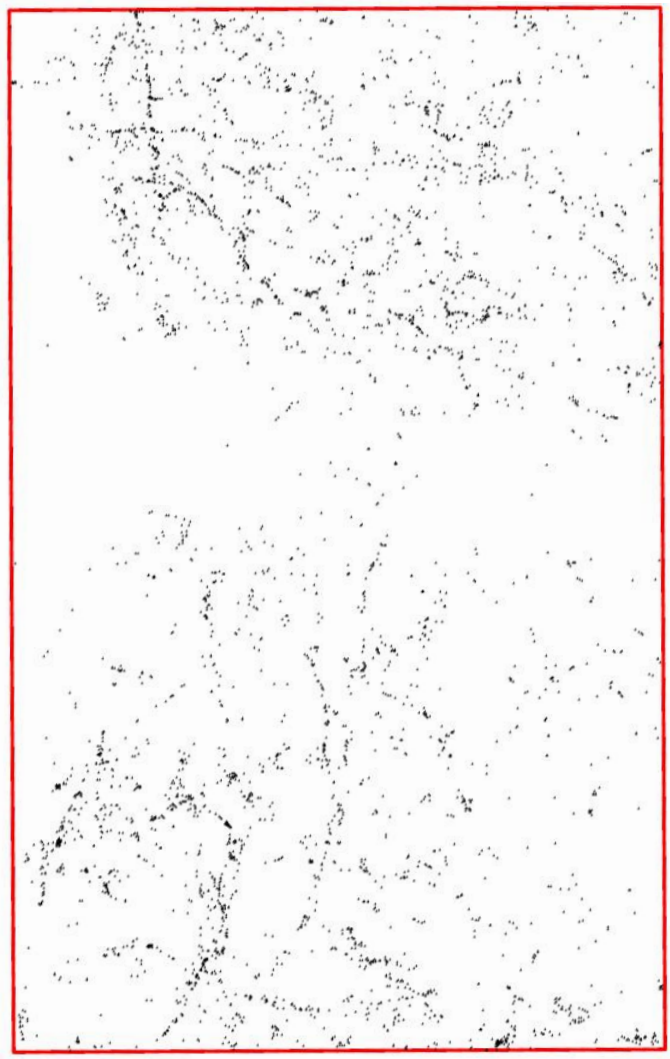
	V1	V2
Min.:	458705	6172511
Max.:	487620	6217922

UTM-koordinaterne er også direkte et udtryk for afstande, således at forskelle i koordinater er afstanden angivet i m. Området i Vestjylland, der afgrænser gravhøjene, er $28915 \times 45411\text{m}^2$ stort og afmærket på et landkort på Figur 1.1.



Figur 1.1: Afmærkning af det $28915 \times 45411\text{m}^2$ store område omkring Skjern å. Scan fra [Johansen, 2009].

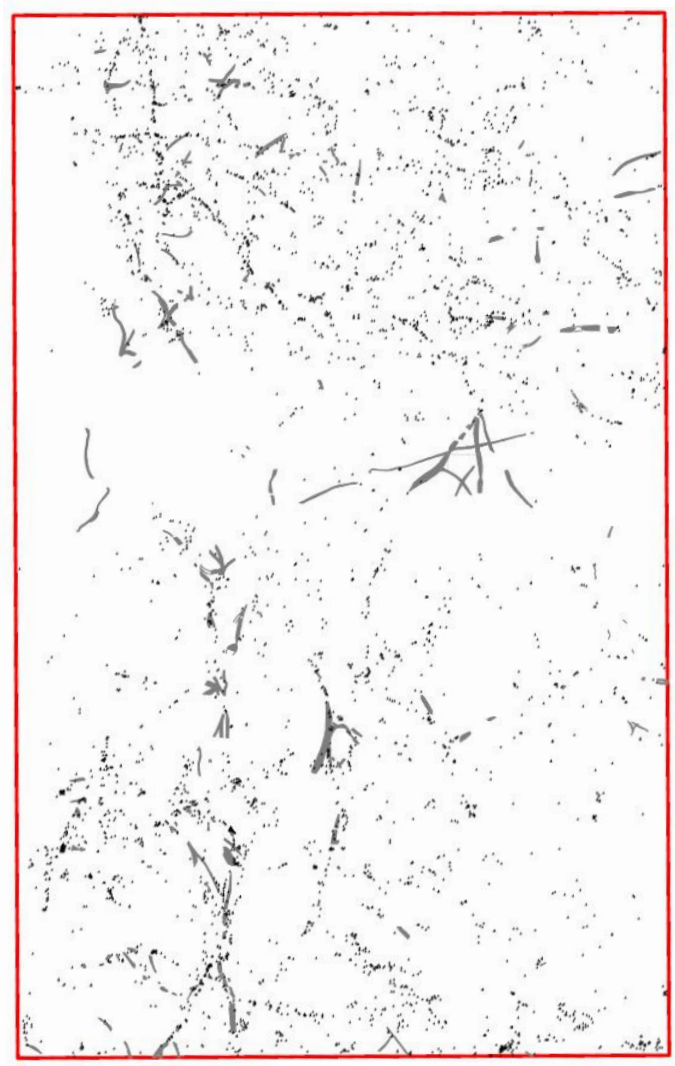
På Figur 1.2 ses markeringer af gravhøjene i området. Det ses på figuren, at der er linjetendenser blandt gravhøjene, som Sophus Müller allerede opdagede omkring år 1900. Desuden indikeres forløbet af Skjern å som det brede bånd på tværs af figuren, hvor intensiteten af gravhøje er mindre end i den resterende del af området.



Figur 1.2: Plot af koordinater for datasættet gravhøje. Scan fra [Johansen, 2009].

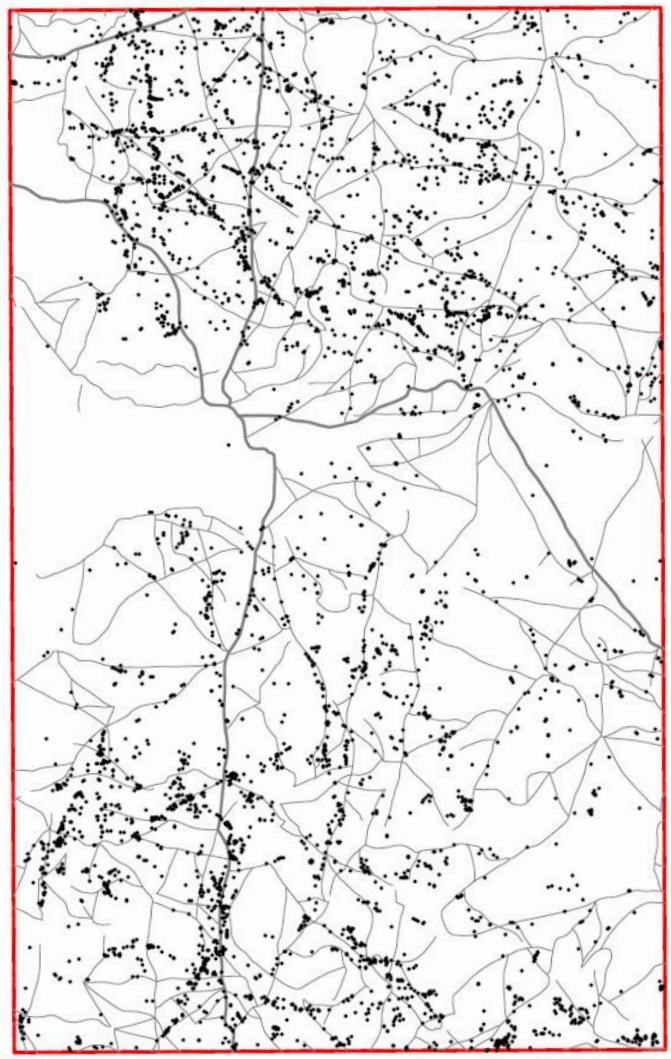
På Figur 1.3 ses gravhøjene sammen med de forhistoriske veje, der er kendte. Det ses, at de historiske veje nogenlunde er sammenfaldne med rækker af gravhøje, og Sophus Müllers teori er altså, at dette skyldes, at gravhøjene blev bygget omkring vejene. Disse historiske veje er en del af 'Høj og Vad'-projektet, men i datasættet

for gravhøjene eksisterer ikke information om, hvor disse veje ligger udover, hvad der kan ses på Figur 1.3.



Figur 1.3: Plot af datasættet for gravhøjene og de forhistoriske veje, som er markeret med mørke bånd. Scan fra [Johansen, 2009].

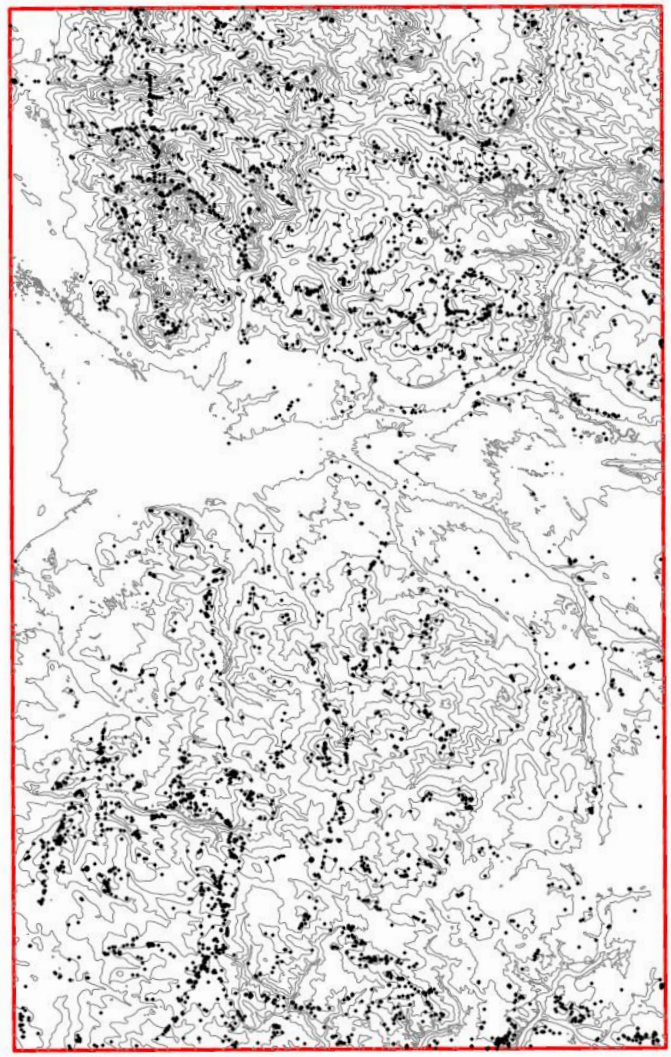
På Figur 1.4 ses gravhøjene sammen med de nutidige veje i området. Det ses, at vejene nogenlunde er sammenfaldne med rækker af gravhøje. Ved sammenligning af Figur 1.3 og Figur 1.4 kan det også ses, at der er stort sammenfald af nutidige og forhistoriske veje. Måske kan det endda formodes, at de nutidige veje hovedsageligt ligger, hvor vejene også lå i oldtiden, og der er således en eventuel sammenhæng mellem veje og gravhøje.



Figur 1.4: Plot af datasættet for gravhøjene og nutidige veje som er markeret med linjer. Scan fra [Johansen, 2009].

På Figur 1.5 ses gravhøjene sammen med højdekurver for området. Det ses tydeligt, hvor Skjern å forløber. Højdekurverne kan have betydning for, hvor gravhøjene ligger, da det ikke er muligt at bygge gravhøje på skråninger og i vådområder.

Yderligere eksisterer der i datasættet informationer om jordbundstyper og friktionen af jordbundstyperne for området. Informationer om jordbundstyperne kan som for højdekurverne have betydning for placeringen af gravhøje. Friktionen af jordbundstyperne giver information om, hvor let det er at færdes på de forskellige



Figur 1.5: Plot af datasættet for gravhøjene og højdekurver. Forløbetaf Skjern å ses tydeligt. Scan fra [Johansen, 2009].

jordbundstyper. Denne information kan benyttes til at vurdere, hvor det er mest sandsynligt, at folk færdedes i oldtiden, og hvor det var muligt at bygge gravhøje.

Der er altså visse geografiske begrænsninger på gravhøjenes placeringer, og disse kan også inddrages i forbindelse med opstilling af punktprocesser til beskrivelse af datasættet for gravhøje.

Problemformulering og projektafgrænsning

Punktmønstre der indeholder linjetendenser, kan ikke nødvendigvis beskrives ved allerede kendte punktprocesser. Det er derfor nødvendigt at opstille nye modeller, der kan beskrive linjetendenser. De foregående afsnit giver et indblik i mulighederne og problemstillingerne i forbindelse med opstilling af en model til beskrivelse af datasættet gravhøje. På baggrund af dette opstilles der følgende problemformulering for projektet:

> ***Der ønskes opstillet en statistisk model til beskrivelse af datasættet gravhøje. Specielt ønskes linjetendenserne i datasættet beskrevet, og dette ønskes gjort ved at formulere en punktprocesmodel, der inkluderer linjer af punkter. Dette ønskes udført ved benyttelse af teori om punkt- og linjeprocesser, hvor linjerne i linjeprocesserne skal repræsentere de forhistoriske veje. Datasættet for gravhøjene skal benyttes til at estimere parametre i modellen og manglende data i form af vejene. Derudover skal der laves modelkontrol for at se, om det er en rimelig antagelse, at gravhøjene skulle ligge langs veje. Informationen om hvor de gamle veje lå, er mangelfuld, men måske er det faktisk muligt at vurdere, hvor vejene kunne have ligget ved at sammenligne de estimerede veje med fund af hjulspor fra oldtiden?***

For at kunne behandle dette problem skal der introduceres en række teoretiske og anvendelsesmæssige begreber, og der opstilles følgende delmål for projektet:

- Opnå viden om punktprocesser og linjeprocesser.
- Tilpasse kendte punktprocesser på datasættet for gravhøje.
- Opstille en passende model, der kan beskrive linjetendenser i datasættet gravhøje og implementering af denne model i `R`.
- Udlede passende summary statistics for punktmønstre med linjetendenser udfra den opstillede model.
- Foretage et modeltjek af den opstillede model.

I selve opstillingen af modellen afgrænses der til et område på $15000 \times 15000\text{m}^2$, som er placeret i nederste venstre hjørne af Figur 1.2 med koordinaterne $[460000, 475000] \times [6175000, 6190000]$. Dette gøres for at arbejde med et kvadratisk område, hvor intensiteten af gravhøje er forholdsvis høj og ligeligt fordelt, og hvor linjetendenserne er specielt tydelige.

Der betragtes ikke vekselvirkning mellem punkter indenfor observationsvinduet og punkter udenfor.

I modellen antages det også, at der ikke eksisterer ikke-observerede gravhøje, altså at der ikke er manglende data i form af ikke-kendte gravhøje. Det eneste ukendte i forbindelse med modelstillingen er således vejene.

I den opstillede model afgrænses der til at punkterne ligger langs rette linjer, der gennemskærer hele observationsvinduet fremfor buede linjer.

Der opstilles kun en model for gravhøjene uden information fra kovariaterne højdeforskelle, jordbundstyper og friktionen af jordbundstyperne.

Rapportstruktur

Rapporten indeholder teori, eksempler på anvendelse af teorien, data-beskrivelse og -behandling, modelopstilling, resultater fra modellen, diskussion og konklusion. For at skabe en overordnet struktur er rapporten derfor inddelt i følgende fire hoveddele:

DEL I Teori

Denne del af rapporten indeholder Kapitel 2–Kapitel 8. Her gives der en introduktion til punktprocesser, linjeprocesser og behandling af punktmønstre.

DEL II Databehandling

Denne del af rapporten indeholder Kapitel 9–Kapitel 12. Her gives der en beskrivelse af data med de opnåede redskaber, der tilpasses eksisterende punktprocesser på data, og der opstilles en model til beskrivelse af datasættet, der inkluderer linjetendenser.

DEL III Epilog

Denne del af rapporten indeholder en opsummering af de opnåede resultater samt konklusion, og afslutningsvis diskuteres forbedringer af modellen og andre måder at anskue problemet på.

DEL IV Appendiks

Denne del af rapporten indeholder appendiks med bl.a. kildekode til den opstillede model.

Del I

Teori

Punktprocesser

I dette kapitel gives en kort introduktion til punktmønstre og rumlige punktprocesser i $\mathbb{R}^d$, $d \in \mathbb{N}$. Desuden gennemgås kort den nødvendige målteori, der danner grundlag for beskrivelsen af punktprocesser. Afslutningsvis gives der en matematisk definition af rumlige punktprocesser, og forskellige resultater vedrørende denne behandles. Kapitlet er skrevet på baggrund af [Møller og Waagepetersen, 2004, Kapitel 1 og 2, Appendiks A og B].

2.1 Introduktion

Et punktmønster kan repræsentere mange forskellige slags hændelser. Eksempler på sådanne hændelser kan være tidspunkter for jordskælv observeret i en bestemt tidsperiode, koordinater for træer i en skov eller placeringen af gravhøje i et landområde. Når der analyseres punktmønstre, betragtes ofte kun en del af punktmønstret, eksempelvis indenfor et begrænset område kaldet et observationsvindue, og beskrivelsen af punktmønstret indenfor observationsvinduet kan så i visse tilfælde generaliseres til hele punktmønstret.

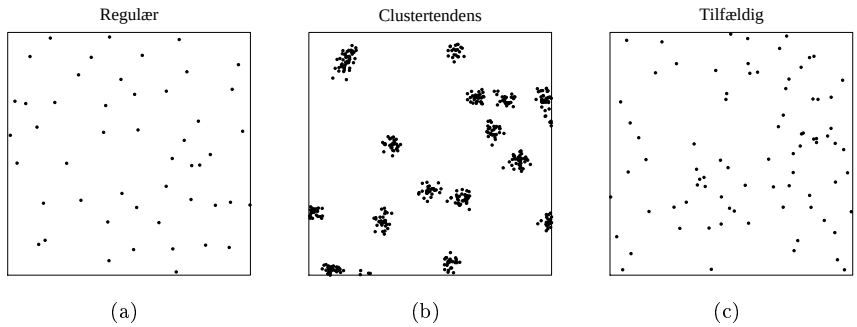
Rumlige punktprocesser eller bare punktprocesser er statistiske modeller, der anvendes til at beskrive rumlige punktmønstre, som i denne sammenhæng er et observeret datasæt. Disse typer af statistiske modeller er dermed anvendelige i forbindelse med analysering af data, der kan betragtes som punktmønstre. En punktproces beskriver, hvordan punkterne i et punktmønster er fordelt i observationsvinduet og med en realisation af en punktproces, menes et specifikt punktmønster. Det vil sige, at forskellige punktmønstre kan opnås fra den samme punktproces. Betegnelsen rumlig henviser til, at en punktproces i realiteten kan beskrive punktmønstre i $\mathbb{R}^d$ for $d \in \mathbb{N}$, men typisk arbejdes der kun med punktprocesser i $\mathbb{R}$, $\mathbb{R}^2$ eller $\mathbb{R}^3$. Ønskes der mere information end blot punkternes position i $\mathbb{R}^d$, kan en punktproces udvides ved at tilknytte et mærke til hvert punkt. En sådan udvidelse kaldes en mærket punktproces, og for gravhøjene kunne et mærke eventuelt repræsentere gravhøjens alder.

Ofte er der vekselvirkning mellem punkterne i et punktmønster i form af tiltrækning eller frastødning mellem dem. Et punktmønster siges at være regulært, hvis punkterne i det er jævnt fordelt over hele observationsvinduet. Regularitet er således et tegn på frastødning mellem punkter i et punktmønster som vist på Figur 2.1(a). Tiltrækning mellem punkter i et punktmønster ses ofte ved at punkter optræder tæt sammen i visse områder. Specielt, hvis der er tegn på, at punkterne klumper sig sammen i grupper, siges punktmønstret at have clustertendenser, hvilket er vist på Figur 2.1(b). På Figur 2.1(c) ses et eksempel på et punktmønster, hvor der ikke er nogen vekselvirkning mellem punkterne i observationsvinduet.

Ved analyse af punktmønstre kan vekselvirkningen mellem punkter give komplikationer, idet punkterne indenfor observationsvinduet sandsynligvis også vekselvirker

med punkter uden for observationsvinduet. Hvis der ikke tages højde for dette i forbindelse med valget af en punktproces, kan dette forårsage såkaldte kanteffekter.

Punktprocesser kan yderligere være homogene eller inhomogene. Hvis intensiteten af punkter i et punktmønster er jævnt fordelt over hele observationsområdet siges punktmønstret at være homogent. Inhomogenitet i et punktmønster betyder tilsvarende, at intensiteten af punkter er større i nogle områder end andre. Inhomogenitet kan skyldes mange ting og hænger ofte sammen med kovariater for punktmønstret, og kan for træerne i en skov eksempelvis skyldes jordbundsforhold. Det er ikke altid muligt visuelt at afgøre, hvorvidt et punktmønster eksempelvis er inhomogent eller har clustertendenser. For at kunne bestemme dette er det nødvendigt at have baggrundsviden om punktmønstret.



Figur 2.1: Eksempler på punktmønstre. (a) Regulært punktmønster. (b) Punktmønster med clustertendenser. (c) Punktmønster uden vekselvirkning mellem punkter.

Afhængig af hvordan punkterne i et punktmønster er fordelt, eksisterer der forskellige anvendelige punktprocesser til beskriveslse af punkternes fordeling. Eksempler på punktprocesser er Poisson-processer, Cox-processer og Markov-processer. Den grundlæggende teori for de forskellige punktprocesser er dog ens og i følgende to afsnit introduceres basale begreber inden for punktproces-teorien.

Punktprocesser kan udvides til linjeprocesser, hvor punkterne er erstattet af linjer, således at en realisering af en linjeproces er et observationsvindue med linjer i forskellige retninger. Linjeprocesser kan ses som punktprocesser, hvor punkterne beskriver linjer. Den generelle teori om punktprocesser kan derfor også benyttes i forbindelse behandling af linjeprocesser.

2.2 Målteori

For at kunne give en matematisk beskrivelse af punktprocesser er det nødvendigt at benytte resultater fra målteorien. I dette afsnit introduceres kort de målteoretiske begreber, der danner fundamentet for dette projekt. Afsnittet er baseret på [Berg og Madsen, 1970, §1-3].

Definition 2.1 (σ-algebra)
Lad $X \subseteq \mathbb{R}^d$. Et system $\mathcal{A}$ af delmængder af en mængde X kaldes en σ-algebra i X, hvis

(i) $X \in \mathcal{A}$.

(ii) $X \backslash A \in \mathcal{A}$, *når* $A \in \mathcal{A}$.

(iii) $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A}$, *når* $A_1, A_2, ... \in \mathcal{A}$.

Systemet givet ved $\{Y | Y \subseteq X\}$ kaldes den største σ-algebra i X, mens systemet bestående af $\{\emptyset, X\}$ kaldes for den mindste σ-algebra i X.

Når der til en mængde X er valgt en σ-algebra $\mathcal{A}$, kaldes parret $(X, \mathcal{A})$ et målbart rum. Mængderne i $\mathcal{A}$ siges at være målelige ($\mathcal{A}$-målelige).

Definition 2.2 (Borel-algebra)
Den mindste σ-algebra i $\mathbb{R}^d$, der indeholder alle åbne mængder i $\mathbb{R}^d$, kaldes for Borel-algebraen i $\mathbb{R}^d$, og noteres $\mathcal{B}(\mathbb{R}^d)$.

Elementerne af Borel-algebraen kaldes Borel-mængder og Borel-algebraen består således af alle Borel-mængder. Jævnfør [Berg og Madsen, 1970, s. 21] kan det antages, at næsten alle mængder $A \subseteq \mathbb{R}^d$ er Borel-mængder.

Definition 2.3 (Mål)
Et mål er en funktion $\mu : \mathcal{A} \to [0, \infty]$ med egenskaberne

(i) $\mu(\emptyset) = 0$.

(ii) $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$, *når* $A_1, A_2, \ldots \in \mathcal{A}$ *er parvis disjunkte.*

En mængde X, hvortil der hører en σ-algebra $\mathcal{A}$ og et mål $\mu(\cdot)$ kaldes et målrum og betegnes $(X, \mathcal{A}, \mu(\cdot))$. Tallet $\mu(A)$ kaldes for målet af A, hvor $A \in \mathcal{A}$.

En egenskab ved mål er

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(A_i), \tag{2.1}$$

for vilkårlige $A_1, A_2, \ldots \in \mathcal{A}$.

Som eksempel på et mål kan nævnes Lebesgue-målet og tællemålet.

Definition 2.4 (Lebesgue-målet)
Målet $\nu(\cdot)$ defineret på Borel-algebraen i $\mathbb{R}^d$, hvis værdi for enhver mængde $A \subseteq \mathbb{R}^d$, er givet ved størrelsen (længden/arealet/volumen) af mængden, kaldes Lebesgue-målet.

Lebesgue-målet er entydigt jf. [Berg og Madsen, 1970, s. 36].

Definition 2.5 (Tællemål)
Funktionen $n(\cdot)$ defineret på alle målelige delmængder af en vilkårlig mængde X ved:

$$n(A) = \begin{cases} \text{antal elementer i } A, & \text{når } A \subseteq X \text{ er endelig,} \\ \infty, & \text{når } A \subseteq X \text{ er uendelig.} \end{cases}$$

er et mål, kaldet tællemålet, i X.

Definition 2.6 (Målelig afbildning)
Lad $(X, \mathcal{A})$ og $(Y, \mathcal{F})$ være målbare rum. En afbildning $\varphi : (X, \mathcal{A}) \to (Y, \mathcal{F})$ kaldes målelig ($\mathcal{A} - \mathcal{F}$-målelig), hvis

$$\varphi^{-1}(F) \in \mathcal{A},$$

for alle $F \in \mathcal{F}$.

Ligesom det oftest kan antages, at alle mængder er Borel-mængder, kan det også i de fleste tilfælde antages, at alle afbildninger er målelige.

Definition 2.7 (Invarians af mål)
Lad $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ være en bijektiv afbildning, hvor både $\varphi(\cdot)$ og $\varphi^{-1}(\cdot)$ er målelige, og lad $\mathbb{R}^d$ være udstyret med Borel-algebraen $\mathcal{B}$, dvs.

$$B \in \mathcal{B} \Leftrightarrow \varphi(B) \in \mathcal{B}.$$

Et mål $\mu(\cdot)$ siges at være invariant under $\varphi(\cdot)$, hvis billedmålet $\varphi(\mu(\cdot))$ er lig med $\mu(\cdot)$, dvs. hvis

$$\mu(\varphi^{-1}(B)) = \mu(B) \text{ for alle } B \in \mathcal{B}.$$

Hvis funktionen $\varphi(\cdot)$ er en translation, er målet invariant under translation, og målet siges så at være stationært. Et eksempel på et invariant mål er Lebesgue-målet.

I den resterende del af rapporten benyttes kun Borel-mængder, dvs. at mængder A, B og C i det følgende vil betegne Borel-mængder, selvom der ikke hver gang gøre opmærksom på dette. På samme måde betegner funktioner målelige afbildninger.

2.3 Punktprocesser

Teorien om punktprocesser til beskrivelse af punktmønstre, vandt for alvor frem i løbet af 1960'erne og 1970'erne med arbejde udført af bl.a. Sir David Cox og Brian D. Ripley. Senere har mange andre været med til at skabe den viden og de resultater, der i dag eksisterer om punktprocesser. Gennem tiderne har der også eksisteret forskellig software til behandling af punktprocesser. Et statistikprogram til behandling af rumlige punktprocesser er `R` [R Development Core Team, 2008] med pakken `spatstat`, specielt udviklet til rumlige punktprocesser.

Ved benyttelse af ovenstående målteori er det muligt at give en matematisk definition af en punktproces.

Definition 2.8 (Punktproces)
Lad $S \subseteq \mathbb{R}^d$, og lad $S \in \mathcal{B}(\mathbb{R}^d)$. En punktproces X på S er en stokastisk tællelig delmængde af S.

Med en stokastisk tællelig delmængde i ovenstående definition menes der, at enhver realisering af X er tællelig.

Laf x være en realisering af en punktproces X på $S \subseteq \mathbb{R}^d$. For ethvert $x \subseteq S$ angiver tællemålet $n(x)$ antallet af elementer tilhørende mængden x jf. Definition 2.5. Tællefunktionen $N : S \to \mathbb{N}_0$ for en punktproces X på S givet ved $N(B) = n(X \cap B) = n(X_B)$ angiver et stokastisk antal af punkter fra X, der falder i $B \subseteq S$, når B er begrænset.

Definition 2.9 (Lokalt endelig)
Lad $B \subseteq S$ og $x \subseteq S$, da siges mængden x at være lokalt endelig, hvis $n(x_B) < \infty$, når B er begrænset.

En punktproces X kaldes lokalt endelig, hvis alle begrænsede delmængder af realiseringer af X indeholder endeligt mange punkter med sandsynlighed en. Dermed tilhører alle realiseringer af X mængden

$$N_{lf} = \{x \subseteq S | n(x_B) < \infty \text{ for alle begrænsede mængder } B \subseteq S\},$$

hvor elementerne i N_{lf} kaldes for de lokalt endelige punktmønstre.

I denne rapport betragtes udelukkende lokalt endelige punktmønstre og punktprocesser.

Indenfor punktprocesser benyttes ofte et mål for det stokastiske antal af punkter, der i middel er indeholdt i en begrænset mængde B.

Definition 2.10 (Intensitetsmål)
Lad $X \subseteq S$ være en punktproces. Intensitetsmålet $\mu(\cdot)$ på S er da givet ved

$$\mu(B) = \mathbb{E}\left[N(B)\right], \quad B \subseteq S.$$

Definition 2.11 (Intensitetsfunktion)
Hvis intensitetsmålet $\mu(\cdot)$ kan skrives som

$$\mu(B) = \int_B \rho(\xi)\ \mathrm{d}\xi, \quad B \subseteq S,$$

hvor $\rho : S \to [0, \infty)$ er lokalt integrabel, så kaldes $\rho(\cdot)$ for intensitetsfunktionen eller blot intensiteten af X.

Udtrykket $\rho(\xi)\mathrm{d}\xi$, hvor $\xi \in X$, kan fortolkes som sandsynligheden for, at der er et punkt i en infinetisimal kugle med centrum i ξ og volumen $\mathrm{d}\xi$.

Det er muligt at identificere enhver punktproces ved det lokalt endelige stokastiske tællemål $N(\cdot)$. Helt generelt vil det sige, at ethvert lokalt endeligt stokastisk tællemål kan ses som en punktproces. [Møller og Waagepetersen, 2004, s. 242].

En punktproces siges at være simpel, hvis der ikke er punkter, der falder oven i hinanden, dvs. at $\mu(\{\xi\})$ enten er nul eller en. I denne rapport betragtes udelukkende simple punktprocesser.

Definition 2.12 (Homogene og inhomogene punktprocesser)
Hvis intensiteten $\rho(\cdot)$ er konstant, siges X at være en homogen punktproces. Ellers kaldes X en inhomogen punktproces.

For en homogen punktproces betegner ρ det gennemsnitlige antal punkter pr. enhedsvolume.

Definition 2.13 (Fordeling af en punktproces)
Den endeligt dimensionale fordeling af de stokastiske tællefunktioner for punktprocessen X, kaldes for fordelingen af punktprocessen X.

Det vil sige, at fordelingen af punktprocessen X er givet ved den simultane fordeling af $N(B_1), \ldots, N(B_m)$ for alle begrænsede $B_1, \ldots, B_m$ med $m \in \mathbb{N}_0$.

Definition 2.14 (Stationaritet og isotropi)
Lad X være en punktproces på $\mathbb{R}^d$.

(i) Punktprocessen X kaldes stationær, hvis dens fordeling er invariant under translation, dvs. hvis X og $X + s = \{\xi + s | \xi \in X\}$ har samme fordeling for ethvert $s \in \mathbb{R}^d$.

(ii) Punktprocessen X kaldes isotropisk, hvis dens fordeling er invariant under rotation omkring origo i $\mathbb{R}^d$, dvs. hvis X og $\mathcal{O}X = \{\mathcal{O}\xi | \xi \in X\}$ har samme fordeling for enhver rotation $\mathcal{O}$ om origo.

Stationaritet og isotropi for en punktproces medfører, at tællemålet og dermed intensitetsmålet er invariant under hhv. translation og rotation.

Ligesom det er muligt at beskrive en punktproces ved at betragte fordelingen af punkterne, er det også muligt at beskrive punktprocessen ved at betragte de mængder, hvor der ikke er nogen punkter.

Definition 2.15 (Voidsandsynlighed)
Mængder på formen

$$F_B = \{x \in N_{lf} | n(x_B) = 0\},$$

hvor B er begrænset, kaldes for voidhændelser. Punktprocessen $X \in F_B$, hvis og kun hvis $N(B) = 0$. Voidsandsynligheden er defineret ved

$$v(B) = \mathbb{P}(N(B) = 0),$$

hvor B er begrænset.

Voidsandsynligheden kan benyttes til at bestemme en punktproces. Dette leder til følgende sætning, hvis bevis kan findes i [Møller og Waagepetersen, 2004, s. 244].

Sætning 2.16 [Møller og Waagepetersen, 2004, Theorem B.1]
En punktproces er entydigt bestemt ud fra voidsandsynligheden.

Den generelle punktproces og enkelte grundlæggende begreber er nu defineret. For at opsummere er der altså flere måder at karakterisere en punktproces på, nemlig ved en lokal endelig stokastisk delmængde, et stokastisk tællemål, en intensitetsfunktion eller en voidsandsynlighed.

I det følgende kapitel betragtes en bestemt type af disse punktprocesser, kaldet for Poisson-processer. Her eksemplificeres den allerede beskrevne punktprocesteori, ved en forholdsvis simpel punktproces.

Poisson-processer

Poisson-processer er en type af modeller til beskrivelse af fuldstændig rumlig tilfældighed i rumlige punktmønstre, dvs. når der ikke er nogen vekselvirkning mellem punkterne. Derudover benyttes de som referenceprocesser, når der studeres summary statistics og mere avancerede punktprocesser som f.eks. Cox-processer.

Kapitlet indledes med en generel definition af Poisson-processen, samt en redegørelse for under hvilke betingelser en Poisson-proces eksisterer. Desuden indeholder kapitlet resultater, der er anvendelige i beskrivelsen af Poisson-linjeprocesser og i modelopstillingen for gravhøje, herunder superposition af Poisson-processer og likelihoodfunktion for et observeret datasæt, hvor punkterne antages at følge en Poisson-proces. Kapitlet er skrevet på baggrund af [Møller og Waagepetersen, 2004, Kapitel 3].

3.1 Definitioner og resultater

I det følgende betragtes punktprocesser, der er defineret på et rum $S \subseteq \mathbb{R}^d$ og specificeret ved intensitetsfunktionen $\rho : S \to [0, \infty)$, som er lokalt integrabel for alle begrænsede $B \subseteq S$. For at kunne definere Poisson-processen er det nødvendigt først at definere en binomialproces, som er en punktproces, der beskriver ensfordelte punkter.

Definition 3.1 (Binomialproces)
Lad $f(\cdot)$ være en tæthedsfunktion på en mængde $B \subseteq S$, og lad $n \in \mathbb{N}$. En punktproces X bestående af n uafhængige og ensfordelte punkter med tæthed $f(\cdot)$ kaldes en binomialproces med n punkter i B med tæthed $f(\cdot)$.
At en punktproces er en binomialproces noteres $X \sim \text{Binomial}(B, n, f(\cdot))$.

Definition 3.2 (Poisson-proces)
En punktproces X på S kaldes en Poisson-proces med intensitetsfunktion $\rho(\cdot)$, hvis følgende egenskaber er opfyldte:

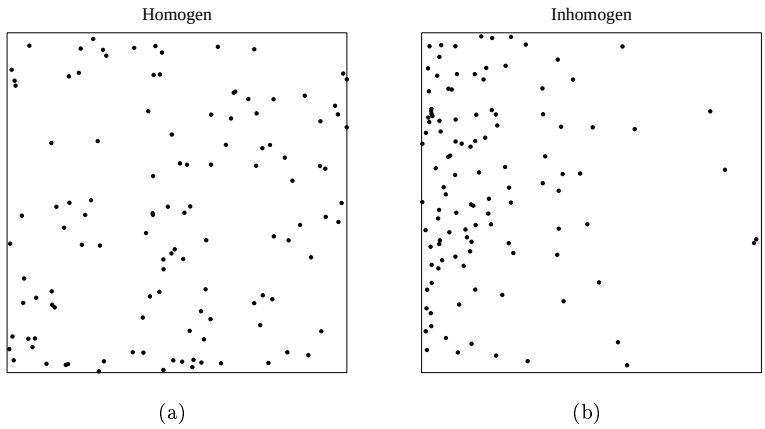
(i) For enhver mængde $B \subseteq S$ med $\mu(B) < \infty$ er antallet af punkter i B Poisson-fordelt med middelværdi $\mu(B)$, dvs. $N(B) \sim poi(\mu(B))$.

(ii) For ethvert $n \in \mathbb{N}$ og $B \subseteq S$ med $0 < \mu(B) < \infty$ gælder der, at betinget med $N(B) = n$ er $X_B \sim \text{Binomial}(B, n, f(\cdot))$, hvor $f(\xi) = \rho(\xi)/\mu(B)$ med $\xi \in B$.

At en punktproces er en Poisson-proces noteres $X \sim \text{Poisson}(S, \rho(\cdot))$.

Hvis ρ er konstant, kaldes Poisson(S, ρ) en homogen Poisson-proces på S med intensitet ρ og ellers kaldes den for en inhomogen Poisson-proces. Hvis en punktproces er Poisson$(S, 1)$ kaldes den for standard Poisson-processen på S. Fra Definition 2.14 ses det, at en homogen Poisson-proces på $\mathbb{R}^d$ er både stationær og isotropisk.

På Figur 3.1 ses to eksempler på realiseringer af Poisson-processer på observationsvinduet $S = [0, 1] \times [0, 1]$. (a) Illustrerer en realisering af en homogen Poisson-proces med intensiteten $\rho = 100$ og (b) viser en realisering af en inhomogen Poisson-proces med intensitetsfunktionen $\rho(x, y) = 500 \exp(-5x)$ for $(x, y) \in S$. Det ses, at punkterne for den homogene Poisson-proces er jævnt fordelt over hele observationsvinduet, mens der for den inhomogene er en tendens til, at punkter ligger længst mod venstre.



Figur 3.1: Eksempler på Poisson-processer på $S = [0, 1] \times [0, 1]$. (a) Homogen Poisson-proces med intensitet $\rho = 100$. (b) Inhomogen Poisson-proces med intensitet $\rho(x, y) = 500 \exp(-5x)$ for $(x, y) \in S$.

Voidsandsynligheden fra Definition 2.15 fås for en Poisson-proces til

$$v(B) = \exp(-\mu(B)) \frac{\mu(B)^0}{0!} = \exp(-\mu(B)), \tag{3.1}$$

da antallet af punkter i en Poisson-proces er Poisson-fordelt.

3.1.1 Eksistens af Poisson-processer

Det ses af det følgende, at en Poisson-proces altid vil eksistere på S. Først vises eksistensen for et begrænset S.

Sætning 3.3 [Møller og Waagepetersen, 2004, s. 16]
Lad X være en punktproces på S med $n(X) \sim poi(\mu(S))$ og $X|n(X) = n \sim Binomial(S, n, \rho(\cdot)/\mu(S))$. Hvis S er begrænset, eller hvis blot $\int_S \rho(\xi)d\xi < \infty$, da vil der altid eksistere en $Poisson(S, \rho(\cdot))$.

Bevis:
Det skal vises, at punkt (i) og (ii) i Definition 3.2 er opfyldt for enhver delmængde $B \subseteq S$. Lad $F \in N_{lf}$ og dermed fås

$$\begin{aligned}
\mathbb{P}(X_B \in F, &n(X_B) = k) \\
&= \mathbb{E}\left[\mathbb{1}[X_B \in F, n(X_B) = k]\right] \\
&= \sum_{n=0}^{\infty} \mathbb{E}\left[\mathbb{1}[X_B \in F, n(X_B) = k] | n(X_S) = n\right] \mathbb{P}(n(X_S) = n) \\
&= \exp(-\mu(B)) \frac{\mu(B)^k}{k!} \int_{B^k} \mathbb{1}[x_1, \ldots, x_k \in F] \prod_{i=1}^{k} \frac{\rho(x_i)}{\mu(B)} \, \mathrm{d}x_1 \ldots \, \mathrm{d}x_k \\
&= \mathbb{P}(n(X_B) = k)\mathbb{P}(X_B \in F | n(X_B) = k),
\end{aligned}$$

hvor andet lighedstegn følger af loven om total middelværdi og fjerde lighedstegn følger af antagelser fra sætningen. Heraf ses det, at $n(X_B) \sim \text{poi}(\mu(B))$ og $X_B | n(X_B) = k \sim \text{Binomial}(B, k, \rho(\cdot)/\mu(B))$. □

Hvis S derimod er ubegrænset, følger det af nedenstående sætning, at der eksisterer en Poisson-proces.

Sætning 3.4 [Møller og Waagepetersen, 2004, Theorem 3.1]
En Poisson-proces, $Poisson(S, \rho(\cdot))$, eksisterer og er entydig bestemt ud fra voidsandsynligheden

$$v(B) = \exp(-\mu(B)),$$

for begrænset $B \subseteq S$.

Bevis:
Lad $\xi \in S$ være et arbitrært punkt, og lad $B_i = \{\eta \in S | i - 1 \leq |\eta - \xi| < i\}$ for $i \in \mathbb{N}$. Mængden S er da en disjunkt forening af de begrænsede B_i. Lad $X = \bigcup_{i=1}^{\infty} X_i$, hvor $X_i \sim \text{Poisson}(B_i, \rho_i(\cdot))$ for $i = 1, 2, \ldots$ er uafhængige, og hvor $\rho_i(\cdot)$ er restriktionen af $\rho(\cdot)$ på B_i. At X eksisterer følger af, at den er en forening af Poisson-processer på begrænsede mængder, som ifølge Sætning 3.3 altid eksisterer.

Dermed er voidsandsynligheden for X, for begrænsede $B \subseteq S$ givet ved

$$\begin{aligned} v(B) =& \mathbb{P}(X \cap B = \emptyset) = \prod_{i=1}^{\infty} \mathbb{P}(X_i \cap B = \emptyset) = \prod_{i=1}^{\infty} \mathbb{P}(N(B_i \cap B) = 0) \\ =& \prod_{i=1}^{\infty} \exp(-\mu(B \cap B_i)) = \exp\left(-\mu\left(\bigcup_{i=1}^{\infty} B \cap B_i\right)\right) = \exp(-\mu(B)). \end{aligned}$$

Ovenstående er på formen af voidsandsynligheden for en Poisson-proces jf. (3.1). At Poisson-processen X er entydig fås af Sætning 2.16. □

Denne sætning kan udvides til uafhængige Poisson-processer, dvs. hvis X_1 og X_2 er uafhængige Poisson-processer på S med intensitetsmål $\mu_1(\cdot)$ og $\mu_2(\cdot)$, så er fordelingen af (X_1, X_2) entydigt bestemt ud fra voidsandsynligheden

$$\begin{aligned} \mathbb{P}(X_1 \cap A = \emptyset, X_2 \cap B = \emptyset) &= \mathbb{P}(X_1 \cap A = \emptyset)\mathbb{P}(X_2 \cap B = \emptyset) \\ &= \exp(-\mu_1(A) - \mu_2(B)) \end{aligned}$$

for begrænsede $A, B \subseteq S$.

3.1.2 Poisson-udvikling

Ofte benyttes følgende sætning i forbindelse med beviser angående Poisson-processer.

Sætning 3.5 (Poisson-udvikling) [Møller og Waagepetersen, 2004, Proposition 3.1]
Lad X være en punktproces på S. Da gælder der

(i) $X \sim$ Poisson$(S, \rho(\cdot))$, hvis og kun hvis der for alle $B \subseteq S$ med $\mu(B) = \int_B \rho(\xi)d\xi < \infty$ og alle mængder $F \subseteq N_{lf}$ gælder, at

$$\mathbb{P}(X_B \in F) = \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_{B^n} \mathbb{1}[\{x_1, \ldots, x_n\} \in F] \prod_{i=1}^{n} \rho(x_i) dx_1 \ldots dx_n, \tag{3.2}$$

hvor integralet for $n = 0$ skal læses som $\mathbb{1}[\emptyset \in F]$.

(ii) Hvis $X \sim$ Poisson$(S, \rho(\cdot))$, så gælder der for positive og målelige funktioner $h : N_{lf} \to [0, \infty)$ og $B \subseteq S$ med $\mu(B) < \infty$, at

$$\mathbb{E}[h(X_B)] = \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_{B^n} h(\{x_1, \ldots, x_n\}) \prod_{i=1}^{n} \rho(x_i) dx_1 \ldots dx_n.$$

Bevis:
(i) Antag, at $X \sim \text{Poisson}(S, \rho(\cdot))$, da gælder det ifølge Definition 3.2, at

$$\begin{aligned}
\mathbb{P}(X_B \in F) &= \sum_{n=0}^{\infty} \mathbb{P}(X \cap B \in F | N(B) = n)\mathbb{P}(N(B) = n) \\
&= \sum_{n=0}^{\infty} \exp(-\mu(B)) \frac{\mu(B)^n}{n!} \int_{(X \cap B \in F)^n} \prod_{i=1}^{n} \frac{\rho(x_i)}{\mu(B)} \, \mathrm{d}x_1 \ldots \, \mathrm{d}x_n \\
&= \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_{B^n} \mathbb{1}[\{x_1, \ldots, x_n\} \in F] \prod_{i=1}^{n} \rho(x_i) \, \mathrm{d}x_1 \ldots \, \mathrm{d}x_n.
\end{aligned}$$

Antag dernæst, at $\mathbb{P}(X_B \in F)$ er givet som i (3.2) for alle $F \subseteq N_{lf}$. Specielt gælder der for $F_B \in F$ jf. Definition 2.15, at

$$\mathbb{P}(X_B \in F_B) = \mathbb{P}(n(X_B) = 0) = \exp(-\mu(B)),$$

og ifølge Sætning 3.4 er X dermed en Poisson-proces med intensitetsmål $\mu(B) = \int_B \rho(\xi) \, \mathrm{d}\xi$.

(ii) Fremgangsmåden er at benytte et resultat fra integrationsteorien, om at enhver positiv, målelig funktion jf. [Berg og Madsen, 1970, s. 45] kan fås som en grænsefunktion af en voksende følge af simple målelige funktioner. I forbindelse med punktprocesser kaldes bevismetoden også standardbeviset. Fra (i) fås

$$\begin{aligned}
\mathbb{E}[\mathbb{1}[X_B \in F]] &= \mathbb{P}(X_B \in F) \\
&= \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_{B^n} \mathbb{1}[\{x_1, \ldots, x_n\} \in F] \prod_{i=1}^{n} \rho(x_i) \, \mathrm{d}x_1 \ldots \, \mathrm{d}x_n.
\end{aligned}$$

Da $h(\cdot)$ er positiv og målelig, gælder at $h_m(x) = \sum_{i=1}^{m} a_{i,m} \mathbb{1}[x \in F_{i,m}]$ og at $\lim_{m \to \infty} h_m(x) = h(x)$. Dermed fås

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{m} a_{i,m} \mathbb{1}[X_B \in F_{i,m}]\right] &= \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \\
&\quad \int_{B^n} h_m(\{x_1, \ldots, x_n\}) \prod_{i=1}^{n} \rho(x_i) \, \mathrm{d}x_1 \ldots \, \mathrm{d}x_n.
\end{aligned}$$

Hernæst beregnes $\mathbb{E}[h(X_B)]$ vha. ovenstående til

$$\begin{aligned}
\mathbb{E}[h(X_B)] &= \lim_{m \to \infty} \mathbb{E}[h_m(X_B)] \\
&= \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_{B^n} \lim_{m \to \infty} h_m(\{x_1, \ldots, x_n\}) \prod_{i=1}^{n} \rho(x_i) \, \mathrm{d}x_1 \ldots \, \mathrm{d}x_n \\
&= \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_{B^n} h(\{x_1, \ldots, x_n\}) \prod_{i=1}^{n} \rho(x_i) \, \mathrm{d}x_1 \ldots \, \mathrm{d}x_n.
\end{aligned}$$

□

3.1.3 Uafhængige Poisson-processer

Betingelse (ii) i Definition 3.2 kan erstattes af den betingelse, at $N(B_1), \ldots, N(B_n)$ er uafhængige for disjunkte mængder $B_1, \ldots, B_n \subseteq S$ og $n \geq 2$. Denne egenskab kaldes uafhængig spredning og vises af det følgende.

Sætning 3.6 [Møller og Waagepetersen, 2004, Proposition 3.2]
Hvis X er en Poisson-proces på S, så er $X_{B_1}, X_{B_2}, \ldots$ uafhængige Poisson-processer, for disjunkte mængder $B_1, B_2, \ldots \subseteq S$.

Bevis:
Dette bevises ved induktion. Antag, at $X \sim \text{Poisson}(S, \rho(\cdot))$, og antag at $F_i \in N_{lf}$, for $i = 1, 2, \ldots$.

Basistrin: Lad $n = 2$, dvs. lad $B_1, B_2 \subseteq S$ være begrænsede og disjunkte mængder, og lad $B = B_1 \cup B_2$. Da er

$$\begin{aligned}
&\mathbb{P}(X_{B_1} \in F_1, X_{B_2} \in F_2) \\
=&\sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_{B^n} \mathbb{1}[X_{B_1} \in F_1, X_{B_2} \in F_2] \prod_{i=1}^{n} \rho(x_i)\, \mathrm{d}x_1 \ldots \mathrm{d}x_n \qquad (3.3) \\
=&\sum_{k=0}^{\infty} \frac{\exp(-\mu(B_1))}{k!} \int_{B_1^k} \mathbb{1}[\{x_1, \ldots, x_k\} \in F_1] \prod_{i=1}^{k} \rho(x_i)\, \mathrm{d}x_1 \ldots \mathrm{d}x_k \\
&\sum_{n=k}^{\infty} \frac{\exp(-\mu(B_2))}{(n-k)!} \int_{B_2^{n-k}} \mathbb{1}[\{x_{k+1}, \ldots, x_n\} \in F_2] \prod_{j=k+1}^{n} \rho(x_j)\, \mathrm{d}x_{k+1} \ldots \mathrm{d}x_n \\
=&\mathbb{P}(X_{B_1} \in F_1)\mathbb{P}(X_{B_2} \in F_2),
\end{aligned}$$

hvor linje (3.3) følger af Sætning 3.5. Heraf ses det, at X_{B_1} og X_{B_2} er uafhængige.

Induktionstrin: Antag, at sætningen er gældende for n disjunkte mængder, dvs. når $B_1, \ldots, B_n \subseteq S$ er begrænsede disjunkte mængder. Lad $B_1, \ldots, B_{n+1} \subseteq S$ være begrænsede disjunkte mængder, og lad $B_1 \cup B_2 \cup \ldots \cup B_n = D$ og $B = B_{n+1} \cup D$, da fås

$$\begin{aligned}
\mathbb{P}(X_D \in F_D, X_{B_{n+1}} \in F_{n+1}) =&\mathbb{P}(X_D \in F_D)\mathbb{P}(X_{B_{n+1}} \in F_{n+1}) \\
=&\mathbb{P}(X_{B_1} \in F_1) \cdots \mathbb{P}(X_{B_n} \in F_n)\mathbb{P}(X_{B_{n+1}} \in F_{n+1}),
\end{aligned}$$

Dermed er $X_{B_1}, \ldots, X_{B_{n+1}}$ uafhængige Poisson-processer for begrænsede disjunkte mængder $B_i \subseteq S$ for $i = 1, 2, \ldots$.

Da enhver delmængde af S er et tælleligt antal foreninger af begrænsede delmængder af S, gælder sætningen for alle disjunkte mængder $B_1, B_2, \ldots \subseteq S$. □

3.1.4 Slivnyak-Meckes sætning

To sætninger for Poisson-processer, der er anvendelige senere, er Slivnyak-Meckes sætninger. Beviset for sætningerne findes i [Møller og Waagepetersen, 2004, s.21].

Sætning 3.7 (Slivnyak-Meckes Sætning) [Møller og Waagepetersen, 2004, Th. 3.2]
Lad $X \sim Poisson(S, \rho(\cdot))$, da gælder for funktioner $h : S \times N_{lf} \to [0, \infty)$, at

$$\mathbb{E}\left[\sum_{\xi \in X} h(\xi, X \backslash \xi)\right] = \int_S \mathbb{E}\left[h(\xi, X)\right] \rho(\xi)\ d\xi.$$

Venstre side af lighedstegnet er endeligt, hvis og kun hvis højre side er.

Sætning 3.8 (Udvidet Slivnyak-Mecke) [Møller og Waagepetersen, 2004, Th. 3.3]
Lad $X \sim Poisson(S, \rho(\cdot))$, da gælder for ethvert $n \in \mathbb{N}$ og enhver funktion $h : S^n \times N_{lf} \to [0, \infty)$, at

$$\mathbb{E}\left[\sum_{\xi_1, \ldots, \xi_n \in X}^{\neq} h(\xi_1, \ldots, \xi_n, X \backslash \{\xi_1, \ldots, \xi_n\})\right]$$
$$= \int_{S^n} \mathbb{E}\left[h(\xi_1, \ldots, \xi_n, X)\right] \prod_{i=1}^{n} \rho(\xi_i)\ d\xi_1 \cdots\ d\xi_n,$$

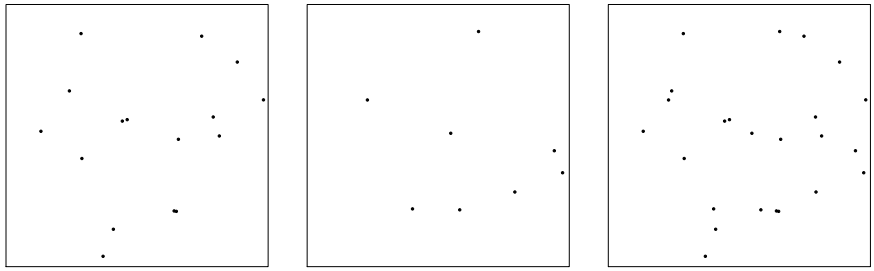
hvor $\neq$ over summationstegnet betyder, at ξ'erne skal være parvis forskellige.

3.1.5 Superposition og uafhængig udtynding

Når der senere opstilles modeller for datasættet gravhøje benyttes resultater om superposition og udtynding af Poisson-processer.

Definition 3.9 (Superposition)
En disjunkt forening $\bigcup_{i=1}^{\infty} X_i$ af punktprocesser $X_1, X_2, \ldots$ kaldes en superposition.

På Figur 3.2 ses eksempel på superposition af to homogene Poisson-processer. Figur (a) og (b) illustrerer to homogene Poisson-processer og figur (c) viser en superposition af (a) og (b).



(a) (b) (c)

Figur 3.2: (a)(b)Realiseringer af homogene Poisson-processer med intensitet hhv. $\rho = 20$ og $\rho = 10$ på $S = [0,1] \times [0,1]$. (c) Superposition af de to Poisson-processer i (a) og (b).

Definition 3.10 (Uafhængig udtynding)
Lad $p : S \to [0,1]$ være en funktion og X en punktproces på S. Punktprocessen givet ved

$$X_{\text{thin}} = \{\xi \in X | R(\xi) \leq p(\xi)\},$$

hvor $R(\xi) \sim unif([0,1])$ for $\xi \in S$, med $R(\xi)$'erne uafhængige af hinanden og af X, kaldes en uafhængig udtynding af X med bevaringssandsynlighed $p(\cdot)$.

På Figur 3.3 ses eksempel på udtynding. (a) er en realisation af en homogen Poisson-proces, og (b) illustrerer en udtynding af (a) med bevarelsessandsynligheden $p = 0.5$.

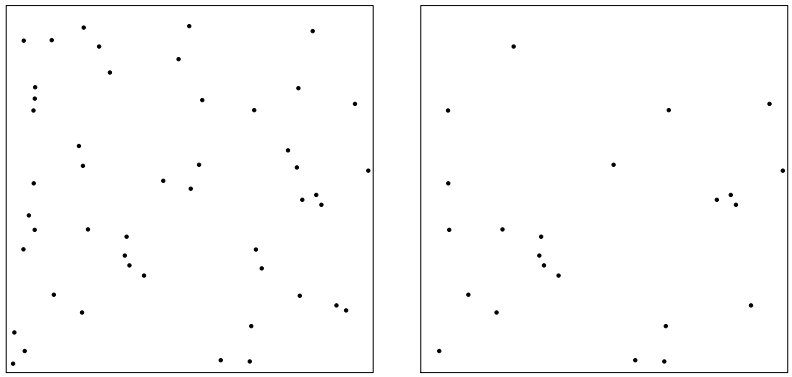
Det ses af følgende to sætninger, at Poisson-processer forbliver Poisson-processer under både superposition og uafhængig udtynding.

Sætning 3.11 [Møller og Waagepetersen, 2004, Proposition 3.6]
Lad $X_i \sim Poisson(S, \rho_i(\cdot))$, $i = 1, 2, \ldots$, være indbyrdes uafhængige og $\rho(\cdot) = \sum_{i=1}^{\infty} \rho_i(\cdot)$ være lokalt integrabel. Da gælder

(i) $X = \bigcup_{i=1}^{\infty} X_i$ er en superposition med sandsynlighed en.

(ii) $X \sim Poisson(S, \rho(\cdot))$.

Bevis:
(i) Betragt en disjunkt inddeling af S i begrænsede mængder B_i, dvs. $S = \bigcup_{i=1}^{\infty} B_i$.



(a) (b)

Figur 3.3: (a) Realisering af homogen Poisson-proces med intensitet $\rho = 50$ på $S = [0,1] \times [0,1]$. (b) Udtynding af (a) med bevarelsessandsynlighed $p = 0.5$.

Da gælder

$$\begin{aligned}\mathbb{P}(X_1 \cap X_2 \neq \emptyset) =& \mathbb{P}(X_1 \cap X_2 \cap S \neq \emptyset) = \mathbb{P}\left(\bigcup_{i=1}^{\infty}(X_1 \cap X_2 \cap B_i) \neq \emptyset\right) \\ \leq& \sum_{i=1}^{\infty} \mathbb{P}(X_1 \cap X_2 \cap B_i \neq \emptyset) = \sum_{i=1}^{\infty}(1 - \mathbb{P}(X_1 \cap X_2 \cap B_i = \emptyset)),\end{aligned}$$

hvor uligheden følger af udtryk (2.1). For ethvert begrænset $B \subseteq S$ gælder det, at $\mathbb{P}(X_1 \cap X_2 \cap B = \emptyset) = 1$, da X_1 og X_2 er uafhængige Poisson-processer. Det er nu vist, at X_1 og X_2 er disjunkte med sandsynlighed en, og dermed er $X = X_1 \cup X_2$ en superposition. Beviset for, at $X = \bigcup_{i=1}^{\infty} X_i$ er en superposition, følger af induktion.

(ii) Det vises nu, at $X = \bigcup_{i=1}^{\infty} X_i$ er en Poisson$(S, \rho(\cdot))$ vha. Sætning 3.4:

$$\begin{aligned}\mathbb{P}(X_B = \emptyset) &= \prod_{i=1}^{\infty} \mathbb{P}(X_i \cap B = \emptyset) = \prod_{i=1}^{\infty} \exp(-\mu_i(B)) = \exp\left(-\sum_{i=1}^{\infty} \int_B \rho_i(\xi)\ \mathrm{d}\xi\right) \\ &= \exp\left(-\int_B \sum_{i=1}^{\infty} \rho_i(\xi)\ \mathrm{d}\xi\right) = \exp(-\mu(B)),\end{aligned}$$

hvilket genkendes som voidsandsynligheden for en Poisson-proces. □

Ved benyttelse af ovenstående sætning fås det, at Figur 3.2(c) er en realisation af en Poisson-proces med intensiteten $\rho = 30$, da den er superposition af Poisson-processer med intensitet hhv. $\rho = 20$ og $\rho = 10$.

Sætning 3.12 [Møller og Waagepetersen, 2004, Proposition 3.7]
Antag, at $X \sim$ Poisson$(S, \rho(\cdot))$ udsættes for en uafhængig udtynding med bevaringssandsynlighed $p(\xi)$ for $\xi \in S$, og lad

$$\rho_{\text{thin}}(\xi) = p(\xi)\rho(\xi), \qquad \xi \in S.$$

Da er X_{thin} og $X \backslash X_{\text{thin}}$ uafhængige Poisson-processer med intensitetsfunktioner henholdsvis $\rho_{\text{thin}}(\cdot)$ og $\rho(\cdot) - \rho_{\text{thin}}(\cdot)$.

Bevis:
Lad $\mu_{\text{thin}}(\cdot)$ være givet ved $\mu_{\text{thin}}(B) = \int_B \rho_{\text{thin}}(\xi)\ \mathrm{d}\xi$. Ifølge Definition 3.10 vil der for et begrænset $A \subseteq S$ gælde, at

$$\begin{aligned}
\mathbb{P}(X_{\text{thin}} \cap A = \emptyset) &= \mathbb{E}[\mathbb{E}[\mathbb{1}[X_{\text{thin}} \cap A = \emptyset]|X]] \\
&= \sum_{n=0}^{\infty} \frac{e^{-\mu(A)}}{n!} \int_{A^n} \mathbb{E}[\mathbb{1}[X_{\text{thin}} \cap A = \emptyset]|X] \prod_{i=1}^{n} \rho(x_i)\ \mathrm{d}x_1 \ldots\ \mathrm{d}x_n \\
&= \sum_{n=0}^{\infty} \frac{e^{-\mu(A)}}{n!} \int_{A^n} \prod_{i=1}^{n} (1 - p(x_i)) \prod_{i=1}^{n} \rho(x_i)\ \mathrm{d}x_1 \ldots\ \mathrm{d}x_n \qquad (3.4) \\
&= e^{-\mu(A)} \sum_{n=0}^{\infty} \frac{1}{n!} \left(\int_A (1 - p(\xi))\rho(\xi)\ \mathrm{d}\xi \right)^n \\
&= e^{-\mu(A)} \exp\left(\int_A (1 - p(\xi))\rho(\xi)\ \mathrm{d}\xi \right) \\
&= e^{-\mu_{\text{thin}}(A)},
\end{aligned}$$

hvor linje (3.4) kommer af, at alle $x_i \in X \cap A$ er frasorteret med sandsynlighed $1 - p(x_i)$ uafhængigt af hinanden.

Ved tilsvarende udregninger fås

$$\mathbb{P}((X \backslash X_{\text{thin}}) \cap B = \emptyset) = \exp(-(\mu(B) - \mu_{\text{thin}}(B))).$$

Altså er både X_{thin} og $X \setminus X_{\text{thin}}$ Poisson-processer da ovenstående voidsandsynligheder er voidsandsynlighederne for en Poisson-proces.

Uafhængigheden fås ved følgende udregning for begrænsede $A, B \subseteq S$.

$$\begin{aligned}
&\mathbb{P}(X_{\text{thin}} \cap A = \emptyset, (X \backslash X_{\text{thin}}) \cap B = \emptyset) \\
=&\mathbb{P}(X \cap A \cap B = \emptyset, X_{\text{thin}} \cap (A \backslash B) = \emptyset, (X \backslash X_{\text{thin}}) \cap (B \backslash A) = \emptyset) \\
=&\mathbb{P}(X \cap A \cap B = \emptyset)\ \mathbb{P}(X_{\text{thin}} \cap (A \backslash B) = \emptyset)\ \mathbb{P}((X \backslash X_{\text{thin}}) \cap (B \backslash A) = \emptyset) \\
=&\exp\left(-\mu(B) - \mu_{\text{thin}}(A) + \mu_{\text{thin}}(B)\right) \\
=&\mathbb{P}(X_{\text{thin}} \cap A = \emptyset)\mathbb{P}((X \backslash X_{\text{thin}}) \cap B = \emptyset).
\end{aligned}$$

□

Af ovenstående sætning følger det, at Figur 3.3(b) er en Poisson-proces med intensiteten $\rho = 0.5 \cdot 50 = 25$.

3.1.6 Mærkede Poisson-processer

Poisson-processen findes også i en mærket udgave, der defineres i det følgende. En mærket punktproces kan eksempelvis fås ved til ethvert punkt at knytte et mærke, der kan indeholde en yderligere information. Som beskrevet i indledningen kunne dette mærke betegne alderen af gravhøje eller størrelsen af gravhøjene. Sætningerne i dette afsnit bevises ikke, men beviserne for sætningerne kan findes i [Møller og Waagepetersen, 2004, s. 25]

Definition 3.13 (Mærket Poisson-proces)
Lad $Y \sim$ Poisson$(T, \varphi(\cdot))$, og lad endvidere mærkerne $\{m_\xi | \xi \in Y\} \subseteq M$ være uafhængige givet Y, hvor M kaldes mærkerummet. Da kaldes $X = \{(\xi, m_\xi) | \xi \in Y\}$ for en mærket Poisson-proces. Endvidere gælder det, at hvis mærkerne er ensfordelte med fordelingen Q, kaldes Q mærkefordelingen.

Det er i nogle tilfælde muligt at beskrive en Poisson-proces ved hjælp af en mærket Poisson-proces af lavere dimension, eller omvendt.

Sætning 3.14 [Møller og Waagepetersen, 2004, Proposition 3.9]
Lad $Y \sim$ Poisson$(T, \varphi(\cdot))$. Lad endvidere $X = \{(\xi, m_\xi) | \xi \in Y\}$ være en mærket Poisson-proces med $m_\xi \in M \subseteq \mathbb{R}^p$, hvor hvert mærke m_ξ betinget med Y har tæthed $p_\xi(\cdot)$, der ikke afhænger af $Y \backslash \xi$. Lad endvidere $\rho(\xi, m) = \varphi(\xi) p_\xi(m)$, da gælder det, at

(i) $X \sim$ Poisson$(T \times M, \rho(\cdot))$,

(ii) hvis tætheden på M defineret ved $\kappa(m) = \int_T \rho(\xi, m) d\xi$ er lokalt integrabel, så er $\{m_\xi | \xi \in Y\} \sim$ Poisson$(M, \kappa(\cdot))$.

Det bør bemærkes, at ikke alle Poisson$(T \times M, \rho(\cdot))$ er mærkede Poisson-processer, og heller ikke alle mærkede Poisson-processer er Poisson-processer.

3.2 Tætheder for Poisson-processer

I det følgende behandles begrebet absolut kontinuitet, der beskriver sammenhængen mellem to punktprocesser, der er defineret på den samme mængde. Begrebet og de efterfølgende resultater benyttes i senere kapitler.

Definition 3.15 (Absolut kontinuitet)
Hvis X og Y er to punktprocesser defineret på det samme rum S, så siges fordelingen af X at være absolut kontinuert mht. fordelingen af Y, hvis og kun hvis der eksisterer en funktion $f : N_{lf} \to [0, \infty]$, således at

$$\mathbb{P}(X \in F) = \mathbb{E}[\mathbb{1}[Y \in F] f(Y)], \qquad F \subseteq N_{lf}.$$

Funktionen $f(\cdot)$ kaldes tætheden for X mht. Y.

Poisson-processer defineret på samme mængde er ikke altid absolut kontinuerte mht. til hinanden, hvilket følgende sætning udtrykker.

Sætning 3.16 [Møller og Waagepetersen, 2004, Proposition 3.8]
Lad $\rho_1, \rho_2 : S \to [0, \infty)$ være intensitetsfunktioner, sådan at $\mu_1(S) = \int_S \rho_1(\xi)\, d\xi$ og $\mu_2(S) = \int_S \rho_2(\xi)\, d\xi$ er endelige, og så $\rho_1(\xi) > 0 \Rightarrow \rho_2(\xi) > 0$. Så er Poisson$(S, \rho_1)$ absolut kontinuert mht. Poisson(S, ρ_2), med tætheden

$$f(y) = \exp(\mu_2(S) - \mu_1(S)) \prod_{\xi \in y} \frac{\rho_1(\xi)}{\rho_2(\xi)}, \tag{3.5}$$

for endelige punktmønstre $y \subset S$.

Bevis:
Lad $X \sim \text{Poisson}(S, \rho_1)$, $Y \sim \text{Poisson}(S, \rho_2)$. Det skal vises, at med tæthedsfunktionen fra (3.5) er Definition 3.15 opfyldt. Ved benyttelse af Sætning 3.5 fås

$$\begin{aligned}
\mathbb{E}[\mathbb{1}[Y \in F] f(Y)] = & \sum_{n=0}^{\infty} \frac{\exp(-\mu_2(S))}{n!} \int_{S^n} \mathbb{1}[y \in F] \exp(\mu_2(S) - \mu_1(S)) \\
& \prod_{i=1}^{n} \frac{\rho_1(y_i)}{\rho_2(y_i)} \prod_{i=1}^{n} \rho_2(y_i)\ \mathrm{d}y_1 \ldots\ \mathrm{d}y_n \\
= & \mathbb{P}(X \in F).
\end{aligned}$$

□

For alle reelle tal $\rho_1, \rho_2 > 0$ glæder, at Poisson$(\mathbb{R}^d, \rho_1)$ er absolut kontinuert mht. Poisson$(\mathbb{R}^d, \rho_2)$, hvis og kun hvis $\rho_1 = \rho_2$ jf. [Møller og Waagepetersen, 2004, s. 25].

En Poisson-proces defineret på en begrænset mængde S er altid absolut kontinuert mht. standard Poisson-processen. For et begrænset S har en Poisson(S, ρ)

tætheden

$$f(x) = \exp(|S| - \mu(S)) \prod_{\xi \in x} \rho(\xi), \tag{3.6}$$

mht. Poisson$(S, 1)$, og med samme fremgangsmåde som i beviset for Sætning 3.16 ses det, at Poisson(S, ρ) er absolut kontinuert mht. Poisson$(S, 1)$.

3.2.1 Likelihoodfunktion for Poisson-proces

I det følgende antages det, at fordelingen af en punktproces X kan specificeres ud fra en tæthed i en parametrisk familie $\{f(\cdot)|\ \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^p$, af tætheder $f(x|\theta)$ mht. standard Poisson-processen Y på $S \subset \mathbb{R}^d$, hvor $|S| < \infty$.

Betragt en realisation x af en punktproces X, hvor tæthedsfunktionen mht. standard Poisson-processen er $f(x|\theta)$. Dermed er likelihoodfunktionen for θ givet ved

$$L(\theta|x) = cf(x|\theta), \tag{3.7}$$

hvor $c \in \mathbb{R}_+$.

Med udgangspunkt i (3.7) er det muligt at definere log-likelihoodfunktionen ved

$$\ell(\theta|x) = \log(f(x|\theta)) + \log(c).$$

I den resterende del af rapporten angives likelihoodfunktionen ofte kun op til proportionalitet.

Lad x være en realisering af Poisson(S, ρ) på en begrænset mængde $S \subseteq \mathbb{R}^d$. Da er likelihoodfunktionen for en Poisson-proces mht. standard Poisson-processen jf. (3.6) og (3.7) givet ved

$$L(\theta|x) = \exp(-\mu(S)) \prod_{\xi \in x} \rho(\xi), \tag{3.8}$$

og ligeledes fås log-likelihoodfunktionen for Poisson-processen til

$$\ell(\theta|x) = -\mu(S) + \log\left(\prod_{\xi \in x} \rho(\xi)\right) = -\mu(S) + \sum_{\xi \in x} \log\left(\rho(\xi)\right).$$

I det følgende kapitel gennemgås summary statistics for punktprocesser. Disse benyttes senere til analyse af datasættet gravhøje, og til at afgøre hvor godt de forskellige opstillede modeller stemmer overens med datasættet.

Summary statistics

I dette kapitel introduceres de forskellige summary statistics, der i senere kapitler bliver brugt til beskrivelse af punktmønstre. Summary statistics benyttes til at give en opsummering af punktmønstre i form af en enkelt graf. Ud fra en summary statistic er det muligt at afgøre, om et punktmønster har større eller mindre clustertendenser end en realisering af en Poisson-proces.

I forbindelse med analyse af et punktmønster, er en vigtig del derfor at estimere forskellige summary statistics fra punktmønstret, og sammenligne dem med de tilsvarende summary statistics beregnet for en Poisson-proces eller estimeret ud fra en realisation af en Poisson-proces.

Summary statistics anvendes ofte til modeltjek, således at en summary statistic beregnes eller estimeres for en opstillet punktproces og sammenlignes med den estimerede fra punktmønstret.

I det følgende gennemgås andenordens egenskaber for punktprocesser, hvorefter summary statistics med udgangspunkt i første og andenordens egenskaber behandles. Afslutningsvis indføres og estimeres summary statistics for andenordens egenskaber baseret på afstande mellem par af punkter. Kapitlet er skrevet på baggrund af [Møller og Waagepetersen, 2004, Kapitel 4].

Gennem hele kapitlet betegner X en rumlig punktproces på $\mathbb{R}^d$ med intensitesfunktion $\rho(\cdot)$ og tællemål $N(\cdot)$.

4.1 Andenordens egenskaber

I dette afsnit behandles første og andenordens egenskaber for det stokastiske tællemål $N(B)$ for mængder $B \subseteq \mathbb{R}^d$, der er beskrevet af intensitetsmålet og det andenordens faktorielle momentmål.

Intensitetsmålet er givet i Definition 2.10 som $\mu(B) = \mathbb{E}[N(B)]$, hvor $B \subseteq S$ og kaldes en førsteordens egenskab. Det andenordens faktorielle momentmål, der kaldes en andenordens egenskab, defineres ved:

Definition 4.1 (Andenordens faktorielle momentmål)
Det andenordens faktorielle momentmål $\alpha^{(2)}(\cdot)$ på $\mathbb{R}^d \times \mathbb{R}^d$ er givet ved

$$\alpha^{(2)}(C) = \mathbb{E}\left[\sum_{\xi,\eta \in X}^{\neq} \mathbb{1}[(\xi,\eta) \in C]\right], \quad C \subseteq \mathbb{R}^d \times \mathbb{R}^d.$$

Definition 4.2 (Andenordens produkttæthed)
Hvis det andenordens faktorielle momentmål $\alpha^{(2)}(\cdot)$ kan udtrykkes ved

$$\alpha^{(2)}(C) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathbb{1}[(\xi,\eta) \in C]\rho^{(2)}(\xi,\eta)\ d\xi d\eta, \quad C \subseteq \mathbb{R}^d \times \mathbb{R}^d,$$

med $\rho^{(2)}(\cdot) \geq 0$, kaldes $\rho^{(2)}(\cdot)$ for andenordens produkttætheden.

Udtrykket $\rho^{(2)}(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta$, hvor $\xi,\eta \in X$ kan tolkes som sandsynligheden for, at X har et punkt i både kuglen med centrum i ξ og volumen $\mathrm{d}\xi$ og kuglen med centrum i η og volumen $\mathrm{d}\eta$.

Følgende sætning benyttes ofte i forbindelse med beviserne i dette kapitel. Til forskel fra Sætning 3.7 og Sætning 3.8, gælder Sætning 4.3 generelt og ikke kun for Poisson-processer.

Sætning 4.3 (Campells sætning) [Møller og Waagepetersen, 2004, Proposition 4.1]
Lad X have andenordens produkttæthed $\rho^{(2)}(\cdot)$. Da gælder det for funktionerne $h_1 : \mathbb{R}^d \to [0,\infty)$ og $h_2 : \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty)$, at

$$\mathbb{E}\left[\sum_{\xi\in X} h_1(\xi)\right] = \int h_1(\xi)\rho(\xi)\ d\xi \qquad \text{og}$$

$$\mathbb{E}\left[\sum_{\xi,\eta\in X}^{\neq} h_2(\xi,\eta)\right] = \int\int h_2(\xi,\eta)\rho^{(2)}(\xi,\eta)\ d\xi d\eta.$$

Bevis:
Fra Definition 2.10 fås

$$\mathbb{E}\left[\sum_{\xi\in X} \mathbb{1}[\xi \in B]\right] = \mathbb{E}\left[N(B)\right] = \int \mathbb{1}[\xi \in B]\rho(\xi)\ \mathrm{d}\xi,$$

og fra Definition 4.2 fås

$$\mathbb{E}\left[\sum_{\xi,\eta\in X}^{\neq} \mathbb{1}[(\xi,\eta) \in C]\right] = \alpha^{(2)}(C) = \int\int \mathbb{1}[(\xi,\eta) \in C]\rho^{(2)}(\xi,\eta)\ \mathrm{d}\xi\mathrm{d}\eta.$$

Ved benyttelse af standardbeviset, omtalt i beviset for Sætning 3.5, følger sætningen. □

For at kunne sammenligne andre punktprocesser med Poisson-processen udfra andenordens produkttætheden, vises følgende.

Sætning 4.4 [Møller og Waagepetersen, 2004, s. 31]
Lad X være en Poisson-proces med intensitet $\rho(\cdot)$. Andenordens produkttætheden er da givet ved

$$\rho^{(2)}(\xi,\eta) = \rho(\xi)\rho(\eta).$$

Bevis:
Antag, at X er en Poisson-proces med intensitet $\rho(\cdot)$, og at $C = A \times B$, hvor $A, B \in \mathbb{R}^d$. Dermed fås det jf. Definition 4.2, at

$$\begin{aligned}\int_B \int_A \rho^{(2)}(\xi,\eta)\ \mathrm{d}\xi\mathrm{d}\eta &= \mathbb{E}\left[\sum_{\xi,\eta\in X}^{\neq} \mathbb{1}\left[\xi \in A, \eta \in B\right]\right] \\ &= \int_{\mathbb{R}^2} \mathbb{E}\left[\mathbb{1}\left[\xi \in A, \eta \in B\right]\right] \rho(\xi)\rho(\eta)\ \mathrm{d}\xi\mathrm{d}\eta \\ &= \int_B \int_A \rho(\xi)\rho(\eta)\ \mathrm{d}\xi\mathrm{d}\eta,\end{aligned}$$

hvor den udvidede Slivnyak-Mecke Sætning 3.8 er benyttet. Dermed er det vist, at $\rho^{(2)}(\xi,\eta) = \rho(\xi)\rho(\eta)$. □

Følgende sætning giver et udtryk for $\rho(\cdot)$ og $\rho^{(2)}(\cdot)$ efter en uafhængig udtynding af en punktproces.

Sætning 4.5 [Møller og Waagepetersen, 2004, Proposition 4.2]
Lad X have intensitetsfunktion $\rho(\cdot)$ og andenordens produkttæthed $\rho^{(2)}(\cdot)$, og lad X_{thin} være en uafhængig udtynding af X med bevaringssandsynlighed $p(\xi)$, hvor $\xi \in \mathbb{R}^d$. Da er intensitetsfunktionen og andenordens produkttætheden for X_{thin} givet ved

$$\rho_{\text{thin}}(\xi) = p(\xi)\rho(\xi) \qquad \text{og} \qquad \rho^{(2)}_{\text{thin}}(\xi,\eta) = p(\xi)p(\eta)\rho^{(2)}(\xi,\eta).$$

Bevis:
For ethvert $B \subseteq \mathbb{R}^d$ gælder det, at

$$\begin{aligned}\mu_{\text{thin}}(B) = \mathbb{E}\left[n(X_{\text{thin}} \cap B)\right] &= \mathbb{E}\left[\sum_{\xi\in X_{\text{thin}}} \mathbb{1}[\xi \in B]\right] \\ &= \mathbb{E}\left[\sum_{\xi\in X} p(\xi)\mathbb{1}[\xi \in B]\right] = \int_B p(\xi)\rho(\xi)\ \mathrm{d}\xi,\end{aligned}$$

hvor sidste lighed kommer af Sætning 4.3. Dermed er $\rho_{\text{thin}}(\xi) = p(\xi)\rho(\xi)$. Tilsvarende fås $\rho^{(2)}_{\text{thin}}(\xi,\eta) = p(\xi)p(\eta)\rho^{(2)}(\xi,\eta)$ ved at betragte $C \subseteq \mathbb{R}^d \times \mathbb{R}^d$. □

4.1.1 Parkorrelationsfunktionen og det andenordens reducerede momentmål

Parkorrelationsfunktionen og det andenordens reducerede momentmål benyttes ofte til at konstruere summary statistics. Parkorrelationen kan bruges til at undersøge, om en punktproces har større eller mindre clustertendenser end Poisson-processen.

Definition 4.6 (Parkorrelationsfunktion)
Hvis både $\rho(\cdot)$ og $\rho^{(2)}(\cdot)$ eksisterer, er parkorrelationsfunktionen defineret ved

$$g(\xi,\eta) = \begin{cases} \frac{\rho^{(2)}(\xi,\eta)}{\rho(\xi)\rho(\eta)} & \text{for } \rho(\xi)\rho(\eta) \neq 0, \\ 0 & \text{for } \rho(\xi)\rho(\eta) = 0. \end{cases}$$

Sætning 4.4 giver, at $g(\cdot) = 1$ for en Poisson-proces. Hvis $g(\cdot) > 1$ for en punktproces, har denne større clustertendenser end en Poisson-proces, og hvis $g(\cdot) < 1$ er en punktproces mere regulær end en Poisson-proces.

Af Sætning 4.5 følger det, at parkorrelationsfunktionen er invariant under uafhængig udtynding af en punktproces, da $g_{\text{thin}}(\cdot) = g(\cdot)$. Hvis X er stationær, er $g(\cdot)$ invariant under translation, således at $g(\xi,\eta) = g(\xi - \eta)$, da $\rho(\cdot)$ er konstant for en stationær punktproces.

En svagere stationaritetsbetingelse er den andenordens intensitetsgenvægtede stationaritet.

Definition 4.7 (Andenordens intensitetsgenvægtet stationaritet)
For $A, B \subseteq \mathbb{R}^d$ og $0 < |A| < \infty$ defineres målet

$$\mathcal{K}(B) = \frac{1}{|A|}\mathbb{E}\left[\sum_{\xi,\eta\in X}^{\neq} \frac{\mathbb{1}[\xi \in A, \eta - \xi \in B]}{\rho(\xi)\rho(\eta)}\right],$$

hvor led i summen er lig 0, hvis $\rho(\xi)\rho(\eta) = 0$. Er målet uafhængig af valget af A kaldes $\mathcal{K}(\cdot)$ det andenordens reducerede momentmål, og X siges at være andenordens intensitetsgenvægtet stationær.

Sammenhængen mellem $g(\cdot)$ og $\mathcal{K}(\cdot)$ illustreres i følgende sætning.

Sætning 4.8 [Møller og Waagepetersen, 2004, s. 33]
Hvis parkorrelationsfunktionen for en punktproces X er translationsinvariant, dvs. $g(\xi,\eta) = g(\xi-\eta)$, er X andenordens intensitetsgenvægtet stationær og

$$\mathcal{K}(B) = \int_B g(\tilde{\eta})\ d\tilde{\eta}, \quad \textit{hvor} \quad \tilde{\eta} = \xi - \eta, \quad B \subseteq \mathbb{R}^d.$$

Bevis:
Fra Definition 4.7 og Sætning 4.3 fås

$$\begin{aligned}\mathcal{K}(B) &= \frac{1}{|A|}\int\int \mathbb{1}[\xi \in A, \xi-\eta \in B] g(\xi,\eta)\ \mathrm{d}\xi \mathrm{d}\eta \\ &= \frac{1}{|A|}\int_B\int_A g(\xi,\xi-\tilde{\eta})\ \mathrm{d}\xi \mathrm{d}\tilde{\eta} = \int_B g(\tilde{\eta})\ \mathrm{d}\tilde{\eta},\end{aligned}$$

hvor sidste lighed fås da $g(\cdot)$ er translationsinvariant. Da $\mathcal{K}(\cdot)$ er uafhængig af valget af A, er X andenordens intensitetsgenvægtet stationær. □

I næste sætning ses det, at $\mathcal{K}(\cdot)$ også er invariant under uafhængig udtynding.

Sætning 4.9 [Møller og Waagepetersen, 2004, Proposition 4.3]
Lad X_{thin} være en uafhængig udtynding af en andenordens intensitetsgenvægtet stationær punktproces X. Da er X_{thin} andenordens intensitetsgenvægtet stationær, og $\mathcal{K}_{\text{thin}}(B) = \mathcal{K}(B)$.

Bevis:
Fra Definition 4.7 og Sætning 4.5 fås

$$\begin{aligned}\mathcal{K}_{\text{thin}}(B) &= \frac{1}{|A|}\mathbb{E}\left[\sum_{\xi,\eta\in X_{\text{thin}}}^{\neq} \frac{\mathbb{1}[\xi \in A, \xi-\eta \in B]}{\rho_{\text{thin}}(\xi)\rho_{\text{thin}}(\eta)}\right] \\ &= \frac{1}{|A|}\mathbb{E}\left[\sum_{\xi,\eta\in X}^{\neq} \frac{\mathbb{1}[\xi \in A, \xi-\eta \in B]p(\xi)p(\eta)}{\rho_{\text{thin}}(\xi)\rho_{\text{thin}}(\eta)}\right] \\ &= \frac{1}{|A|}\mathbb{E}\left[\sum_{\xi,\eta\in X}^{\neq} \frac{\mathbb{1}[\xi \in A, \xi-\eta \in B]}{\rho(\xi)\rho(\eta)}\right] \\ &= \mathcal{K}(B).\end{aligned}$$

□

4.2 Andenordens summary statistics

I dette afsnit indføres K og L-funktioner, og der foretages ikke-parametrisk estimation af disse.

Ved at betragte $\mathcal{K}(B)$, hvor $B = b(0,r)$ er en kugle med centrum i 0 og radius r, indføres definitionerne på K og L-funktionerne.

Definition 4.10 (K og L-funktionerne)
For en andenordens intensitetsgenvægtet stationær punktproces X på $\mathbb{R}^d$ defineres K og L-funktionerne for $r > 0$ ved

$$K(r) = \mathcal{K}(b(0,r)) \quad \text{og} \quad L(r) = \left(\frac{K(r)}{\omega_d}\right)^{\frac{1}{d}},$$

hvor $\omega_d = \pi^{d/2}/\Gamma(1+d/2)$ er det d-dimensionale volumen af enhedskuglen $b(0,1)$.

Ovenstående definition giver ifølge Definition 4.7 for en stationær punktproces X med intensitet ρ, at

$$K(r) = \mathbb{E}\left[\frac{1}{\rho^2|A|}\sum_{\xi,\eta\in X}^{\neq} \mathbb{1}[\xi \in A, |\xi-\eta| \le r]\right].$$

Funktionen $K(\cdot)$ beskriver i det stationære tilfælde hvor stor tiltrækning, der er mellem punkter, således at $\rho K(r)$ beskriver det forventede antal punkter inden for afstanden r om et tilfældigt punkt $\xi \in X$.

I visse situationer er det lettere at arbejde med $L(r) - r$, da den er lettere at afbillede og genkende end $K(r)$. For små $r > 0$ gælder der for Poisson-processen, at $L(r) - r = 0$, hvilket ses ved

$$L(r) = \left(\frac{1}{\omega_d|A|}\mathbb{E}\left[\sum_{\xi,\eta\in X}^{\neq} \frac{\mathbb{1}[\xi \in A, |\xi-\eta| \le r]}{\rho(\xi)\rho(\eta)}\right]\right)^{\frac{1}{d}} = \left(\frac{1}{\omega_d|A|}|A|\omega_d r^d\right)^{\frac{1}{d}} = r.$$

Hvis $L(r) - r > 0$, er der større clustertendenser end ved Poisson-processen, og hvis $L(r) - r < 0$, er processen mere regulær end Poisson-processen.

Ofte er det ikke muligt at beregne en summary statistic, og det er derfor nødvendigt at estimere disse, hvilket gøres i Afsnit 4.2.2. Hertil skal der benyttes estimater af intensitetfunktionen, hvilket derfor gennemgås først.

4.2.1 Ikke-parametrisk estimation af intensitetsfunktionen

Intensitetsfunktionen for et observeret punktmønster er ofte ukendt og skal derfor estimeres. Ved estimering behandles kun tilfælde, hvor et enkelt punktmønster $X_W = x$ er observeret i et begrænset vindue $W \subset \mathbb{R}^d$ med $0 < |W| < \infty$.

I det homogene tilfælde er det ikke-parametriske estimat

$$\hat{\rho} = \frac{n(x)}{|W|}, \tag{4.1}$$

et oplagt unbiased estimat af intensitetsfunktionen, da det udtrykker det gennemsnitlige antal af punkter pr. areal.

Til det inhomogene tilfælde skal begrebet kerne benyttes.

Definition 4.11 (Kerne)
Funktionen $k(\cdot)$ kaldes en kerne, hvis $\xi \mapsto k(\xi - c)$ er en tæthed for alle $c \in \mathbb{R}^d$.

I det inhomogene tilfælde er et muligt ikke-parametrisk kerneestimat af intensitetsfunktionen givet ved

$$\hat{\rho}_b(\xi) = \sum_{\eta \in x} \frac{k_b(\xi - \eta)}{c_{W,b}(\eta)}, \qquad \xi \in W,$$

hvor $k_b(\cdot)$ er en kerne med bredde $b > 0$, dvs. $k_b(\xi) = k(\xi/b)/b^d$. Da W er begrænset, skal $k_b(\cdot)$ skaleres med

$$c_{W,b}(\eta) = \int_W k_b(\xi - \eta)\ \mathrm{d}\xi.$$

Dermed sikres det, at $k_b(\cdot)$ integrerer til en på W. Estimatet af intensitetsfunktionen afhænger normalt af valget af b, mens valget af $k(\cdot)$ er mindre vigtigt.

Sætning 4.12 [Møller og Waagepetersen, 2004, Lemma 4.1]
Lad

$$\hat{\rho}_b(\xi) = \sum_{\eta \in x} \frac{k_b(\xi - \eta)}{c_{W,b}(\eta)} \qquad \text{med } \xi \in W,$$

hvor $c_{W,b}(\eta) = \int_W k_b(\xi - \eta) d\xi$. Da er estimatet $\int_W \hat{\rho}_b(\xi) d\xi$ et unbiased estimat af $\mu(W)$.

Bevis:
Ved benyttelse af Campells Sætning 4.3 fås, at

$$\mathbb{E}\left[\int_W \hat{\rho}_b(\xi)\ \mathrm{d}\xi\right] = \int_W \mathbb{E}\left[\sum_{\eta\in X} \frac{k_b(\xi-\eta)}{c_{W,b}(\eta)}\right]\ \mathrm{d}\xi = \int_W \rho(\eta)\ \mathrm{d}\eta = \mu(W).$$

□

4.2.2 Ikke-parametrisk estimation af $\mathcal{K}$, K og L-funktionerne

Når et ikke-parametrisk estimat af $\mathcal{K}(\cdot)$ ønskes, kan følgende lemma benyttes, hvor beviset kan findes i [Møller og Waagepetersen, 2004, s.37].

Sætning 4.13 [Møller og Waagepetersen, 2004, Lemma 4.2]
Lad $|W \cap W_\xi| > 0$ for alle $\xi \in B$, hvor $W_\xi = \{\xi + \eta | \eta \in W\}$ er en translation af W. Lad ligeledes X være andenordens intensitetsgenvægtet stationær. Da er

$$\hat{\mathcal{K}}(B) = \sum_{\xi,\eta\in x}^{\neq} \frac{\mathbb{1}[\eta-\xi\in B]}{\rho(\xi)\rho(\eta)|W\cap W_{\eta-\xi}|} \tag{4.2}$$

et unbiased estimat af $\mathcal{K}(B)$.

Når $\mathcal{K}(\cdot)$ estimeres i Sætning 4.13, vil der, hvis f.eks. $B = b(0,r)$, og W er rektangulær, kræves, at r er mindre end den mindste side i W, så $|W \cap W_{\tilde{\eta}}| \neq 0$.

Sætning 4.13 giver et unbiased estimat af $\mathcal{K}(B)$, hvis $\rho(\cdot)$ er kendt. I praksis er $\rho(\cdot)$ dog ikke kendt, så $\rho(\xi)\rho(\eta)$ i (4.2) må erstattes af estimatet $\widehat{\rho(\xi)\rho(\eta)}$. Det giver estimatet

$$\hat{\mathcal{K}}(B) = \sum_{\xi,\eta\in x}^{\neq} \frac{\mathbb{1}[\eta-\xi\in B]}{\widehat{\rho(\xi)\rho(\eta)}|W\cap W_{\eta-\xi}|}, \tag{4.3}$$

som er biased, da $\widehat{\rho(\eta)\rho(\xi)} \neq \rho(\eta)\rho(\xi)$.

I det homogene tilfælde er estimatet af $K(r)$

$$\hat{K}(r) = \frac{1}{\rho^2}\sum_{\xi,\eta\in x}^{\neq} \frac{\mathbb{1}[|\eta-\xi|<r]}{|W\cap W_{\eta-\xi}|}$$

unbiased, da A i Definition 4.7 kan vælges til $W \cap W_{\eta-\xi}$, så

$$\mathbb{E}\left[\hat{K}(r)\right] = \mathbb{E}\left[\frac{1}{\rho^2}\sum_{\xi,\eta\in X}^{\neq}\frac{\mathbb{1}[|\eta-\xi|<r]}{|W\cap W_{\eta-\xi}|}\right] = K(r).$$

Imens estimatet af $L(r)$,

$$\hat{L}(r) = \left(\frac{\hat{K}(r)}{\omega_d}\right)^{\frac{1}{d}},$$

generelt er biased, da

$$\mathbb{E}\left[\hat{L}(r)\right] = \mathbb{E}\left[\left(\frac{\hat{K}(r)}{\omega_d}\right)^{\frac{1}{d}}\right] \neq \mathbb{E}\left[\left(\frac{\hat{K}(r)}{\omega_d}\right)\right]^{\frac{1}{d}} = L(r).$$

4.2.3 Kantkorrektion

Kantkorrektionsfaktorer tager højde for, at punkterne i observationsvinduet kan vekselvirke med punkter udenfor observationsvinduet, dvs. punkter der ikke er observeret. Eksempelvis er vægten $|W \cap W_{\eta-\xi}|^{-1}$ i ligning (4.2) en kantkorrektionsfaktor.

Et simpelt alternativ er minus sampling. Dette opnås ved for $r > 0$ at lade

$$W_{\ominus r} = \{\eta \in W | b(\eta, r) \subseteq W\}$$

betegne mængden af punkter i W, hvor afstanden fra et punkt til grænsen af W er større end r. Ved benyttelse af minus sampling vil

$$\hat{\mathcal{K}}(B) = \sum_{\xi\in x,\eta\in x\cap W_{\ominus r}}^{\neq}\frac{\mathbb{1}[\eta-\xi\in B]}{\rho(\xi)\rho(\eta)|W_{\ominus r}|} \tag{4.4}$$

være et unbiased estimat af $\mathcal{K}(B)$, der dog bliver biased, hvis $\rho(\xi)\rho(\eta)$ erstattes af et estimat. Der er et tab af information i dette estimat, da det ikke er alle par, hvor $\eta - \xi \in B$, som medregnes, i modsætning til summen i (4.3), hvor alle observerede par af punkter er inkluderet.

I tilfælde hvor der er givet tilstrækkeligt data, og hvor meget høje vægte $|W \cap W_{\eta-\xi}|^{-1}$ er givet til nogle par af punkter i (4.3), kan det være at foretrække at bruge (4.4). Ellers vil få data få stor betydning, hvis de ligger langt fra hinanden.

4.2.4 Ikke-parametrisk estimat af parkorrelationsfunktion

For estimation af parkorrelationsfunktionen antages det, at $g(\xi,\eta) = g(|\xi-\eta|)$, der kun afhænger af afstanden mellem de to punkter, er isotropisk. Et muligt estimat

er

$$\hat{g}(r) = \frac{1}{\sigma_d r^{d-1}} \sum_{\xi,\eta \in x}^{\neq} \frac{k_b(r - |\eta - \xi|)}{\widehat{\rho(\xi)\rho(\eta)}|W \cap W_{\eta-\xi}|},$$

hvor $k_b(\xi) = k(\xi/b)/b$, $\xi \in \mathbb{R}$, for en kerne $k(\cdot)$ og bredde $b > 0$, og hvor det er antaget, at $|W \cap W_\xi| > 0$ for $|\xi| \leq r$. Estimatet $\hat{g}(r)$ forventes at være tæt på $g(r)$ for tilpas lille b, hvis $\widehat{\rho(\xi)\rho(\eta)}$ er tæt på $\rho(\xi)\rho(\eta)$.

4.3 Afstandsbaserede summary statistics

I dette afsnit indføres tre summary statistics kaldet F, G og J-funktioner. Disse adskiller sig fra K og L-funktionerne, ved at være baseret på afstande mellem par af punkter og ved kun at være defineret for stationære punktprocesser. Afsnittet afsluttes med at estimere disse funktioner.

Definition 4.14 (F, G og J-funktionerne)
Lad X være en stationær punktproces. Den tomme mængde funktion $F(\cdot)$ defineres til at være fordelingsfunktionen for afstanden fra et fast punkt $\eta \in \mathbb{R}^d$ til det nærmeste punkt i X, dvs. at

$$F(r) = \mathbb{P}(X \cap b(\eta, r) \neq \emptyset), \quad r > 0.$$

Den nærmeste nabo funktion $G(\cdot)$ defineres for en arbitrær mængde $A \subset \mathbb{R}^d$ med $0 < |A| < \infty$, ved

$$G(r) = \frac{1}{\rho|A|}\mathbb{E}\left[\sum_{\xi \in X \cap A} \mathbb{1}\left[(X \backslash \xi) \cap b(\xi, r) \neq \emptyset\right]\right], \quad r > 0.$$

J-funktionen defineres ved

$$J(r) = \frac{1 - G(r)}{1 - F(r)} \quad \text{for} \quad F(r) < 1.$$

G-funktionen kan fortolkes som fordelingen af afstanden fra et punkt i X til den nærmeste nabo. På grund af stationariteten afhænger $G(\cdot)$ ikke af valget af A.

Sætning 4.15 [Møller og Waagepetersen, 2004, s. 36]
For en stationær Poisson-proces på $\mathbb{R}^d$ med intensitet ρ gælder der for $r > 0$, at

$$F(r) = G(r) = 1 - \exp(-\rho\omega_d r^d) \qquad \text{og} \qquad J(r) = 1.$$

Bevis:
Ved benyttelse af Definition 4.14 og Sætning 4.3 fås

$$\begin{aligned} G(r) &= \frac{1}{|A|}\int_A \mathbb{P}((X\backslash\xi)\cap b(\xi,r)\neq\emptyset)\ \mathrm{d}\xi = \frac{1}{|A|}\int_A \mathbb{P}(X\cap b(\eta,r)\neq\emptyset)\ \mathrm{d}\xi \\ &= \mathbb{P}(X\cap b(\eta,r)\neq\emptyset) = F(r). \end{aligned}$$

Ved benyttelse af, at volumet af en d-dimensional kugle $b(0,r)\subset\mathbb{R}^d$ er givet ved $|b(0,r)| = (\pi r^2)^{d/2}/\Gamma(1+d/2)$, fås

$$\begin{aligned} \mathbb{P}(X\cap b(\xi,r)\neq\emptyset) &= 1-\mathbb{P}(X\cap b(\xi,r)=\emptyset) = 1-\exp(-\mu(b(0,r))) \\ &= 1-\exp\left(-\rho\frac{(\pi r^2)^{d/2}}{\Gamma(1+\frac{d}{2})}\right) = 1-\exp(-\rho\omega_d r^d). \end{aligned}$$

At $J(r)=1$ følger af, at $G(r)=F(r)$. □

For små værdier af $r>0$ indikerer $F(r)<G(r)$ eller $J(r)<1$ clustertendenser, og $F(r)>G(r)$ eller $J(r)>1$ en tendens til frastødning blandt punkterne i en punktproces. Bemærk at $J(r)=1$ ikke betyder, at en punktproces er en Poisson-proces.

4.3.1 Ikke-parametrisk estimation af F-, G- og J-funktionerne

En mulig estimation af $F(\cdot)$ og $G(\cdot)$ er udledt af minus sampling. Lad $d(\xi,B) = \inf\{|\xi-\eta|\ |\eta\in B\}$ være den korteste afstand fra et punkt $\xi\in\mathbb{R}^d$ til en mængde $B\subset\mathbb{R}^d$. Lad $I\subset W$ betegne et endeligt regulært gitter af punkter. Da er

$$\hat{F}(r) = \frac{1}{n(I\cap W_{\ominus r})}\sum_{\xi\in I\cap W_{\ominus r}} \mathbb{1}[d(\xi,x)\leq r]$$

et unbiased estimat af $F(r)$.

Egenskaben unbiased er normalt uopnåelig for mange estimatorer i rumlig statistik, men istedet er de ofte ratio-unbiased, dvs. at de er på formen $\hat{\theta} = Y/Z$, hvor $\theta = \mathbb{E}[Y]/\mathbb{E}[Z]$. Eksempelvis er

$$\hat{G}(r) = \frac{1}{\hat{\rho}|W_{\ominus r}|}\sum_{\xi\in x\cap W_{\ominus r}} \mathbb{1}[d(\xi,x\backslash\xi)\leq r]$$

et ratio-unbiased estimat af $G(r)$, for $|W_{\ominus r}|>0$. Givet estimaterne for $F(r)$ og $G(r)$ fås estimatet

$$\hat{J}(r) = \frac{1-\hat{G}(r)}{1-\hat{F}(r)}$$

af $J(r)$, for $\hat{F}(r)<1$.

4.4 Envelopes for summary staistics

I dette afsnit gennemgås kort, hvorledes summary statistics kan anvendes til modeltjek af punktprocesser. Et modeltjek udføres ved at sammenligne en summary statistic $T(r)$ fra en model, med estimatet $T_0(r)$ udregnet fra data for en given afstand $r > 0$. Notationen for en summary statistics er i dette afsnit $T(r)$ for at specificere, at metoden gælder for alle gennemgåede summary statistics. Sammen med plottet af en summary statistic kan der angives et konfidensinterval for denne for enhver værdi af r.

For at approksimere $T(r)$ ud fra modellen simuleres n uafhængige og ensfordelte punktmønstre, og $T_1(r), \ldots, T_n(r)$ estimeres derudfra. Ved hjælp af den emperiske fordeling af $T_1(r), \ldots, T_n(r)$ kan der estimeres enhver kvantil for fordelingen af $T(r)$ med enhver præcision.

Er beregningerne af $T_1(r), \ldots, T_n(r)$ tidskrævende, kan der benyttes envelopes i stedet. Envelopes viser sig specielt anvendelig ved små n. Ordnes estimaterne, så $T_{(1)}(r) \leq \ldots \leq T_{(n)}(r)$ vil

$$\mathbb{P}(\ T(r) \leq T_{(1)}(r) \text{ eller } T_{(n)}(r) \leq T(r)\ ) \leq \frac{2}{n+1}$$

med lighed, hvis $T_{(1)}(r) < \cdots < T_{(n)}(r)$. Ved modeltjek bekræftes valget af en model dermed, hvis $T_{(1)} \leq T_0(r) \leq T_{(n)}$, da $T_0(r)$ så holder sig indenfor envelopegrænserne.

Bemærk at selvom $T_1(r), \ldots, T_n(r)$ er uafhængige og ensfordelte, vil vektorerne $(T_1(r), \ldots, T_n(r))$ for forskellige værdier af r være afhængige, hvilket kan give besværligheder ved sammenligning af resultater for forskellige værdier af r.

Sættes $n = 39$ fås der ifølge ovenstående den 2.5%-nedre envelope og den 97.5%-øvre envelope svarende til, at området mellem disse envelopes er et 95%-konfidensinterval.

I dette kapitel har summary statistics været benyttet til at afgøre, om en given punktproces har mere eller mindre clustertendenser end en Poisson-proces. På grund af denne egenskab benyttes en Poisson-proces som referenceproces, når forskellige punktprocesser sammenlignes. I Kapitel 8 ses eksempler på benyttelse af summary statistics, og ligeledes benyttes summary statistics i forbindelse med databeskrivelse og modelopstillinger i Kapitel 9 og Kapitel 10. I det følgende kapitel gennemgås en anden type punktproces.

Cox-processer

Den tidligere beskrevne Poisson-proces er ofte en for simpel model til at beskrive et punktmønster. Ved benyttelse af den gennemgåede teori angående Poisson-processer er det dog muligt at opnå punktprocesser, der er i stand til at beskrive andre typer af punktmønstre. Til konstruktion af punktprocesser, der kan beskrive en tiltrækning eller clustertendenser blandt punkter introduceres nu Cox-processer.

I dette kapitel gives en introduktion til forskellige Cox-processer. Specielt beskrives clusterprocesser og Neyman-Scott processer, så som Thomas- og Matérn-processen. Kapitlet er skrevet på baggrund af [Møller og Waagepetersen, 2004, Kapitel 5].

5.1 Definitioner og resultater

I dette afsnit introduceres forskellige centrale begreber vedrørende Cox-processer. Først defineres den generelle Cox-proces og dernæst betragtes intensitetsmålet og egenskaber ved Cox-processen.

Definition 5.1 (Cox-proces)
Lad $S \subseteq \mathbb{R}^d$, og lad $Z = \{Z(\xi) | \xi \in S\}$, hvor $Z(\xi)$ er en ikke-negativ stokastisk variabel for alle $\xi \in S$, således at funktionen $\xi \mapsto Z(\xi)$ med sandsynlighed en er lokalt integrabel. Hvis $X|Z$ er en Poisson-proces på S med intensitetsfunktion Z, kaldes X for en Cox-proces drevet af Z.

Da $X|Z$ er en Poisson-proces, kan egenskaberne fra Kapitel 3 benyttes, når den betingede fordeling betragtes.

Sætning 5.2 (Intensitetsfunktionen) [Møller og Waagepetersen, 2004, s. 60]
Lad $X \subseteq S$ og $\xi, \eta \in S$. Hvis X er en Cox-proces drevet af Z, så er intensitetsfunktionen for Cox-processen givet ved

$$\rho(\xi) = \mathbb{E}[Z(\xi)],$$

og andenordens produkttætheden er givet ved

$$\rho^{(2)}(\xi, \eta) = \mathbb{E}[Z(\xi)Z(\eta)].$$

Bevis:
Lad X være en Cox-proces på S med $B \subseteq S$ og $C \subseteq S \times S$. Dermed fås det ved

benyttelse af Definition 2.10, at intensitetsmålet er

$$\mu(B) = \mathbb{E}[N(X \cap B)] = \mathbb{E}[\mathbb{E}[N(X \cap B)|Z]] = \mathbb{E}\left[\int_B Z(\xi)\ \mathrm{d}\xi\right] = \int_B \mathbb{E}[Z(\xi)]\ \mathrm{d}\xi,$$

hvor andet lighedstegn følger af [Olofsson, 2005, s.194]. Dermed er $\rho(\xi) = \mathbb{E}[Z(\xi)]$. Ved at anvende andenordens momentmålet fra Definition 4.2 ses det, at

$$\begin{aligned}
\alpha^{(2)}(C) &= \mathbb{E}\left[\mathbb{E}\left[\sum_{\xi,\eta\in X}^{\neq} \mathbb{1}[(\xi,\eta)\in C]\,\middle|\, Z\right]\right] \\
&= \mathbb{E}\left[\int\int \mathbb{1}[(\xi,\eta)\in C] Z(\xi)Z(\eta)\ \mathrm{d}\xi\mathrm{d}\eta\right] \\
&= \int\int \mathbb{1}[(\xi,\eta)\in C]\ \mathbb{E}[Z(\xi)Z(\eta)]\ \mathrm{d}\xi\mathrm{d}\eta,
\end{aligned}$$

hvor andet lighedstegn følger af Sætning 4.3, og af at andenordensprodukttætheden er produktet af de enkelte intensiteter for Poisson-proces. Det vil sige, at $\rho^{(2)}(\xi,\eta) = \mathbb{E}[Z(\xi)Z(\eta)]$ for Cox-processen X. □

Dermed fås det fra Sætning 5.2 og Definition 4.6, at parkorrelationsfunktionen for en Cox-proces er givet ved

$$g(\xi,\eta) = \frac{\mathbb{E}[Z(\xi)Z(\eta)]}{\mathbb{E}[Z(\xi)]\mathbb{E}[Z(\eta)]}. \tag{5.1}$$

Oftest gælder det for Cox-processer, at $g(\cdot)$ er større end en. Det vil sige, at Cox-processer kan benyttes til at beskrive tiltrækning og clustertendenser mellem punkter, jf. Afsnit 4.1. Tegn på tiltrækning mellem punkter i et punktmønster ses også ud fra følgende sætning.

Sætning 5.3 (Overspredning) [Møller og Waagepetersen, 2004, s. 60]
Lad $B \subseteq S$, så gælder det for Cox-processer, at

$$Var[N(B)] \geq \mathbb{E}[N(B)].$$

Bevis:
Lad X være en Cox-proces på S, og $B \subseteq S$. Ved en omskrivning af $\mathrm{Var}[N(B)]$ fra [Olofsson, 2005, s.197] fås det, at

$$\mathrm{Var}[N(B)] = \mathbb{E}[\mathrm{Var}[N(B)|Z]] + \mathrm{Var}\left[\mathbb{E}[N(B)|Z]\right] \geq \mathbb{E}[\mathrm{Var}[N(B)|Z]] = \mathbb{E}\left[N(B)\right],$$

da $N(B)|Z \sim \mathrm{poi}\left(\int_B Z(u)\ \mathrm{d}u\right)$. □

Sætning 5.4 [Møller og Waagepetersen, 2004, s. 58]
Funktionen $\xi \mapsto Z(\xi)$ er lokalt integrabel med sandsynlighed en, hvis $\rho(\xi) = \mathbb{E}[Z(\xi)]$ eksisterer og er lokalt integrabel.

Bevis:
Antag, at $\rho(\xi) = \mathbb{E}[Z(\xi)]$ eksisterer og er lokalt integrabel, dvs.

$$\infty > \int_B \mathbb{E}[Z(\xi)] \, \mathrm{d}\xi = \mathbb{E}\left[\int_B Z(\xi) \, \mathrm{d}\xi\right].$$

Heraf fås $\int_B Z(\xi) \, \mathrm{d}\xi < \infty$, dvs. lokalt integrabel. □

Med restriktionen af Cox-processen ved X_B, hvor $B \subseteq S$ og $|B| < \infty$, er tætheden af X_B mht. standard Poisson-processen, for et endeligt punktmønster $x \subset B$, jf. (3.6) givet ved

$$f(x) = \int f(x|Z)f(Z) \, \mathrm{d}Z = \mathbb{E}[f(x|Z)] = \mathbb{E}\left[\exp\left(|B| - \int_B Z(\xi) \, \mathrm{d}\xi\right) \prod_{\xi \in x} Z(\xi)\right]. \tag{5.2}$$

Jævnfør [Møller og Waagepetersen, 2004, s.61] er integralet i ovenstående svær at beregne og middelværdien er derfor ofte ukendt.

Da $g(\cdot) > 1$ for Cox-processer vises det i det følgende, at disse kan konstrueres til at beskrive clustertendenser.

5.2 Generelt om clusterprocesser

Til at beskrive clustertendenser mellem punkter i en punktproces kan clusterprocesser benyttes. I disse processer kan punktmønstret betragtes som en foreningsmængde af delpunktmønstre. Delpunktmønstrene antages at forgrene sig ud fra forskellige punkter, kaldet for centerpunkter eller mødrepunkter. Mængden af alle centerpunkter danner en punktproces.

Definition 5.5 (Centerproces)
En punktproces C bestående af centerpunkter c kaldes en centerproces.

Ethvert centerpunkt $c \in C$ kan tillægges en punktproces, der beskriver punktmønstret omkring $c \in C$. Denne punktproces kaldes for et cluster.

Definition 5.6 (Cluster)
Lad C være en centerproces. Punktprocessen hørende til et $c \in C$ kaldes et cluster og betegnes X_c.

For at iagttage hele punktmønstret skal foreningsmængden af de forskellige clustre betragtes. Dette giver en ny punktproces kaldet en clusterproces.

Definition 5.7 (Clusterproces)
Lad C være en centerproces med tilhørende clustre X_c. Foreningsmængden af clustrene $X = \bigcup_{c \in C} X_c$ kaldes en clusterproces.

Hvis C yderligere er en Poisson-proces, kaldes X en Poisson-clusterproces. Hvis C er en centerproces med clustre X_c, så er foreningsmængden $X = \bigcup_{c \in C} X_c$ en Cox-proces, hvis og kun hvis X_c er uafhængige Poisson-processer. Dermed gælder det for Cox-processer, at centerpunkterne ikke kan være en del af clustrene, da X_c givet $c \in C$ ellers ikke er uafhængige Poisson-processer.

Clusterprocesser anvendes i det efterfølgende til at beskrive forskellige typer af Cox-processer. Bemærk, at ikke alle clusterprocesser er Cox-processer og omvendt.

5.3 Neyman-Scott processer

I det følgende introduceres visse Neyman-Scott processer som også er Cox-processer. Disse er anvendelige i forbindelse med beskrivelse af clustertendenser i punktmønstre.

Lad C være en stationær Poisson-proces på $\mathbb{R}^d$ med intensitet $\kappa > 0$. Lad endvidere X_c med $c \in C$ være uafhængige Poisson-processer på $\mathbb{R}^d$, hvor X_c har intensitetsfunktionen

$$\rho_c(\xi) = \alpha k(\xi - c),$$

hvor funktionen $k(\cdot)$ er en kerne, der er den samme for alle $c \in C$, og $\alpha > 0$. Da X_c er en Poisson-proces, beskriver parameteren α det gennemsnitlige antal punkter hørende til X_c, da kernen integrerer til en.

Foreningsmængden af X_c, dvs. $X = \bigcup_{c \in C} X_c$, er en Neyman-Scott proces med clustercentre i C og clustre X_c, hvor c ikke er en del af clustrene. Endvidere er X en Cox-proces på $\mathbb{R}^d$ drevet af

$$Z(\xi) = \sum_{c \in C} \alpha k(\xi - c),$$

ifølge Sætning 3.11, da Z er lokalt integrabel. Intensiteten af X er ifølge Sætning 5.2 givet ved

$$\mathbb{E}[Z(\xi)] = \alpha\mathbb{E}\left[\sum_{c\in C} k(\xi - c)\right] = \alpha \int k(\xi - c)\kappa \, \mathrm{d}c = \alpha\kappa. \tag{5.3}$$

Af ovenstående ses det, at X er stationær, da både α og κ er konstante. Endvidere er X isotropisk, hvis $k(\xi) = k(|\xi|)$ er invariant under rotation.

Ligeledes er

$$\begin{aligned}
\mathbb{E}[Z(\xi)Z(\eta)] &= \alpha^2\mathbb{E}\left[\sum_{c\in C} k(\xi - c) \sum_{c\in C} k(\eta - c)\right] \\
&= \alpha^2\mathbb{E}\left[\sum_{c,\tilde{c}\in C}^{\neq} k(\xi - c)k(\eta - \tilde{c})\right] + \alpha^2\mathbb{E}\left[\sum_{c\in C} k(\xi - c)k(\eta - c)\right] \\
&= \alpha^2 \int\int k(\xi - c)k(\eta - \tilde{c})\kappa^2 \, \mathrm{d}c\mathrm{d}\tilde{c} + \alpha^2 \int k(\xi - c)k(\eta - c)\kappa \, \mathrm{d}c \\
&= \alpha^2\kappa^2 + \alpha^2\kappa \int k(\xi - c)k(\eta - c) \, \mathrm{d}c.
\end{aligned}$$

Dermed fås parkorrelationen jf. (5.1) til $g(\xi, \eta) = 1 + \kappa^{-1} \int k(\xi - c)k(\eta - c) \, \mathrm{d}c$, og da $k(\cdot)$ er stationær, fås det ved substitution af $u = \xi - c$ og $v = \eta - \xi$, at

$$g(v) = 1 + \frac{\int k(u)k(u + v) \, \mathrm{d}u}{\kappa}. \tag{5.4}$$

Det vil sige, at Neyman-Scott beskriver clustertendenser, da $k(\cdot) > 0$ og $\kappa > 0$.

5.3.1 Thomas-processer

Et eksempel på en Neyman-Scott proces som også er en Cox-proces, er Thomas-processen. Ved Thomas-processen er kernerne givet ved

$$k(\xi) = \frac{1}{(2\pi\omega^2)^{\frac{d}{2}}} \exp\left(-\frac{|\xi|^2}{2\omega^2}\right), \tag{5.5}$$

hvilket er tætheden for en d-dimensional normalfordeling $N_d(0, \omega I_d)$ med forventningsværdi nul og varians $\omega^2 > 0$. Dette skal forstås som, at normalfordelingerne har udgangspunkt i hvert $c \in C$ og notationen med middelværdien 0 benyttes, da kernerne betragtes hver for sig.

Jævnfør (5.5) er Thomas-processen isotropisk, og dermed kan det vises vha. (5.4), at parkorrelationsfunktionen for Thomas-processen bliver

$$g(\xi) = 1 + \frac{1}{\kappa\,(4\pi\omega^2)^{\frac{d}{2}}} \exp\left(-\frac{|\xi|^2}{4\omega^2}\right). \tag{5.6}$$

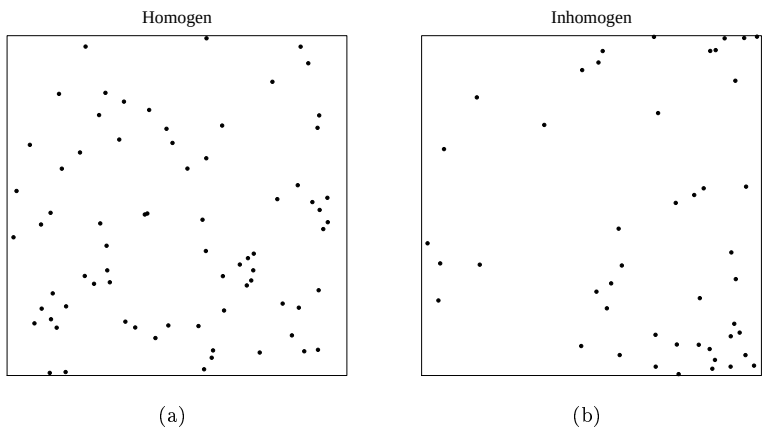
Da $g(\xi) \geq 1$ ses det, at Thomas-processen har clustertendenser.

For $\mathbb{R}^2$ kan K-funktionen for Thomas-processsen fås til

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

På Figur 5.1(a) ses et plot af en homogen Thomas-proces på $S = [0,1] \times [0,1]$. Parametrene for realisationen af Thomas-processen er: $\kappa = 10$, der beskriver intensiteten af Poisson-centerprocessen, $\omega = 0.2$ der beskriver spredningen af kernerne og $\alpha = 5$, der beskriver det forventede antal af punkter omkring et centerpunkt.

På Figur 5.1(b) ses et plot af en inhomogen Thomas-proces med parametrene: $\kappa = 10$, $\sigma = 0.2$ og $\alpha(x,y) = 5\exp(2x-1)$ for $(x,y) \in S$. Teorien for inhomogene Thomas-processer behandles dog ikke, og det bør bemærkes at resultater som eksempelvis (5.3) ikke gælder, da α og κ ikke nødvendigvis er konstanter.



Figur 5.1: Eksempler på realiseringer Thomas-processer på $S = [0,1]\times[0,1]$. (a) Realisering af homogen Thomas-proces med parametre $\kappa = 10, \omega = 0.2$ og $\alpha = 5$. (b) Realisering af inhomogen Thomas-proces med parametre $\kappa = 10$, $\omega = 0.2$ og $\alpha(x,y) = 5\exp(2x-1)$ for $(x,y) \in S$.

5.3.2 Matérn-processer

Endnu et eksempel på en Neyman-Scott-proces der også er en Cox-proces, er Matérn-processen. I en Matérn-proces er kernerne givet ved en ligefordelt tæthed i kuglen $b(0,R)$, således at

$$k(\xi) = \frac{\mathbb{1}[|\xi| < R]}{\omega_d R^d},$$

hvoraf det ses, at $k(\cdot)$ er invariant under rotation, og en Matérn-proces er altså isotropisk. Jævnfør [Stoyan og Stoyan, 1994, s.311-312] er parkorrelationsfunktio-

nen for Matérn-processen i $\mathbb{R}^2$ givet ved

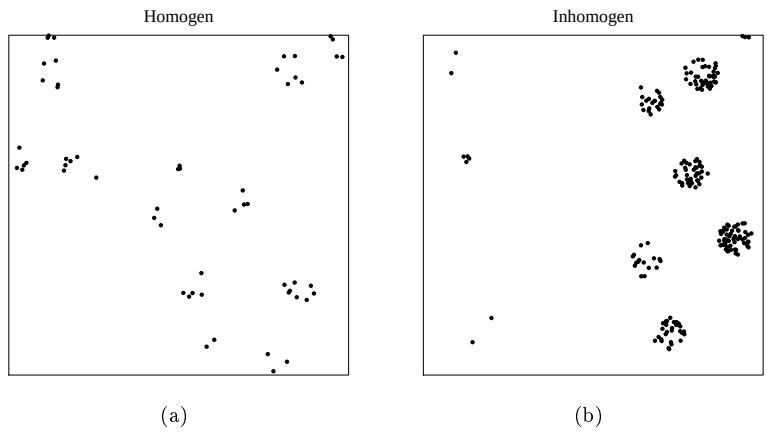
$$g(\xi) = \begin{cases} 1 + \frac{2}{\pi^2 R^2 \kappa}\left(\arccos\left(\frac{\xi}{2R} - \frac{\xi}{2R}\sqrt{1 - \frac{\xi^2}{4R^2}}\right)\right) & \text{for } 0 \le \xi \le 2R, \\ 1 & \text{for } \xi > 2R, \end{cases}$$

og K-funktionen i $\mathbb{R}^2$ er da

$$K(r) = \pi r^2 + \frac{1}{\kappa}\begin{cases} 2 + \frac{1}{\pi}\left((8z^2 - 4)\arccos z - 2\arcsin z + 4z\sqrt{(1-z^2)^3} - 6z\sqrt{1-z^2}\right) \\ 1 \end{cases}$$

for hhv. $r \le 2R$ og $r > 2R$ og med $z = r/2R$.

På Figur 5.2(a) ses et plot af en homogen Matérn-proces på $S = [0,1] \times [0,1]$. Parametrene for realiseringen af Matérn-processen er $\kappa = 10$, der beskriver intensiteten af Poisson-centerprocessen, $R = 0.05$ der er radiusparameteren af clusterne og $\alpha = 4$ der er det forventede antal af punkter hørende til hvert cluster. På Figur 5.2(b) ses en inhomogen Matérn-proces med parametre $\kappa = 10$, $R = 0.05$ og $\alpha(x,y) = 4\exp(5x - 1)$ for $(x,y) \in S$. Teorien for inhomogene Matérn-processer behandles ikke.



Figur 5.2: Eksempler på realiseringer af Matérn-processer på $S = [0,1] \times [0,1]$. (a) Realisering af homogen Matérn-proces med $\kappa = 10$, $R = 0.05$ og $\alpha = 4$. (b) Realisering af inhomogen Matérn-proces med $\kappa = 10, R = 0.05$ og $\alpha(x,y) = 4\exp(5x - 1)$ for $(x,y) \in S$.

5.4 Inferens for Cox-processer

Som nævnt i (5.2) er tætheden for en Cox-proces mht. standard Poisson-processen ofte ukendt, og det samme er derfor likelihoodfunktionen. Dette gør traditionel

parametrisk inferens for Cox-processer vanskeligt. En alternativ metode til parametrisk inferens er minimumkontrastestimation, hvor parameterestimatet opnås ved sammenligning af summary statistics. Denne metode er også brugbar ved inferens for Poisson-processen.

5.4.1 Minimumkontrastestimation

Lad nu $T(r)$ være en given summary statistic. Ideen med minimumkontrastestimation er at lade estimatet af de ukendte parametre være de parametre, der giver mindst kontrast imellem den teoretiske værdi $T(r)$ og den ikke-parametriske estimation $\hat{T}(r)$ fra data. Med mindst kontrast menes, at en given afstandsfunktion $d(\hat{T}, T)$ minimeres. Ofte benyttes

$$d(\hat{T}, T) = \int_{r_1}^{r_2} \left(\hat{T}(r) - T(r)\right)^2 \, \mathrm{d}r,$$

hvor $0 \leq r_1 < r_2$ er data-specifikke værdier. En hvilken som helst summary statistic kan i princippet bruges, dog bør der vælges en summary statistic, som er veldefineret for punktprocessen, som beskriver vigtige aspekter af data, og som ikke er for tung beregningsmæssigt.

Minimumkontrastestimation anvendes i det følgende på Thomas-processen fra Afsnit 5.3.1 i tilfældet $d = 2$.

Thomas-processen

Som summary statistic for Thomas-processen betragtes parkorrelationsfunktionen, der er givet i (5.6). De ukendte parametre er κ, ω og α. Ved benyttelse af minimumkontrastestimation fås der for $0 \leq a_1 < a_2$, at

$$\begin{aligned}
&\int_{a_1}^{a_2} (\hat{g}(r) - g(r))^2 \, \mathrm{d}r \\
&\quad = \int_{a_1}^{a_2} (\hat{g}(r) - 1 - (g(r) - 1))^2 \, \mathrm{d}r \\
&\quad = \int_{a_1}^{a_2} (\hat{g}(r) - 1)^2 \, \mathrm{d}r + \int_{a_1}^{a_2} (g(r) - 1)^2 \, \mathrm{d}r - 2\int_{a_1}^{a_2} (\hat{g}(r) - 1)(g(r) - 1) \, \mathrm{d}r \\
&\quad = \int_{a_1}^{a_2} (\hat{g}(r) - 1)^2 \, \mathrm{d}r + A(\omega^2)\frac{1}{\kappa^2} - 2B(\omega^2)\frac{1}{\kappa}, \qquad (5.7)
\end{aligned}$$

hvor

$$A(\omega^2) = \int_{a_1}^{a_2} ((g(r) - 1)\kappa)^2 \, \mathrm{d}r \qquad \text{og} \qquad B(\omega^2) = \int_{a_1}^{a_2} (\hat{g}(r) - 1)(g(r) - 1)\kappa \, \mathrm{d}r.$$

Ved at differentiere (5.7) mht. κ og sætte lig med nul fås parameterestimatet til

$$\hat{\kappa} = \frac{A(\omega^2)}{B(\omega^2)}.$$

Ved at benytte dette estimat af κ og udelade første led i (5.7), da det er konstant mht. κ og ω^2, fås estimatet af ω^2 til

$$\hat{\omega}^2 = \arg\min_{\omega^2} \left\{ \frac{A(\omega^2)}{\hat{\kappa}^2} - \frac{2B(\omega^2)}{\hat{\kappa}} \right\} = \arg\min_{\omega^2} \left\{ -\frac{B(\omega^2)^2}{A(\omega^2)} \right\} = \arg\max_{\omega^2} \left\{ \frac{B(\omega^2)^2}{A(\omega^2)} \right\},$$

hvor der med $\arg\min_{\omega^2}\{\cdot\}$ menes den værdi af ω^2, der minimerer udtrykket.

Dog mangler parameteren α fra Thomas-processen stadig at blive estimeret. Dette kan gøres ved at indsætte estimatet af intensitetsfunktionen fra Afsnit 4.2.1, dvs. $\hat{\rho} = n(x)/|W|$, i $\rho = \alpha\kappa$ fra Afsnit 5.3, og der opnås, at $\hat{\alpha} = \hat{\rho}/\hat{\kappa}$, hvorved alle parametre i Thomas-processen er estimeret.

Da Cox-processerne er anvendelige i forbindelse med punktmønstre med clustertendenser, kan de være brugbare i forbindelse med beskrivelse af datasættet for gravhøje, da plots af datasættet udviser visse clustertendenser. Da clustertendenserne for datasættet gravhøje virker til at fordele sig langs linjer, frem for direkte clustere, kan Cox-processer ikke nødvendigvis beskrive datasættet i tilstrækkelig grad.

I følgende kapitel vises det, hvordan simulering af punktprocesser er muligt. Mange af de allerede kendte punktprocesser, der har været gennemgået i dette kapitel og i Kapitel 3 er allerede implementeret i `R`. Det vil sige, at det ved benyttelse af funktioner i `R` er muligt at estimere parametre udfra datasættet og simulere fra en bestemt punktproces. For mere komplicerede punktprocesser er funktionerne ikke implementeret i `R`, så i næste kapitel vises det, hvordan det er muligt at simulere punktprocesser udfra en tæthed mht. standard Poisson-processen. Dette bliver senere brugbart, når der skal simuleres fra den opstillede model til beskrivelse af linjetendenserne i datasættet gravhøje.

Simulering af punktprocesser

En metode til at kunne simulere en punktproces med en unormeret tæthed mht. standard Poisson-processen er markovkæde Monte Carlo (MCMC). Hvis X betegner den punktproces, der ønskes simuleret, er metoden en MCMC-algoritme, der genererer en markovkæde, $Y_0, Y_1, \ldots$, sådan at for et passende stort m vil realisationer af $Y_m, Y_{m+1}, \ldots$, kunne betragtes som realisationer af X. Formålet er altså at danne en markovkæde, hvis fordeling netop konvergerer mod fordelingen af punktprocessen X, når $m \to \infty$. Den normerede tæthed for X kaldes targettætheden, og kan eksempelvis for en Poisson-proces ses som $f(\cdot)$ fra Afsnit 3.2.

I det følgende gennemgås MCMC-algoritmer til simulering af markovkæder. De MCMC-algoritmer, der betragtes, er alle Metropolis-Hastings algoritmer (MH-algoritmer). Kapitlet er skrevet på baggrund af [Møller og Waagepetersen, 2004, Kapitel 7] og [Berthelsen og Møller, 2004].

6.1 Metropolis-Hastings algoritmer

I det følgende betegner X en punktproces kendetegnet ved en unormeret tæthed $h(\cdot)$ mht. Poisson$(B, 1)$, hvor $B \subset \mathbb{R}^d$ og $0 < |B| < \infty$. I dette afsnit ønskes det at simulere fra den normerede tæthed $f(\cdot)$ mht. standard Poisson-processen, men da denne kan være ukendt, simuleres der fra den unormerede tæthed $h(\cdot) = cf(\cdot)$.

Først gennemgås en algoritme hvor antallet af punkter er konstant, og senere betragtes en algoritme hvor antallet af punkter kan variere.

6.1.1 For punktprocesser med et fast antal punkter

Antag, at punktprocessen X indeholder n punkter, dvs. $n(X) = n$, hvor $n \geq 1$ og lad den unormerede betingede tæthed $h_n(x) = h(x|n(x) = n)$. Lad de n punkter $X_1, \ldots, X_n$ i X være ordnede, og betragt den unormerede tæthed $\pi(\cdot)$ for $(X_1, \ldots, X_n)$, der er givet ved

$$\pi(x_1, \ldots, x_n) = h_n(\{x_1, \ldots, x_n\}).$$

Antag, at $\pi(x_1, \ldots, x_n)$ er defineret for alle $(x_1, \ldots, x_n) \in B^n$, dvs. nogle af punkterne tillades at være ens. Dette medfører, at simulering af mængden af ordnede punkter $\bar{X} = (X_1, \ldots, X_n)$ ud fra $\pi(\cdot)$ er ækvivalent med at simulere $X = \{X_1, \ldots, X_n\}$ ud fra den unormerede tæthed $h_n(\cdot)$.

For alle $\bar{x} = (x_1, \ldots, x_n) \in B^n$ og for alle $i \in \{1, \ldots, n\}$ angiver $q_i(\bar{x}, \cdot)$ en given tæthedsfunktion på B, der kaldes forslagstætheden. Det vil sige, at forslagstætheden genererer et forslag med udgangspunkt i $\bar{x}$. Lad $x = \{x, \ldots, x_n\}$ og $\xi \in B$, da

er Hastings-forholdet defineret ved

$$r_i(\bar{x},\xi) = \frac{h_n((x\backslash\{x_i\})\cup\{\xi\})q_i((x_1,\ldots,x_{i-1},\xi,x_{i+1},\ldots,x_n),x_i)}{h_n(x)q_i(\bar{x},\xi)}, \qquad (6.1)$$

hvor $a/0 = 1$ for $a \geq 0$. Sandsynligheden for at acceptere et forslag fra forslagstætheden er givet ved $\alpha_i(\bar{x},\xi) = \min\{1, r_i(\bar{x},\xi)\}$, og kaldes acceptsandsynligheden.

Med udgangspunkt i en given begyndelsestilstand er det ved benyttelse af den følgende algoritme muligt at simulere en markovkæde med et fast antal punkter.

Algoritme 6.1 (Move-algoritme) [Møller og Waagepetersen, 2004, Alg. 7.1]
Lad $\pi(Y_0) > 0$, hvor Y_0 er en kendt begyndelsestilstand.
For $m = 0, 1, \ldots$ givet, at $Y_m = \bar{x} = (x_1, \ldots, x_n) \in B^n$, genereres Y_{m+1} ved

1) *Generer $I_m \sim unif(\{1,\ldots,n\})$ og $R_m \sim unif([0,1])$.*
2) *Givet $I_m = i$ generer $\xi_m \sim q_i(\bar{x},\cdot)$.*
3) *Givet $I_m = i$, lad*

$$Y_{m+1} = \begin{cases} (x_1,\ldots,x_{i-1},\xi_m,x_{i+1}\ldots,x_n), & \text{hvis } R_m \leq r_i(\bar{x},\xi_m), \\ \bar{x} & \text{ellers.} \end{cases}$$

□

Hvordan Y_0 genereres afhænger normalt af det specifikke problem, eksempelvis kan det være en realisation af en binomialproces.

Algoritmen udvælger altså i 1) et punkt ligefordelt blandt punkterne i punktprocessen. Til dette punkt genereres et forslag i 2) og i 3) vurderes det, hvorvidt forslaget skal accepteres eller forkastes ved brug af R_m.

6.1.2 For punktprocesser med skiftende antal punkter

En fødsels-døds-algoritme er en algoritme, der genererer en markovkæde ved enten at tilføje (fødsel) eller slette (død) et punkt fra den nuværende punktproces X. Lad $h(\cdot)$ være en unormeret tæthed mht. Poisson$(B,1)$. Lad endvidere $x \in N_{lf}$, hvor N_{lf} er mængden af lokalt endelige punktmønstre jf. Kapitel 2. Lad nu x betegne den nuværende tilstand af kæden, lad $p(x)$ være en sandsynlighed for at foreslå en fødsel af et punkt ξ, og lad $q_b(x,\xi)$ være tæthedsfunktionen på B, for placeringen af ξ. Derudover lad $1 - p(x)$ være sandsynligheden for at foreslå død af et punkt η. Hvis $x \neq \emptyset$, lad $q_d(x,\eta)$ være en given diskret fordeling på x, der udvælger η, og hvis $x = \emptyset$ gøres intet.

For et fødselsforslag $x \mapsto x \cup \xi$ er acceptsandsynligheden givet ved

$$\alpha_b(x,\xi) = \min\{1, r_b(x,\xi)\},$$

med hastingsratioet

$$r_b(x,\xi) = \frac{h(x \cup \xi)(1 - p(x \cup \xi))q_d(x \cup \xi, \xi)}{h(x)p(x)q_b(x,\xi)}, \tag{6.2}$$

hvor $a/0 = 1$ for $a \geq 0$. For et dødsforslag $x \mapsto x \backslash \eta$ er acceptsandsynligheden givet ved

$$\alpha_d(x,\eta) = \min\{1, r_d(x,\eta)\}$$

med hastingsratioet

$$r_d(x,\eta) = \frac{h(x\backslash\eta)p(x\backslash\eta)q_b(x\backslash\eta,\eta)}{h(x)(1-p(x))q_d(x,\eta)}, \tag{6.3}$$

hvor igen $a/0 = 1$ for $a \geq 0$.

Den følgende algoritme genererer en markovkæde med et varierende antal punkter.

Algoritme 6.2 (Birth-Death-algoritme) [Møller og Waagepetersen, 2004, Alg. 7.4]
Lad Y_0 være en given begyndelsestilstand. For $m = 0, 1, \ldots,$ givet $Y_m = x \in N_f$, genereres Y_{m+1} ved

1) *Generer $R'_m \sim unif([0,1])$ og $R''_m \sim unif([0,1])$.*
2) *Hvis $R'_m \leq p(x)$, generer da $\xi_m \sim q_b(x,\cdot)$, og lad*

$$Y_{m+1} = \begin{cases} x \cup \xi_m, & \text{hvis } R''_m \leq r_b(x,\xi_m), \\ x & \text{ellers.} \end{cases}$$

3) *Hvis $R'_m > p(x)$*
   a) *og hvis $x = \emptyset$, lad da $Y_{m+1} = x$*
   b) *ellers generer $\eta_m \sim q_d(x,\cdot)$, og lad*

$$Y_{m+1} = \begin{cases} x\backslash\eta_m, & \text{hvis } R''_m \leq r_d(x,\eta_m), \\ x & \text{ellers.} \end{cases}$$

□

For denne algoritme kan Y_0 f.eks. være det tomme punktmønster eller en realisation af en Poisson-proces.

Denne algoritme vælger i 1) et $\mathbb{R}'_m$ til at afgøre, om der skal fødes eller slettes. I 2) og 3) genereres forslag for fødsel og død og det vurderes, om der skal fødes eller slettes på baggrund af et R''_m.

Algoritme 6.2 kan kombineres med Move-algoritmen fra Algoritme 6.1 således at der både skabes, slettes og flyttes punkter. Denne algoritme kaldes en Birth-Death-Move-algoritme eller blot BDM-algoritme.

Algoritme 6.3 (BDM-algoritme) [Møller og Waagepetersen, 2004, Alg. 7.5]
Lad $0 \leq q < 1$, *for* $m = 0, 1, \ldots$ *givet* Y_m,

1) *Generer* $R'''_m \sim unif([0,1])$.
2) *Hvis* $R'''_m \leq q$, *generer da* Y_{m+1} *som i Algoritme 6.1.*
3) *Hvis* $R'''_m > q$, *generer da* Y_{m+1} *som i Algoritme 6.2.*

□

Her trækkes der i 1) et tal, der afgør, om det er Algoritme 6.1 eller Algoritme 6.2, der skal køres.

Markovkæderne genereret af algoritmerne fra dette afsnit kan vises at konvergere mod en invariant tæthed. Dette betyder, at de kan benyttes til MCMC-simulering af en punktproces med en unormeret tæthed mht. standard Poisson-processen. Generelt gælder det for MH-algoritmerne, at hvis forslagssandsynligheden er symmetrisk, går disse ud i Hastings-forholdet, og MH-algoritmen reducerer dermed til, hvad der kaldes en Metropolis-algoritme.

Algoritme 6.1 er også anvendelig i forbindelse med estimation af ukendte parametre for f.eks. datasættet gravhøje. Da det ved benyttelse af Algoritme 6.1 er muligt at simulere fra tætheder med ukendte normeringskonstanter, er det muligt at simulere forskellige parametre for en given model. For datasættet gravhøje kunne det f.eks. være en intensitet. Den direkte anvendelse af denne algoritme beskrives nærmere i forbindelse med opstillingen af modellen for gravhøje.

I det foregående har der hidtil kun været beskrevet egenskaber ved og simulering af punktprocesser. For at kunne beskrive datasættet gravhøje er det nødvendigt at kombinere punktprocesteori med linjeprocesser, hvorfor næste kapitel omhandler disse linjeprocesser.

Linjeprocesser

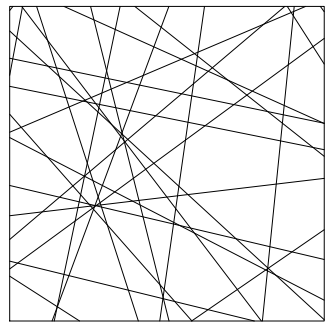
Punktprocesser er anvendelige til at beskrive punktmønstre, der indeholder tiltrækning, clustertendenser eller frastødning mellem punkter. Det er muligt at udvide disse processer, således at de er anvendelige i forbindelse med beskrivelse af punktmønstre med linjetendenser, som f.eks. datasættet for gravhøje. Til at beskrive forskellige typer af linjetendenser i punktmønstre, introduceres i dette kapitel linjeprocesser. En linjeproces kan forestilles som en punktproces, defineret på et rum, der kaldes repræsentationsrummet, hvor punkterne repræsenterer linjer i et rum bestående af linjer. Disse typer af punktprocesser kan analogt til de tidligere beskrevne punktprocesser benyttes til opstilling af modeller for mønstre af linjer, som eksempelvis tøjfibre eller mineralårer.

Der beskrives udelukkende linjeprocesser i $\mathbb{R}^2$, men disse kan med modifikationer generaliseres til $\mathbb{R}^3$. Desuden beskrives specielle typer af linjeprocesser, så som Poisson-linjeprocesser og Cox-linjeprocesser, og visse af begreberne fra Kapitel 2 om punktprocesser overføres til linjeprocesser. Kapitlet er hovedsageligt skrevet på baggrund af [Daley og Vere-Jones, 2008, Kap. 15] og [Stoyan et al., 1995, Kap. 8]. Derudover er dette kapitel baseret på [Kingman, 1993, Kap. 7], [Miles, 1964], [Santaló, 1976, Kap. 3] og [Stoyan og Stoyan, 1994, App. H].

7.1 Definitioner og resultater

I det følgende afsnit gives en beskrivelse af linjer i $\mathbb{R}^2$, således at linjeprocesser kan defineres. Derudover opnås forskellige nyttige resultater for linjeprocesser, tilsvarende dem der blev opnået for punktprocesser.

På Figur 7.1 ses et system af vilkårlige rette linjer i et område af $\mathbb{R}^2$. Mønstret linjerne danner kaldes et linjemønster, og som for punktmønstre og punktprocesser kan et linjemønster ses som en realisering af en linjeproces.



Figur 7.1: System af vilkårlige rette linjer i et område i $\mathbb{R}^2$.

Enhver ret linje ℓ i xy-planen kan skrives på formen

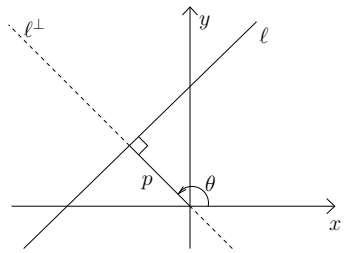
$$ax + by = c, \tag{7.1}$$

hvor parametrene $a, b, c \in \mathbb{R}$ entydigt bestemmer linjen ℓ.

En ækvivalent måde at beskrive en ret linje i $\mathbb{R}^2$ på, er ved en orientering θ samt den vinkelrette afstand p fra origo. Denne beskrivelse kan foretages på flere måder afhængig af om afstande regnes med fortegn eller ej. Hvis ℓ er en vilkårlig linje i $\mathbb{R}^2$, så eksisterer der en entydig linje $\ell^\perp$ gennem origo og vinkelret på ℓ. Afstanden fra origo til skæringspunktet mellem de to linjer er da p. Dette er illustreret på Figur 7.2, og to mulige beskrivelser af linjerne er:

(i) Afstanden p er positiv, og om vinklen θ mellem $\ell^\perp$ og x-aksen gælder der, at $\theta \in (0, 2\pi]$, hvor θ måles mod uret.

(ii) Afstanden p er positiv for skæringspunktet i øvre halvplan $\{(x, y)|y > 0\}$ og positiv x-akse $\{(x, 0)|x > 0\}$, og ellers er den negativ. For vinklen med x-aksen gælder her $\theta \in (0, \pi]$. Igen regnes θ mod uret.



Figur 7.2: Illustration der viser parametrisering af en linje ℓ ved (p, θ).

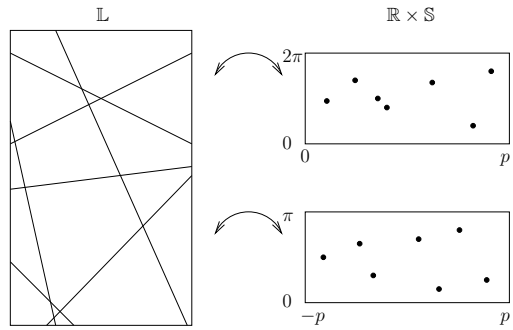
Lad nu $\mathbb{L}$ være mængden bestående af alle rette linjer i xy-planen, dvs. $\ell \in \mathbb{L}$, hvor $\mathbb{L}$ kaldes linjerummet. Af det ovenstående fremgår det, hvorledes de rette linjer i $\mathbb{L}$ kan parametriseres som punkter (p, θ). Med ovenstående repræsentation fra (i) af linjer i $\mathbb{R}^2$ er der en 1:1-korrespondance mellem $\mathbb{L}$ og

$$\mathbb{R}_+ \times (0, 2\pi] = \{(p, \theta)|0 \leq p < \infty, 0 < \theta \leq 2\pi\},$$

hvor $\mathbb{R}_+ \times (0, 2\pi]$ kaldes repræsentationsrummet for linjerne. Tilsvarende fås repræsentationsrummet for parametriseringen (ii) til

$$\mathbb{R} \times (0, \pi] = \{(p, \theta)|p \in \mathbb{R}, 0 < \theta \leq \pi\}.$$

Det vil sige, at der nu er to repræsentationer af linjer, enten som punkter i et repræsentationsrum eller som linjer i $\mathbb{L}$. I det efterfølgende benævnes begge repræsentationsrum $\mathbb{R} \times \mathbb{S}$, hvor $\mathbb{R}$ blot repræsenterer $\mathbb{R}_+$ når $\mathbb{S} = (0, 2\pi]$. Figur 7.3 illustrerer 1:1-korrespondancen mellem en delmængde af $\mathbb{L}$ og den tilhørende delmængde af repræsentationsrummet.



Figur 7.3: Illustration af sammenhængen mellem en delmængde af linjerummet og den tilhørende delmængde af repræsentationsrummet.(Bemærk, linjerne er ikke de viste punkter i denne figur.)

Definitionen af en linjeproces gives udfra teorien om at linjer både kan beskrives i linjerummet og i repræsentationsrummet. Da der er en 1:1-korrespondance mellem linjer i linjerummet og punkterne på repræsentationsrummet defineres linjeprocesser på to ækvivalente måder.

Definition 7.1 (Linjeproces)
Lad $\mathbb{L}$ være mængden af alle rette linjer i $\mathbb{R}^2$.

i) En stokastisk tællelig delmænge $Y \subseteq \mathbb{L}$ kaldes en linjeproces på $\mathbb{L}$.

ii) En linjeproces på $\mathbb{L}$ er en punktproces på repræsentationsrummet $\mathbb{R} \times \mathbb{S}$, hvor hvert punkt $(p, \theta) \in \mathbb{R} \times \mathbb{S}$ repræsenterer en linje i $\mathbb{L}$.

I punkt i) betegner en tællelig delmængde Y af $\mathbb{L}$, at antallet af linjer i Y er tælleligt.

Som for punktprocesserne kan linjeprocesserne defineres til at være lokalt endelige jf. Definition 2.9, dvs. at kun endeligt mange linjer rammer hver begrænset mængde. En linjeproces er en vilkårlig delmængde af $\mathbb{R} \times \mathbb{S}$, og linjeprocessen er altså lokalt endelig, når delmængden er lokalt endelig på $\mathbb{R} \times \mathbb{S}$, dvs. når delmængden er en punktproces på $\mathbb{R} \times \mathbb{S}$. I det følgende betragtes udelukkende lokalt endelige linjemønstre og linjeprocesser. Da linjeprocesser på $\mathbb{L}$ er ækvivalente med punktprocesser på $\mathbb{R} \times \mathbb{S}$, kan teorien for linjeprocesser ses som specialtilfælde af teorien for punktprocesser og i det følgende vises sammenhængen mellem stationaritet af de to rum.

7.1.1 Stationaritet og isotropi

Med stationaritet og isotropi af linjeprocesser menes invarians af fordelingerne under hhv. translation og rotation i planen, hvor der med fordelingen af en linjeproces menes fordelingen af den tilhørende punktproces i repræsentationsrummet.

Definition 7.2 (Stationaritet og isotropi)
En linjeproces $Y = \{\ell_1, \ell_2, \ldots, \ell_n\}$ er stationær, hvis den har samme fordeling som $Y_T = \{T\ell_1, T\ell_2, \ldots, T\ell_n\}$ for enhver translation T. Tilsvarende gælder ved isotropi for enhver rotation O.

Stationaritet og isotropi af linjeprocesser er altså analogt til stationaritet og isotropi af punktprocesser. Fordelingen af en linjeproces er defineret i repræsentationsrummet, men translationer og rotationer af linjer foretages i linjerummet. Det er derfor nødvendigt at undersøge effekten af translationer og rotationer af linjerne i $\mathbb{L}$ på punkterne i $\mathbb{R} \times \mathbb{S}$.

En translation af en linjeproces med afstand s langs x-aksen medfører

$$(p, \theta) \to (p + s\cos(\theta), \theta),$$

hvilket svarer til en forvridning (shear) af repræsentationsrummet. En rotation af en linjeproces med vinkel γ omkring origo medfører

$$(p, \theta) \to (p, \theta + \gamma), \tag{7.2}$$

hvor additionen af vinkler regnes modulo 2π.

En transformation $T_{(s,\gamma)}$, der er en translation med længde s i retningen γ, af et linjemønster ændrer ikke på vinklen θ, men den medfører

$$(p, \theta) \to (p + s\cos(\theta - \gamma), \theta). \tag{7.3}$$

Denne transformation laver altså en forvridning af repræsentationsrummet.

Da fordelingen af en linjeproces er defineret i repræsentationsrummet medfører stationaritet og isotropi af en linjeproces, at de sædvanlige stationaritets- og isotropibetingelser for en punktproces på $\mathbb{R}^2$ skal erstattes med, at punktprocessen i repræsentationsrummet skal være invariant under hhv. (7.2) og (7.3).

En linjeproces er altså stationær, hvis den tilhørende punktproces i repræsentationsrummet $Y' = \{(p_1, \theta_1), (p_2, \theta_2), \ldots, (p_n, \theta_n)\}$ har den samme fordeling som den transformerede punktproces $Y'_{T_{(s,\gamma)}} = \{(p_1 + s\cos(\theta_1 - \gamma), \theta_1), (p_2 + s\cos(\theta_2 - \gamma), \theta_2), \ldots, (p_n + s\cos(\theta_n - \gamma), \theta_n)\}$ for ethvert $s \in \mathbb{R}$ og $\gamma \in (0, 2\pi]$. En linjeproces er tilsvarende isotrop, hvis $Y' = \{(p_1, \theta_1), (p_2, \theta_2), \ldots, (p_n, \theta_n)\}$ har den samme fordeling som $Y'_R = \{(p_1, \theta_1 + \gamma), (p_2, \theta_2 + \gamma), \ldots, (p_n, \theta_n + \gamma)\}$ for ethvert $\gamma \in (0, 2\pi]$, hvor addition af vinkler regnes modulo 2π.

Intensiteten af en linjeproces bestemmes udfra intensiteten af punktprocessen på repræsentationsrummet, som det vises i det følgende afsnit. I dette projekt behandles udelukkende stationære linjeprocesser og her er intensiteten konstant. Når

intensiteten er konstant, er punktprocessen på repræsentationsrummet stationær, også under forvridninger, og der opstår således ikke problemer med oversættelsen af stationaritet.

7.1.2 Typer af linjeprocesser

Ligesom der eksisterer forskellige typer af punktprocesser, findes der også forskellige typer linjeprocesser. Typen af en linjeproces bestemmes udfra punktprocessen på $\mathbb{R} \times \mathbb{S}$.

Definition 7.3 (Poisson-linjeproces)
Hvis punktprocessen på $\mathbb{R} \times \mathbb{S}$ er en Poisson-proces, kaldes linjeprocessen på $\mathbb{L}$ for en Poisson-linjeproces.

Beregningen af intensiteten for en linjeproces er forskellig for de to repræsentationsrum, og den afhænger af typen af linjeproces, og dermed af typen af punktprocessen på repræsentationsrummet. Det er ikke altid indlysende, hvordan en intensitet for en linjeproces beregnes. I det følgende forklares det, hvorledes intensiteten af en Poisson-linjeproces kan beregnes, ved benyttelse af repræsentationsrummet $\mathbb{R} \times (0, \pi]$.

For en Poisson-linjeproces følger linjerne en fuldstændig rumlig tilfældighed, hvilket betyder, at de vinkelrette afstande udgør en Poisson-proces på $\mathbb{R}$, med en intensitet som i det følgende benævnes λ. Vinklerne hørende til linjerne er uafhængige og ligefordelte over $(0, \pi]$. Det vil sige, at punkterne i repræsentationsrummet kan betragtes som en mærket Poisson-proces, hvor vinklerne er mærkerne jf. Definition 3.13. Sætning 3.14 om mærkede Poisson-processer benyttes nu ved at betragte vinklerne som mærker hørende til hvert punkt på Poisson$(\mathbb{R}, \lambda)$, hvor hvert mærke er fordelt med tæthed $p = 1/\pi$. Sætning 3.14 giver nu, at punkterne er Poisson$(\mathbb{R} \times (0, \pi], \lambda/\pi)$ i repræsentationsrummet. Den konstante intensitet λ kaldes i det efterfølgende intensiten af Poisson-linjeprocessen, og den tilhørende intensitet af punktprocessen er altså λ/π.

Det samme gælder for repræsentationsrummet $\mathbb{R}_+ \times (0, 2\pi]$. Her vil intensiteten af Poisson-processen, som de vinkelrette afstande udgør, være en Poisson-proces på $\mathbb{R}_+$, men til gengæld med intensitet 2λ. Samtidig er $p = 1/2\pi$, således intensiteten i repræsentationsrummet igen er λ/π.

Som oftest kendes intensiteten for Poisson-linjeprocessen ikke, men ved at estimere intensiteten af Poisson-processen i repræsentationsrummet kan den bestemmes. Estimeres intensiteten af Poisson-processen til ρ, kan intensiteten af Poisson-linjeprocessen fås til $\rho\pi$. Intensiteten af en Poisson-linjeproces kan være svær at fortolke, da den ikke har nogen logisk forklaring som eksempelvis antal linjer pr. areal, men til gengæld har den altså en tæt relation til intensiteten af den tilhørende Poisson-proces.

Definition 7.4 (Cox-linjeprocesser)
Hvis punktprocessen på $\mathbb{R} \times \mathbb{S}$ er en Cox-proces, kaldes linjeprocessen på $\mathbb{L}$ for en Cox-linjeproces.

Da linjeprocesser er beskrevet ud fra punktprocesser på repræsentationsrummet, er linjeprocesser ligesom punktprocesserne bestemt ud fra mål på repræsentationsrummet.

7.1.3 Intensitetsmål for linjeprocesser

Ved at betragte intensitetsmålene for linjeprocesser er det muligt at opnå nogle resultater vedrørende eksempelvis det forventede antal linjer, der rammer en konveks mængde (Definition A.1 i Appendiks A). Beviserne laves hovedsageligt for repræsentationsrummet $\mathbb{R} \times (0, \pi]$, men resultaterne gælder også for $\mathbb{R}_+ \times (0, 2\pi]$.

Det forventede antal af punkter fra en punktproces X i $A \subseteq \mathbb{R}^2$ er jf. Definition 2.10 givet ved

$$\mathbb{E}[N(A)] = \mu(A).$$

I dette afsnit skal følgende lighed benyttes

$$\begin{aligned}\mathbb{E}\left[\sum_{\xi \in X} h(\xi)\right] &= \int h(\xi)\rho(\xi)\mathrm{d}\xi = \int h(\xi)\mu(\mathrm{d}\xi) = \int h(\xi)\mathbb{E}[N(\mathrm{d}\xi)] \\ &= \int \mathbb{E}[h(\xi)N(\mathrm{d}\xi)] = \mathbb{E}\left[\int h(\xi)N(\mathrm{d}\xi)\right],\end{aligned}$$

hvor første lighedstegn følger af Sætning 4.3.

Beviset for den følgende sætning kan findes i [Daley og Vere-Jones, 2008, s. 473].

Sætning 7.5 [Stoyan et al., 1995, Theorem 8.1]
Lad intensitetsmålet $\mu(\cdot)$ være et lokalt endeligt mål på $\mathbb{R} \times \mathbb{S}$, som er invariant under forvridningerne fra (7.3). Da er $\mu(\cdot)$ på formen

$$\mu(dp \times d\theta) = \nu(dp)G(d\theta),$$

hvor $G(\cdot)$ er et endeligt mål på $\mathbb{S}$, og $\nu(\cdot)$ er Lebesgue-målet på $\mathbb{R}$.

Da $G(\cdot)$ er et endeligt mål, kan dette normeres ved

$$\Pi(\mathrm{d}\theta) = \frac{G(\mathrm{d}\theta)}{\int_{\mathbb{S}} G(\mathrm{d}\theta)} = \frac{G(\mathrm{d}\theta)}{G(\mathbb{S})},$$

hvor $\Pi(\mathrm{d}\theta)$ kan tolkes som sandsynligheden for at en vilkårlig linje har orienteringen θ. Hvis f.eks. vinklen θ er ligefordelt, vil

$$\Pi(\mathrm{d}\theta) = \frac{\mathrm{d}\theta}{2\pi} \qquad \text{eller} \qquad \Pi(\mathrm{d}\theta) = \frac{\mathrm{d}\theta}{\pi},$$

afhængigt af hvilket repræsentationsrum der betragtes.

Sætning 7.6 [Stoyan et al., 1995, Theorem 8.1]
Lad intensitetsmålet $\mu(\cdot)$ være et lokalt endeligt mål på $\mathbb{R} \times \mathbb{S}$, som er invariant under rotationerne fra (7.2) og forvridningerne fra (7.3). Da er $\mu(\cdot)$ på formen

$$\mu(dp \times d\theta) = c\nu(dp \times d\theta),$$

hvor $0 \leq c < \infty$ og $\nu(\cdot)$ er Lebesgue-målet på $\mathbb{R}^2$.

Det vil sige, at i dette tilfælde svarer intensitetsmålet $\mu(\cdot)$ altså til Lebesgue-målet på $\mathbb{R} \times \mathbb{S}$ ganget en konstant.

7.1.4 Linjemål

For enhver linje ℓ_i i en konveks mængde $A \subseteq \mathbb{R}^2$ eksisterer der en veldefineret længde $\nu(\ell_i \cap A)$. Enhver mængde af linjer i A kan ses som en realisering af en given linjeproces, og linjemålet defineres jf. [Daley og Vere-Jones, 2008, s.474] ved

$$Z(A) = \sum_i \nu(\ell_i \cap A).$$

Det vil sige, at linjemålet er summen af længden af linjerne indeholdt i A.

Sætning 7.7 [Daley og Vere-Jones, 2008, Proposition 15.3.III]
Lad $Z(\cdot)$ være linjemålet hørende til en stationær linjeproces Y på $\mathbb{R}^2$. Så er $Z(\cdot)$ et stationært mål på $\mathbb{R}^2$, og hvis $\mu(\cdot)$ er et lokalt endeligt intensitetsmål på $\mathbb{R} \times \mathbb{S}$ for Y, så har $Z(\cdot)$ intensiteten

$$m := G(\mathbb{S}) = \int_{(0,1] \times \mathbb{S}} \mu(dp \times d\theta).$$

Bevis:
Linjemålet $Z(\cdot)$ er jf. Kapitel 2 stationært, da linjeprocessen Y er stationær. Lad nu længden af en linje i en begrænset mængde $A \subseteq \mathbb{R}^2$ være udtrykt ved koordinaterne (p, θ) fra repræsentationsrummet, således at

$$\nu_A(p, \theta) = \nu(\ell(p, \theta) \cap A),$$

hvor $\ell(p,\theta)$ er en linje i linjerummet. Så kan linjemålet udtrykkes vha. punkprocessen på repræsentationsrummet ved

$$Z(A) = \int_{\mathbb{R}\times\mathbb{S}} \nu_A(p,\theta) N(\mathrm{d}p \times \mathrm{d}\theta),$$

hvor $N(\cdot)$ er tællemålet for punktprocessen på $\mathbb{R}\times\mathbb{S}$. Der fås, at den forventede længde af linjerne i A er

$$\begin{aligned}\mathbb{E}[Z(A)] &= \int_{\mathbb{R}\times\mathbb{S}} \nu_A(p,\theta)\mu(\mathrm{d}p \times \mathrm{d}\theta) \\ &= \int_{\mathbb{S}}\int_{\mathbb{R}} \nu_A(p,\theta)\nu(\mathrm{d}p)G(\mathrm{d}\theta)\end{aligned}$$

hvor andet lighedstegn følger af Sætning 7.5.

For et fastholdt θ er integralet over $\mathbb{R}$ svarende til arealet af A, da $\nu_A(\cdot)$ kun betegner længden af linjen indenfor A. Dermed fås det, at

$$\mathbb{E}[Z(A)] = \int_{\mathbb{S}} G(\mathrm{d}\theta)\nu(A) = m\nu(A),$$

svarende til intensiteten af linjeprocessen ganget arealet af A. □

For et givet linjemønster kan et alternativ til linjemålet $Z(\cdot)$ være antallet af linjer, der rammer en konveks mængde jf. Definition A.1. Antallet af linjer, der rammer en konveks mængde er dog ikke et mål jf. Definition 2.3, da det ikke er additivt for disjunkte mængder, da samme linje kan gennemskære flere forskellige konvekse mængder.

Sætning 7.8 [Daley og Vere-Jones, 2008, Proposition 15.3.IV]
Lad Y være en stationær linjeproces på $\mathbb{R}^2$ med intensiteten m og lad sandsynligheden for, at en vilkårlig linje har en bestemt orientering være givet ved $\Pi(\cdot)$. Lad endvidere V være en bestemt linje i $\mathbb{R}^2$ med koordinaterne (p_V, θ_V) og lad X_V være punktprocessen på V skabt af skærringspunkterne med linjerne i linjeprocessen. Så er X_V en stationær punktproces på V med intensitet

$$m_V = m\int_{\mathbb{S}} |\cos(\theta - \theta_V)|\Pi(\ d\theta).$$

Bevis:
Da punktprocessen X_V er genereret af den stationære linjeproces Y, er der en sammenhæng mellem antallet linjer, der skærer V og linjeprocessen Y. Jævnfør [Daley og Vere-Jones, 2008, s. 475] gælder det, at for enhver begrænset målelig funktion $h(\cdot)$ med begrænset støtte jf. Definition A.2 i $\mathbb{R}$ fås

$$\int_{\mathbb{R}} h(x)N_V(\mathrm{d}x) = \int_{\mathbb{R}\times\mathbb{S}} h(x(p,\theta))N(\mathrm{d}p \times \mathrm{d}\theta),$$

hvor $x = x(p, \theta)$ er afstanden fra et fastholdt origo på V til skæringspunktet med linjen givet ved (p, θ). Da Y er stationær, kan linjen V vælges til at gå gennem origo for linjeprocessen Y, og stationariteten af Y medfører også, at X_V er stationær. Dermed er afstanden fra origo til et skæringspunkt med en linje givet ved

$$x(p, \theta) = \frac{p}{\cos(\theta - \theta_V)}. \tag{7.4}$$

Det forventede antal af linjer, der skærer V, er da givet ved

$$\begin{aligned} \mathbb{E}\left[\int_{\mathbb{R}} h(x) N_V(\mathrm{d}x)\right] =& \mathbb{E}\left[\int_{\mathbb{R}\times\mathbb{S}} h(x(p,\theta)) N(\mathrm{d}p \times \mathrm{d}\theta)\right] \\ =& \int_{\mathbb{R}\times\mathbb{S}} h(x(p,\theta)) \mu(\mathrm{d}p \times \mathrm{d}\theta). \end{aligned} \tag{7.5}$$

Ved benyttelse af Sætning 7.5 og Sætning 7.7 og ved indsættelse af (7.4) fås det, at

$$\begin{aligned} \int_{\mathbb{R}\times\mathbb{S}} h(x(p,\theta)) \mu(\mathrm{d}p \times \mathrm{d}\theta) &= \int_{\mathbb{R}\times\mathbb{S}} h\left(x(p,\theta)\right) \nu(\mathrm{d}p) m \Pi(\mathrm{d}\theta) \\ &= \int_{\mathbb{R}\times\mathbb{S}} h\left(\frac{p}{\cos(\theta - \theta_V)}\right) m \nu(\mathrm{d}p) \Pi(\mathrm{d}\theta). \end{aligned}$$

Substitueres der med $u = p/\cos(\theta - \theta_V)$ og dermed $\nu(\mathrm{d}p) = |\cos(\theta - \theta_V)| \nu(\mathrm{d}u)$, hvor numerisktegnet opstår da $p \in \mathbb{R}$, fås

$$\int_{\mathbb{S}} \int_{\mathbb{R}} h(u) |\cos(\theta - \theta_V)| m \nu(\mathrm{d}u) \Pi(\ \mathrm{d}\theta) = \int_{\mathbb{R}} m \int_{\mathbb{S}} |\cos(\theta - \theta_V)| \Pi(\ \mathrm{d}\theta) h(u) \nu(\mathrm{d}u).$$

Venstresiden af (7.5) kan skrives som

$$\int_{\mathbb{R}} h(x) \mathbb{E}[N_V(\mathrm{d}x)] = \int_{\mathbb{R}} h(x) m_V \nu(\mathrm{d}x),$$

da X_V er stationær. Ved at betragte integranderne over $\mathbb{R}$ er det muligt at identificere

$$m_V = m \int_{\mathbb{S}} |\cos(\theta - \theta_V)| \Pi(\mathrm{d}\theta).$$

□

Er linjeprocessen Y fra Sætning 7.8 endvidere isotropisk er m_V uafhængig af V og er for repræsentationsrummet $\mathbb{R}_+ \times (0, 2\pi]$ givet ved

$$m_V = m \int_0^{2\pi} |\cos(\theta - \theta_V)| \frac{\mathrm{d}\theta}{2\pi} = \frac{2m}{\pi},$$

da orienteringen af linjerne er ligefordelte. Bemærk at for repræsentationsrummet $\mathbb{R} \times (0, \pi]$ vil resultatet være det samme, da der så skal benyttes $\Pi(\theta) = \mathrm{d}\theta/\pi$.

Når linjeprocessen er en Poisson-linjeproces med intensitet λ, bliver den punktproces, der skabes på linjen V en Poisson-proces på $\mathbb{R}$ med intensitet $m_V = 2\lambda/\pi$, jf. [Miles, 1964, Theorem 2].

Sætning 7.9 [Daley og Vere-Jones, 2008, Proposition 15.3.IV]
Lad Y være en stationær og isotropisk linjeproces med intensiteten m og lad $A \subseteq \mathbb{R}^2$ være en kompakt konveks mængde, hvor $N(A)$ er antallet af forskellige linjer, der skærer mængden A. Da er det forventede antal af linjer, der skærer mængden A givet ved

$$\mathbb{E}[N(A)] = \frac{m h_1(\partial A)}{\pi},$$

hvor $h_1(\partial A)$ er længden af perimeteren for A.

Bevis:
Antag, at A er et konvekst polygon bestående af sidelængderne $S_1, \ldots S_n$, således at

$$h_1(\partial A) = \sum_{i=1}^{n} S_i.$$

Ved benyttelse af Sætning 7.1.4 fås det, at det forventede antal af skæringspunkter fra linjeprocessen Y og linjestykket V_i er givet ved

$$\mathbb{E}[N(V_i)] = \frac{2m}{\pi} S_i$$

Det vil sige, at det forventede antal skæringspunkter med perimeteren af A er

$$\mathbb{E}[N(h_1(\partial A))] = \mathbb{E}\left[N\left(\bigcup_{i=1}^{n} V_i\right)\right] = \mathbb{E}\left[\sum_{i=1}^{n} N(V_i)\right] = \frac{2m}{\pi}\sum_{i=1}^{n} S_i = \frac{2m}{\pi} h_1(\partial A)$$

Da antallet af skæringspunkter for en linje med en konveks mængde er to på nær, når linjen tangerer randen af A, men dette sker med sandsynligheden nul, fås det, at forventede antal af linjer, der skærer mængden A er

$$\mathbb{E}[N(A)] = \frac{m}{\pi} h_1(\partial A).$$

Jævnfør [Daley og Vere-Jones, 2008, s.476] kan det vises at gælde for alle konvekse mængder, og ikke blot for polygoner. □

Specielt for en stationær og isotropisk Poisson-linjeproces med intensitet λ gælder der, at det forventede antal af linjer, der skærer en konveks mængde $A \subseteq \mathbb{R}^2$ er

Poisson-fordelt med parameteren $\lambda h_1(\partial A)/\pi$. At dette antal er Poisson-fordelt ses ved, at antallet af linjer, der skærer A, er antallet af punkter, der ligger indenfor et bestemt område af repræsentationsrummet. Af definitionen på Poisson-proces Definition 3.2 fås det at antallet af punkter indenfor en sådan mængde er Poisson-fordelt med intensitetsmålet som parameter, hvilket i dette tilfælde er $\lambda h_1(\partial A)/\pi$. For en cirkel $b(o,r)$ er antallet af linjer fra en Poisson-linjeproces med intensitet λ, der skærer denne, altså $N(b(o,r)) \sim \text{poi}(2\lambda r)$.

7.1.5 Længden af linjer

Resultaterne i dette afsnit benyttes senere i forbindelse med modelopstillingen. Det vises bl.a., hvordan der omregnes mellem de to rum, og hvordan der beregnes længder af linjer.

Med repræsentationen af linjer ved (p,θ) kan (7.1) skrives

$$y(x) = x\tan(\alpha) + \frac{p}{\sin(\theta)}, \tag{7.6}$$

hvor $\alpha = \pi/2 + \theta$. For at beregne længden af linjer indenfor et kvadratisk område af $\mathbb{R}^2$ betragtes en linje givet ved (p,θ). Ved omskrivelse af (7.6) fås

$$x(y) = \frac{y}{\tan\alpha} + \frac{p}{\sin(\theta)\tan(\alpha)},$$

hvor igen $\alpha = \pi/2 + \theta$.

Lad $[d \times d]$ afgrænse et område i $\mathbb{R}^2$ og lad origo være i midten af dette kvadrat. Skal længden af en linje indenfor det afgrænsede område beregnes, findes først skæringpunkterne med linjerne, der udgør kvadratets sider, ved

$$(-d/2, y(-d/2)),\ (d/2, y(d/2)),\ (x(-d/2), -d/2) \ \text{ og } \ (x(d/2), d/2).$$

Punkterne sorteres derefter i stigende rækkefølge efter x-værdien, og skæringspunkterne med den største og mindste x-værdi fravælges og afstanden mellem de to resterende punkter beregnes på almindelig vis.

Haves der en realisation af en linjeproces kendes p og θ for de enkelte linjer ikke. Disse kan dog opnås ved at finde linjernes skæringspunkter med observationsvinduet, således linjens ligning er udtrykt ved 2 punkter (x_1, y_1) og (x_2, y_2). Herefter kan α og θ fås til

$$\alpha = \tan\left(\frac{y_2 - y_1}{x_2 - x_1}\right) \qquad \text{og} \qquad \theta = \arctan\left(\frac{y_2 - y_1}{x_2 - x_1}\right) - \frac{\pi}{2}. \tag{7.7}$$

Den vinkelrette afstand p fås ved omskrivning af (7.6) til

$$p = \left(y_1 - \frac{(y_2 - y_1)x_1}{(x_2 - x_1)}\right)\sin(\theta), \tag{7.8}$$

hvor θ fra (7.7) kan sættes ind, således at der opnås et udtryk, der kun er bestemt af koordinaterne for de to punkter på linjen.

Afstanden fra et vilkårligt punkt $(x, y) \in \mathbb{R}^2$ til linjen givet ved (p, θ) kan ved geometriske betragtninger fås til

$$d(x) = |p - x\cos(\theta) - y\sin(\theta)|. \tag{7.9}$$

7.2 Likelihood for Poisson-linjeproces

Ved benyttelse af Afsnit 3.2 er det muligt at opstille en likelihoodfunktion for Poisson-linjeprocessen. Resultaterne fra Afsnit 3.2 kan overføres direkte til repræsentationsrummet $\mathbb{R} \times \mathbb{S}$. Det vil sige, at en Poisson-proces defineret på repræsentationsrummet altid er absolut kontinuert mht. standard Poisson-processen på repræsentationsrummet, og dermed vil tætheden herfor være på formen (3.6).

Det vil sige, at for en Poisson-linjeproces med intensitet λ på $A \subseteq \mathbb{R}^2$ er likelihoodfunktionen for parameteren θ givet en realisering $x = \{x_1, \ldots, x_n\}$ i repræsentationsrummet

$$L(\theta|x) = \exp\left(-\mu(B)\right) \prod_{j=1}^{n} \frac{\lambda}{\pi}. \tag{7.10}$$

I forbindelse med datasættet for gravhøje er linjerne ukendte, og likelihoodfunktionen givet ved (7.10) er derfor ikke anvendelig. I sådanne tilfælde skal der opskrives en anden likelihoodfunktion, hvor linjerne repræsenterer det ukendte data.

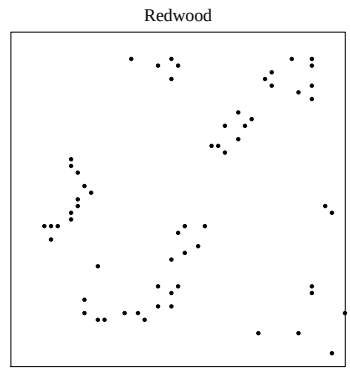
I dette kapitel er en teoretisk beskrivelse af linjeprocesser blevet gennemgået. Disse linjeprocesser skal i det senere benyttes til at opstille modeller, der kan beskrive punktmønstre med linjetendenser. Nu er den teori, der er nødvendig i forbindelse med opstilling af en model for gravhøje, blevet gennemgået, nemlig teori om punktprocesser og linjeprocesser. I det følgende kapitel anvendes denne teori på punktmønste af forskellige art, og dette giver et eksempel på, hvorledes analyse og valg af model kan foretages for punktmønstre

Eksempel på punkt og linjeprocesser

I dette kapitel introduceres modelopstillingen til beskrivelse af to datasæt, hhv. datasættet redwood fra `spatstat`-pakken i `R` [R Development Core Team, 2008] og et datasæt bestående af linjemønster-data. Dette gøres for at vise fremgangsmåden for opstillingen af en model, og hvilke overvejelser der bør gøres i denne forbindelse. Endvidere gives et eksempel på hvordan et modeltjek kan udføres. De fleste funktioner, der benyttes i kapitlet, er funktioner, der allerede er implementeret i `spatstat`-pakken i `R`.

8.1 Eksempel på punktproces

I dette afsnit opstilles modeller til at beskrive datasættet redwood fra `spatstat`-pakken. Dette datasæt er plottet på Figur 8.1. På figuren ses 62 punkter i et observationsvindue på $[0,1] \times [-1,0]$.



Figur 8.1: Plot af datasættet for redwood.

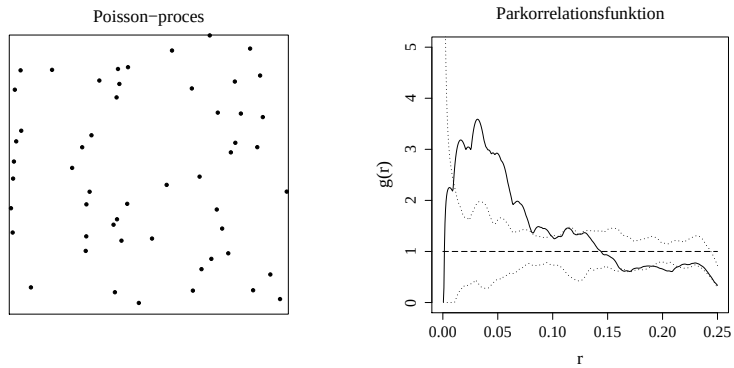
Af figuren ses det, at der er tegn på, at punkterne optræder i clustre, men det er svært at afgøre med sikkerhed. Første skridt i at tilpasse en model på punktmønstret, er at forsøge at beskrive punktmønstret ved en homogen Poisson-proces.

Poisson-proces

Ved benyttelse af funktionen `ppm`$(\cdot)$, der anvender estimatet fra (4.1), estimeres intensiteten af den homogene Poisson-proces til 62. På Figur 8.2(a) ses et plot af en realisering af en Poisson-proces med intensiteten 62 på $[0,1] \times [-1,0]$. Ved sammen-

ligning af Figur 8.1 og Figur 8.2(a) ses det, at clustertendenserne fra datasættet ikke er eksisterende for realisationen af Poisson-processen. Det ses på plottet, at denne realisering af Poisson-processen ikke beskriver punktmønstret i tilstrækkelig grad. For mere præcist at vurdere om punktprocessen stemmer overens med data, kan teorien fra Kapitel 4 omkring summary statistics benyttes.

På Figur 8.2(b) ses et plot af parkorrelationsfunktionen fra datasættet og parkorrelationsfunktionen for modellen med øvre- og nedre-envelopes. Den øvre- og nedre-envelope for Poisson-processen er estimeret ud fra 39 simuleringer jf. Afsnit 4.4 ved benyttelse af funktionen `envelope` $(\cdot)$.



(a) (b)

Figur 8.2: (a) Poisson-proces med intensitet 62 på $[0,1] \times [-1,0]$. (b) Parkorrelationsfunktionen for redwood datasættet. Den fuldtoptrukne linje er estimatet fra data, mens (- - -) er den teoretiske parkorrelationsfunktion for Poisson-processen. Linjerne beskrevet ved $(\cdots)$ er envelopes for Poisson-processen.

Ved sammenligning af den estimerede parkorrelationsfunktion fra datasættet med de simulerede envelopes for Poisson-processen ses det, at parkorrelationsfunktionen ligger et stykke over den øvre envelope. Dette er igen en indikation på, at det ikke er muligt at beskrive datasættet i tilstrækkelig grad ved benyttelse af en homogen Poisson-proces.

Da det er parkorrelationsfunktionen, der er benyttet som summary statistics i Figur 8.2(b), og da den estimerede parkorrelationsfunktion fra data ligger over parkorrelationsfunktionen fra Poisson-processen, kan det antages, at redwood datasættet udviser større clustertendenser end Poisson-processen. Men at dette er tilfældet burde ikke komme som en større overraskelse efter at have sammenlignet Figur 8.1 og Figur 8.2(a).

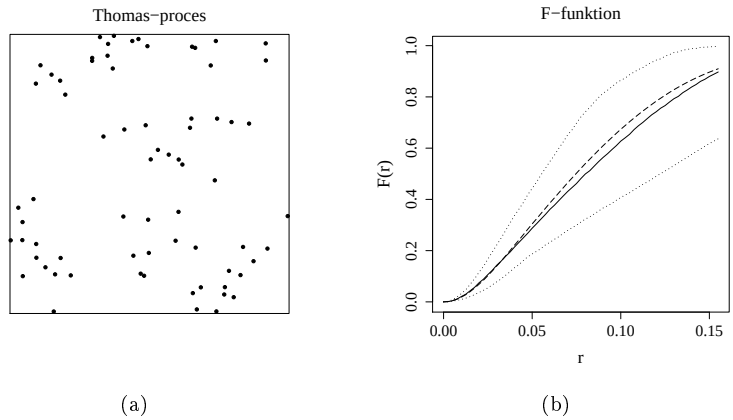
Til beskrivelse af clustertendenser kan der vælges mellem flere forskellige typer af f.eks. Cox-processer, som eksempelvis Thomas-processen fra Afsnit 5.3.1 og Matérn-processen fra Afsnit 5.3.2.

Thomas-proces

I Thomas-processen er centerprocessen en Poisson-proces, mens kernerne er givet ved en normalfordeling. Ved benyttelse af funktionen `thomas.estK`$(\cdot)$ er det muligt at estimere de tre parametre, der indgår i Thomas-processen. Disse parametre er κ, der angiver intensiteten af Poisson-processen af moderpunkter, ω der angiver standardafvigelsen af punkterne fra cluster-centrene, og α der angiver det forventede antal af punkter i hvert cluster. `R` estimerer parametrene ved benyttelse af minimumkontrastestimation jf. Afsnit 5.4 ud fra K-funktionen. For datasættet redwood fås parametrene til

$$\kappa = 23.54, \qquad \omega = 4.70 \cdot 10^{-2}, \qquad \text{og} \qquad \alpha = 2.63.$$

Det er nu muligt at lave en realisering af Thomas-processen ved benyttelse af funktionen `rThomas`$(\cdot)$. På Figur 8.3(a) ses en simulering af Thomas-processen med de estimerede parametre.



Figur 8.3: (a) Realisering af Thomas-proces med parametre $\kappa = 23.54$, $\omega = 4.70 \cdot 10^{-2}$ og $\alpha = 2.63$. (b) F-funktionen for Thomas-processen. Den fuldtoptrukne linje er estimatet fra data, mens (- - -) er middelværdien af de 39 simuleringer for F-funktion af Thomas-processen. Linjerne beskrevet ved $(\cdots)$ er envelopes for Thomas-processen.

Ved sammenligning af datasættet for redwood og realiseringen af Thomas-processen, dvs. Figur 8.1 og Figur 8.3(a) ses det, at der er en nogenlunde overensstemmelse mellem fordelingen af punkterne, selvom clustertendenserne for redwood-data ses at være tydeligst. Igen betragtes en summary statistics for punktmønstret samt modellen. På Figur 8.3(b) ses et plot af F-funktionen for både datasættet og Thomas-processen. Det er F-funktionen, der er valgt til sammenligning af model og data og ikke K-funktionen, da `R` har brugt denne til at estimere parametrene i Thomas-processen. På Figur 8.3(b) ses det, at den estimerede F-funktion fra datasættet ligger passende indenfor den nedre og øvre envelope af Thomas-processen. Desuden er F-funktionen fra datasættet og middelværdien for envelopes fra Thomas-processen

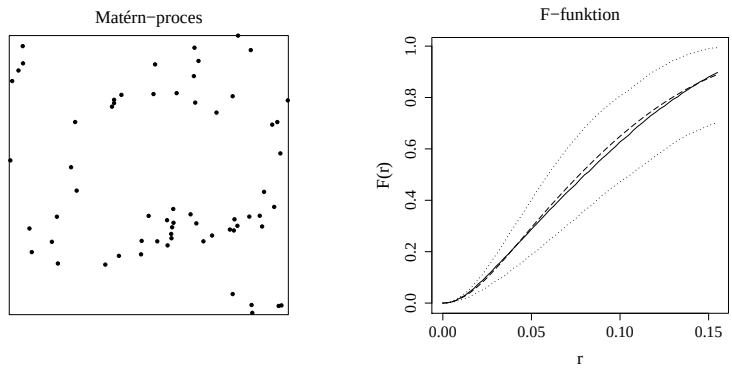
næsten sammenfaldende. Det vil sige, at modelantagelsen om at punkterne følger Thomas-prossen, kan betragtes som værende korrekt.

Matérn-proces

I Matérn-processen indgår parametrene κ, der angiver intensiteten for Poisson-processen af moderpunkter, R der er radius parameter for clusterne og α der er det forventede antal punkter pr. cluster. Ved brug af funktionen `matclust.estK`$(\cdot)$ estimeres parametrene til

$$\kappa = 24.56, \qquad R = 8.65 \cdot 10^{-2}, \qquad \text{og} \qquad \alpha = 2.52,$$

igen ved benyttelse af minimumkontrastestimation ud fra K-funktionen. Ved indsættelse af parametrene i `rMatClust`$(\cdot)$ kan Figur 8.4(a) opnås, og ved at plotte F-funktionen for data samt envelopes for Matérn-processen opnås Figur 8.4(b).



(a) (b)

Figur 8.4: (a) Realisering af Matérn-proces med parametre $\kappa = 24.56$, $R = 8.65 \cdot 10^{-2}$ og $\alpha = 2.52$. (b) F-funktionen for Matérn-processen. Den fuldtoptrukne linje er estimatet fra data, mens (- - -) er middelværdien af F-funktion for Matérn-processen. Linjerne beskrevet ved $(\cdots)$ er envelopes for Matérn-processen.

Som det ses på Figur 8.4 kan Matérn-processen også benyttes til at beskrive datasættet redwood, da estimatet fra data ligger indenfor envelopes for modellen.

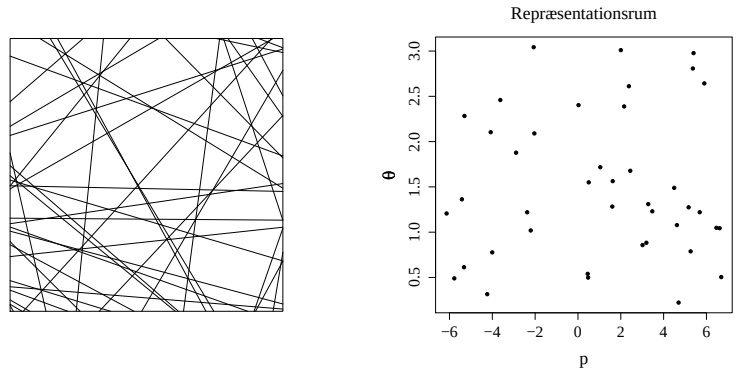
Der er nu to modeller til beskrivelse af datasættet redwood, Thomas-processen og Matérn-processen. De to modeller er i og for sig lige gode til at beskrive datasættet, og de adskiller sig kun i udtrykkene af kernerne.

8.2 Eksempel på linjeproces

I dette afsnit benyttes linjeprocesser som beskrevet i Kapitel 7 til at beskrive linjemønstre. På Figur 8.5(a) ses et observeret linjemønster på observationsvinduet $[-5,5] \times [-5,5]$. Dette kunne f.eks. være tøjfibre, men i dette tilfælde er det blot en realisering af en linjeproces.

Der er to måder at betragte linjemønstret på, enten som linjer indenfor observationsvinduet, eller som en realisering af en punktproces på repræsentationsrummet. Da der er en 1:1-korrespondance mellem de to rum, er det lige meget hvilket et af dem der betragtes. Linjerne betragtes som en realisering af en punktproces på repræsentationsrummet, som er opnået ved oversættelse af linjerne ud fra (7.7) og (7.8). På Figur 8.5(b) ses det til Figur 8.5(a) hørende punktmønster i repræsentationsrummet.

Det forsøges nu at tilpasse en homogen Poisson-proces til punktmønstret på Figur 8.5(b) ved benyttelse af funktionen `ppm`$(\cdot)$. Dermed opnås et estimat af intensiteten for Poisson-processen til 0.92 bestemt ud fra punktmønstret på Figur 8.5(b). Da der er en 1:1-korrespondance mellem repræsentationsrummet og linjerummet, kan intensiteten jf. Kapitel 7 for det ene rum opnås, ved at oversætte intensiteten fra det andet rum.

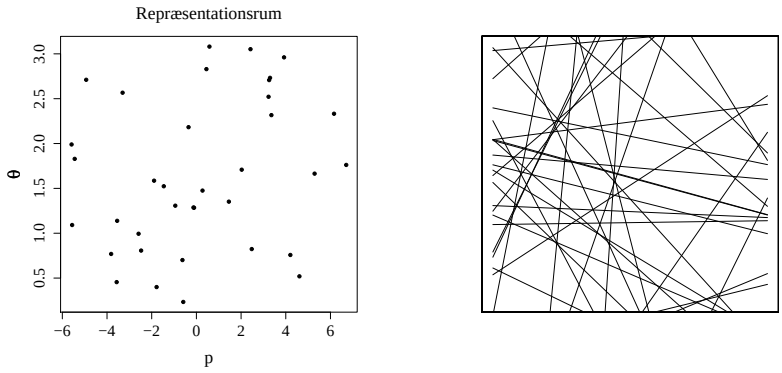


(a) (b)

Figur 8.5: (a) Observeret linjemønster på $[-5,5] \times [-5,5]$. (b) Punktmønster i repræsentationsrummet for linjemønstret givet ved (a).

På Figur 8.6(a) ses en simulering af Poisson-processen med den estimerede intensitet fra Figur 8.5(b). På Figur 8.6(b) ses punkterne fra repræsentationsrummet overført til linjerummet. Ved en visuel sammenligning ses det, at der er en nogenlunde overensstemmelse mellem punktmønstrene på Figur 8.5(b) og Figur 8.6(a), og ved betragtning af Figur 8.5(a) og Figur 8.6(b) ses det, at der også er en overensstemmelse mellem observeret data og en realisering af den opstillede model (dette

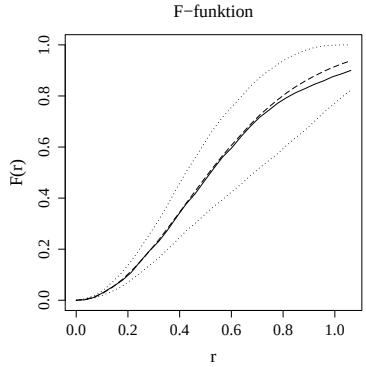
burde ikke komme som nogen overraskelse, da punktmønstret og linjemønstret er sammenhørende).



(a) (b)

Figur 8.6: (a) Realisering af Poisson-proces på repræsentationsrum med intensiteten 0.92. (b) Punkterne fra repræsentationsrummet oversat til linjeproces.

Summary statistics for linjeprocessen kan opnås ved at betragte repræsentationsrummet fremfor linjerummet. På Figur 8.7 ses et plot af den estimerede F-funktion fra repræsentationsrummet givet ved Figur 8.5(b) samt envelopes for F-funktionen fra Poisson-processen. Det ses, at der er en overensstemmelse mellem F-funktionen fra data og fra modellen.



Figur 8.7: F-funktionen for observeret repræsentationsrum og simlueret repræsentationsrum samt envelopes. Den fuldtoptrukne linje er estimatet fra data, mens (- - -) er middelværdien af F-funktion for Poisson-processen. Linjerne beskrevet ved $(\cdots)$ er envelopes for Matérn-processen.

Dette kapitel har ved eksempler illustreret store dele af den forudgående teori og vist, hvordan dataanalyse, modelvalg og modeltjek foretages. Dette kapitel er også slutningen på den del af rapporten, der beskæftiger sig med generel teori angående punkt og linjeprocesser.

I næste del af rapporten fokuseres der på datasættet gravhøje og opstilling af model for dette. Som det senere erfares er det nødvendigt med en mere kompliceret model til beskrivelse af gravhøjenes placering. Derfor indgår der løbende nødvendig teori i tilknytning til dette.

I indledningen blev datasættet for gravhøjene præsenteret og kort beskrevet. Der startes i næste kapitel med en yderligere databeskrivelse og analyse, da der nu kendes analyseredskaber som eksempelvis summary statistics.

Del II

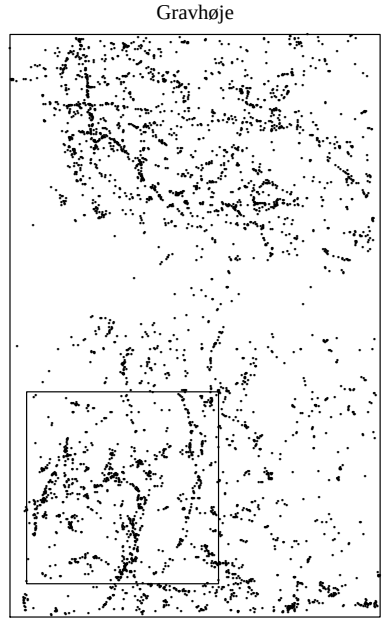
Databehandling

Databeskrivelse

I dette kapitel introduceres datasættet for gravhøjes placering omkring Skjern å, og der foretages en indledende dataanalyse. Dette gøres for at skabe et overblik over, hvad datasættet indeholder, og med henblik på at kunne opstille en model til beskrivelse af datasættet for gravhøjene. Til den indledende dataanalyse benyttes programmet R [R Development Core Team, 2008].

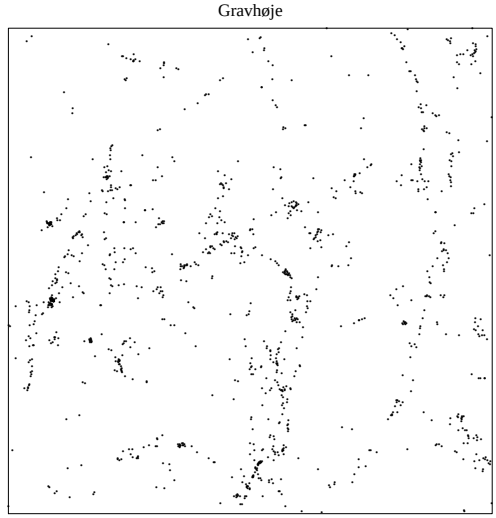
9.1 Databeskrivelse

Datasættet for gravhøjenes placering består af geografiske x- og y-koordinater, der beskriver gravhøjes placering i området $[458705, 487620] \times [6172511, 6217922]$ i UTM-koordinater, som det ses på Figur 9.1. Det ses, at punktmønstret af gravhøjene er inhomogent indenfor dette område, hvilket skyldes geografiske forhold som f.eks. forløbet af Skjern å, højdeforskelle i landskabet og fordelingen af befolkningen på det tidspunkt, hvor gravhøjene blev oprettet. På Figur 9.1 indikeres Skjern å som de brede bånd uden gravhøje horisontalt i midten af figuren.



Figur 9.1: Plot af datasættet for gravhøjene, der er spredt over området $[458705, 487620] \times [6172511, 6217922]$. Kvadratet afgrænser området $[460000, 475000] \times [6175000, 6190000]$.

Nord og syd for Skjern å ses der områder med større homogenitet blandt punkterne. Et sådant område kunne være indenfor koordinaterne $[460000, 475000] \times [6175000, 6190000]$, som på Figur 9.1 er markeret med et kvadrat. Dette område er forstørret og plottet på Figur 9.2. Her ses det, at de forskellige punkter ligger forholdsvis jævnt fordelt over området, og samtidig er der tydelige linjetendenser. Desuden ses der mindre clustertendenser, hovedsageligt i forbindelse med linjetendenserne, og der ses områder med næsten ingen gravhøje. Yderligere gælder der for dette område, at der ikke er nogle punkter, der ligger oveni hinanden, hvilket er tjekket med funktionen `duplicated.ppp`$(\cdot)$ i `R`, således at punktmønstret er simpelt.



Figur 9.2: Plot af gravhøje i området $[460000, 475000] \times [6175000, 6190000]$.

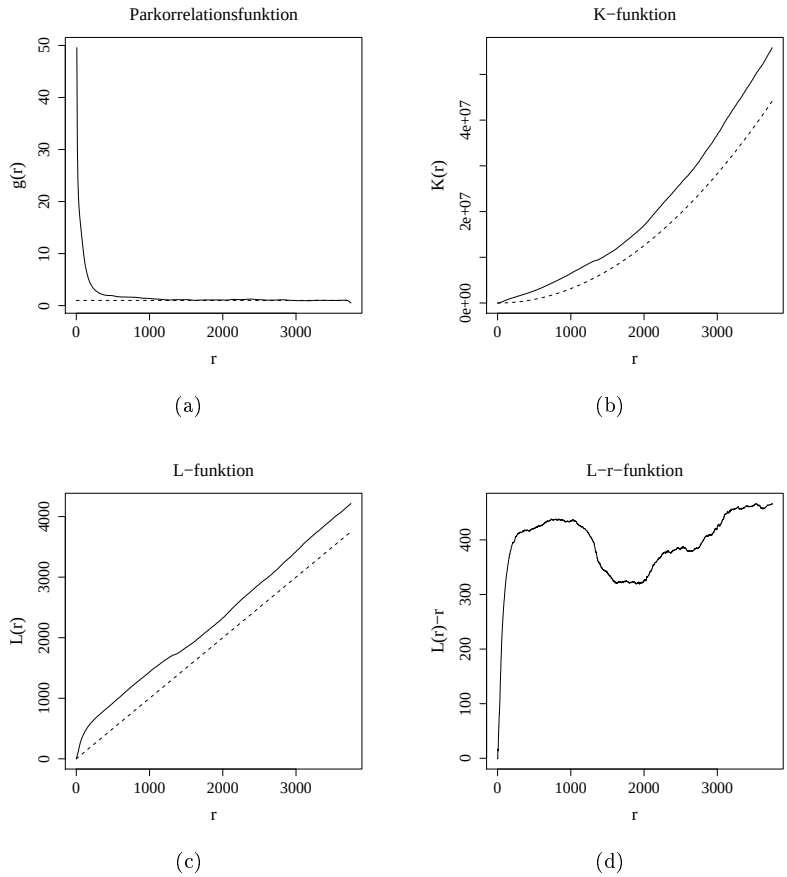
Denne homogenitet af punkterne på området gør, at det kan være fordelagtigt at arbejde med dette område. Da det netop er linjetendenserne, der ønskes beskrevet i den senere opstillede model er dette område anvendelig. I det følgende er det således kun dette område, der omtales 'datasættet gravhøje'.

9.1.1 Summary statistics

For at få et førstehåndsindtryk af punkternes indbyrdes vekselvirkning betragtes forskellige summary statistics for datasættet. Dette kan give en indikation af, om der punkterne imellem er en større eller mindre grad af tiltrækning eller clustertendens end en homogen Poisson-proces. Dermed giver det også en indikation af hvilken model, der kunne tænkes at passe på datasættet.

I Figur 9.3(a) ses et estimat af parkorrelationsfunktionen for datasættet fra Figur 9.2. Det ses, at for små r er $g(r) > 1$, hvilket vil sige, at punktmønstret udviser

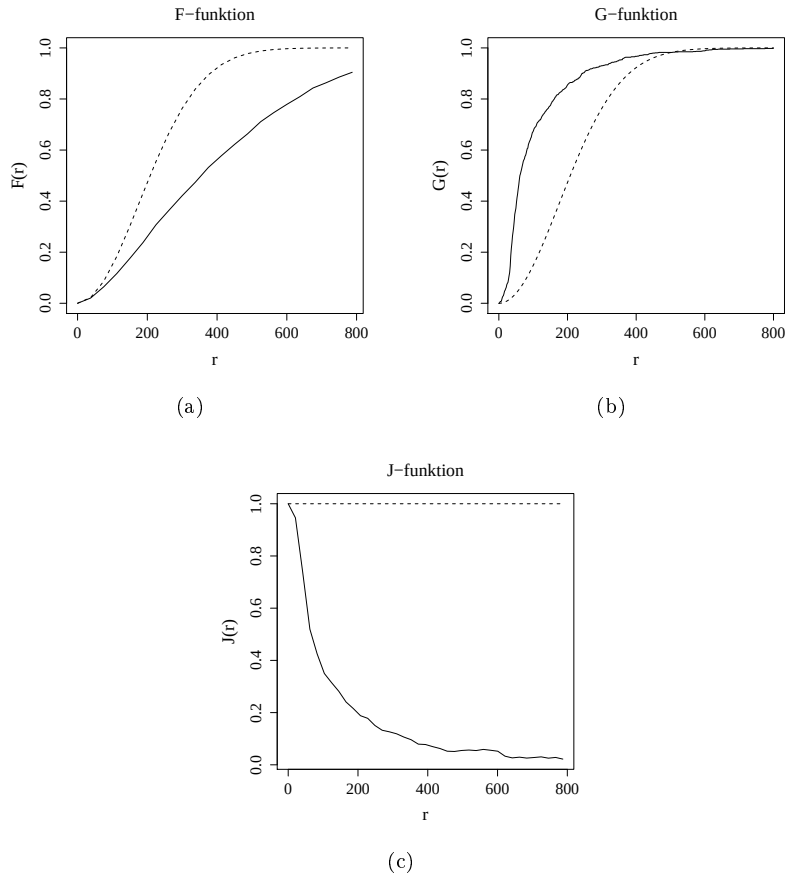
større tiltrækning eller clustertendenser jf. Afsnit 4.1.1 end en homogen Poisson-proces.



Figur 9.3: Summary statistics for gravhøjene. Den fuldtoptrukne linje er estimatet fra data, mens (- - -) er for Poisson-processen. (a) Estimat af parkorrelationsfunktionen. (b) Estimat af K-funktionen. (c) Estimat af L-funktionen. (d) Funktionen $L(r) - r$, hvor $L(r)$ er estimatet fra (c).

I Figur 9.3(b) og Figur 9.3(c) ses et estimat af K- og L-funktionen for punktmønstret. Indikationen af at punktmønstret udviser tiltrækning eller clustertendenser for små r, bliver jf. Afsnit 4.2 bekræftet ved at betragte $L(r) - r$ givet i Figur 9.3(d), da $L(r) - r$ er større end nul for små r.

En anden måde at beskrive punktmønstret på er ved benyttelse af de afstandsbaserede summary statistics givet ved F-, G- og J-funktionen jf. Afsnit 4.3. Disse ses plottet på Figur 9.4.



Figur 9.4: Afstandsbaserede summary statistics. Den fuldtoptrukne linje er estimatet fra data, mens (- - -) er for Poisson-processen. (a) Estimat af F-funktionen. (b) Estimat af G-funktionen. (c) Estimat af J-funktionen.

På Figur 9.4(a) og Figur 9.4(b) ses estimater af F- og G-funktionen for datasættet. Det ses, at for små r er $F(r) < G(r)$, hvilket indikerer en tiltrækning eller clustertendens i forhold til Poisson-processen, som også ses ved, at $J(r) < 1$ i Figur 9.4(c).

De forskellige summary statistics indikerer altså, at der er tiltrækning eller clustertendenser i punktmønstret. Når der skal opstilles en model til beskrivelse af datasættet, er det derfor nærliggende at benytte eksempelvis Cox-processer som beskrevet i Kapitel 5. I det følgende kapitel forsøges det at tilpasse punktprocesser til beskrivelse af datasættet gravhøje ved allerede implementerede funktioner i `R`.

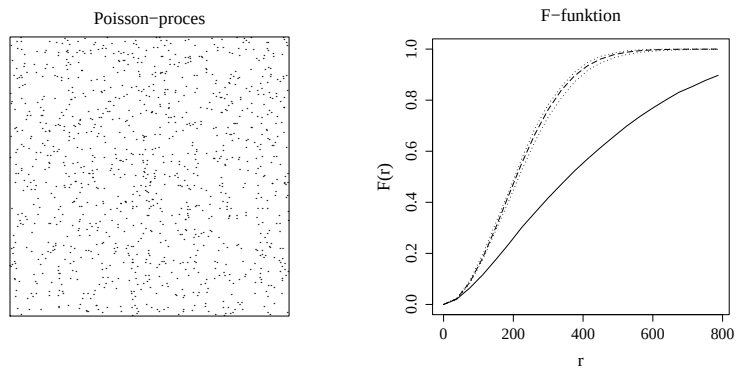
Punktproces-modeller for gravhøje

I dette kapitel opstilles punktprocesser til beskrivelse af datasættet gravhøje. Dette gøres på baggrund af den viden om punktmønstret, der er opnået i Kapitel 9. Punktprocesserne der tilpasses datasættet er nogle af de allerede behandlede punktprocesser fra Kapitel 3 og Kapitel 5. Fremgangsmåden er tilsvarende den fra Kapitel 8.

Alle punktprocesserne opstilles på området $[460000, 475000] \times [6175000, 6190000]$ og laves ved brug af `R` [R Development Core Team, 2008].

10.1 Tilpasning af Poisson-proces

I forrige kapitel blev det vist, at der var tiltrækning eller clustertendenser blandt punkterne i punktmønstret ved, at $g(r) > 1$ for små r. I første omgang antages det dog, at gravhøjene følger en fuldstændig rumlig tilfældighed, dvs. at der benyttes en homogen Poisson-proces til beskrivelse af data.



(a) (b)

Figur 10.1: (a) Realisation af homogen Poisson-proces med intensitet $5.1 \cdot 10^{-6}$. (b) F-funktion for Poisson-proces. Den fuldtoptrukne linje er estimatet fra data, mens (- - -) er den teoretiske værdi af F-funktionen for Poisson-processen. Linjerne beskrevet ved $(\cdots)$ er envelopes for Poisson-processen.

Ved benyttelse funktionen `ppm`$(\cdot)$ er det ud fra data muligt at estimere intensiteten af den homogene Poisson-proces til $\rho = 5.1 \cdot 10^{-6}$. På Figur 10.1(a) ses en realisering af en homogen Poisson-proces, med den estimerede intensitet.

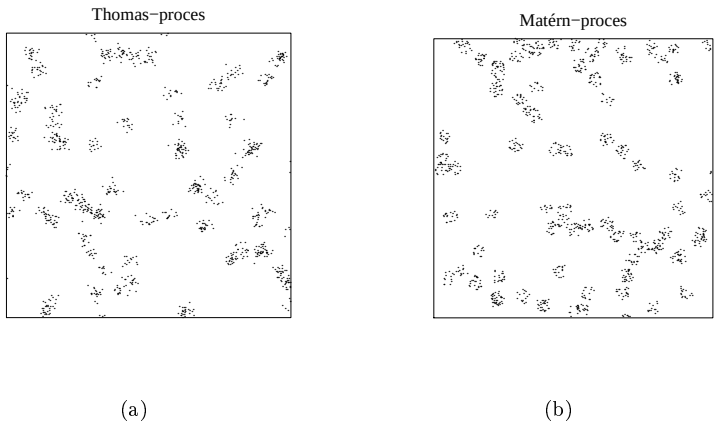
Ved en visuel sammenligning af Figur 9.2 og Figur 10.1(a) ses det, at punkterne fra datasættet har større tiltrækning eller clustertendenser i forhold til realiseringen af den homogene Poisson-proces. På realiseringen ses der ingen linjetendenser, ingen mindre clustertendenser og ingen større områder uden punkter i modsætning til datasættet. Dermed bekræftes formodningen fra Kapitel 9, om at datasættet har større tiltrækning eller clustertendenser end en homogen Poisson-proces.

På Figur 10.1(b) ses den teoretiske F-funktion for den homogene Poisson-proces, samt envelopes for denne og F-funktionen for datasættet. Envelopes er beregnet på baggrund af 39 realiseringer af Poisson-processen svarende til et konfidensinterval på 95% jf. Afsnit 4.4. Det ses, at der ikke er god overenstemmelse mellem envelopes for Poisson-processen og F-funktionen estimeret fra datasættet.

I stedet for at beskrive datasættet med en model, der tager udgangspunkt i den homogene Poisson-proces, vil det være mere hensigtmæssigt at benytte en punktproces med større tiltrækning eller clustertendenser, som eksempelvis Cox-processen.

10.2 Tilpasning af Cox-proces

Der eksisterer flere typer af Cox-processer som beskrevet i Kapitel 5. På Figur 10.2(a) ses en realisering af en Thomas-proces. Denne Thomas-proces er opstillet ved først at estimere parametre vha. funktionen `kppm`$(\cdot)$, der anvender minimumkontrastestimation. Herved estimeres parametrene til $(\kappa = 6.58 \cdot 10^{-7}, \omega = 1.11, \alpha = 7.75)$. Herefter er der genereret en realisering af Thomas-processen med disse parametre.

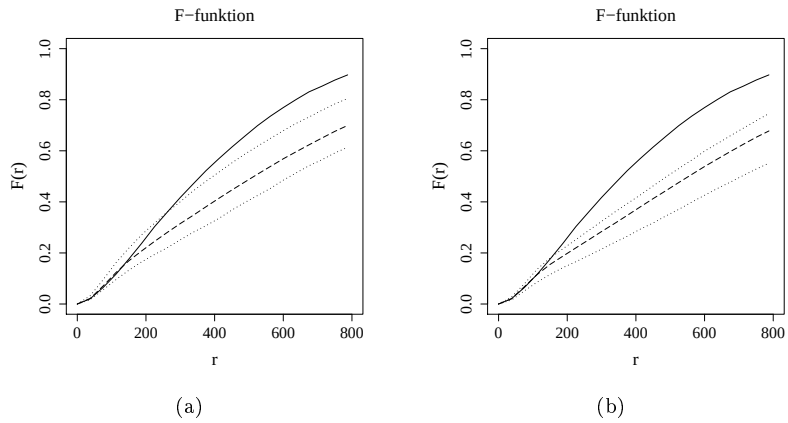


Figur 10.2: Realiseringer af punktprocesser, hvor parametre er estimeret fra data ved benyttelse af minimumkontrastestimation. (a) Thomas-proces. (b) Matérn-proceser.

Ved sammenligning af Figur 9.2 og realiseringen af modellen ses det, at der ikke er en overenstemmelse mellem de to. Det ses på realiseringen, at clustertendenserne er for tydelige, linjetendenserne er ikke til stede, og der mangler områder med lavere intensitet.

På Figur 10.2(b) ses en realisering af en Matérn-proces tilpasset vha. minimumkontrastestimation og med parametre $(\kappa = 6.58 \cdot 10^{-7}, R = 2.69, \alpha = 7.75)$. Ud fra en visuel vurdering af en realisering af Matérn-processen og plots af data, ses der at være samme mangler i realiseringen som ved Thomas-processen.

På Figur 10.3 sammenlignes F-funktionen estimeret fra data med envelopes for F-funktionerne estimeret fra Thomas- og Matérn-processen. Det ses, at for små r udviser Thomas- og Matérn-processen samme clustertendenser som datasættet, da F-funktionen ligger indenfor envelopes. Dette kan eventuelt skyldes, at for små r kan der ikke skelnes mellem linjetendenser og clustertendenser. Endvidere ses det, at for større r stemmer de to punktprocesser ikke overens med datasættet.



Figur 10.3: (a) F-funktionen for Thomas-processen. (b) F-funktionen for Matérn-processen. Den fuldtoptrukne linje er estimatet fra data, mens (- - -) er den forventede værdi af F-funktionen for hhv. Thomas- og Matérn-processen. Linjerne beskrevet ved $(\cdots)$ er envelopes for hhv. Thomas- og Matérn-processen.

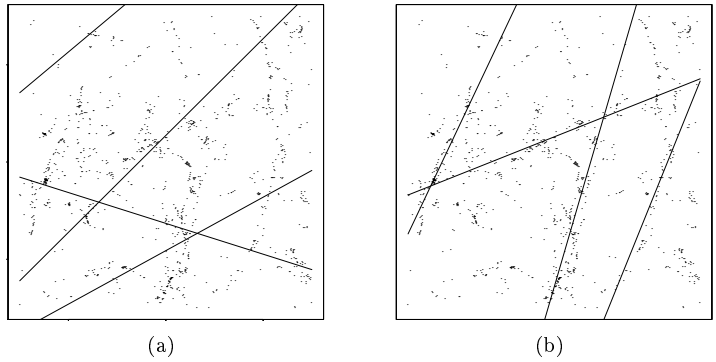
Samtlige af de opstillede modeller, hvad enten de er tilpasset med clusterprocesser eller ej, formår ikke i tilstrækkelig grad at skabe linjer af gravhøje, som dem datasættet antydes at indeholde. For kunne beskrive linjetendenserne er det derfor nødvendigt at udvikle punktprocesser specielt til beskrivelse af linjetendenser, hvilket gøres i følgende kapitel.

Model med linjetendenser for gravhøje

I dette kapitel opstilles en model til beskrivelse af punktmønstret for gravhøje. Forskellen fra de opstillede modeller i Kapitel 10 er, at der her benyttes teori om linjeprocesser til at opnå en model, der kan genskabe linjetendenser.

Intentionen er bl.a. at benytte en linjeproces, og betragte hver linje i linjeprocessen som udgangspunkt for en punktproces på et område langs linjen. Selve linjeprocessen vil her optræde som en skjult linjeproces, som blot benyttes til at skabe linjetendenserne. Denne metode vil give linjetendenser blandt punkterne, men eftersom linjerne i en linjeproces er rette, vil der sandsynligvis ikke kunne skabes samme buede linjetendenser, som i datasættet for gravhøje.

Som eksempel er der på Figur 11.1(a) placeret 4 tilfældige linjer på punktmønstret. Disse følger ikke linjetendenserne i datasættet. På Figur 11.1(b) er der derimod placeret 4 linjer på punktmønstret, som nogenlunde ligger langs nogle af linjetendenserne i datasættet. Det kan forestilles, at hvis der simuleres punktprocesser omkring linjerne, vil disse kunne optræde som linjetendenser, der ligger de samme steder som i datasættet. Afhængig af hvilke antagelser der gøres vedrørende interaktionsafstanden mellem en linje og punkterne, skulle der måske have været flere linjer eller måske kunne de 4 linjer være placeret mere optimalt. Ved at opstille en passende model er det muligt at placere linjerne i de områder, der er linjetendenser, og estimere antallet af linjerne.



Figur 11.1: Linjer på punktmønstret for gravhøje. (a) 4 tilfældige linjer. (b) 4 linjer der følger linjetendenserne.

Først opstilles modellen til at beskrive punktmønstret. Dette gøres på baggrund af de teorier, der eksisterer angående gravhøje langs veje fra indledningen. Derefter implementeres modellen i R. Modeltjek af den opstillede model foretages i næste kapitel.

11.1 Opstilling af model

For at opstille en model til beskrivelse af linjetendenserne i datasættet for gravhøje, skal forskellige antagelser og betragtninger foretages. Målet med dette afsnit er at opnå en likelihoodfunktion for modellen, der afhænger af ukendte parametre og ukendte linjer, der kan estimeres fra datasættet.

Fjernes gravhøjene fra Figur 11.1(b) ses det, at de fire linjer antydes at ligge fuldstændigt vilkårligt uden nogen form for vekselvirkning. Det vil sige, at den bagvedliggende linjeproces, der benyttes til at skabe linjetendenserne, kan antages at være en homogen Poisson-linjeproces på observationsvinduet $[-7500, 7500] \times [-7500, 7500]$, som i det følgende benævnes A. Der er valgt dette observationsvindue, da en linjeproces er defineret på et område centreret omkring origo. Punkterne i datasættet forskydes til at ligge i dette observationsvindue. Kantlængden af kvadratet A noteres d.

Intensiteten for Poisson-linjeprocessen er ukendt og sættes til λ, dvs. der skal opstilles en likelihoodfunktion til estimation af λ. Når linjerne udgør en Poisson-linjeproces, kan disse som bekendt repræsenteres som punkter i et repræsentationsrum, hvor punkterne udgør en Poisson-proces jf. Definition 7.3. Lad repræsentationsrummet $S \subseteq \mathbb{R}^2$ være givet ved $S = [0, d/\sqrt{2}] \times [0, 2\pi]$, da de mulige vinkler er $[0, 2\pi]$ og da den maksimale vinkelrette afstand fra centrum til et af hjørnerne af A er $d/\sqrt{2}$. Dette valg af repræsentationsrum kan give linjer, der ligger udenfor A, eksempelvis vil en linje med $p = d/\sqrt{2}$ og en meget lille vinkel ligge uden for A. Det vil fremgå af en senere forklaring, at dette ikke giver problemer i forbindelse med beskrivelse af punktmønstre med linjetendenser. Hvis intensiteten af Poisson-linjeprocessen er λ, er intensiteten af Poisson-processen λ/π på repræsentationsrummet. Jævnfør (7.10) bliver likelihoodfunktionen for λ

$$\begin{aligned} L(\lambda|y) &= \exp\left(-\int_0^{2\pi}\int_0^{\frac{d}{\sqrt{2}}} \frac{\lambda}{\pi}\,\mathrm{d}p\,\mathrm{d}\theta\right)\prod_{i=1}^{n}\frac{\lambda}{\pi} \\ &= \left(\frac{\lambda}{\pi}\right)^n \exp\left(-\frac{2\lambda d}{\sqrt{2}}\right), \end{aligned} \tag{11.1}$$

hvor $y = \{y_1, \ldots, y_n\}$ repræsenterer linjeprocessen i form af punktmønstret på repræsentationsrummet. Da linjeprocessen er en homogen Poisson-proces på repræsentationsrummet ses det, at $\mu(A) = 2\lambda d/\sqrt{2}$ i det ovenstående, hvilket svarer til intensiteten af Poisson-processen på $\mathbb{R}$ ganget intervallet for den vinkelrette afstand jf. Afsnit 7.1.2. Log-likelihoodfunktionen for λ bliver

$$\log(L(\lambda|y)) = n\log\left(\frac{\lambda}{\pi}\right) - \frac{2\lambda d}{\sqrt{2}}.$$

Den skjulte linjeproces er således beskrevet, og der skal nu konstrueres punktprocesser omkring linjerne i denne. Da gravhøjene fra datasættet ikke ligger direkte på linjer, men antydes at fordele sig i områder omkring linjerne jf. Figur 9.2, er det nødvendigt at tilføje en fordeling af punkterne indenfor disse områder til modellen.

Fordelingen af punkter omkring linjerne, ses på Figur 11.1(b) at være forskellig fra linje til linje for de fire valgte linjer. Fordelingen af punkter om en linje ses at variere, idet intensiteten af punkter kan være forholdsvis stor i nogle områder og næsten nul i andre områder, som det ses for linjen længst til venstre på Figur 11.1(b).

I denne model antages punktprocesserne omkring linjerne at kunne beskrives ved Poisson-processer. Punkterne hørende til en given linje kan projiceres ind på linjen, og da punktprocessen omkring linjen er en Poissonproces, medfører dette, at punktprocessen på linjen også er en Poisson-proces. Intensiteten for Poisson-processen på linjen betegnes β. Afstandene fra linjen til hvert af punkterne kan ses som uafhængige mærker tilknyttet punkterne på linjen, og punkterne og mærkerne kan ses som en mærket Poisson-proces jf. Definition 3.13. For at beregne intensiteten af Poisson-processerne omkring linjerne benyttes Sætning 3.14, hvor tætheden for mærkerne er en tæthedsfunktion $f(\cdot)$ vinkelret på linjerne. Denne tæthedsfunktion skal være en funktion af afstanden fra linjen, og et passende valg er en symmetrisk tæthedsfunktion omkring linjen. Afhængig af hvordan denne tæthedsfunktion vælges, kan den være bestemmende for interaktionsafstanden fra linjen, dvs. hvor langt punkterne hørende til en linje maksimalt kan ligge, og om der er flere punkter tæt på linjen i forhold til længere væk fra linjen. Eksempler er normalfordelingen eller en ligefordeling.

Den punktproces, der ligger omkring en bestemt linje er med de ovenstående indførelser givet ved en Poisson-proces med intensitet $\beta f(\cdot)$ jf. Sætning 3.14. Tæthedsfunktionen $f(\cdot)$ er som nævnt en funktion af den vinkelrette afstand fra linjen. Fra (7.9) fås det, at den vinkelrette afstand fra et vilkårligt punkt $(x_1, x_2) \in \mathbb{R}^2$ til en linje (p_i, θ_i) kan skrives som

$$d_i(x) = |p_i - x_1 \cos(\theta_i) - x_2 \sin(\theta_i)|. \tag{11.2}$$

I henhold til datasættet om gravhøjene betragtes punktet (x_1, x_2) som koordinat for en gravhøj, mens linjen (p_i, θ_i) er den ukendte vej, som det antages at gravhøjene ligger langs.

Ved benyttelse af (11.2) er intensiteten af Poisson-processen omkring linje i givet ved

$$\rho_i(x) = \beta f(d_i(x)) = \beta f(|p_i - x_1 \cos(\theta_i) - x_2 \sin(\theta_i)|).$$

I datasættet for gravhøje, er der lokaliseringer af gravhøje, der ikke ser ud til at have nogen direkte forbindelse med vejene. Sådanne punkter skal derfor også kunne beskrives af modellen. Disse punkter kaldes baggrundspunkter, og beskrives ved en punktproces, som kaldes baggrundsprocessen. Baggrundsprocessen antages at være en Poisson-proces, og den tildeles intensiteten

$$\rho_0 = \beta_0.$$

Intensiteten af punkterne i baggrundsprocessen afhænger i høj grad af antallet af linjer og intensiteten omkring linjerne, således at hvis modellen ikke indeholder nogen linjer, skal punkterne i modellen kun stamme fra baggrundsprocessen.

Den samlede intensitet for punktprocessen til beskrivelse af datasættet består af bidrag fra punktprocesserne om de n linjer, som modellen antages at indeholde og baggrundsprocessen. Da både de enkelte punktprocesser om linjerne og baggrundsprocessen er Poisson-processer, fås det fra Sætning 3.11, at superpositionen af disse også er det, og den samlede intensitet består af summen af de enkelte intensiteter. I Sætning 3.11 skal de enkelte Poisson-processer være defineret på det samme område. Dette giver umiddelbart ikke problemer i forbindelse med denne model, da Poisson-processerne om linjerne er defineret på hele A, idet intensiteterne for dem er defineret på hele A. Desuden skal Poisson-processerne være disjunkte, hvilket gælder med sandsynlighed en. Den samlede intensitet bliver altså

$$\rho(x) = \sum_{i=0}^{n} \rho_i(x) = \beta_0 + \beta \sum_{i=1}^{n} f(d_i(x)). \tag{11.3}$$

Datasættet gravhøje indeholder 1147 punkter, hvilke i det følgende betegnes $x = \{x_1, \ldots, x_k\}$. Likelihoodfunktionen for en Poisson-proces med den ovenstående intensitet er da jf. (3.8) givet ved

$$\begin{aligned} &L(\beta, \beta_0, R|x, y) \\ &= \exp\left(-\int_A \rho(x)\ \mathrm{d}x\right) \prod_{j=1}^{k} \rho(x_j) \\ &= \exp\left(-\int_A \left(\beta_0 + \beta \sum_{i=1}^{n} f(d_i(x))\right) \mathrm{d}x\right) \prod_{j=1}^{k} \left(\beta_0 + \beta \sum_{i=1}^{n} f(d_i(x_j))\right) \end{aligned} \tag{11.4}$$

I (11.4) skal der angives en specifik tæthed for $f(d_i(x))$. I første omgang betragtes integralet

$$\int_A \rho(x)\ \mathrm{d}x = \int_A \beta_0\ \mathrm{d}x + \beta \sum_{i=1}^{n} \int_A f(d_i(x))\ \mathrm{d}x. \tag{11.5}$$

I det efterfølgende antages det, at punkterne ligger i en maksimal afstand af $R \in (0, \infty)$ fra linjerne. Tætheden vinkelret på linjen vælges som ligefordelt på $[-R, R]$. Ved en sammenligning med datasættet for gravhøje ville dette betyde, at urfolket, der benyttede vejene, ville oprette en gravhøj og begrave de døde i en maksimal afstand R fra vejene.

Med ovenstående antagelser fås det, at tætheden $f(\cdot)$ er

$$\begin{aligned} f(d_i(x)|R) &= \frac{1}{2R} \mathbb{1}[d_i(x) < R] \\ &= \frac{1}{2R} \mathbb{1}[|t_i - x_1 \cos(\theta_i) - x_2 \sin(\theta_i)| < R]. \end{aligned} \tag{11.6}$$

Denne begrænsning af hvor punkterne kan ligge i forhold til en given linje afgrænser et rektangel, hvor den ene kantlængde er $2R$ og den anden kantlængde er længden

af linjen. Betegnes længden af linje i ved l_i fås det fra (11.6), at

$$\begin{aligned}\int_A f(d_i(x)|R)\ \mathrm{d}x &= \frac{1}{2R}\int_A \mathbb{1}[|p_i - x_1\cos(\theta_i) - x_2\sin(\theta_i)| < R]\ \mathrm{d}x \\ &\approx \frac{1}{2R}2Rl_i = l_i, \end{aligned} \tag{11.7}$$

hvor integralet af indikatorfunktionen er approksimeret med arealet af rektanglen. Hvis A f.eks. er et kvadrat, er det ikke sikkert at hele rektanglet, givet ved $2R \times l_i$ ligger indenfor A, hvis f.eks. linjestykket går diagonalt igennem A, vil endefladerne af rektanglet ikke ligge indenfor A. Dette antages dog at være en passende approksimation af arealet af rektanglet indenfor A.

Dette valg af tæthedsfunktionen $f(\cdot)$ sætter begrænsninger for linjetendenserne. Eksempelvis vil linjetendenserne optræde som bånd af punkter, hvor intensiteten af punkterne er konstant over hele båndet. Der vil altså ikke optræde områder, hvor intensiteten er aftagende med afstanden fra en linje, som det antydes i datasættet for gravhøje. Det forventede antal af punkter i A fås ved indsættelse af (11.7) i (11.5) til

$$\int_A \rho(x)\ \mathrm{d}x \approx \beta_0 d^2 + \beta\sum_{i=1}^{n} l_i.$$

Det vil sige, at likelihoodfunktionen for et fastholdt linjemønster fra (11.4) bliver

$$\begin{aligned} L(\beta,\beta_0,R|x,y) = &\exp\left(-d^2\beta_0 - \beta\sum_{i=1}^{n} l_i\right) \\ &\times \prod_{j=1}^{k}\left(\beta_0 + \frac{\beta}{2R}\sum_{i=1}^{n}\mathbb{1}[|p_i - x_1\cos(\theta_i) - x_2\sin(\theta_i)| < R]\right). \end{aligned} \tag{11.8}$$

Antages det, at likelihoodfunktionen for parameteren λ givet ved (11.1) og likelihoodfunktionen for parametrene β, β_0 og R fra (11.8) er uafhængige, fås likelihoodfunktionen for λ, β, β_0 og R ved at tage produktet af de to:

$$\begin{aligned} L(\lambda,\beta,\beta_0,R|x,y) =& L(\lambda|y)L(\beta,\beta_0,R|x,y) \\ =& \left(\frac{\lambda}{\pi}\right)^n \exp\left(-\frac{2\lambda d}{\sqrt{2}}\right)\exp\left(-d^2\beta_0 - \beta\sum_{i=1}^{n} l_i\right) \\ &\times \prod_{j=1}^{m}\left(\beta_0 + \frac{\beta}{2R}\sum_{i=1}^{n}\mathbb{1}[|p_i - x_1\cos(\theta_i) - x_2\sin(\theta_i)| < R]\right). \end{aligned} \tag{11.9}$$

Udtryk (11.9) er likelihoodfunktionen for modellen til beskrivelse af punktmønstret for gravhøjene. Med de antagelser der er gjort undervejs, kan denne model forhåbentligt genskabe de linjetendenser, som ses for gravhøjene. Som en opsummering er modellen opstillet på baggrund af følgende parametre:

- λ der beskriver intensiteten af Poisson-linjeprocessen i linjerummet.
- β der beskriver intensiteten af Poisson-processen på linjerne.
- β_0 der beskriver intensiteten af Poisson-baggrundsprocessen.
- R der beskriver den maksimale afstand fra en linje til et punkt.

Disse parametre skal estimeres ud fra datasættet for gravhøjene. Dette gøres ved benyttelse af bayesiansk inferens og MH-algoritmer. Modellen bygger dog på, at linjemønstret er kendt, hvilket ikke er tilfældet ved gravhøjene, derfor skal linjerne også estimeres udfra datasættet. Linjerne kan vælges ved at betragte datasættet eller de kan estimeres med en BDM-algoritme som beskrevet i Kapitel 6. For at kunne estimere parametrene og linjerne fra data, skal modellen implementeres i `R`, hvilket gøres i følgende afsnit.

11.2 Implementering af model

I den opstillede model er både parametrene og linjerne ukendte, og disse skal dermed estimeres udfra datasættet. Dette gøres på to forskellige metoder ved benyttelse af MH-algoritmer fra Kapitel 6. `R`-koden for implementeringen af modellen findes i Appendiks B.

11.2.1 Parametre

Til at estimere parametrene λ, β, β_0 og R i modellen ud fra datasættet benyttes bayesiansk statistik. Indenfor den bayesianske statistik betragtes parameteren ζ, der ønskes estimeret, som en stokastisk variabel, og hvis der haves en forhåndsviden omkring ζ, beskrives denne ved en tæthedsfunktion $f(\zeta)$, der kaldes en prior.

Lad x være en realisering af en punktproces på repræsentationsrummet S, der følger en tæthed fra den parametriske familie, og lad likelihoodfunktionen for x være givet ved (3.7). Det er ved benyttelse af prioren samt Bayes formel muligt at opnå et udtryk for posterior-fordelingen for ζ givet data x ved

$$f(\zeta|x) \propto f(\zeta)L(\zeta|x). \tag{11.10}$$

Som det ses i (11.10) giver forskellige valg af priors forskellige posteriorfordelinger. Det viser sig som regel, at når blot der indsamles tilstrækkeligt data, vil posteriorfordelinger for to forskellige priors ikke afvige væsentlig fra hinanden. I det følgende benyttes der uegentlige priors, dvs. $f(\zeta) \propto 1$.

Det er en MH-algoritme som Algoritme 6.1, der skal benyttes til opdatering af parametrene. Definer nu parameter-vektoren $\zeta = (\lambda, \beta, \beta_0, R)$, og lad parmetervektoren $\zeta' = (\lambda', \beta', \beta_0', R')$ være et forslag. Forslagstætheden er under antagelse

af, at parametrene er uafhængige givet ved

$$q(\zeta, \zeta') = q(\lambda, \lambda')q(\beta, \beta')q(\beta_0, \beta_0')q(R, R').$$

Forslagsfordelingerne vælges for alle fire parametre til at være normalfordelte. Endimensionelle normalfordelinger er symmetriske, således at $q(\zeta, \zeta') = q(\zeta', \zeta)$ går ud i Hastings-forholdet, som reducerer til

$$H(\zeta, \zeta') = \frac{\pi(\zeta')}{\pi(\zeta)},$$

og MH-algoritmen reduceres til en Metropolis-algoritme jf. Kapitel 6.

I selve implementeringen af Metropolis-algoritmen i `R` betragtes dog i stedet logaritmen til Hastings-forholdet, idet beregningerne da bliver numerisk stabile. Dette gør ingen forskel, da logaritme-funktionen er en voksende funktion. Logaritmen til Hastings-forholdet bliver for denne model

$$\begin{aligned}\log(H(\zeta, \zeta')) =& n \log\left(\frac{\lambda'}{\lambda}\right) + \frac{2d}{\sqrt{2}}(\lambda - \lambda') + d^2(\beta_0 - \beta_0') + (\beta - \beta')\sum_{j=1}^{n} l_i \\ &+ \sum_{i=1}^{k} \log\left(\frac{\beta_0' + \frac{\beta'}{2R'}\sum_{j=1}^{n} \mathbb{1}\left[|p_i - x_{1j}\cos(\theta_i) - x_{2j}\sin(\theta_i)| < R'\right]}{\beta_0 + \frac{\beta}{2R}\sum_{j=1}^{n} \mathbb{1}\left[|p_i - x_{1j}\cos(\theta_i) - x_{2j}\sin(\theta_i)| < R\right]}\right)\end{aligned}$$

Da parametrene antages at være uafhængige, kan der opstilles et Hastings-forhold for hver enkelt parameter, hvor alt der ikke afhænger af denne parameter forkortes ud, hvilket forsimpler ovenstående Hastings-forhold betragteligt.

Lad $U_i \sim \text{unif}(0, 1)$, og parameteren udvikler sig nu som

$$\zeta_{m+1} = \begin{cases} \zeta_m', & \text{hvis} \quad \log(U_{m+1}) \leq \log(H(\zeta_m, \zeta_m')) \\ \zeta_m, & \text{ellers,} \end{cases}.$$

for $m = 1, 2, \ldots$, og en kendt begyndelsestilstand ζ_0. For passende store m konvergerer Metropolis-algoritmen, og da vil simuleringer af ζ_m kunne betragtes som simuleringer af ζ.

11.2.2 Linjer

For at opnå en passende beskrivelse af datasættet skal der tilføjes eller fjernes linjer i vist omfang, således at antallet af linjer er i overensstemmelse med antallet af forhistoriske veje. Da vejene antages at følge placeringen af gravhøjene, skal linjerne kunne placeres så de er i overensstemmelse med linjetendenserne blandt gravhøjene, dvs. at modellen også skal kunne flytte linjerne. For at estimere linjerne ud fra datasætte benyttes derfor en BDM-algoritme som gennemgået i Kapitel

6. Denne algoritme kan både skabe nye linjer, slette eksisterende linjer og flytte eksiterende linjer, således at når algoritmen køres længe nok vil antallet af linjer og selve linjerne ligge nogenlunde konstant. BDM-algoritmen gør formentlig også, at der ikke ligger linjer udenfor observationsvinduet, da der ikke er gravhøje i dette område.

For at benytte denne algoritme skal Hastings-forholdene for hhv. flytning, fødsel og død beregnes. Fra (6.1) haves Hastings-forholdet for flytning. Her kan den unormerede tæthedsfunktion $h_n(\cdot)$ ses som likelihoodfunktionen fra (11.9), hvor alt der ikke afhænger direkte af linjerne forkortes ud. Forslagstætheden for fødsel vælges på baggrund af [Geyer og Møller, 2004] til at være ligefordelt over hele området, dvs. $q_i(\cdot,\cdot) = 1/|S|$, hvor der divideres med $|S|$, da likelihoodfunktionen for linjerne er opstillet i repræsentationsrummet.

Lad nu $x = (x_1, \ldots, x_n)$ betegne det nuværende linjemønster, lad ξ være forslaget for x_i, og lad $x' = (x \backslash x_i) \cup \xi$, da fås Hastings-forholdet for flytning fra (6.1) til

$$\begin{aligned} r_f(x,\xi) =& \frac{h_n(x')\frac{1}{|S|}}{h_n(x)\frac{1}{|S|}} \\ =& \exp\left(-\beta\left(\sum_{i=1}^n l_i' - \sum_{i=1}^n l_i\right)\right) \\ &\times \prod_{j=1}^k \left(\frac{\beta_0 + \frac{\beta}{2R}\sum_{j=1}^n \mathbb{1}\left[|p_i' - x_{1j}\cos(\theta_i') - x_{2j}\sin(\theta_i')| < R\right]}{\beta_0 + \frac{\beta}{2R}\sum_{j=1}^n \mathbb{1}\left[|p_i - x_{1j}\cos(\theta_i) - x_{2j}\sin(\theta_i)| < R\right]}\right) \end{aligned}$$

hvor p_i', θ_i', l_i' alle tilhører x'.

Igen benyttes logaritmen til Hastings-forholdet i `R`, og denne bliver

$$\begin{aligned} \log(r_f(x,\xi)) =& -\beta\sum_{i=1}^n (l_i' - l_i) \\ &+ \sum_{i=1}^k \log\left(\frac{\beta_0 + \frac{\beta}{2R}\sum_{j=1}^n \mathbb{1}\left[|p_i' - x_{1j}\cos(\theta_i') - x_{2j}\sin(\theta_i')| < R\right]}{\beta_0 + \frac{\beta}{2R}\sum_{j=1}^n \mathbb{1}\left[|p_i - x_{1j}\cos(\theta_i) - x_{2j}\sin(\theta_i)| < R\right]}\right) \end{aligned}$$

For skabelse eller sletning af en linje skal (6.2) og (6.3) bestemmes. Den unormerede tæthedsfunktion $h(\cdot)$ kan igen ses som likelihoodfunktionen fra (11.9), hvor alt der ikke afhænger direkte af linjerne er forkortet ud. De øvrige udtryk, der indgår i Hastings-forholdene, vælges på baggrund af [Geyer og Møller, 2004]. Sandsynligheden for en fødsel, dvs. at der skabes en ny linje, vælges til at være konstant lig p over hele repræsentationsrummet S. Tæthedsfunktionen på $|S|$ for placeringen af en fødsel $q_b(\cdot,\cdot)$ vælges til at være $1/|S|$, hvilket betyder, at det er lige sandsynligt at skabe et punkt på hele området. Ligeledes vælges $q_d(\cdot,\cdot)$ til $1/n$, hvilket betyder, at alle punkter er lige sandsynlige for at blive fjernet.

Lad nu $x = (x_1, \ldots, x_n)$ betegne det nuværende linjemønster, lad ξ være forslaget for fødsel og η forslaget for død, og lad x' betegne det nye punktmønster i begge

tilfælde. Derved fås Hastings-forholdene til

$$
\begin{aligned}
r_b(x,\xi) =& \frac{h(x\cup\xi)(1-p)\frac{1}{n+1}}{h(x)p\frac{1}{|S|}} = \frac{h(x\cup\xi)(1-p)|S|}{h(x)p(n+1)} \\
=& \frac{\lambda}{\pi}\exp(-\beta l(\xi)) \\
& \times \prod_{j=1}^{k}\left(\frac{\beta_0 + \frac{\beta}{2R}\sum_{j=1}^{n+1}\mathbb{1}\left[|p_i' - x_{1j}\cos(\theta_i') - x_{2j}\sin(\theta_i')| < R\right]}{\beta_0 + \frac{\beta}{2R}\sum_{j=1}^{n}\mathbb{1}\left[|p_i - x_{1j}\cos(\theta_i) - x_{2j}\sin(\theta_i)| < R\right]}\right),
\end{aligned}
$$

og

$$
\begin{aligned}
r_d(x,\eta) =& \frac{h(x\backslash\eta)p\frac{1}{|S|}}{h(x)(1-p)\frac{1}{n}} = \frac{h(x\backslash\eta)pn}{h(x)(1-p)|S|} \\
=& \left(\frac{\lambda}{\pi}\right)^{-1}\exp(+\beta l(\eta)) \\
& \times \prod_{j=1}^{k}\left(\frac{\beta_0 + \frac{\beta}{2R}\sum_{j=1}^{n-1}\mathbb{1}\left[|p_i' - x_{1j}\cos(\theta_i') - x_{2j}\sin(\theta_i')| < R\right]}{\beta_0 + \frac{\beta}{2R}\sum_{j=1}^{n}\mathbb{1}\left[|p_i - x_{1j}\cos(\theta_i) - x_{2j}\sin(\theta_i)| < R\right]}\right).
\end{aligned}
$$

Det er med det ovenstående nu muligt at estimere parametrene og linjerne ud fra datasættet.

11.3 Model med linjer fra BDM-algoritme

For at implementere algoritmen i `R`, fastsættes først begyndelsesværdierne for parameteren $\zeta = (\lambda, \beta, \beta_0, R)$. Begyndelsesværdierne vælges først forholdsvis tilfældigt, og valget optimeres ved at vælge nye værdier efter én simulering. I første omgang er begyndelsesværdierne valgt til $\zeta' = (0.001, 0.001, 5\cdot 10^{-6}, 250)$.

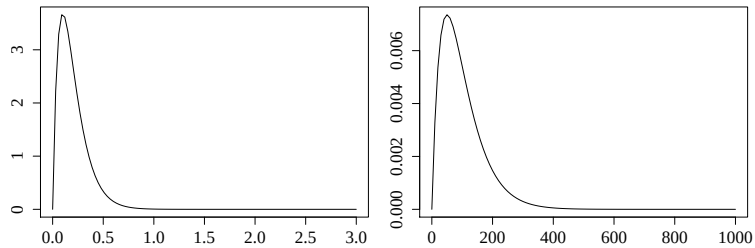
Begyndelsesværdien for λ er valgt på baggrund af linjetendenserne i Figur 9.2. I første omgang antages det, at der skal benyttes ca. 30 linjer til at beskrive linjetendenserne, hvor der tages højde for, at der måske skal bruges flere linjer til at beskrive en linjetendens, da linjetendenserne ofte buer. Dette betyder også, at antallet af punkter i repræsententationsrummet er 30, og $\mathbb{E}[\#\text{punkter}] = \rho\cdot 2\pi\cdot 15000/\sqrt{2} = 30$, således at $\rho \approx 0.0004$ og dermed at $\lambda = \rho\cdot\pi \approx 0.001$. Begyndelseværdierne for β er valgt uden de store overvejelser, da der ikke vides noget om denne parameter. Begyndelsesværdien for β_0 er valgt udfra, at hvis $\lambda, \beta, R = 0$, så ville alle punkter stamme fra baggrundsprocessen. Altså $\mathbb{E}[\#\text{punkter}] = \beta_0 \cdot 15000^2 = 1147$, hvorudfra β_0 bestemmes til ca. $5\cdot 10^{-6}$. Begyndelsesværdien for R er valgt på baggrund af, at linjetendenserne i punktmønstret vurderes til at være 500 meter brede. De valgte begyndelsesværdier indsættes i Hastings-forholdet for parametrene som de kendte begyndelsestilstande.

For at sikre at posterioren ikke bliver uegentlig vælges der ikke-uegentlige priorer til parametrene. Det er antaget, at parametrene apriori er uafhængige, og da der ikke

haves nogen forhåndsviden om fordelingen af parametrene vælges der forholdsvis flade priors i form af gammatætheder. Prioerne for λ, β og β_0 vælges ens, og de vælges så størstedelen af sandsynlighedsmassen ligger i nærheden af de valgte begyndelsesværdier. Der vælges derfor gammatætheder med skalaparameter 0.1 og formparameter 2

$$f(\lambda) = \frac{\lambda}{2} \exp\left(-\frac{\lambda}{0.1}\right).$$

Formen af en gammatæthed med disse parametre kan ses på Figur 11.2, hvor det ses, at med de valgte parametre ligger størstedelen af sandsynlighedsmassen mellem 0 og 0.5, som også er der λ, β og β_0 forventes at ligge.



Figur 11.2: Priortætheder for parametrene. Til venstre prior for λ, β og β_0 og til højre prior for R.

Prioren for R vælges til en gammatæthed med skalaparameter 50 og formparameter 2,

$$f(R) = \frac{R}{2^{50}\Gamma(50)} \exp\left(-\frac{R}{50}\right).$$

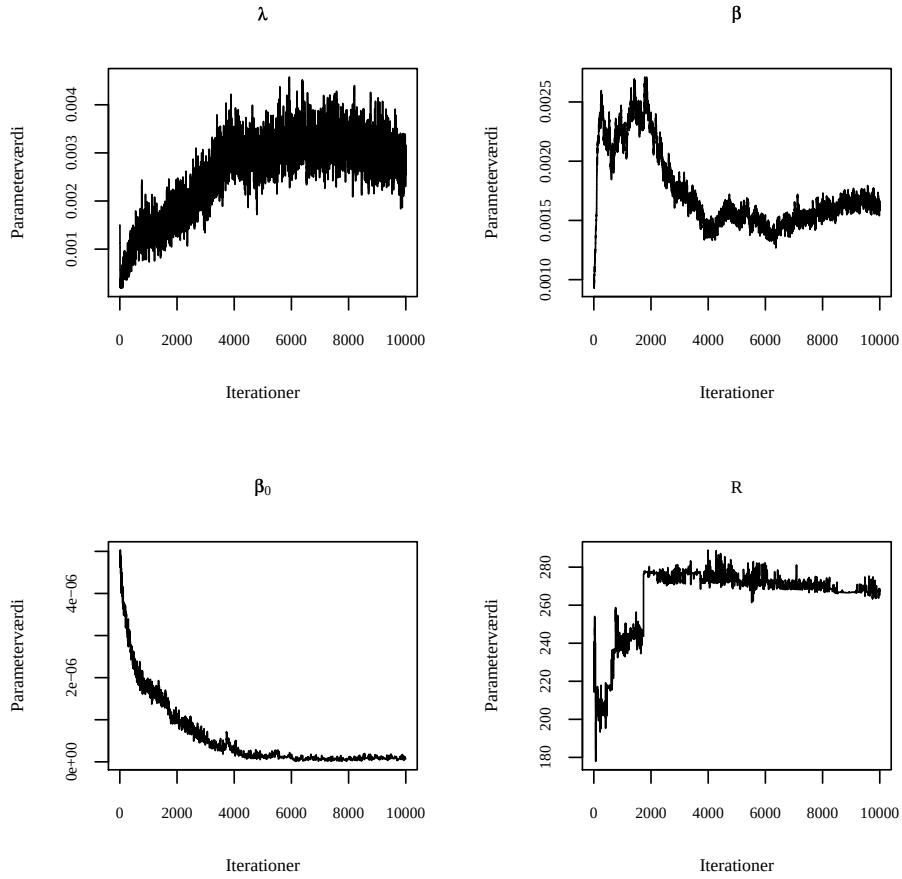
da denne fordeling giver en større varians. Formen af denne priortæthed ses på Figur 11.2.

De ovenstående priortætheder ganges på likelihoodfunktionen fra (11.9), og de ganges også på Hastings-forholdet. (Som det senere ses har disse priortætheder ikke den store betydning på posterioren, og det er tjekket, at fire uegentlige priorer ($f(\cdot) \propto 1$) giver samme form af posteriorfordelingen.)

Til estimeringen af linjerne vælges $q = 0.8$ i Algoritme 6.3 og $p = 0.8$ i Algoritme 6.2. Der er forsøgt med forskellige værdier af q og p, men de valgte værdier synes at give den hurtigste konvergens. Som begyndelsestilstand for BDM-algortimen vælges en Poisson-proces på repræsentationsrummet med lav intensitet svarende til det næsten-tomme punktmønster. Der vælges det næsten-tomme punktmønster fremfor det tomme punktmønster, da beregningen af Hastings-forholdene for nul linjer ikke er muligt. Desuden kræver Hastings-forholdet for død minimum to linjer. Dette giver ikke problemer, da der ses at være flere end nul linjetendenser, og da BDM-algoritmen automatisk flytter disse få linjer efter behov.

Der foretages først en kørsel på 10000 iterationer for at kontrollere begyndelsesværdierne for parametrene og for eventuelt at kunne optimere disse. Plottes de

itererede værdier for hhv. λ, β, β_0 og R mod iterationsnummeret fås udviklingen af værdierne i kæden, dvs. et traceplot, som på Figur 11.3.

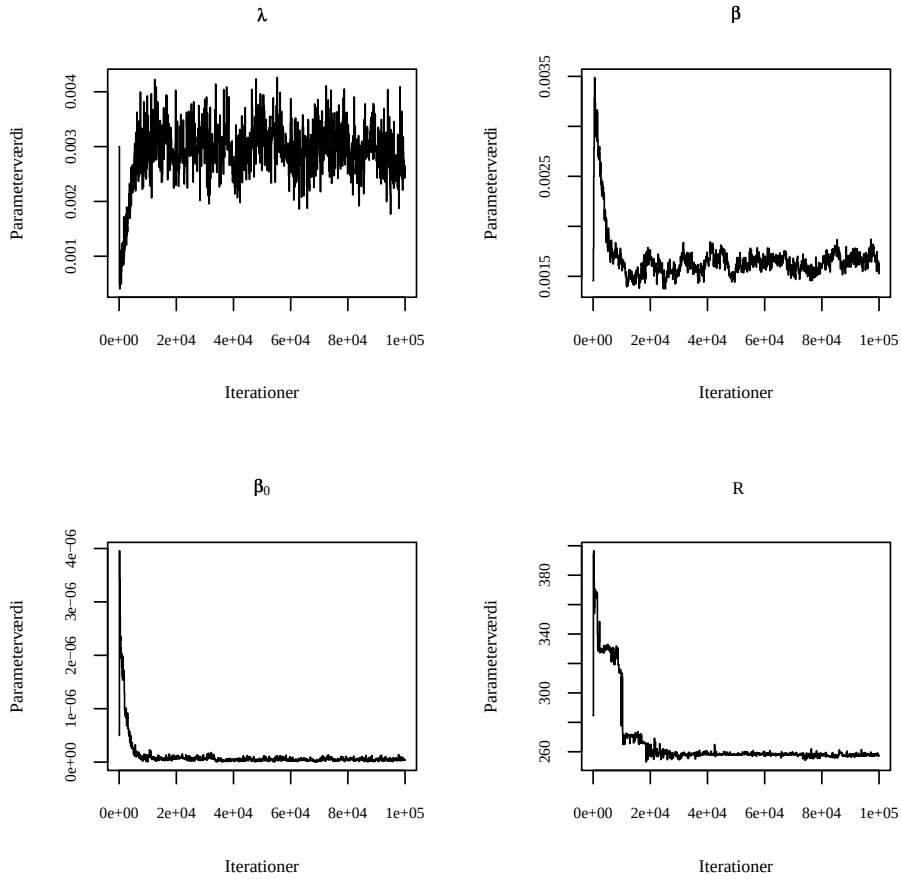


Figur 11.3: Traceplot af λ, β, β_0 og R med 10000 iterationer.

Af denne forholdsvise korte kørsel ses det, at parametrene ikke helt er konvergerede endnu, men kørslen giver en indikation af, mod hvilket område de enkelte parametre konvergerer.

På baggrund af denne kørsel vælges der nye startværdier på $\zeta' = (0.003, 0.0015, 5 \cdot 10^{-7}, 270)$. Med disse begyndelsesværdier foretages en kørsel på 100000 iterationer, og traceplots for denne ses på Figur 11.4. Bemærk, at der i dette plot er foretaget en udtynding i parameterværdierne, så kun hver 100. parameterværdi er plottet, da dette giver mere informative traceplots.

Acceptsandsynlighederne for parametrene styres af varianserne af forslagsfordelingerne og er på hhv. $(0.65, 0.81, 0.07, 0.10)$. Jævnfør [Berthelsen og Møller, 2004]



Figur 11.4: Traceplot af hver 100. værdi af λ, β, β_0 og R fra simulering med 100000 iterationer.

bør acceptsandsynlighederne ligge i området 0.2-0.4 for at opnå en optimal konvergens. Som det ses, ligger acceptsandsynlighederne for parametrene ikke i dette interval, de er faktisk væsentligt højere eller væsenligt lavere. Det har været forsøgt at få acceptsandsynlighederne til at ligge i intervallet 0.2-0.4, men dette har ikke kunnet gøres bedre end det ovenstående, da en ændring af variansen i forslagsfordelingen for en af parametrene kan få betydning for acceptsandsynligheden for andre parametre. Disse acceptsandsynligheder kan altså have indvirkning på konvergensen. I [Møller og Waagepetersen, 2004, s. 202] opnås en acceptsandsynlighed på 0.57, uden der kommenteres nærmere på denne, og derfor betragtes acceptsandsynlighederne som værende acceptable.

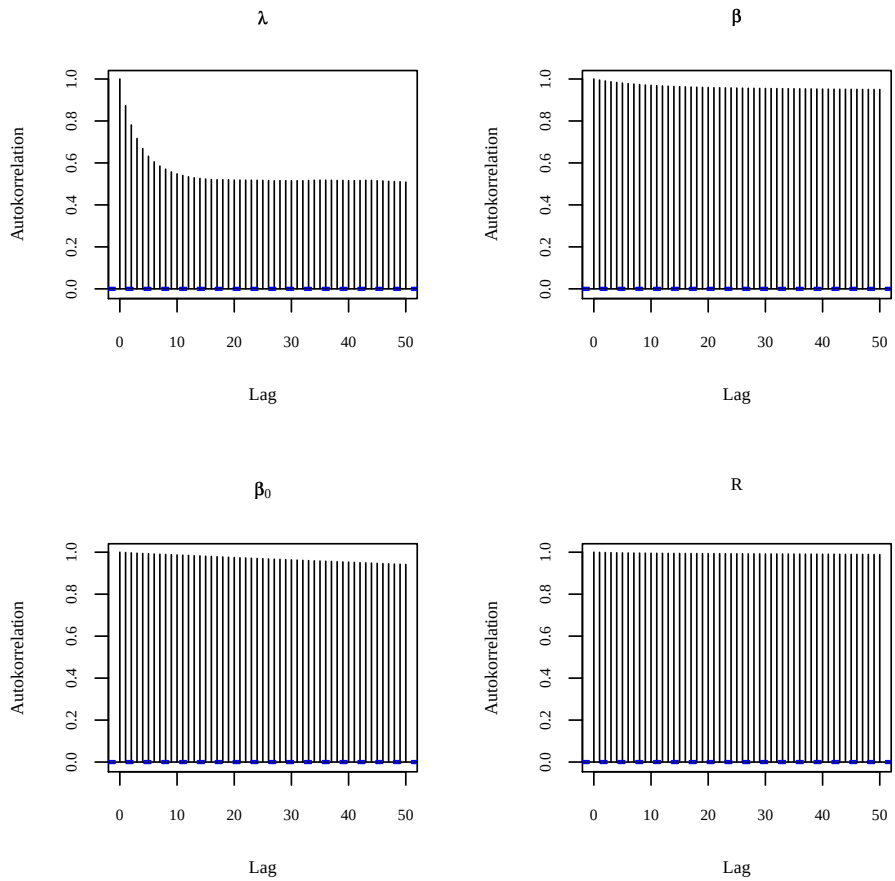
For at vurdere, hvor hurtigt de forskellige parametre konvergerer, altså hvor hurtigt en parameterværdi er uafhængig af foregående parameterværdier, betragtes

en metode, der kan afgøre, hvor hurtigt en markovkæde går mod den invariante fordeling. Metoden består i at plotte autokorrelationen for de forskellige parametre. For en reel funktion $k(\cdot)$ med en endelig varians givet ved

$$\sigma^2 = \mathrm{Var}[k(X_j)],$$

er lag-m-autokorrelationen κ_m defineret ved

$$\kappa_m = \mathrm{corr}(k(X_j), k(X_{j+m})) \quad \text{for } m = 0, 1, \ldots,$$

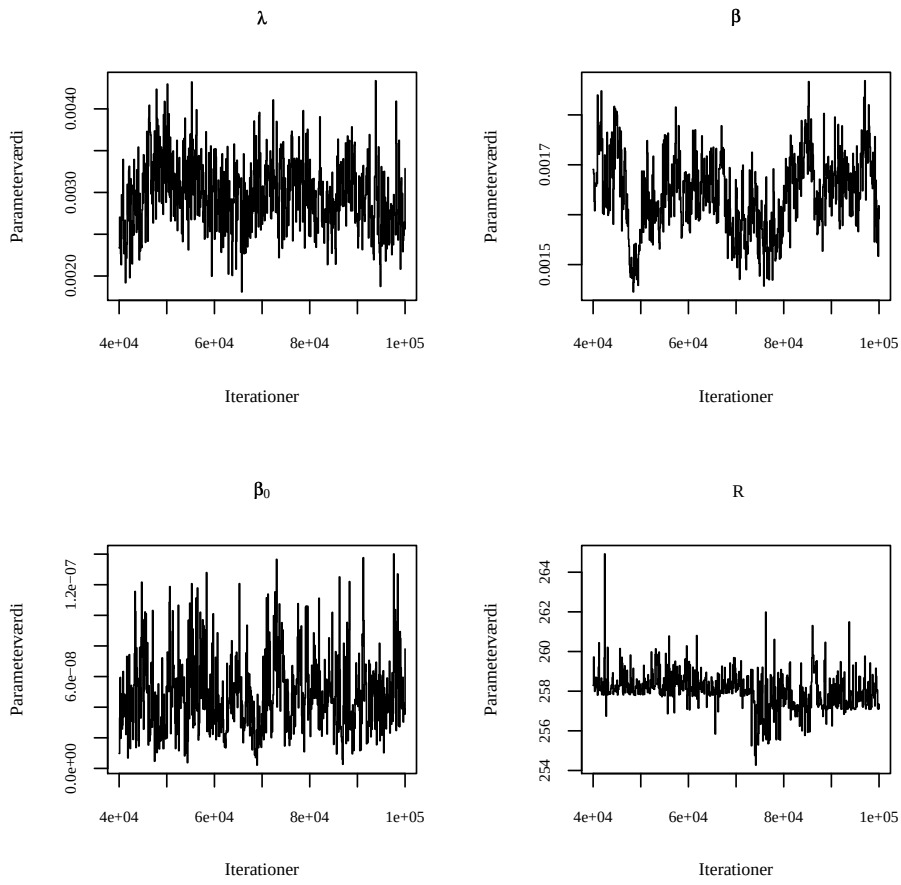


Figur 11.5: Autokorrelationsplot for hhv. λ, β, β_0 og R.

hvor indekset m kaldes for lag. Under passende betingelser vil $\kappa_m \to 0$ for $m \to \infty$. Autokorrelationen er altså et udtryk for, hvor afhængig en parameter er af de foregående værdier.

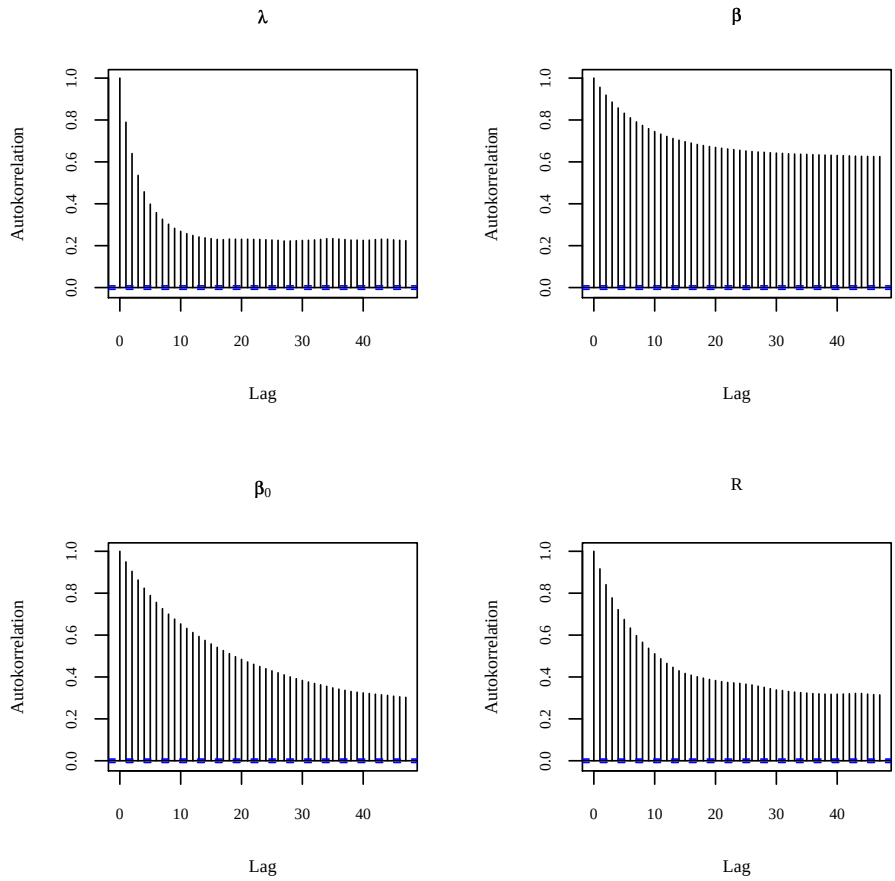
På Figur 11.5 ses autokorrelationen af parametrene fra Figur 11.4. Det ses på figuren, at alle fire parametre er autokorrelerede. Dette kan skyldes, at der jf. Figur 11.4 bør foretages et burn-in for at opnå den korrekte posteriorfordeling af parametrene. Det ses af Figur 11.4, at alle parametre er konvergeret i fordeling efter ca. 40000 iterationer og der laves et burn-in på dette.

På Figur 11.6 ses traceplots for hver 100. værdi efter burn-in på 40000. Efter dette burn-in ses det tydligere, at parametrene er konvergerede.



Figur 11.6: Traceplot af hver 100. værdi af λ, β, β_0 og R med 100000 iterationer og et burn-in 40000.

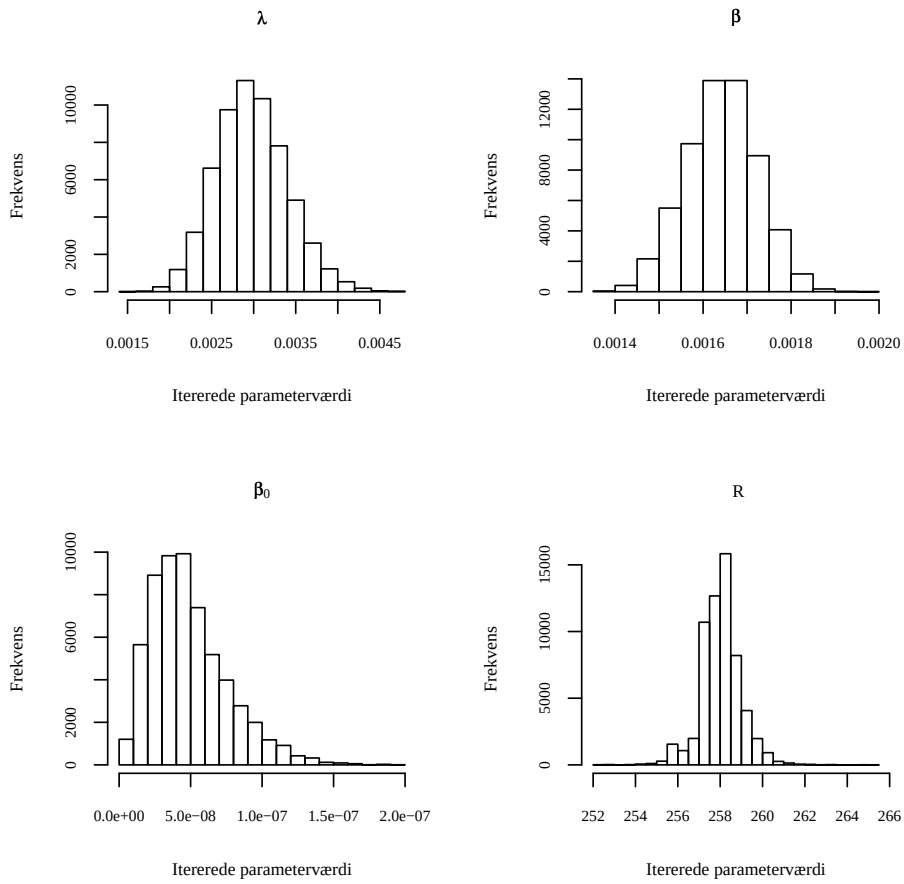
Autokorrelationsplot efter burn-in på Figur 11.7 viser, at parametrene konvergerer hurtigere end før, men uafhængigheden mellem nuværende og tidligere parameterværdier indtræffer lidt langsomt.



Figur 11.7: Autokorrelationsplot for λ, β, β_0 og R med 100000 iterationer og et burn-in 40000.

Betragtes histogrammer for hhv. λ, β, β_0 og R, efter burn-in på 40000 fås posteriorfordelingen for disse. På Figur 11.8 er histogrammerne for de fire forskellige parametre afbilledet.

Det ses, at λ er centreret omkring ca. 0.006, β er centreret omkring ca. 0.00165. Disse to posteriorfordelinger ligner begge normalfordelinger. Parameteren β_0 minder mere om en gammafordeling, hvor størstedelen af værdierne ligger i intervallet 0-1.5$\cdot 10^{-7}$. Parameterværdien for R ligner tildels også en normalfordeling centreret omkring 258. I Tabel 11.1 er forskellige informationer om parametrene opsummeret.

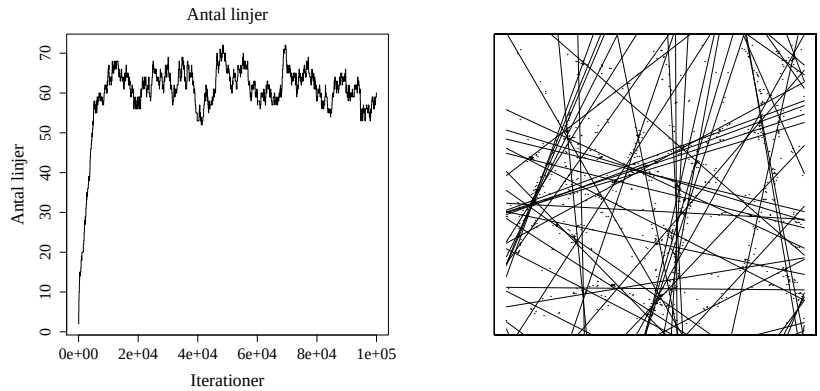


Figur 11.8: Histogrammer for hhv. λ, β, β_0 og R efter burn-in.

	λ	β	β_0	R
Min.	0.001442929	0.001355540	2.160517e-09	252.1521
1. fraktil	0.002678878	0.001587348	2.909111e-08	257.4182
Median	0.002959206	0.001644079	4.400968e-08	258.0388
Middelværdi	0.002977638	0.001642109	4.901079e-08	258.0329
3. fraktil	0.003250586	0.001697454	6.284899e-08	258.5278
Maks.	0.004756489	0.001978465	1.932323e-07	265.2808

Tabel 11.1: Beskrivende information om parametrene λ, β, β_0 og R fra simulering med 100000 iterationer og efter burn-in på 40000.

En vigtig del af modellen er også, at antallet af linjer er nogenlunde konstant ellers vil parameterværdierne ikke konvergere. At antallet af linjer er konstant, betyder også, at det sandsynligvis er nogle optimale linjer, der er estimeret. Udviklingen i antallet af linjer ses på Figur 11.9(a), og dette ses at være konstant omkring de 61 linjer, der er opnået til slut.



(a) (b)

Figur 11.9: (a) Udviklingen af antallet af linjer for hver 100. iteration, som ses at være stabiliseret omkring ca. 61 efter 10000 iterationer. (b) Linjerne på punktmønstret.

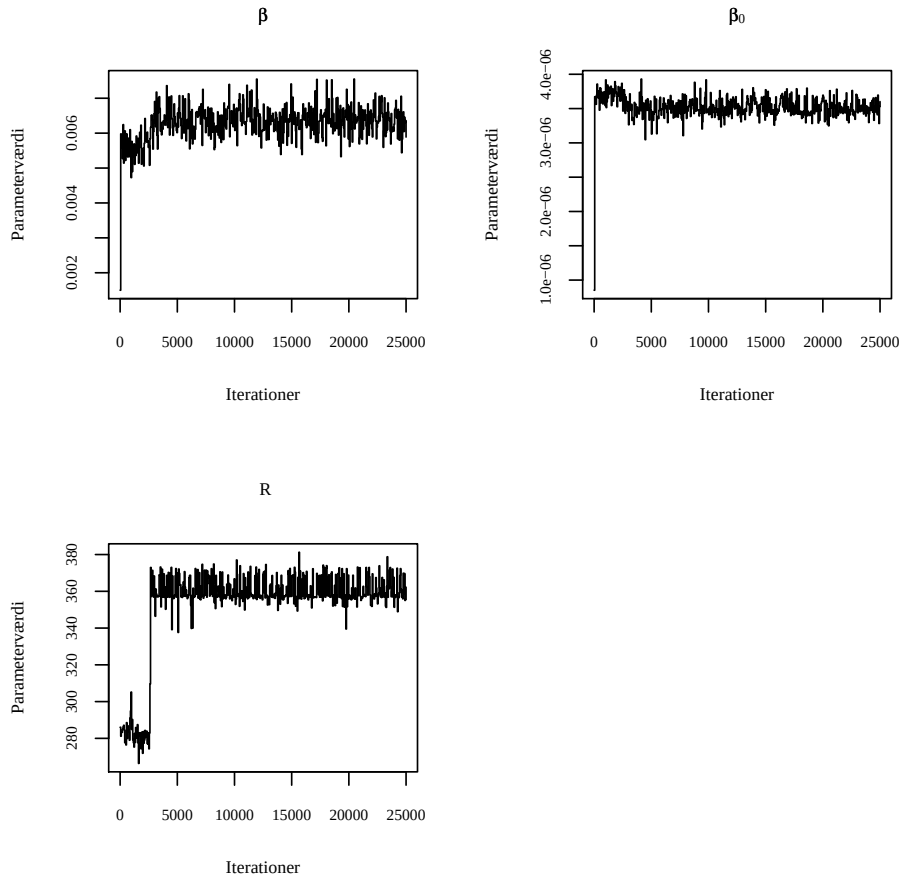
De ukendte parametre fra modellen er nu estimeret. I næste kapitel foretages et modeltjek af modellen med de estimerede linjer og parametre. Men allerede her kan det undersøges om de opnåede linjer ligger langs linjetendenserne. På Figur 11.9(b) er linjerne plottet sammen med det oprindelige punktmønster. Antallet af linjer virker lidt voldsomt, men dette skyldes, at de buede linjetendenser estimeres med rette linjer, og derfor forsøger modellen at tilpasse linjetendenserne ved brug af flere linjer. På figuren ses det, at de fleste af linjerne ser ud til at følge linjetendenserne. Der er dog også visse linjer, der ikke umiddelbart ser ud til at have nogen tilknytning til linjetendenserne i datasættet. Punkter i tilknytning til disse lidt afvigende linjer kunne være beskrevet ved baggrundsprocessen fremfor punktprocessen om den pågældende linje.

Af plottet ses det dog, at modellen passer rimeligt, men for at konstatere om den er anvendelig, er det nødvendigt at fortage et decideret modeltjek, hvilket gøres i næste kapitel. Først undersøges det, om det er muligt at opnå en bedre konvergens af parametrene.

11.4 Model med passende fast linjemønster

Modellen i det forrige afsnit beskriver en linjetendens i datasættet med flere linjer, da de forhistoriske veje ofte er buede, og dermed ikke kan beskrives ved en

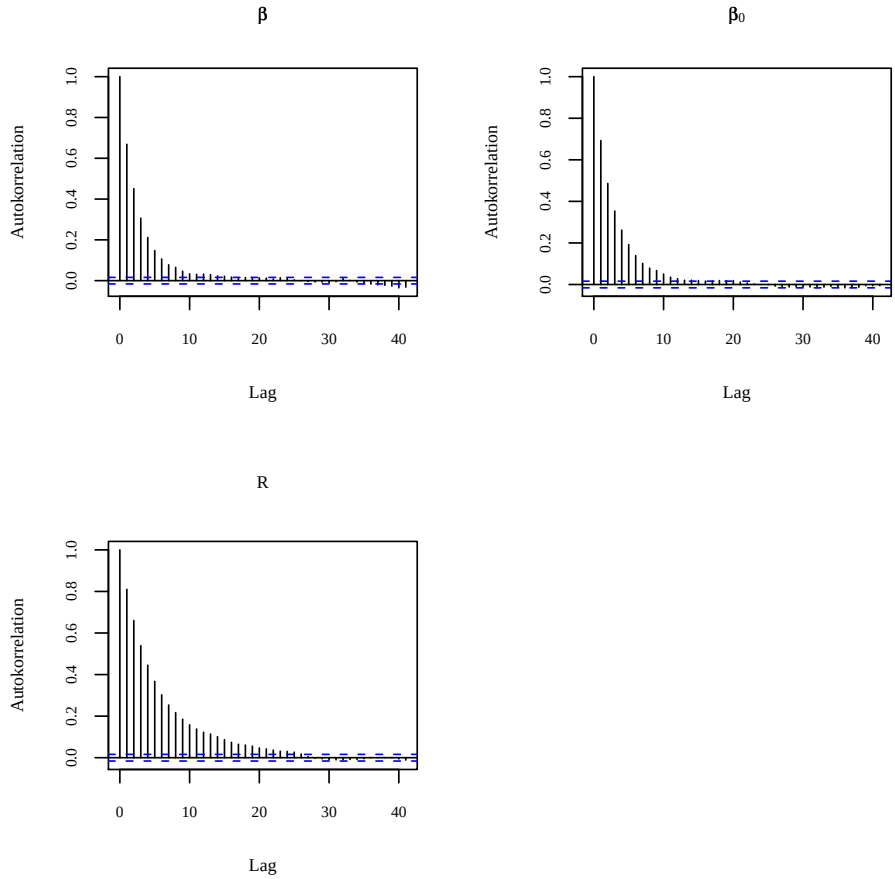
enkelt ret linje. Samtidig er linjerne i den forrige model ikke fastholdte, og der er mulighed for at tilføje, slette og flytte linjer. Dette kan være en af årsagerne til at autokorrelationen viser, at parametrene ikke konvergerer specielt godt. I dette afsnit undersøges det, om det er BDM-algoritmen med de varierende linjer, der medfører den dårlige konvergens. Dette gøres for at sikre, at det ikke er selve modellen, der er problemet i forbindelse med konvergensen, men at det nærmere er flytningen af linjerne, og at der tilføjes og slettes i henhold til BDM-algoritmen. For at optimere konvergensen kan der vælges nogle faste linjer. Når parametrene opdateres udfra disse faste linjer, er der mulighed for at opnå en bedre konvergens.



Figur 11.10: Traceplot af hver 50. værdi af β, β_0 og R med 25000 iterationer.

Som det fastholdte linjemønster kan der eksempelvis vælges 10 tilfældige linjer og tilpasse disse til datasættets linjetendenser med en Move-algoritme, eller der kunne benyttes de opnåede linjer fra forrige afsnit. En ulempe ved at benytte Move-algoritmen er, at alle de 10 linjer i princippet kan komme til at ligge næsten

oven i hinanden og beskrive den samme linjetendens. Selvom dette ikke burde give problemer for konvergensen, vælges linjerne fra Figur 11.1(b), som de faste linjer.



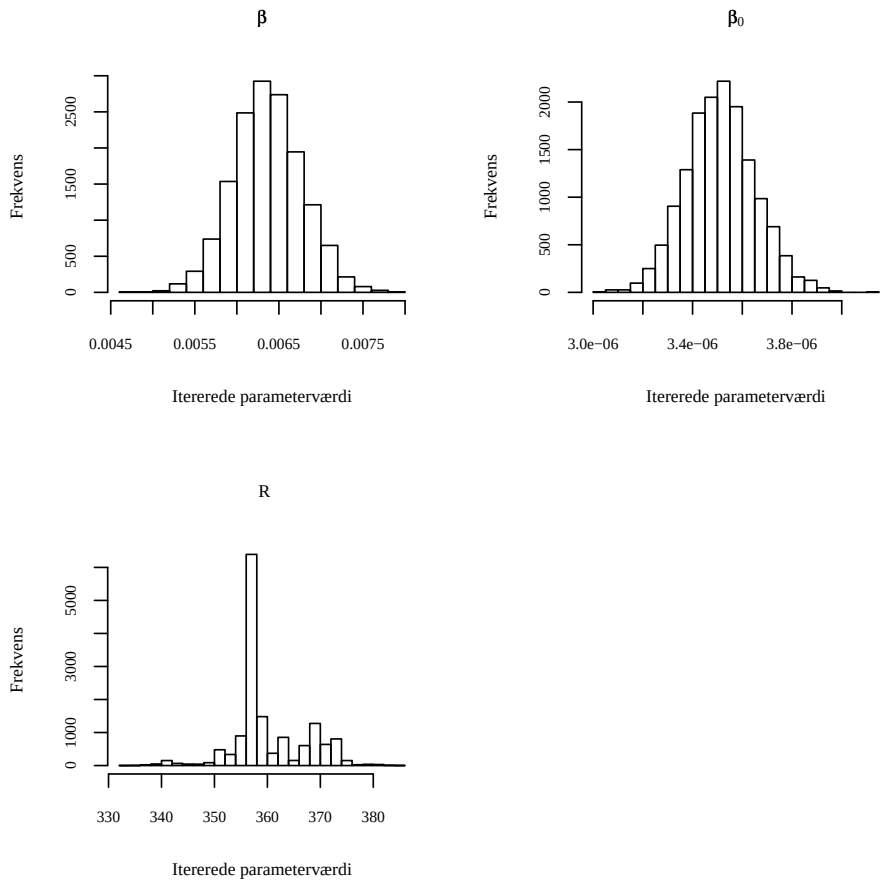
Figur 11.11: Autokorrelationsplots for β, β_0 og R med 25000 iterationer og efter burn-in på 10000.

I henhold til forrige afsnit svarer dette til, at det kun er parametrene, der skal opdateres. Men når linjemønstret allerede er kendt, optræder der ingen likelihood-funktionen for Poisson-linjeprocessen i (11.3). Det er således kun parametrene β, β_0 og R, der skal estimeres fra datasættet.

På Figur 11.10 ses traceplots af hver 50. værdi for de fire parametre for en kørsel med 25000 iterationer. Parametrene virker konvergerede efter ca. 5000-10000 iterationer og der laves burn-in på 10000. Begyndelsesværdierne for denne kørsel er valgt i nærheden af tidligere opnåede værdier, men disse burde ikke have betydning, når linjerne er faste.

På figuren ses der et besynderligt spring i traceplottet for parameteren R ved omkring 3000 iterationer, og der ses samtidig en lille ændring for β og β_0. Forklaringen på dette spring vides ikke. Ved at betragte værdierne i området ses det, at ændringen ikke foregår i et spring men mellem ca. 50-100 iterationer, og det store spring i værdier virker derfor ikke så besynderligt.

På Figur 11.11 ses autokorrelationen af parametrene, efter der er fortaget burn-in. I forhold til modellen med de varierende linjer ses det, at parametrene hurtigt bliver uafhængige af tidligere værdier. Grunden til denne hurtige konvergens skyldes altså, at linjerne ligger fast.



Figur 11.12: Histogrammer for β, β_0 og R med 25000 iterationer og efter burn-in på 10000.

Histogrammer for parameterværdierne ses på Figur 11.12. Posteriorfordelingerne for parameterne β og β_0 ligner begge normalfordelinger. Posteriorfordelingen for

R er mere udefinerbar, men den har en tendens til at ligge omkring 258. I Tabel 11.2 er forskellige informationer om parametrene opsummeret.

	β	β_0	R
Min.	0.004750056	3.046487e-06	333.0582
1. fraktil	0.006093235	3.418108e-06	356.7417
Median	0.006361246	3.508204e-06	357.5262
Middelværdi	0.006367706	3.513129e-06	360.1118
3. fraktil	0.006634981	3.600954e-06	363.9512
Maks.	0.007978382	4.113962e-06	384.5382

Tabel 11.2: Beskrivende information om parametrene β, β_0 og R fra simulering med 25000 iterationer og efter burn-in på 10000.

Med fastlagte linjer er det altså muligt at opnå en væsentlig bedre konvergens af parametrene. Det er altså formentlig BDM-algoritmen, der er skyld i, at parameterværdierne ikke konvergerer.

For alle parametrene gælder det dog, at de ikke konvergerer mod de samme værdier som for de varierende linjer. Dette skyldes, at der kun er 4 linjer imod 61 linjer, og at en linjetendens i datasættet nu kun beskrives ved en enkelt linje. Dette gør, at R og β bliver større, da det samme antal punkter i datasættet nu skal beskrives ved kun 4 linjer. Parameteren β_0 bliver også væsentligt større, da der er flere punkter, der ikke har forbindelse til de fire linjer.

Der har i dette kapitel været opstillet en model til beskrivelse af linjetendenser i datasættet for gravhøje. Modellen indeholder fire ukendte parametre og ukendte linjer, som alle er blevet estimeret udfra datasættet for gravhøje. Det kan umiddelbart være svært at vurdere om disse parametre stemmer overens med datasættet, så for at undersøge om den opstillede model kan beskrive datasættet i tilstrækkelig grad, skal der foretages et modeltjek. Dette gøres i næste kapitel, hvor der også beregnes summary statistics for den opstillede model.

Modeltjek for model med linjetendenser

I dette kapitel foretages to modeltjek af modellen fra Kapitel 11. Det første modeltjek er for modellen fra Afsnit 11.3, hvor de estimere linjer benyttes. Det andet modeltjek foretages ved benyttelse af λ, således at antallet af linjer og placeringen af linjer også er stokastisk.

En første indikation af om en model giver en god beskrivelse af data fås ved at plotte en realisering af modellen og sammenligne den visuelt med punktmønstret for data. Udfra denne realisering er det muligt at opnå en fornemmelse af om modellen er acceptabel.

Et decideret modeltjek er dog nødvendig for mere præcist at afgøre om modellen giver en passende beskrivelse af datasættet. Som tidligere beskrevet foretages et sådant modeltjek for punktproces-modeller ofte ved at sammenligne summary statistics for data med summary statistics fra modellen.

Jævnfør Kapitel 4 og Kapitel 5 beregnes en specifik summary statistic udfra modellen. I det følgende afsnit beregnes en summary statistic for den opstillede model, der inddrager de forskellige modelparametre, der er blevet introduceret undervejs.

12.1 Beregning af summary statistic

I dette afsnit beregnes parkorrelationsfunktionen for modellen opstillet i Afsnit 11.1. Alt notation i dette afsnit er som i afsnittet for modelopstillingen.

Modellen består af en baggrundsproces, der er en Poisson-proces med intensitet β_0, en skjult Poisson-linjeproces med n linjer og intensitet λ, og omkring hver af disse linjer er der en Poisson-proces med intensitet $\beta f(\cdot)$.

Den samlede model er da en Poisson-proces med intensitet $\beta_0 + \beta \sum_{i=1}^{n} f(d_i(x))$ givet linjerne. Det vil altså sige, at givet linjemønstret er punkterne en Poisson-proces, og dermed udgør punkterne en Cox-proces jf. Definition 5.1, hvilket anvendes i følgende sætning. I forhold til sprogbrugen i Kapitel 5 følger punkterne en Cox-proces drevet af $\beta_0 + \beta \sum_{i=1}^{n} f(d_i(x))$.

Sætning 12.1
Intensiteten af punkterne fra modellen er

$$\rho(x) = \beta_0 + \beta\lambda.$$

Bevis:
Antag, at $y = \{\ell_1, \ldots, \ell_n\}$ er en realisation af en Poisson-linjeproces Y på linjerummet. Da er den betingede intensitet af Poisson-processen $X|(Y = y)$ jf. udtryk (11.3) givet ved

$$Z(x) = \sum_{i=0}^{n} \rho_i(x) = \beta_0 + \beta \sum_{i=1}^{n} f(d_i(x)),$$

hvor det for den opstillede model er antaget, at $f(d_i(x)|R) = \frac{1}{2R}\mathbb{1}[d_i(x) < R]$. Det vil sige, at punkterne udgør en Cox-proces drevet af $Z(\cdot)$.

For et givet punkt x fås antallet af linjer, der har en afstand mindre end R til x ved

$$\sum_{i=1}^{n} f(d_i(x)|R) = \frac{1}{2R} N(B(x,R)),$$

hvor $N(B(x,R))$ er antallet af linjer, der går gennem cirklen $B(x,R)$.

Da linjeprocessen er stationær, kan origo for den betragtede cirkel vælges vilkårligt, dvs. at $B(0,R)$ kan vælges fremfor $B(x,R)$. Da en cirkel i $\mathbb{R}^2$ er en konveks mængde, kan det forventede antal af linjer, der gennemskærer cirklen, fås til intensiteten af linjeprocessen λ ganget med perimeteren af cirklen divideret med π jf. Sætning 7.9. Dermed fås intensiteten af punkterne jf. Sætning 5.2 ved

$$\begin{aligned}\rho(x) = \mathbb{E}[Z(x)] = \beta_0 + \beta\mathbb{E}\left[\sum_{i=1}^{n} f(d_i(x))\right] &= \beta_0 + \frac{\beta}{2R}\mathbb{E}[N(B(0,R))] \\ = \beta_0 + \frac{\beta}{2R}\frac{\lambda 2\pi R}{\pi} &= \beta_0 + \beta\lambda. \end{aligned} \tag{12.1}$$

□

Det bemærkes at i ovenstående sætning afhænger intensiteten af punkterne kun af parametrene λ, β og β_0, hvilket netop er de parametre, der er bestemmende for intensiteten af punkterne for modellen.

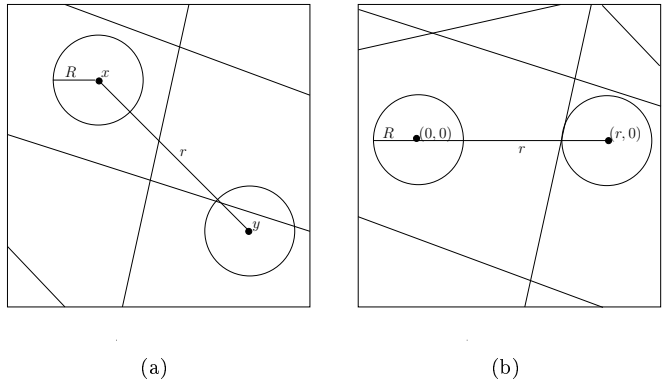
Sætning 12.2
Lad punkterne $x, y \in \mathbb{R}^2$ og lad $R > 0$. Så er det gennemsnitlige antal linjer fra modellen, der rammer både $B(x,R)$ og $B(y,R)$ givet ved

$$\lambda_1 = \begin{cases} \frac{2\lambda}{\pi}\left(\sqrt{\frac{r^2-4R^2}{r^2}}r + 2R\sin^{-1}\left(\frac{2R}{r}\right) - r\right), & \text{for } r > 2R, \\ \frac{2\lambda}{\pi}(\pi R - r), & \text{for } r \leq 2R, \end{cases}$$

hvor $r = |x - y|$.

Bevis:
Da linjeprocessen er stationær og isotropisk, kan punkterne x og y flyttes til $(0,0)$

og $(0, r)$. På Figur 12.1 ses en illustration af cirklernes translation og rotation. Figurerne er kun illustrative, og linjerne er blot tilfældigt placeret i linjerummet.

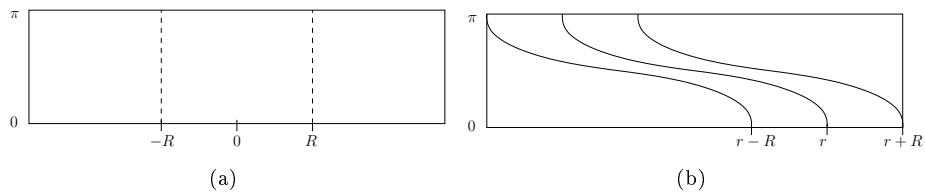


Figur 12.1: Illustration af translation og rotation. (a) Før flytningen og (b) efter flytningen (Linjerne er kun illustrative.)

For de linjer, der rammer $B((0,0), R)$, er $|p_i| < R$, svarende til et afgrænset område i repræsentationsrummet, som det ses på Figur 12.2(a). De linjer der rammer cirklen $B((0,r), R)$, er dem, der rammer indenfor R af $(0, r)$ for et givet θ. Illustration af dette ses på Figur 12.3. Ved benyttelse af sinusrelationen fås det, at

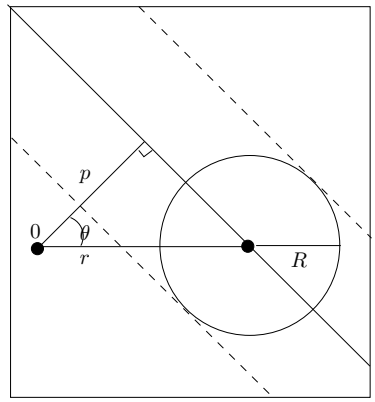
$$\frac{\sin(\pi/2 - \theta)}{x} = \frac{\sin(\pi/2)}{r},$$

og dermed $\theta = \cos^{-1}(x/r)$. Området i repræsentationsrummet for linjerne i cirklen $B((0,r), R)$ ses på Figur 12.2(b).



Figur 12.2: Repræsentationsrum for de linjer der skrærer en given cirkel i linjerummet. (a) Cirklen $B((0,0), R)$. (b) Cirklen $B((0,r), R)$.

Linjerne, der rammer begge områder, ligger i overlappene mellem de to områder af repræsentationsrummet, dvs. overlappene mellem Figur 12.2(a) og (b). Der betragtes tre tilfælde og arealerne i de tre tilfælde beregnes til



Figur 12.3: Illustration af de linjer der skærer cirklen $B((0,r),R)$.

(i) $r > 2R$:

$$\begin{aligned} A_1 &= \int_{-R}^{R} \left(\cos^{-1}\left(\frac{t-R}{r}\right) - \cos^{-1}\left(\frac{t+R}{r}\right)\right) \, \mathrm{d}t \\ &= 2\left(\sqrt{\frac{r^2-4R^2}{r^2}}r + 2R\sin^{-1}\left(\frac{2R}{r}\right) - r\right). \end{aligned}$$

(ii) $R < r \leq 2R$:

$$\begin{aligned} A_1 =& 2\int_{0}^{r-R} \left(\cos^{-1}\left(\frac{t-R}{r}\right) - \cos^{-1}\left(\frac{t+R}{r}\right)\right) \, \mathrm{d}t \\ &+ 2\int_{r-R}^{R} \left(\cos^{-1}\left(\frac{t-R}{r}\right) - 0\right) \, \mathrm{d}t \\ =& 2(\pi R - r). \end{aligned}$$

(iii) $0 \leq r \leq R$:

$$\begin{aligned} A_1 &= 2\int_{0}^{R-r} (\pi - 0) \, \mathrm{d}t + 2\int_{R-r}^{R} \left(\cos^{-1}\left(\frac{t-R}{r}\right) - 0\right) \, \mathrm{d}t \\ &= 2(\pi R - r). \end{aligned}$$

Fra Definition 3.2 fås det, at antallet af linjer der rammer begge cirkler, dvs. punkter indenfor A_1, er Poisson-fordelt med middelværdi

$$\lambda_1 = \frac{\lambda}{\pi} A_1. \tag{12.2}$$

Når de forskellige arealer indsættes i (12.2) fås sætningen. □

Med de to foregående sætninger er det nu muligt at beregne parkorrelationsfunktionen for modellen.

Sætning 12.3
Parkorrelationsfunktionen for en punktproces med intensitet givet ved Sætning 12.1 er

$$g(r) = \frac{\beta_0^2 + 2\beta\beta_0\lambda + \frac{\beta^2}{4R^2}(\lambda_1 + (2R\lambda)^2)}{\beta_0^2 + 2\beta\beta_0\lambda + \beta^2\lambda^2}.$$

Bevis:
Da det er parkorrelationsfunktionen fra (5.1), der skal beregnes betragtes i første omgang andenordens produkttætheden

$$\begin{aligned}\mathbb{E}[Z(x)Z(y)] &= \mathbb{E}\left[\left(\beta_0 + \frac{\beta}{2R}\sum \mathbb{1}[d_i(x) < R]\right)\left(\beta_0 + \frac{\beta}{2R}\sum \mathbb{1}[d_i(y) < R]\right)\right] \\ &= \beta_0^2 + 2\beta_0\beta\lambda + \frac{\beta^2}{4R^2}\mathbb{E}\left[\sum_i \mathbb{1}[d_i(x) < R]\sum_i \mathbb{1}[d_i(y) < R]\right].\end{aligned}$$

Middelværdien i det sidste led i ovenstående kan omskrives til

$$\mathbb{E}[(x_1 + x_2)(x_1 + x_3)],$$

hvor x_1 betegner antallet af linjer, der både skærer indenfor R af x og indenfor R af y, x_2 betegner antallet af linjer, der kun skærer indenfor R af x og x_3 betegner antallet af linjer, der kun skærer indenfor R af y.

Fra (12.2) er x_1 Poisson-fordelt med parameter λ_1. Ligeledes er x_2 og x_3 Poisson-fordelte med parametrene som benævnes hhv. λ_2 og λ_3. Af stationaritet følger det, at $\lambda_2 = \lambda_3$, og $x_1 + x_2$ er antallet af punkter indenfor afstand R af x. Dermed fås det, at $\lambda_1 + \lambda_2 = 2R\lambda$, og

$$\begin{aligned}\mathbb{E}[(x_1 + x_2)(x_1 + x_3)] =& \lambda_1 + \lambda_1^2 + 2\lambda_1\lambda_2 + \lambda_2^2 = \lambda_1 + (\lambda_1 + \lambda_2)^2 \\ =& \lambda_1 + (2R\lambda)^2,\end{aligned}$$

da $\mathbb{E}[x_1^2] = \lambda_1 + \lambda_1^2$ for Poisson-fordelingen. Indsættes dette i ovenstående sammen med (12.1) er sætningen bevist. □

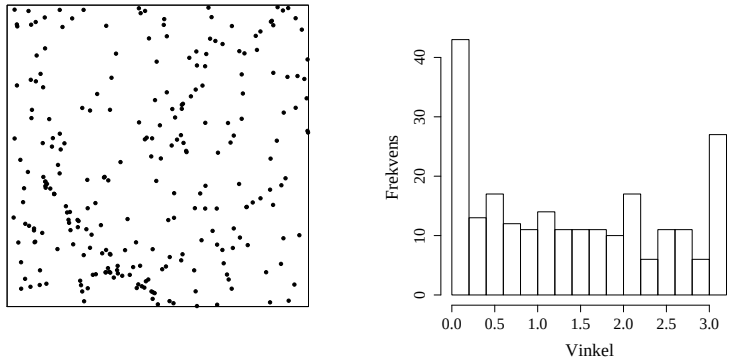
Der er nu opstillet en parkorrelationsfunktion, der inkluderer de forskellige parametre fra modellen. Det vil sige, at det nu er muligt at beregne en parkorrelationsfunktion for en model givet et sæt af parametre. En anden mulig summary statistics er fordelingen af vinklerne.

12.1.1 Summary statistics ved vinkler

En mulig summary statistic for et punktmønster med linjetendenser er at beskrive vinklerne mellem linjestykkerne fra et givent punkt til dets to nærmeste naboer. Indeholder punktmønstret tilstrækkeligt mange linjetendenser, vil det forventes, at der er en overvægt af vinkler nær nul og vinkler nær π. Dette ses ved at betragte tre punkter, der ligger næsten på en linje. Her vil der fås to vinkler nær nul og en vinkel nær π. Hvis punkterne i et datasæt derimod ligger tilfældigt, vil vinklerne være mere ligefordelte.

På Figur 12.4(a) ses et datasæt indeholdende tydelige linjetendenser og på Figur 12.4(b) ses den tilsvarende fordeling af vinklerne. Her ses det, at hovedvægten af vinklerne fordeles omkring nul og π, hvilket indikerer linjetendenser i datasættet. De øvrige vinkler kommer af knap så tydelige linjetendenser.

I `R` benyttes funktionen `nndist`$(\cdot)$ til at finde afstanden mellem de nærmeste naboer og funktionen `nnwhich`$(\cdot)$ til at afgøre hvilket nabopunkt afstanden findes til, hvorved vinklen kan opnås. Denne summary statistics tager i det beregningsmæssige ikke hensyn til eventuelle kanteffekter. Da det er den nærmeste naboafstand, der betragtes ved beregning af vinklerne, kan kanteffekter undgås ved at vælge observationsvinduet en smule mindre end det faktiske observationsvindue.



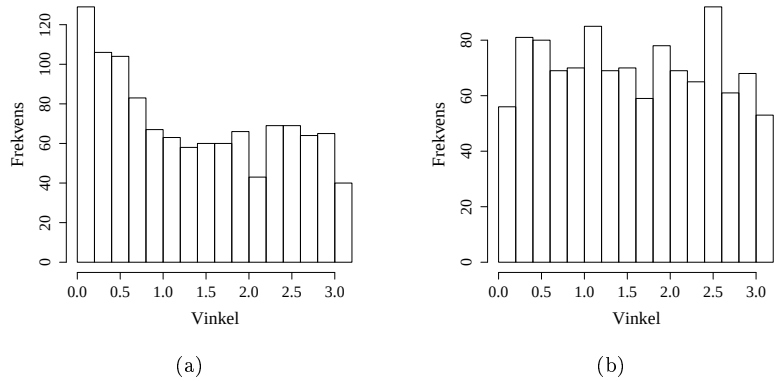
(a) (b)

Figur 12.4: (a) Realisering af punktproces med linjetendenser. (b) Vinkelfordelingen for (a).

Denne type summary statistics kan benyttes til at kontrollere, hvorvidt punkter i et givet datasæt ligger som linjer af punkter efterfulgt af hinanden. Hvis linjetendenserne optræder som brede bånd af punkter, er linjetendenserne i vinkelfordelingen knap så tydelig som på Figur 12.4(b).

På Figur 12.5(a) ses et plot at vinkelfordelingen mellem et punkt og de to nærmeste nabopunkter fra datasættet gravhøje. Det ses på plottet, at datasættet indholder linjetendenser, da der er en højere repræsentation af vinkler omkring 0 end ved de

andre vinkler. Dog er der ikke en højere repræsentation af vinkler nær π. Der er ikke nogen tydelig indikation af, at punkterne skulle ligge langs linjer, men dette kan som sagt skyldes, at linjetendenserne i datasættet optræder som bredere bånd jf. Figur 9.2.



Figur 12.5: (a) Vinkelfordelingen for datasættet for gravhøje. (b) Vinkelfordelingen for en homogen Poisson-proces med $\rho = 5.1 \cdot 10^{-6}$ på $[-7500, 7500] \times [-7500, 7500]$.

På Figur 12.5(b) ses et plot af vinkelfordelingen for en homogen Poisson-proces med intensitet $\rho = 5.1 \cdot 10^{-6}$ på $[-7500, 7500] \times [-7500, 7500]$. Det ses, at vinklerne mellem et punkt og dets nabopunkter tilnærmelsesvist er ligefordelt, og at der for en homogen Poisson-proces ikke er tegn på linjetendenser.

Selvom det ved denne summary statistics kan være svært, at vurdere hvor godt en punktproces stemmer overens med data, benyttes den alligevel senere til sammenligning af realiseringer af modellen fra det foregående og datasættet.

12.2 Modeltjek af model fra Afsnit 11.3

I dette afsnit udføres et modeltjek af modellen fra Afsnit 11.3. Som før beskrevet er der flere forskellige måder at udføre et modeltjek for en given model på. I dette afsnit udføres modeltjekket ved først at sammenligne datasættet og en realisering af modellen visuelt. Dernæst sammenlignes en estimeret parkorrelationsfunktion fra datasættet med envelopes fra modellen og efterfølgende sammenlignes vinkelfordelingerne, som blev præsenteret i Afsnit 12.1.1.

Der udføres to modeltjek. Et modeltjek af modellen, hvor linjerne er de estimerede linjer fra BDM-algoritmen og et modeltjek, hvor linjerne udgør en Poisson-linjeproces med den opnåede intensitet λ. For at udføre modeltjekket skal der, trækkes parametre $(\lambda, \beta, \beta_0, R)$ fra posterior-fordelingerne, der blev fundet i Afsnit 11.3.

Som det tidligere er set er parametrene afhængige af hinanden, da f.eks. forholdsvis store R, β og λ medfører, at β_0 er lille. Derfor bør der trækkes et sæt af sammenhørende parameterværdier. Det vil sige, at alle parametre der trækkes er fra samme iteration. Til modeltjek trækkes der 39 sæt af sammenhørende parametre.

Parametrene til realiseringer af modellen udtrækkes til

$$\lambda = 2.33 \cdot 10^{-3}, \quad \beta = 1.68 \cdot 10^{-3}, \quad \beta_0 = 1.65 \cdot 10^{-8}, \quad R = 258.29, \quad (12.3)$$

og sættet af parametre $(\lambda, \beta, \beta_0, R)$ er gennemgående i begge modeltjek.

12.2.1 Modeltjek med linjer fra BDM-algoritmen

Det første modeltjek er for modellen med de 61 opnåede linjer fra BDM-algoritmen.

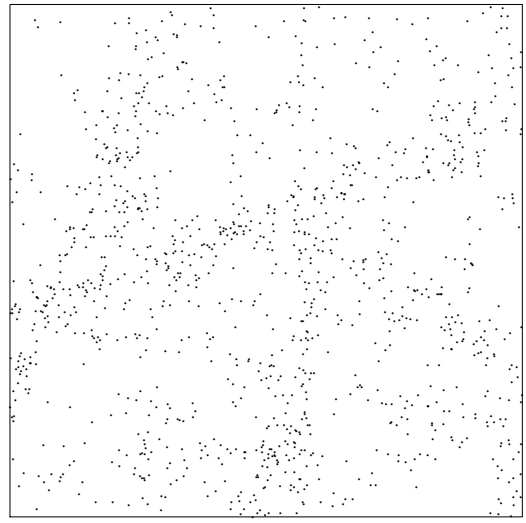
Visuel sammenligning

På Figur 12.6 ses et plot af en realisering af modellen fra Afsnit 11.3 med fastholdte linjer og ovenstående sæt af parametre. På plottet ses 1193 punkter. Plottet er lavet ved først at simulere baggrundsprocessen ved en Poisson-proces med intensitet β_0. Derefter simuleres der for hver linje en Poisson-proces på observationsvinduet, som så udtyndes jf. Definition 3.10 ved at fjerne alle punkter, der er længere væk end afstanden R fra linjen. Sluttelig laves en superposition jf. Definition 3.9 af baggrundsprocessen og de enkelte Poisson-processer om linjerne. På denne måde undgås det at få punkter, der ligger udenfor observationsvinduet.

Det ses, at realiseringen af modellen stemmer nogenlunde overens med Figur 9.2. I realiseringen af modellen er der områder, hvor intensiteten af punkter er lavere end andre steder, hvilket er i overensstemmelse med plottet af gravhøjene. På plottet af realiseringen af modellen ses linjetendenser, men de fremkommer dog ikke i lige så tydelig grad som på plottet for gravhøje. På plottet af gravhøjene er der desuden områder med mindre clustertendenser, hvilket ikke ses på realiseringen af modellen, da dette ikke er medtaget i modelopstillingen.

Det kan være svært at vurdere udfra Figur 12.6, om modellen stemmer overens med data. En mulig indikation på om modellen kunne beskrive data er ved at se på antallet af punkter i realiseringerne af modellen. Hvis der laves 39 realiseringer af modellen og antallet af punkter beregnes fås det, at den nedre-envelope for antallet af punkter er 1010 og den øvre-envelope er 1306. Da datasættet gravhøje indeholder 1147 punkter, ses det, at antal punkter i datasættet ligger indenfor grænserne af de to envelopes. Dette er også, hvad der ville forventes, da modellen er opstillet på baggrund af datasættet for gravhøjene.

På Figur 12.6 ses det, at punkterne ligger nogenlunde i de samme områder som på datasættet, dvs. linjetendenserne er de samme steder som ved datasættet. Dette er i overensstemmelse med modellen, da linjerne er fastlagt ved BDM-algoritmen, og derfor ligger de steder, hvor der er linjetendenser i datasættet. I datasættet for gravhøje ses det også, at linjetendensernes bredde varierer, hvorimod realiseringen



Figur 12.6: Realisation af model for gravhøje med 61 fastholdte linjer fra BDM-algoritmen og det valgte parametersæt.

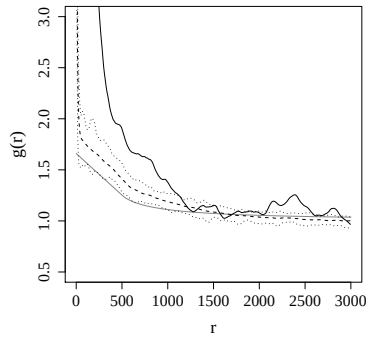
af modellen er for et konstant R. Den visuelle vurdering af modellen må dog antages at være i overensstemmelse med datasættet.

Parkorrelationsfunktion

På Figur 12.7 ses et plot af parkorrelationsfunktionen. Den fuldtoptrukne linje er estimatet fra datasættet. Linjerne beskrevet ved $(\cdots)$ er envelopes for modellen, beregnet ud fra 39 realiseringer med 39 forskellige sæt af parametre, og (- - -) er middelværdien af de 39 realiseringer af modellen. Den grå linje betegner den teoretiske værdi af parkorrelationsfunktionen for modellen. Den teoretiske parkorrelationsfunktion fra Sætning 12.3 er beregnet udfra gennemsnittet af de enkelte parameterværdier, som kan aflæses i Tabel 11.1.

Som før beskrevet udviser datasættet tiltrækning eller clustertendenser for små r, hvilket kommer til udtryk ved parkorrelationsfunktionen estimeret fra data. Modellen beskriver ikke datasættet i tilstrækkelig grad, da den estimerede parkorrelationsfunktion fra datasættet ikke ligger indenfor de to envelopes fra modellen. Modellen ses dog at beskrive tiltrækning eller clustertendenser, bare ikke samme type som datasættet udviser.

Denne uoverenstemmelse mellem model og data kan skyldes, at punkterne fordeles som brede bånd i modellen fremfor tyndere bånd som det antydes på plots af datasættet. En anden årsag kan være, at modellen ikke indeholder de mindre clustertendenser, som der ses i datasættet. Det ses, at den teoretisk beregnede parkorrelationsfunktion visse steder ligger udenfor envelopes for modellen. Dette



Figur 12.7: Parkorrelationsfunktionen. Den fuldtoptrukne linje er estimatet fra datasættet gravhøje. Linjerne beskrevet ved $(\cdots)$ er envelopes for modellen, og (- - -) er middelværdien af de 39 simuleringer. Den grå linje betegner den teoretiske værdi af parkorrelationsfunktionen.

kan skyldes, at linjerne i dette modeltjek er fastholdte og opnået ved benyttelse af BDM-algoritmen, og der er derfor ikke så meget variation i de 39 realiseringer.

Sammenligning af vinkler

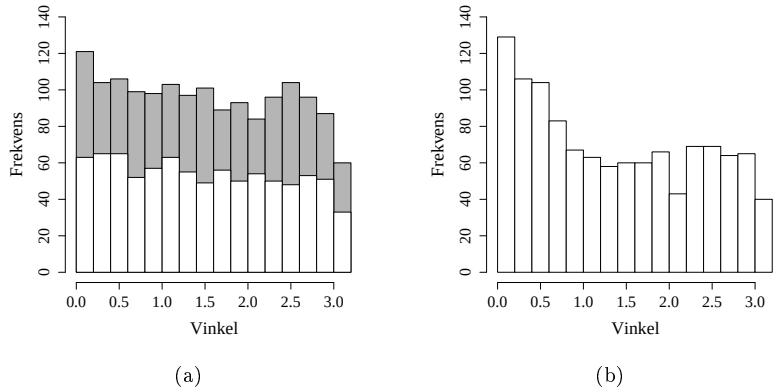
På Figur 12.8(a) ses et plot af envelopes for vinkelfordelingen. Figuren er lavet ved at lave 39 realiseringer af modellen, hvor vinklerne mellem linjestykkerne for et givet punkt og dets to nærmeste naboer er beregnet. Derefter er den største og mindste frekvens valgt for hver af intervallerne, hvilket netop udgør den øvre- og nedre-envelope. De forskellige envelopes skal ses som top og bund af det grå område på figuren.

På Figur 12.8(b) ses et histogram over vinkelfordelingen mellem et punkt i datasættet gravhøje og dets to nærmeste naboer (Samme plot som Figur 12.5(a)).

Ved at sammenligne Figur 12.8(a) og (b) ses det, at datasættet ligger indenfor envelopes ved størstedelen af intervallerne, dog er datasættet ikke indenfor envelopes i det første interval, dvs. intervallet $[0, 0.2)$, hvilket betyder, at modellen ikke kan beskrive de linjetendenser, der forekommer i datasættet. At modellen ikke indikerer linjetendenser i vinkelfordelingen, kan igen skyldes, at linjetenderserne i realiseringer af modellen fremkommer som brede bånd af punkter, og at der er flere linjer oveni hinanden, fremfor at punkterne skulle ligge i smallere bånd.

12.2.2 Modeltjek med linjer fra Poisson-linjeproces

I dette afsnit udføres et modeltjek med linjer fra en Poisson-linjeproces med den estimerede intensitet λ. Modeltjekket udføres på samme måde som før, dvs. visuelt, ved parkorrelationsfunktionen og ved fordelingen af vinklerne, blot med den



Figur 12.8: (a). Grænserne af det grå område er svarende til øverste og nederst envelope for vinkelfordelingen af modellen. (b) Vinkelfordelingen af datasættet gravhøje (Samme plot som Figur 12.5(a)).

ændring at linjerne bestemmes udfra parameteren λ. Sættet af parametre, der benyttes i realiseringen, er fra (12.3).

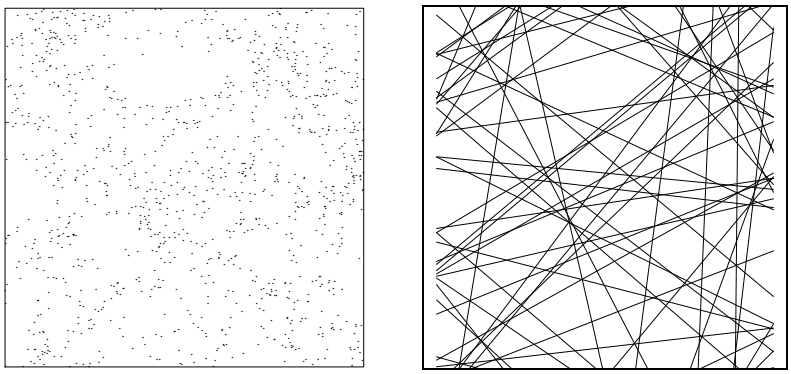
Visuel sammenligning

På Figur 12.9(a) ses et plot af en realisering af modellen. På figuren ses 1133 punkter, der stammer fra baggrundprocessen og fra fordelingen af punkter langs 63 linjer, der er plottet på Figur 12.9(b).

Ved sammenligning af datasættet gravhøje fra Figur 9.2 og Figur 12.9(a) ses det, at der er en nogenlunde overenstemmelse mellem realisationen af modellen og datasættet gravhøje. Ved sammenligning af Figur 12.6 og Figur 12.9(a) ses det, at linjetendenserne ikke er nær så fremherskende på Figur 12.9(a) som på Figur 12.6. Dette skyldes, at på Figur 12.6 er linjerne tilpasset datasættet ved benyttelse af BDM-algoritmen, og dermed ligger linjerne tættere. Hvorimod at på Figur 12.9(a) er linjerne tilfældige spredt udover hele observationsvinduet, da de følger en Poisson-linjeproces med konstant intensitet, som ses på Figur 12.9(b).

Ved at lave 39 simuleringer af modellen, fås det, at den nedre-envelope for antallet af punkter er 643 og den øvre-envelope er 1440. Hvilket betyder, at da datasættet gravhøje indeholder 1147 punkter, ligger dette indenfor envelopes for modellen. Ved sammenligning af antallet af punkter med modellen for de fastholdte linjer fra Afsnit 12.2.1, ses det, at der i denne model er en større variation i antallet af punkter. Dette skyldes, at linjerne ikke ligger fast, og dermed kan give en større variation i de 39 realiseringerne. Nedre- og øvre-envelope for antallet af linjer fra modellen er hhv. 46 og 84, men da der ikke haves noget linjedata for datasættet

gravhøje, er det ikke muligt at afgøre, hvor godt modellen beskriver data udfra dette.



(a) (b)

Figur 12.9: (a) Realisering af model for gravhøje med linjer fra Poisson-linjeproces. (b) Realisering af Poisson-linjeprocessen, der ligger bag plottet (a).

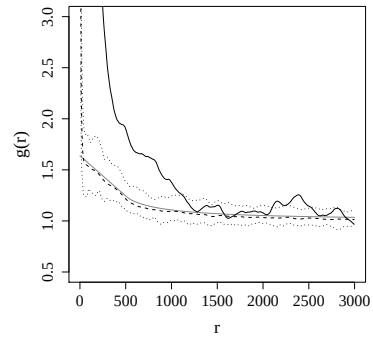
Parkorrelationsfunktion

På Figur 12.10 ses et plot af parkorrelationsfunktionen. Den fuldtoptrukne linje er estimeret ud fra datasættet, mens envelopes, markeret med $(\cdots)$ er beregnet udfra de 39 realiseringer af modellen. Det ses på figuren, at endnu engang ligger den estimerede parkorrelationsfunktion ikke indenfor envelopes fra modellen. Dog ses det, at modellen er i stand til at beskrive clustertendenser for små r, men dog ikke i tilstrækkelig grad til at kunne beskrive datasættet gravhøje.

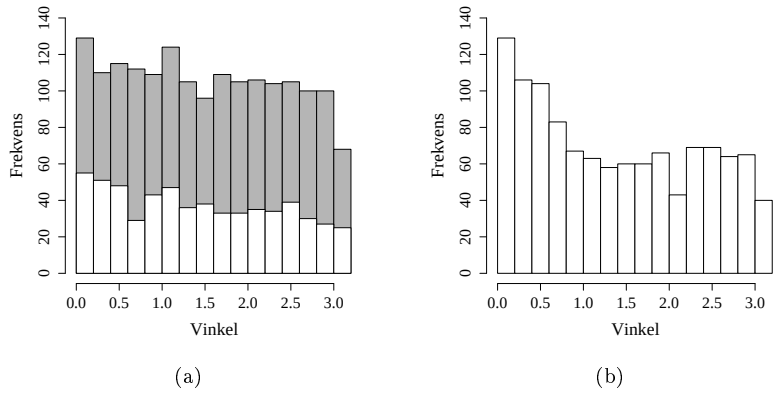
Ved sammeligning af envelopes fra Figur 12.7 og parkorrelationsfunktionen for denne model, ses det at envelopes for denne model er bredere end for Figur 12.7. Dette skyldes, ligesom ved de bredere envelopes for antallet af punkter, at linjerne i dette tilfælde ikke ligger fast.

Vinkelfordelingen

På Figur 12.11(a) ses et plot af envelopes for vinkelfordelingen af 39 simuleringer fra datasættet. Den nedre- og øvre-envelope afgrænses af det grå område på figuren. Ved sammenligning af de to figurer på Figur 12.11 ses det igen, at den opstillede model ikke er i stand til at beskrive den helt samme type af linjetendenser, som datasættet gravhøje indeholder. Igen ses det, at envelopes for vinkelfordelingen ligger højere og lavere end i modellen med de fastholdte linjer.



Figur 12.10: Parkorrelationsfunktionen. Den fuldtoptrukne linje er estimatet fra datasættet gravhøje. Linjerne beskrevet ved $(\cdots)$ er envelopes for modellen og (- - -) er den forventede værdi af de 39 simuleringer. Den grå linje betegner den teoretiske værdi af parkorrelationsfunktionen.



Figur 12.11: (a). Grænserne af det grå område er svarende til øverste og nederst envelope for vinkelfordelingen af modellen. (b) Vinkelfordelingen af datasættet.

12.2.3 Opsummering af modeltjek

Der har nu været foretaget et modeltjek af modellen fra Afsnit 11.3 både med de linjer, der var opnået fra BDM-algoritmen, og med linjer genereret som en realisering af en Poisson-linjeproces med intensitet λ.

Realiseringen af modellen med de estimerede linjer viste sig at ligne punktmønstret mest. Det skyldes, at linjerne er tilpasset vha. BDM-algoritmen, og dermed ligger de steder, hvor linjetendenserne også forekommer i datasættet. I modellen med

det varierende antal linjer, følger linjerne en Poisson-linjeproces og gennemskærer derfor området mere tilfældigt. Ved sammenligning af summary statistics viste begge modeller ikke de store overensstemmelser med datasættet.

Eftersom modellen med de estimerede linjer fra BDM-algoritmen umiddelbart beskriver datasættets linjetendenser bedst, må det konkluderes, at det er de estimerede linjer, der skal benyttes sammen med de estimerede parametre til beskrivelse af datasættet. Dette gælder specielt, når de enkelte linjer sammenlignes med forhistoriske veje. Når dette gøres antages modellen at være acceptabel.

Del III

Epilog

Afrunding

Der er nu opstillet en model til beskrivelse af linjetendenserne i datasættet gravhøje. For at kunne gøre dette har det været nødvendigt med overvejelser omkring punktprocesser og linjeprocesser, herunder summary statistics og simulering. Til opstillingen af modellen har teorien om punktprocesser og linjeprocesser været benyttet til at lave linjeprocesser, baggrundsprocesser og punktprocesser langs linjer.

13.1 Konklusion

Hovedformålet med denne rapport er at opstille modeller til beskrivelse af datasættet gravhøje. I første omgang forsøges det at tilpasse traditionelle punktprocesser i form af Poisson-, Thomas- og Matérn-processen til datasættet. Efter et modeltjek er det klart, at de forskellige punktprocesser ikke er i stand til at beskrive datasættet i tilstrækkelig grad, eftersom linjetendenserne fra datasættet ikke er gengivet ved realiseringer af modellen.

For at kunne beskrive datasættet er det nødvendigt at inkludere mekanismer, der kan producere linjetendenser. I indledningen er forskellige teorier om, at gravhøje angiveligt skulle være placeret langs forhistoriske veje beskrevet. Det vides ikke med sikkerhed, hvor disse veje præcist har gået, men placeringen af gravhøjene er kendt. Vejteorierne er benyttet i forbindelse med opstillingen af en model til beskrivelse af linjetendenserne af gravhøjene. Modellen er opstillet under antagelse om, at punkterne fordeles langs rette linjer, hvor disse linjer forestilles som forhistoriske veje. Dette er en lidt grov antagelse, da linjetendenserne i datasættet tydeligt ses at være buede som på Figur 9.2.

De rette linjer antages i modellen at følge en Poisson-linjeproces med intensitet λ. Omkring hver linje er der antaget at være en Poisson-proces med intensiteten $\beta f(\cdot)$, hvor $f(\cdot)$ er en ligefordelt tæthed på $[-R, R]$. I modellen beskrives punkter, der ikke har nogen direkte forbindelse med linjerne ved en Poisson-proces med intensiteten β_0. Med de antagelser er det muligt at opstille en likelihoodfunktion for parametrene, der er givet ved (11.9)

$$L(\lambda, \beta, \beta_0, R|x, y) = \left(\frac{\lambda}{\pi}\right)^n \exp\left(-\frac{2\lambda d}{\sqrt{2}}\right) \exp\left(-d^2\beta_0 - \beta \sum_{i=1}^{n} l_i\right) \\ \times \prod_{j=1}^{m} \left(\beta_0 + \frac{\beta}{2R} \sum_{i=1}^{n} \mathbb{1}[|p_i - x_1 \cos(\theta_i) - x_2 \sin(\theta_i)| < R]\right),$$

hvor x og y betegner hhv. gravhøje og linjer.

De fire parametre λ, β, β_0 og R i likelihoodfunktionen er ukendte, og det samme er placeringen og antallet af linjer, dvs. $y = (p, \theta)$. Ved implementering af en BDM-

algoritme i R, som beskrevet i Algoritme 6.3 og implementeret i Appendiks B, er det muligt at estimere parametrene og linjerne udfra datasættet.

Ved benyttelse af den implementerede BDM-algoritme estimeres antallet af linjer, der gennemskærer observationsvinduet til ca. 60. Dette er dog ikke et udtryk for, at der er ca. 60 veje, da der går flere linjer til beskrivelse af en enkelt vej. Dette skyldes, at modellen forsøger at beskrive de buede linjetendenser med flere rette linjer. På Figur 13.1(a) ses de estimerede linjer fra datasættet, sammen med gravhøjene. Det er antagelsen om, at linjetendenserne kan beskrives ved rette linjer, der er skyld i, at antallet af linjer er højt. Desuden ses det på figuren, at visse linjer ikke ligger langs linjetendenserne i datasættet, men nærmere er placeret for at beskrive vilkårlige punkter, som i princippet kunne være beskrevet ved en højere intensitet af baggrundsprocessen.

Gravhøjene er oprettet over en periode på 4000 år, men betragtes alle som samtidige. Det kan forestilles, at befolkningen i området har lavet nye veje og nedlagt gamle veje i et vist omfang i denne lange tidsperiode. Det vil sige, at antallet af veje, der går gennem landskabet, ikke beskriver antallet af veje til et bestemt tidspunkt, men over hele tidsperioden, hvor der blev oprettet gravhøje. Dette kan være en yderligere årsag til de mange linjer, men det vil også kunne være skyld i, at der estimeres veje i områder, hvor der aldrig har eksisteret veje.

Da parametrene λ, β og β_0 er intensiteter for Poisson-linjeprocesser og Poisson-processer er disse kun interessante i forbindelse med plots af modellen. Den parameter, der er lettest at fortolke, er R, da denne angiver den vinkelrette afstand som gravhøjene maksimalt kan ligge fra linjerne. Parameteren R er estimeret til at ligge omkring 260m, svarende til at urfolket, der har benyttet vejene, har oprettet gravhøje i en maksimal afstand af ca. 260m fra vejen. Denne afstand kan virke stor i forhold til, at en gravhøj i denne afstand skulle have en relation til vejene. Da vejene sandsynligvis har ligget i de områder, hvor befolkningen boede, og hvor de begravede de døde, kan det godt tænkes at parameteren R er stor, da befolkningen kun boede i nærheden af vejene og ikke på vejen. Desuden kan det store antal af linjer, der ligger ovenpå hinanden, have betydning for parameteren R, og derfor skal fortolkningen af R som maksimumsafstand tages med forbehold.

Ved realiseringer af modellen med de estimerede parametre og estimerede linjer er det muligt at opnå punktmønstre, der tilnærmelsesvist indeholder de samme linjetendenser som datasættet gravhøje, som det ses på Figur 12.6. Selvom linjetendenserne synes at være beskrevet i tilstrækkelig grad ved realiseringer, ses det på plot af parkorrelationsfunktionen og vinkelfordelingen, at der ikke er helt samme type af clustertendenser mellem punkter fra modellen i forhold til datasættet. Dog ses der på plots af parkorrelationen, at modellen til en vis grad beskriver samme type vekselvirkning som data. Overordnet vurderes modellen at være acceptabel.

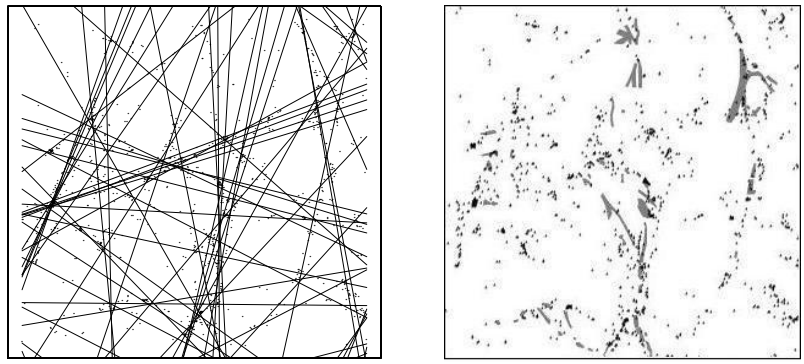
En anden tilgangsvinkel til modellen er at lade placeringen og antallet af linjer være stokastisk ved benyttelse af den estimerede λ til at skabe linjerne, fremfor at benytte de estimerede linjer fra BDM-algoritmen. Ved linjerne fra BDM-algoritmen, hvor flere linjer anvendes til beskrivelse af en enkelt linjetendens, vil dette medføre, at λ er for stor i forhold til det egentlige antal linjetendenser i

datasættet. Ved realisering af en Poisson-linjeproces med denne λ, vil dette betyde mange linjer, der ligger spredt over hele observationsvinduet, hvilket udvisker linjetendenserne. Til at beskrive datasættet kan det dermed konkluderes, at de estimere linjer fra BDM-algoritmen skal bruges for at få passende realiseringer.

En generel mangel ved modellen er, at den ikke er i stand til at gengive de mindre clustertendenser, der er ved linjetendenserne i datasættet. Dette kommer dog ikke som nogen overraskelse, da modellen ikke inddrager disse clustertendenser, men kun fokuserer på linjetendenserne.

Da det ved at opstille en model, der inkluderer linjetendenser, er muligt at beskrive datasætttet til en vis grad, kan det konkluderes at gravhøjene følger linjetendenser. Det vides selvfølgelig ikke med sikkerhed om, gravhøjene følger veje, eller hvilke andre, f.eks. religøse eller kulturelle årsager, der skyldes, at gravhøjene ligger langs linjer. For at undersøge om linjetendenserne stammer fra veje, kan de estimerede veje sammenlignes med de forhistoriske veje. Eftersom de estimerede veje ikke er særligt veldefinerede, er dette besværligt, og sammenligningen kan derfor kun foretages på et overordnet plan. Kun de steder, der beskrives ved flere linjer, betragtes som veje, og vejene antages at ligge mellem grupperne af de estimerede linjer. Ved sammenligning skal det også bemærkes, at det nogle steder kun er en halv linje, der beskriver en linjetendens.

På Figur 13.1(a) ses de estimerede linjer og på (b) ses fund af forhistoriske veje inden for området. Ved sammenligning ses det, at mange af de fundne forhistoriske veje ligger i samme områder som de estimerede veje. Generelt ses det, at de steder der er fundet flest rester fra forhistoriske veje, er i de områder, hvor modellen anslår, at der skulle have ligget veje.



(a) (b)

Figur 13.1: Veje indenfor observationsområde. (a) Estimerede linjer (Samme plot som Figur 11.9(b)). (b) Forhistoriske veje. Scan fra [Johansen, 2009].

Overordnet kan det konkluderes, at den opstillede model er i stand til at beskrive datasættet for gravhøje til en vis grad. Dette vurderes udfra realiseringer, model-

tjek og sammenligning af veje. Afhængig af hvor præcis information, der ønskes fra modellen, er modellen derfor acceptabel. Med hensyn til teorien om gravhøje langs veje kan det konkluderes, at dette er en sandsynlig teori. Modellen giver dog ikke præcise informationer om, hvor vejene kunne have ligget, men den giver indikationer af i hvilke områder, der kunne have været veje.

Målet for projektet er hermed opnået, da problemformuleringen er blevet behandlet, og de opstillede delmål er undervejs blevet opfyldt. Modellen levner dog plads til forbedringer, hovedsageligt mht. beskrivelse af de buede veje og mindre clustertendenser, hvilket beskrives i følgende perspektivering.

13.2 Perspektivering

I det foregående blev det vist, at det ved benyttelse af modellen, der inkluderer linjetendenser, var muligt at beskrive datasættet gravhøje. I dette afsnit introduceres hvilke muligheder, der er i modellen og hvilke muligheder, der er i forbindelse med udvidelse af modellen.

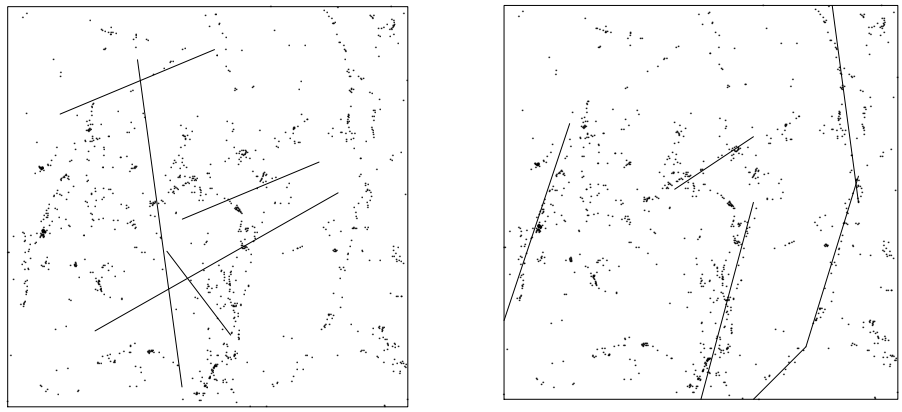
Modellen tilpasser linjetendenserne i datasættet gravhøje ved rette linjer, der gennemskærer hele observationsvinduet. De enkelte linjetendenser i datasættet beskrives ved flere rette linjer, da linjetendenserne er buede. I stedet for at linjetendenserne beskrives ved rette linjer, kan det forestilles, at linjetendenserne beskrives bedre af mindre linjestykker. En enkelt linjetendens i datasætet vil da kunne bestå af en sammensætning af linjestykker med forskellig orientering og længde.

Den opstillede model er i stand til at danne linjetendenser, der består af mindre linjestykker fremfor hele linjer. Dette kan gøres ved at dele observationsvinduet op i mindre observationsvinduer, indenfor hvilke linjetendenserne er tilnærmelsesvist rette. Derefter kan disse mindre vinduer betragtes som nye observationsvinduer. Ved benyttelse af BDM-algoritmen er det muligt at tilpasse linjestykker til gravhøjene i hver af de små observationsvinduer og estimere de forskellige parametre. Ved at sammensætte de forskellige små observationsvinduer til det store observationsvindue opnås en beskrivelse af datasættet gravhøje med linjetendenser.

Der virker umiddelbart som en bedre beskrivelse af linjetendenserne at benytte linjestykker i stedet for hele linjer, da linjestykkerne ved sammensætning ville kunne beskrive buede veje. Det bør dog bemærkes at ved sammensætningen af de små observationsvinduer, kan der opstå blinde veje, da linjetendenserne indenfor det lille observationsvindue stadig er beskrevet af rette linjer.

En anden måde at beskrive de buede linjetendenser på er ved en type processer kaldet linjestykkeprocesser, som kort er omtalt i [Barndorff-Nielsen et al., 1999]. I disse processer er linjerne givet ved linjestykket mellem to punkter i $\mathbb{R}^2$, og linjestykkerne kan eventuelt være sammenhængende. På Figur 13.2(a) ses tilfældige linjestykker, der er fordelt over datasættet gravhøje. Som for linjeprocesserne er det eventuelt muligt at lave en BDM-algoritme, der kan tilpasse disse linjestykker

til linjetendenserne. Dermed kan der skabes et billede som Figur 13.2(b). Dette billede er dog skabt ved manuelt at tilpasse linjestykker til datasættet gravhøje, og dermed er linjestykkerne placeret udfra en personlig vurdering af linjetendenserne. Men det ses, hvordan det er muligt at tillade, at vejene skifter retning.



(a) (b)

Figur 13.2: (a) Tilfældige linjestykker på datasættet. (b) Linjestykker der ligger langs linjetendenserne.

I konklusionen er det beskrevet, at da alle gravhøje bliver betragtet som identiske, kan modellen kun beskrive linjetendenser over ca. 4000 år og ikke kortere tidsperioder. Hvis vejene til en kortere tidsperiode ønskes bestemt, kan gravhøjene inddeles i tidsperioder, der angiver, hvornår gravhøjene blev oprettet. Ved at tilføje et mærke til de gravhøje, der tilhører en bestemt tidsperiode, er det muligt at betragte de mærkede gravhøje som en mærket punktproces. Dermed er det muligt at beskrive gravhøje for en bestemt tidsperiode ved en punktproces, hvorved det også er muligt at beskrive linjetendenserne for gravhøjene hørende til denne bestemte tidsperiode.

Ved at betragte parkorrelationsfunktionen estimeret fra datasættet og modellen fås det, at vekselvirkningen mellem punkterne ikke er den samme for modellen som for datasættet. Eksempelvis er det ikke muligt for modellen at genskabe den samme type af clustertendens i forbindelse med linjetendenserne som for datasættet. En måde at opnå en anden vekselvirkning mellem punkterne er ved at ændre den måde, hvorpå punkterne fordeles vinkelret på linjerne. En mulig måde at gøre dette på er ved at beskrive tætheden vinkelret på linjerne ved anden fordeling end ligefordelingen. Dette kunne f.eks. være en endimensionel normalfordeling vinkelret på linjerne, hvilket ville betyde at intensiteten af punkterne ville være større inde omkring selve linjen og aftagende vinkelret væk fra linjen. Dette ville formentlig kunne betyde en anden interaktion mellem punkter, og der ville måske kunne opnås en parkorrelationsfunktion for modellen, der er i bedre overenstemmelse med

data. En anden tæthedsfunktion vil dog ikke medføre de mindre clustertendenser. Hvis disse skal beskrives, er det nødvendigt at inddrage en clusterproces som eks. Thomas- eller Matérn-processen fra Kapitel 5.

I modellen beskrives linjetendenserne som identiske bånd af punkter i den forstand, at de er beskrevet ved den samme intensitet $\beta f(\cdot)$ for Poisson-processen, og at alle bånd er af bredde $2R$. På datasættet gravhøje ses det, at det ikke er alle linjetendenser, der indeholder det samme antal punkter eller er af samme bredde. For at kunne beskrive disse tendenser i datasætte kan der betragtes inhomogene punktprocesser, istedet for de homogene som modellen indeholder. Ved de inhomogene punktprocesser kan der medtages forskellige kovarianter som jordbundstype, højdeforskelle osv., som er beskrevet i indledningen, hvilket formentlig også ville kunne give en bedre beskrivelse af datasættet. Det kan være svært at inkludere disse kovariater, da disse skal oversættes fra linjerummet, hvor gravhøjene ligger, til repræsentationsrummet for linjerne, når modellen opstilles. Desuden kunne baggrundprocessen også beskrives ved en inhomogen Poisson-proces istedet for den homogene Poisson-proces som modellen indeholder.

Som det ses i det foregående, er der flere muligheder for at forbedre den opstillede model. Hvis nogle af de ovenstående forslag blev inddraget i modellen, ville en bedre model for datasættet måske kunne opnås.

Del IV

Appendiks

Matematiske definitioner

I dette appendiks er der angivet to definitioner, der er brugt i forbindelse med rapporten.

Definition A.1 (Konveks mængde)
Lad S være en mængde i $\mathbb{R}^n$. S er en konveks mængde, hvis der for alle $x_1, x_2 \in S$ gælder, at $\lambda x_1 + (1-\lambda)x_2 \in S$, for alle $\lambda \in [0,1]$.

Det vil sige, at to punkter indeholdt i en konveks mængde kan forbindes med en ret linje, der er indeholdt i den konvekse mængde.

Definition A.2 (Støtte)
For en funktion $f : \mathbb{R}^k \to \mathbb{R}$ defineres støtten til at være mængden

$$supp(f) = \overline{\{x \in \mathbb{R}^k | f(x) \neq 0\}},$$

hvor $\overline{\{\cdot\}}$ betyder at mængden er lukket. Støtten af en funktion er altså en mængde i $\mathbb{R}^k$, hvor udenfor funktionen er nul.

R-kode

Dette appendiks indeholder kildekoden, der er benyttet i rapporten. Kildekoden er skrevet i R [R Development Core Team, 2008]. Det er BDM-algoritmen, der er benyttet til estimering af parametre og linjer i Afsnit 11.3, som er gengivet her. Kommandoerne benyttet til at generere figurene igennem rapporten er forholdsvis simple og kan findes i `spatstat`-manualen [Baddeley, 2008]. Koden kan eventuelt bruges som inspiration til eget arbejde med punktmønstre med R. Undervejs i koden er der forklaringer, og for yderligere forklaringer af de forskellige funktioner, der benyttes i koden, kan kommandoen `help()` benyttes i R.

B.1 Model BDM

Dette afsnit indeholder R-koden til implementering af BDM-algoritmen fra Afsnit 11.3. Algoritmen følger fremgangenmåden i Algoritme 6.3. Sideløbendende med BDM-algoritmen opdateres parametrene med en Metropolis-algoritme.

```
##########################################################################
### BDM - algoritme  ######################################################
##########################################################################
rm(list=ls())                # clear all
m=100000                     # antal iterationer af algoritmen
library(spatstat)
barrows = read.table("barrowscoor.dat")
bar=subset(barrows, 460000<V1 & V1<475000 & 6175000<V2 & V2<6190000)
bb=as.matrix(bar)
b=matrix(nrow=1147, ncol=2)
b[,1]= bb[,1]-467500
b[,2]= bb[,2]-6182500

# Initialisering af konstanter
d=15000                                # bredde af vindue
k=1147                                 # antal punkter
W=owin(c(0,2*pi), c(0,d/sqrt(2)))      # observationsvindue på strippen
Y0=rpoispp(0.00003,win=W)              # det næsten-tomme punktmønster
n=Y0$n                             # antal linjer

# startværdier af parametre
lambda=0.006
beta=0.0015
beta0=0.5e-6
R=270

# Angivelse af hvordan parametrene skal gemmes
trace1=numeric(length=m)
trace2=numeric(length=m)
trace3=numeric(length=m)
trace4=numeric(length=m)
trace5=numeric(length=m)
```

```

# funktion til beregning af længden af en linje
li=function(theta,p){alpha= theta + pi/2
                    yf=function(x){tan(alpha)*x+p/sin(theta)}
                    xf=function(y){y/tan(alpha)-p/(sin(theta)*tan(alpha))}

                    #Skæringspunkter mellem x og observationsvinduet
                    x11=c(-d/2,yf(-d/2))
                    x12=c(d/2, yf(d/2))
                    x21=c(xf(-d/2),-d/2)
                    x22=c(xf(d/2),d/2)
                    k=c(x11,x12,x21,x22)

                    #Sortér x-værdier
                    mt=matrix(nrow=4,ncol=2)
                    mt[1,]=x11
                    mt[2,]=x12
                    mt[3,]=x21
                    mt[4,]=x22
                    sor=mt[order(mt[,1]),]

                    #fjern første og sidste række af matrix
                    sor1=sor[-1,]
                    sor2=sor1[-3,]

                    li= sqrt((sor2[1,]-sor2[2,])%*%(sor2[1,]-sor2[2,]))
                    return(li)
                    }

# udtryk der skal bruges til acceptsandsynligheder
a=0
aa=0
aaa=0
aaaa=0
a1=0
a2=0
a3=0
a4=0
a5=0
a6=0
a7=0
a8=0
a9=0

##########################################################################
####### Hastingsratioer#################################################
##########################################################################
lh.move=function(Y0,Y1){
                        sum1=matrix(nrow=k,ncol=Y1$n)
                        for(j in 1:k){sum1[j,]=c(abs(Y1$y
                                                 -b[j,1]*cos(Y1$x)
                                                 -b[j,2]*sin(Y1$x))<R)}
                        sum11=numeric(length=k)
                        for(w in 1:k){sum11[w]=sum(sum1[w,])}
                        sum11=beta/(2*R)*sum11
                        sum11=sum11+beta0

                        sum2=matrix(nrow=k,ncol=Y0$n)
```

```
                        for(j in 1:k){sum2[j,]=c(abs(Y0$y
                                                 -b[j,1]*cos(Y0$x)
                                                 -b[j,2]*sin(Y0$x))<R)}
                        sum22=numeric(length=k)
                        for(w in 1:k){sum22[w]=sum(sum2[w,])}
                        sum22=beta/(2*R)*sum22
                        sum22=sum22+beta0

                        r.ratio =(-beta*(sum(ll)-sum(l))
                                        + sum(log(sum11))
                                        - sum(log(sum22)))
                        return(r.ratio)
}


lh.fod=function(Y0,Y1){
                        sum3=matrix(nrow=k,ncol=Y1$n)
                        for(j in 1:k){sum3[j,]=c(abs(Y1$y
                                                 -b[j,1]*cos(Y1$x)
                                                 -b[j,2]*sin(Y1$x))<R)}
                        sum33=numeric(length=k)
                        for(w in 1:k){sum33[w]=sum(sum3[w,])}
                        sum33=beta/(2*R)*sum33
                        sum33=sum33+beta0

                        sum4=matrix(nrow=k,ncol=Y0$n)
                        for(j in 1:k){sum4[j,]=c(abs(Y0$y
                                                 -b[j,1]*cos(Y0$x)
                                                 -b[j,2]*sin(Y0$x))<R)}
                        sum44=numeric(length=k)
                        for(w in 1:k){sum44[w]=sum(sum4[w,])}
                        sum44=beta/(2*R)*sum44
                        sum44=sum44+beta0


                        rb = (log((1-pfod)/pfod)
                              +log((2*pi*d/sqrt(2))/(Y0$n+1))
                              +log((lambda)/(pi))
                              -beta*li(s1$x,s1$y)
                              + sum(log(sum33))
                              - sum(log(sum44)))
                        return(rb)
}

lh.dod=function(Y0,Y2){
                    sum5=matrix(nrow=k,ncol=Y2$n)
                    for(j in 1:k){sum5[j,]=c(abs(Y2$y
                                             -b[j,1]*cos(Y2$x)
                                             -b[j,2]*sin(Y2$x))<R)}
                    sum55=numeric(length=k)
                    for(w in 1:k){sum55[w]=sum(sum5[w,])}
                    sum55=beta/(2*R)*sum55
                    sum55=sum55+beta0

                    sum6=matrix(nrow=k,ncol=Y0$n)
                    for(j in 1:k){sum6[j,]=c(abs(Y0$y
                                             -b[j,1]*cos(Y0$x)
                                             -b[j,2]*sin(Y0$x))<R)}
```

```
                    sum66=numeric(length=k)
                    for(w in 1:k){sum66[w]=sum(sum6[w,])}
                    sum66=beta/(2*R)*sum66
                    sum66=sum66+beta0

                     rd = (log(pfod/(1-pfod))
                           +log(Y0$n/(2*pi*d/sqrt(2)))
                           +log(pi/(lambda))
                           +beta*li(Y0$x[s2],Y0$y[s2])
                           + sum(log(sum55))
                           - sum(log(sum66)))
                    return(rd)
}

lh.lambda=function(Y0){
                       LH.lambda=(log(lambda.prime)-log(lambda)
                                  -(lambda.prime-lambda)/0.1
                                  +Y0$n*log(lambda.prime/lambda)
                                  -2*d/sqrt(2)*(lambda.prime-lambda))
                       return(LH.lambda)
}

lh.beta=function(Y0){

         sum7=matrix(nrow=k,ncol=Y0$n)
         for(j in 1:k){sum7[j,]=c(abs(Y0$y
                                  -b[j,1]*cos(Y0$x)
                                  -b[j,2]*sin(Y0$x))<R)}
         sum77=numeric(length=k)
         for(w in 1:k){sum77[w]=sum(sum7[w,])}
         sum77=beta.prime/(2*R)*sum77
         sum77=sum77+beta0

         sum10=matrix(nrow=k,ncol=Y0$n)
         for(j in 1:k){sum10[j,]=c(abs(Y0$y
                                   -b[j,1]*cos(Y0$x)
                                   -b[j,2]*sin(Y0$x))<R)}
         sum100=numeric(length=k)
         for(w in 1:k){sum100[w]=sum(sum10[w,])}
         sum100=beta/(2*R)*sum100
         sum100=sum100+beta0

         LH.beta = (log(beta.prime)-log(beta)-(beta.prime-beta)/0.1
                   +(beta-beta.prime)*sum(l)
                   + sum(log(sum77))
                   - sum(log(sum100)))
         return(LH.beta)
}

lh.beta0=function(Y0){
         sum8=matrix(nrow=k,ncol=Y0$n)
         for(j in 1:k){sum8[j,]=c(abs(Y0$y
                                  -b[j,1]*cos(Y0$x)
                                  -b[j,2]*sin(Y0$x))<R)}
         sum88=numeric(length=k)
         for(w in 1:k){sum88[w]=sum(sum8[w,])}
         sum88=beta/(2*R)*sum88
         sum88=sum88+beta0.prime
```

```
        sum10=matrix(nrow=k,ncol=Y0$n)
        for(j in 1:k){sum10[j,]=c(abs(Y0$y
                                  -b[j,1]*cos(Y0$x)
                                  -b[j,2]*sin(Y0$x))<R)}
        sum100=numeric(length=k)
        for(w in 1:k){sum100[w]=sum(sum10[w,])}
        sum100=beta/(2*R)*sum100
        sum100=sum100+beta0

        LH.beta0=(log(beta0.prime)-log(beta0)-(beta0.prime-beta0)/0.1
                 +d^2*(beta0-beta0.prime)
                 + sum(log(sum88))
                 - sum(log(sum100)))
        return(LH.beta0)
}

lh.R=function(Y0){
        sum9=matrix(nrow=k,ncol=Y0$n)
        for(j in 1:k){sum9[j,]=c(abs(Y0$y
                                 -b[j,1]*cos(Y0$x)
                                 -b[j,2]*sin(Y0$x))<R.prime)}
        sum99=numeric(length=k)
        for(w in 1:k){sum99[w]=sum(sum9[w,])}
        sum99=beta/(2*R.prime)*sum99
        sum99=sum99+beta0

        sum10=matrix(nrow=k,ncol=Y0$n)
        for(j in 1:k){sum10[j,]=c(abs(Y0$y
                                  -b[j,1]*cos(Y0$x)
                                  -b[j,2]*sin(Y0$x))<R)}
        sum100=numeric(length=k)
        for(w in 1:k){sum100[w]=sum(sum10[w,])}
        sum100=beta/(2*R)*sum100
        sum100=sum100+beta0

        LH.R=(log(R.prime)-log(R)-(R.prime-R)/50
              +sum(log(sum99))
              - sum(log(sum100)))
        return(LH.R)
}

##########################################################################
###### Algoritmen  #######################################################
##########################################################################
for(i in 1:m){
qq=0.8             # til at finde ud af hvilken algoritme der skal køres
pfod=0.8           # sandsynlighed for fødsel

####
#1.#
####
u=runif(1,0,1)

####  ##############################################
#2.#  ##### MOVE
####  ##############################################
      if(u<qq){
```

```
                if(Y0$n==0){ # fordi et ikke-eksisterende punkt
                                # ikke kan flyttes
                                Y0=Y0
                                a1=a1+1}

                if(Y0$n>0){
                #1.
                Rm=runif(1,0,1)

                #2.
                Im=sample(1:Y0$n,1)
                s=runifpoint(1,W)
                Y1=Y0
                Y1$x[Im]=s$x
                Y1$y[Im]=s$y

                l=numeric(length=Y0$n)
                for(q in 1:Y0$n){l[q]=li(Y0$x[q],Y0$y[q])}

                ll=numeric(length=Y1$n)
                for(q in 1:Y1$n){ll[q]=li(Y1$x[q],Y1$y[q])}

                #3.
                if(log(Rm) < lh.move(Y0,Y1) ){
                                                Y0=Y1
                                                a2=a2+1
                                                 }

                if(log(Rm) > lh.move(Y0,Y1) ){
                                                Y0=Y0
                                                a3=a3+1
                                                }
                            }
                } # Her slutter MOVE

####   ##############################################
#3.#   ##### BIRTH-DEATH
####   ##############################################
      if(u>qq){
               #1.
               uu = runif(1,0,1)
               uuu = runif(1,0,1)

               #2.
               if(uu<pfod){##############################################
                           ##### fødsel
                           ##############################################
                           if(Y0$n==0){ # fordi der ikke kan beregnes
                                         # Hasting-forhold med nul linjer
                                           Y0=Y0
                                           a4=a4+1}

                           if(Y0$n>0){ #lh.fod(Y0,Y1)

                                          s1 = runifpoint(1, W)
                                          Y1=superimpose(Y0,s1)

                                  if(log(uuu)<lh.fod(Y0,Y1)){
```

```
                                                                Y0=Y1
                                                                a5=a5+1
                                                                }
                                        if(log(uuu)>lh.fod(Y0,Y1)){
                                                                 Y0=Y0
                                                                 a6=a6+1
                                                                 }
                                                }
                                } # Her slutter fødsel

        #3.
        if(uu>pfod){##############################################
                    ##### død
                    ##############################################
                    #a:
                    if(Y0$n<1.5){ # fordi der ikke kan beregnes
                                    # Hasting-forhold med nul linjer
                                    Y0=Y0
                                    a7=a7+1}
                    #b:
                    if(Y0$n>1.5){ #lh.dod(Y0,Y2)
                                    s2=sample(1:(Y0$n-1),1)
                                    Y2=Y0[-s2]

                                    if(log(uuu)<lh.dod(Y0,Y2)){
                                                                Y0=Y2
                                                                a8=a8+1
                                                                }
                                    if(log(uuu)>lh.dod(Y0,Y2)){
                                                                Y0=Y0
                                                                a9=a9+1
                                                                }
                                    }
                            } # Her slutter død
                    }# Her slutter BD

  ##########################################################################
  ##### Opdatering af parametre  ##########################################
  ##########################################################################
     # forslag til parametrene fra normalfordelinger
     lambda.prime = rnorm(1, lambda,0.0009)
     beta.prime = rnorm(1,beta,0.00003)
     beta0.prime = rnorm(1,beta0,5e-07)
     R.prime = rnorm(1,R,10)

     # vektor med længde af linjer efter BDM-algoritmen
     l=numeric(length=Y0$n)
     for(q in 1:Y0$n){l[q]=li(Y0$x[q],Y0$y[q])}

     # ligefordelte tal til at afgøre om forslag skal accepteres
     uuuu = runif(1,0,1)
     uuuuu = runif(1,0,1)
     uuuuuu = runif(1,0,1)
     uuuuuuu = runif(1,0,1)

      # Alle vores parmetre skal være positive
      if(lambda.prime>0){#lh.lambda(Y0)
      if(log(uuuu)< lh.lambda(Y0)){lambda = lambda.prime
```

```
                                  a=a+1}}

      if(beta.prime>0){#lh.beta(Y0)
      if(log(uuuuu)< lh.beta(Y0)){beta = beta.prime
                                 aa=aa+1}}

      if(beta0.prime>0){#lh.beta0(Y0)
      if(log(uuuuuu)< lh.beta0(Y0)){beta0 = beta0.prime
                                   aaa=aaa+1}}

      if(R.prime>0){#lh.beta(Y0)
      if(log(uuuuuuu)< lh.R(Y0)){R = R.prime
                                aaaa=aaaa+1}}

  # Parametrene gemmes i vektorer
  trace1[i]=lambda
  trace2[i]=beta
  trace3[i]=beta0
  trace4[i]=R
  trace5[i]=Y0$n

  print(i)              # viser hvor langt algoritmen er kommet
  flush.console()
}
a1        # antal gange algoritmen er gået ind i Move med nul linjer
a2        # antal gange algoritmen er gået ind i Move og flyttet
a3        # antal gange algoritmen er gået ind i Move og blevet stående
a4        # antal gange algoritmen er gået ind i Fødsel med nul linjer
a5        # antal gange algoritmen er gået ind i Fødsel og flyttet
a6        # antal gange algoritmen er gået ind i Fødsel og blevet stående
a7        # antal gange algoritmen er gået ind i Død med under 2 linjer
a8        # antal gange algoritmen er gået ind i Død og flyttet
a9        # antal gange algoritmen er gået ind i Død og blevet stående

a/m         # acceptsandsynlighed af lambda
aa/m        # acceptsandsynlighed af beta
aaa/m       # acceptsandsynlighed af beta0
aaaa/m      # acceptsandsynlighed af R
```

Litteratur

[Baddeley, 2008] Baddeley, A. (October 2008). Analysing spatial point patterns in R. Workshop Notes, Version 3.

[Barndorff-Nielsen et al., 1999] Barndorff-Nielsen, O. E., Kendall, W. S. og van Lieshout, M. N. M. (1999). *Stohastic Geometry. Likelihood and Computations.* Chapman & Hall/CRC, 1. udgave. ISBN 0-8493-0396-6.

[Berg og Madsen, 1970] Berg, C. og Madsen, T. G. (1970). *Mål- og integrationsteori.* Universitetsbogladen, København.

[Berthelsen og Møller, 2004] Berthelsen, K. K. og Møller, J. (2004). A short diversion into the theory of Markov chains, with a view to Markov chain Monte Carlo methods. Noter til Ph.D.-kursus.

[Daley og Vere-Jones, 2008] Daley, D. J. og Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure.* Springer, 2. udgave. ISBN 978-0-387-21337-8.

[Egeberg, 2004] Egeberg, T. (2004). Høje og hjulspor i tusindvis. Færdsel i det vestjyske landskab. *FRAM - Fra Ringkøbing Amts Museer*, side 44 - 51.

[Egeberg, 2005] Egeberg, T. (2005). De første hjulspor. *FRAM - Fra Ringkøbing Amts Museer*, side 39 - 44.

[Geyer og Møller, 2004] Geyer, C. J. og Møller, J. (2004). Simulation Procedures and Likelihood Inference for Spatial Point Processes. *Board of the Foundation of the Scandinavian Journal of Statistics*, 21(4):359-373.

[Johansen, 2009] Johansen, K. L. (2009). Upubliseret privat emailkorrespondance modtaget sammen med datasættet i feb. 2009.

[Johansen et al., 2004] Johansen, K. L., Laursen, S. T. og Holst, M. K. (2004). Spatial patterns of social organization in the Early Bronze Age of South Scandinavia. *Journal of Anthropological Archaeology*, 23.

[Kingman, 1993] Kingman, J. F. C. (1993). *Poisson Processes.* Oxford University Press. ISBN 0-19-853693-3.

[Miles, 1964] Miles, R. E. (1964). Random Polygons Determined by Random Lines in a Plane. *Proceedings of the National Academy of Sciences of the United States of America, Vol. 52, No. 4*, side 901 - 907.

[Møller og Rasmussen, 2009] Møller, J. og Rasmussen, J. G. (2009). Modelling point patterns with linear structures. *The MIRIAM Project Series.*

[Møller og Waagepetersen, 2004] Møller, J. og Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes.* Chapman & Hall/CRC, 1. udgave. ISBN 1-58488-265-4.

[Olofsson, 2005] Olofsson, P. (2005). *Probability, Statistics, and Stochastic Processes.* John Wiley & Sons, 1. udgave. ISBN 0-471-67969-0.

[Online Leksikon, 2006] Online Leksikon, G. (2006). Gravhøj. http://www.gyldendalsleksikon.dk/leksikon/leksikon.htm.

[R Development Core Team, 2008] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Santaló, 1976] Santaló, L. A. (1976). *Integral Geometry and Geometric Probability.* Addison-Wesley publishing, 2. udgave. ISBN 0-201-13500-0.

[Stoyan et al., 1995] Stoyan, D., Kendall, W. S. og Mecke, J. (1995). *Stochastic Geometry and its Applications.* John Wiley and Sons, 2. udgave. ISBN 0-471-95099-8.

[Stoyan og Stoyan, 1994] Stoyan, D. og Stoyan, H. (1994). *Fractals, Random Points and Point Fields.* Wiley, 1. udgave. ISBN 0-471-93757-6.