Rumlige punktprocessor

Aalborg University
Mathematics

AALBORG UNIVERSITY
STUDENT REPORT

Department of Mathematical Sciences
Mathematics
Skjernvej 4A
9220 Aalborg Øst
http://math.aau.dk

Titel:
Rumlige punktprocesser

Projekt:
Kandidatprojekt

Periode:
Efterår 2024

Deltager:
Daniel Bursac Hasanovic

Vejleder:
Anne Marie Svane

Sider: 51

Afleveret: 18. december 2024

Abstract:

This project investigates how point processes can be used as a tool to model and analyze spatial patterns. After an introduction to the basic concepts, including the Poisson process, the focus is on both the theory behind point processes and their practical application.
The project looks on K- and L-functions to analyze structures in datasets and determine whether there are signs of clustering or repulsion among the points. The introduced methods are tested on the dataset redwoodfull, which contains tree positions and is available in Rstudio. Here, the Poisson process, Strauss process, and Matérn Cluster process are compared to evaluate how well they describe the selected dataset. In the project, it can be seen that the Matérn Cluster Process is the best at describing the chosen dataset.

Table of Contents

1 Indledning

Punktprocesser bruges til at beskrive og kigge på, hvordan punkter fordeler sig i et område, og til at forstå, hvilke mønstre der opstår. For eksempel kan man bruge dem til at undersøge, hvordan træer står placeret i en skov, og hvad der påvirker, hvor de står og hvor de vokser. Hvis jorden er dårlig, spreder træerne sig ofte for at få mest muligt ud af de få ressourcer, der er. Omvendt, hvis der kun er god jord nogle steder, samler træerne sig tit i grupper, som kaldes klustre, for at konkurrere om de gode områder.

Punktprocesser kan også bruges til mange andre ting. For eksempel er de gode til at kigge på, hvordan sygdomme spreder sig i en befolkning, og til at finde de steder, hvor smitten eventuelt er startet. Det kan også hjælpe med at forstå, hvordan smitten bevæger sig rundt, og hvilke områder der skal have ekstra fokus for at stoppe den og holde den under kontrol.

Derudover kan punktprocesser også anvendes mere samfundsfagligt. For eksempel kan de bruges til at studere kriminalitet i et område, hvor man kan analysere mønstre og finde områder med høj risiko for kriminalitet. Når man har fundet de her mønstre, kan man bedre arbejde forebyggende og planlægge, hvordan man fordeler ressourcerne for at bekæmpe kriminaliteten.

De mange måder som man kan bruge punktprocesser på, gør dem netop til et rigtig godt værktøj både i forskning og til mere praktiske ting.

Dette projekt handler om at forklare, hvad punktprocesser er, og hvordan de virker. Det bliver gjort ved at kigge på teorien bag og ved at bruge et rigtigt datasæt til at vise og afprøve teorien. Projektet fokuserer mest på Poisson punktprocessen, som til sidst bliver sammenlignet med Strauss punktprocessen og Matérn cluster, der er en specialtilfælde af Cox punktproces. Programmet RStudio, og specifikke funktioner deri, bliver brugt til at finde $L-$ og K -funktioner for de forskellige punktprocesser, og så bliver de brugt til at se, hvor godt de passer til det selvvalgte datasæt.

2 | Grundlæggende egenskaber

Følgende afsnit tager udgangspunkt i [1] og [2].

I dette afsnit vil punktprocesser samt relevant teori bliver introduceret.

En punktprocss er en model, der kan bruges til at beskrive tilfældige mønstre af punkter i tid og rum. Punktprocesser kan defineres i flere dimensioner. Der gives nu et illustrativt eksempel på en punktproces i en dimension.

Figur 2.1: Punktproces i en dimension

I Figur 2.1 kan det ses det første eksempel, nemlig en punktproces i en dimension. Her kan linjen repræsentere en tidslinje på et kundeservicecenter, og hvert punkt kan repræsentere et opkald, der modtages på et tilfældigt tidspunkt.
Punktprocesser kan også blive mere komplekse og det bliver de i takt med at vi arbejder i højere dimensioner.Når man arbejder med to dimensionelle punktprocesser og opefter kaldes de *rumlige punktprocesser*. Her introduceres et illustrativt eksempel på en to dimensionel rumlig punktproces.

Figur 2.2: Punktprocess i to dimensioner

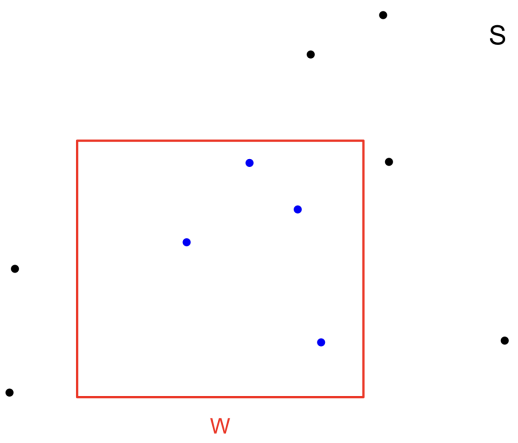
Forskellen mellem en punktproces og en rumlig punktproces ligger hovedsageligt i antallet

af dimensioner og anvendelsesområdet. En punktproces i en dimension beskriver tilfældige hændelser langs en tidslinje, hvor hver hændelse kun har en position. I Figur 2.1 er denne position angivet som tid. En rumlig punktproces, foregår i to eller flere dimensioner og bruges til at modellere tilfældige mønstre af punkter i et rumligt område, som for eksempel positioner af objekter i et geografisk rum. Rumlige punktprocesser giver derfor mulighed for at analysere fordelingen af punkter i fysisk rum fremfor kun over tid.

For eksempel kan en rumlig punktproces repræsentere placeringen af træer i en skov, spredning af en bestemt sygdom blandt en population eller positionerne af stjerner på himlen. Ser man på Figur 2.2 kan det eksempelvis illustrere et billede af en lille del af stjernehimlen, hvor hvert punkt repræsentere en stjerne. Dette leder op til uformel definitionen af rumlige punktprocessor.

En rumlig punktprocess X er en tilfældig, tællelig delmængde af et rum S.

Det vil ofte være sådan, at S vil være hele $\mathbb{R}^d$, men da $\mathbb{R}^d$ er ubegrænset, og det er umuligt at arbejde (og observere) med en uendelig datamængde, så vil man i praksis gerne have begrænset det på en eller anden måde, så dataen bliver mere håndgriblig. Dette kan man gøre med et *observationsvindue* W hvor $W \subseteq S$. Intuitivt kan S i astronomi ses som hele himlen, mens observationsvinduet W kan ses som det område, som et teleskop kan observere. På den måde kan man begrænse det område man arbejder på.



Figur 2.3: Et observationsvindue W illusteret på hele S. Her er W angivet med en rød firkant, og punkterne vi kan observere er angivet med blå farve.

I Figur 2.3 kan observationsvinduet ses angivet som firkantet, men observationsvinduet kan have flere forskellige former.

En ting man også gerne vil begrænse, er antal af punkter eller data i et observationsvindue. Derfor er man også interesseret i, at punktkonfigurationerne opfylder egenskaben *lokal endlighed*.

Definition 2.1. Lokal endelig punktkonfiguration

En punktkonfiguration $x \subseteq S$ siges at være lokal endelig, hvis ethvert begrænset område $B \subseteq S$ indholder et endeligt antal punkter. Her er x lokal endelig, hvis $n(x_B) < \infty$ for alle begrænsede delmængeder $B \subseteq S$, hvor $n(x_B)$ angiver kardinaliteten af x_B og $x_B = x \cap B$.

Dette sikrer, at der kun er et endeligt antal punkter i ethvert begrænset område af S. Ses på Figur 2.3 så vil $n(x_W) = 4$ da der kun er fire punkter indenfor observationsvinduet.
Nu kan rummet N_{lf}, som indeholder alle sådanne lokalt endelige punktkonfigurationer introduceres.

Definition 2.2. Rummet for lokalt endelige punktkonfigurationer
En *lokalt endelig punktkonfiguration* er en specifik samling af punkter $x \subseteq S$, som opfylder egenskaben om lokal endlighed. Rummet af alle lokalt endelige punktkonfigurationer betegnes N_{lf}:

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

Her kræves det, at vores punktprocesser altid giver os endelige mængder.

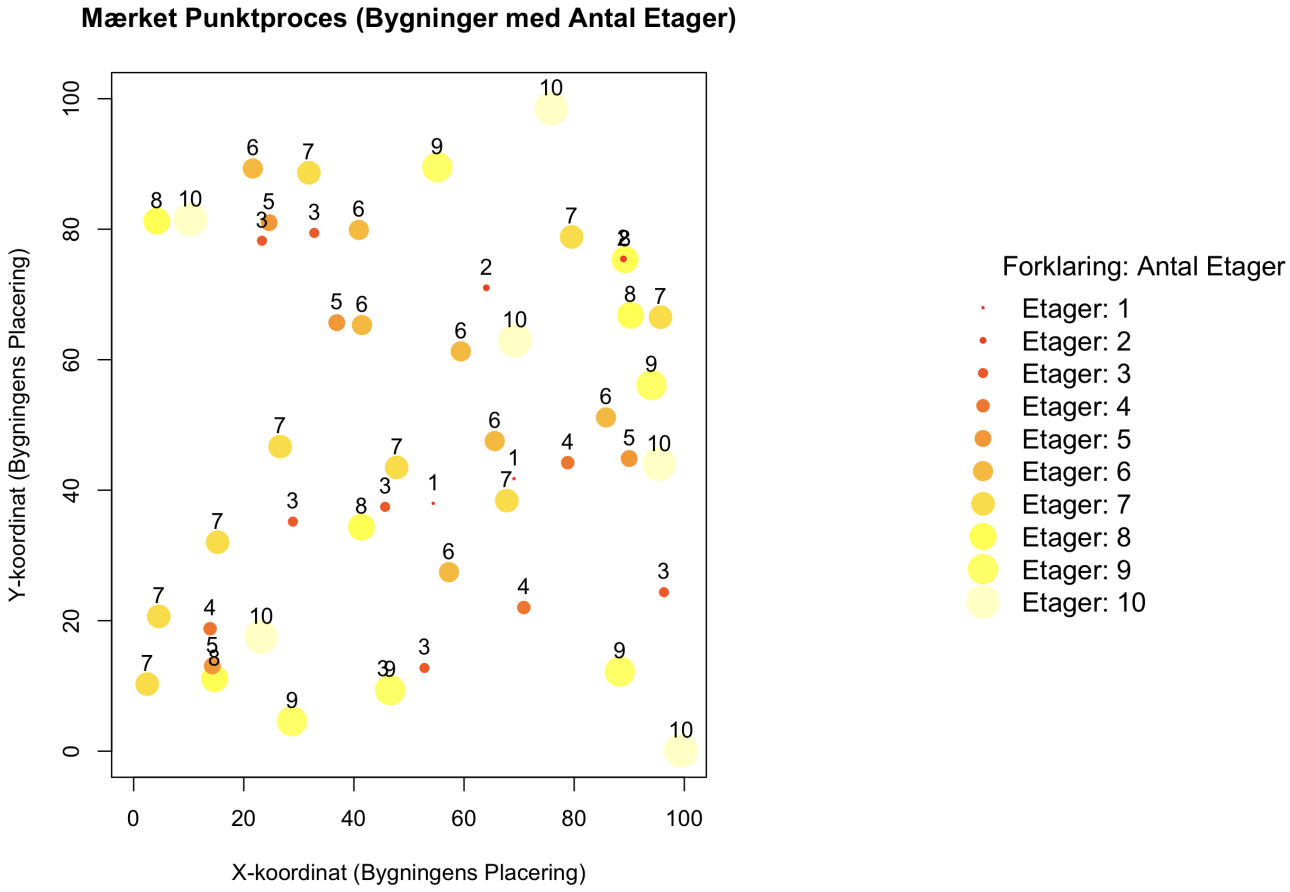
I de punktprocesser der indtil videre er blevet introduceret, får man kun information om placeringen af et punkt i et rum eller en tidslinje. I nogle tilfælde kan man have interesse i at markere dem, enten med en vægt eller en anden form for værdi og på den måde være mere informativ. Disse værdier kan eksempelvis være størrelser,typer eller former. Denne type af punktprocesser kaldes for *mærket punktprocesser*.

Definition 2.3. Mærket punktproces
Lad X være en punktproces på et rum $S \subseteq \mathbb{R}^d$. Lad M være et *mærketrum*. Så defineres en mærket punktproces som en proces, hvor hver punkt $x \in X$ har et tilknyttet mærke $m_x \in M$. Denne mærkede punktproces kan derfor opskrives som $Y = \{(x, m_x)|x \in X\}$

For at give en intuitiv forståelse introduceres følgende eksempel.

Eksempel 2.4.
Et datasæt kan for eksempel illustrere en storby og dens bygninger, hvor punkterne

repræsenterer bygningernes placering. Der kunne også være yderligere data, såsom antallet af etager i hver bygning. Dette eksempel viser en sådan situation.



Figur 2.4: Mærket punktproces hvor hver punkt repræsenterer en bygning og en højde

I figur 2.4 repræsenterer hver cirkel placeringen af en bygning i byen, mens cirklens størrelse og farve viser bygningens højde i antal etager. Jo højere bygningen er, desto større og lysere er cirklen. Denne mærket punktproces illustrerer, hvordan både position og ekstra information, såsom højde, kan repræsenteres i en samlet model. Ved at tilføje vægtning til punktprocessen bliver det muligt at analysere både bygningernes fordeling, men også deres højde.

For at karakterisere rumlige punktprocesser kan man benytte begrebet *void sandsynligheder*. Void sandsynligheder beskriver sandsynligheden for, at et givet område er tomt, dvs. at der ikke findes nogen punkter fra punktprocessen i området. Denne void sandsynlighed

er givet ved

$$v(B) = P(N(B) = 0)$$

hvor $B \subseteq S$ og $N(B) = n(X_B)$. For *Poisson punktprocesser*, som først vil blive introduceret i næste afsnit, gælder det, at void sandsynligheden er givet ved

$$\nu(B) = \exp(-\mu(B))\frac{\mu(B)^0}{0!} = \exp(-\mu(B)) \tag{2.1}$$

, hvor $B \subseteq$ S er begrænset. Dette leder op til følgende sætning.

Sætning 2.5. Entydighed af punktprocesser via void sandsynligheder
En punktproces X er entydigt bestemt af sine void sandsynligheder $v(B) = P(N(B) = 0)$ for alle begrænsede $B \subseteq S$.

Denne sætning vil ikke blive bevist i projektet.

Den næste del af projektet vil undersøge Poisson punktprocesser.

3 | Poisson processer

Følgende afsnit tager udgangspunkt i [1] og [2]

Før vi kan definere en Poisson punktproces, er det vigtigt at introducere nogle vigtige begreber, der beskriver fordelingen og tætheden af punkter i et rum. Disse begreber hjælper os med at forstå, hvordan punkter fordeles og forventes at optræde inden for et bestemt område.

Definition 3.1. Intensitetsfunktion

Lad $S \subseteq \mathbb{R}^d$ være det rum, hvor en punktproces er defineret. En *intensitetsfunktion* $\rho : S \to [0, \infty)$, beskriver tætheden af punkter i rummet. Intensitetsfunktionen er *lokalt integrabel*, hvilket betyder, at

$$\int_B \rho(\xi)\,\mathrm{d}\xi < \infty$$

for alle begrænsede delmængder $B \subseteq S$.

Intensitetsfunktionen ρ beskriver den gennemsnitlige tæthed af punkter i rummet S. Den gør det muligt at måle, hvor sandsynligt det er at finde punkter i forskellige dele af rummet.

Definition 3.2. Intensitetsmål

Givet en intensitetsfunktion ρ for en punktproces defineres det tilhørende *intensitetsmål* μ. Intensitetsmålet er det forventet antal punkter i en mængde og er givet ved $\mu(B) = \mathbb{E}[N(B)]$. Den har en intensitetsfunktion hvis målet kan beskrives ved at integrere en intensitetsfunktion over en mængde B, altså

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

hvor $B \subseteq S$. Intensitetsmålet $\mu(B)$ er *lokalt endeligt*, dvs. $\mu(B) < \infty$ for alle begrænsede $B \subseteq S$, samt *diffust*, hvilket betyder, at $\mu(\{\xi\}) = 0$ for alle $\xi \in S$.

Eksempelvis kan intensitetsmålet fortælle om det forventede antal træer i et bestemt skovområde eller det forventede antal bygninger i et kvarter.

Nu introduceres en *binomial punktproces* for at kunne definere en Possion punktproces. En binomial punktproces er et eksempel på en simpel punktproces, hvor et fast antal punkter placeres tilfældigt inden for et område i overensstemmelse med en givet tæthedsfunktion.

Definition 3.3. Binomial punktproces
Lad f være en tæthedsfunktion på en mængde $B \subseteq S$, og lad $n \in \mathbb{N}$ (hvor $\mathbb{N} = \{1, 2, 3, \dots\}$. En punktproces X, der består af n uafhængigt og identisk fordelte punkter med tæthed f, kaldes en binomial punktproces af n punkter i B med tæthed f. Vi skriver dette som

$$X \sim \text{binomial}(B, n, f).$$

Nu hvor binomial punktproces er blevet introduceret, altså en punktproces der har et fast antal punkter, introduceres nu Poisson punktprocessen, der er karakteriseret ved, at antallet af punkter i ethvert område er tilfældig, og altså ikke fast, og følger samtidig en Poisson fordeling.

Definition 3.4. Poisson punktproces
En punktproces X på S kaldes en Poisson punktproces med intensitetsfunktion ρ, hvis følgende egenskaber er opfyldt:

1. For enhver mængde $B \subseteq S$ med $\mu(B) < \infty$, er antallet af punkter $N(B)$ i B Poisson fordelt med middelværdi $\mu(B)$. Det vil sige, $N(B) \sim \text{Po}(\mu(B))$. Sandsynligheden for at finde præcist k punkter i B er givet ved:

   $$P(N(B) = k) = \frac{\mu(B)^k}{k!} exp(-\mu(B).$$

   Hvis $\mu(B) = 0$, er $N(B) = 0$.

2. For ethvert $n \in \mathbb{N}$ og enhver mængde $B \subseteq S$ med $0 < \mu(B) < \infty$, givet at $N(B) = n$, er punkterne i B, X_B, fordelt som en binomial punktproces med n punkter og tæthed $f(\xi) = \frac{\rho(\xi)}{\mu(B)}$.

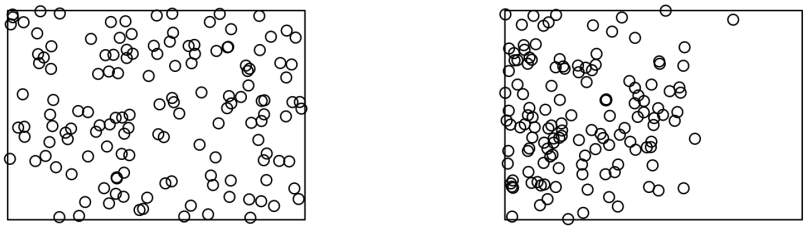
Her gælder så, at $X \sim \text{Po}(S, \rho)$.

Her skal punkt 1 forstås som fordeling af antal af punkter i et område B, hvor punkterne følger en Possion fordeling med middelværdien $\mu(B)$.
Her skal punkt 2 forstås som, hvordan punkterne fordels i B, når man har n antal punkter.

Nu hvor Poisson punktproces er introduceret, så kan punktprocessen enten være *homogen* eller *inhomogen*

Definition 3.5. Homogen og inhomogen
Hvis ρ er konstant, kaldes processen Po(S, ρ) for en homogen Poisson proces på S med intensiteten ρ. Hvis ρ derimod varierer over S, siges processen at være inhomogen. Desuden kaldes Po($S, 1$) for *standard* Poisson punktprocessen eller enheds Poisson proces på S.

Det vil altså være sådan, at hvis processen er homogen, så vil punkterne være jævnt fordelt over hele observationsvinduet, mens en inhomogen proces vil være koncentreret i specifikke områder. Følgende eksempel illustrerer disse to.



Figur 3.1: Simulation af en homogen Poisson punktproces (venstre) og simulation af en inhomogen Poisson punktproces(højre).

Som det ses på Figur 3.1, er punkterne i den homogene punktproces jævnt spredt over området, hvilket indikerer, at ρ er konstant. I den inhomogene proces er punkterne derimod mere kompakte og tættere samlet i visse områder, hvilket indikerer, at ρ varierer over S. Intuitivt kan vi tænke på en homogen proces som en skov, hvor træerne vokser tilfældigt og jævnt spredt ud. Omvendt kan en inhomogen proces sammenlignes med en storby, hvor bygninger er tæt koncentreret omkring centrum og gradvist bliver færre, jo længere vi bevæger os væk fra byen.

Nu introduceres yderligere vigtige egenskaber for en Poisson punktproces for eksistens og uafhængighed.

Sætning 3.6.

Lad $X \sim \text{Po}(S, \rho)$. Da gælder følgende:

(i) Punktprocessen X følger fordelingen $\text{Po}(S, \rho)$ hvis og kun hvis for alle $B \subseteq S$ med $\mu(B) = \int_B \rho(\xi)\, d\xi < \infty$ og for alle $F \subseteq N_{\text{lf}}$, har vi

$$P(X_B \in F) = \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B \cdots \int_B \mathbf{1}(\{x_1, \ldots, x_n\} \in F) \prod_{i=1}^{n} \rho(x_i)\, dx_1 \ldots dx_n, \tag{3.1}$$

hvor integralet for $n = 0$ læses som $\mathbf{1}(\emptyset \in F)$.

(ii) Hvis $X \sim \text{Po}(S, \rho)$, så gælder for funktioner $h : N_{\text{lf}} \to [0, \infty)$ og $B \subseteq S$ med $\mu(B) < \infty$:

$$E[h(X_B)] = \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B \cdots \int_B h(\{x_1, \ldots, x_n\}) \prod_{i=1}^{n} \rho(x_i)\, dx_1 \ldots dx_n. \tag{3.2}$$

Bevis. Først viser vi, at lighed ved Sætning 3.6 punkt (i) er opfyldt. Her antages det, at $X \sim Po(S, \rho)$. Fra loven om total sandsynlighed gælder følgende omskrivning af venstre side:

$$P(X_B \in F) = \sum_{n=0}^{\infty} P(X_B \in F \mid N(B) = n) P(N(B) = n). \tag{3.3}$$

For at udregne $P(X_B \in F \mid N(B) = n)$ anvendes at udfaldene er uafhængige, samt at sandsynligheden kan findes ved at integrere tætheden.

Det følger derfor, at følgende lighed gælder

$$P(X_B \in F \mid N(B) = n) = \int_B \cdots \int_B \mathbf{1}\{(x_1, \ldots, x_n) \in F\} \mu(B)^{-n} \prod_{i=1}^{n} \rho(x_i)\, dx_1 \cdots dx_n.$$

hvor indikatorfunktionen sikrerer at punkterne er i F.
Herved kan Ligning 3.3 omskrives til

$$\sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B \cdots \int_B \mathbf{1}\{(x_1, \ldots, x_n) \in F\} \prod_{i=1}^{n} \rho(x_i)\, dx_1 \cdots dx_n$$

hvilket også stemmer overens med punkt (i) i Sætning 3.6.

Nu skal det vises den anden vej. Her antages nu, at ligning (i) 3.6 gælder. Lad $F_k = \{x \in N_{lf} \mid N(B) = k$. Så gælder følgende omskrivning:

$$P(N(B) = k) = P(X \in F_k) \tag{3.4}$$

$$= \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_{B^n} \mathbf{1}\{(x_1, \ldots, x_n) \in F_k\} \prod_{i=1}^{n} \rho(x_i)\, dx_1 \cdots dx_n \tag{3.5}$$

Indikatorfunktionen $\mathbf{1}(\{x_1, \ldots, x_n\} \in F_k)$ sikrer, at summen over n kun har ét gyldigt led. Det er leddet hvor $n = k$. For alle andre værdier af $n \neq k$ vil indikatorfunktionen være 0, og disse led bidrager derfor ikke til summen. Derfor kan Ligning 3.5 omskrives til:

$$\frac{\exp(-\mu(B))}{k!} \int_{B^k} \prod_{i=1}^{k} \rho(x_i)\, dx_1 \cdots dx_k = \frac{\exp(-\mu(B))}{k!} \left(\int_B \rho(u)\, du \right)^k = \exp(-\mu(B)) \frac{\mu(B)^k}{k!}$$

I første lighed gælder det, at vi kan omskrive integralet. Da vi har k punkter i B, og hvert punkt er uafhængigt og fordelt i henhold til intensiteten ρ, kan vi skrive integralet som et produkt af integraler over hvert punkt. Den næste lighed stammer fra Definition 3.2, hvor integralet kan omskrives til intensitetsmålet.

På baggund af dette gælder det at $N(B) \sim Po(\mu(B))$.
Vi har

$$P(X_B \in F, N(B) = k) = \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_{B^n} \mathbf{1}\{(x_1, \ldots, x_n) \in F, N(B) = k\} \prod_{i=1}^{n} \rho(x_i)\, dx_1 \cdots dx_n$$

Her sikrer indikatorfunktionen igen, at vi kun har leddet hvor $n = k$ med. Derfor kan vi yderligere reducere til:

$$= \frac{\exp(-\mu(B))}{k!} \int_{B^k} \mathbf{1}\{(x_1, \ldots, x_k) \in F\} \prod_{i=1}^{k} \rho(x_i)\, dx_1 \cdots dx_k,$$

Nu kan vi finde den betingede sandsynlighed $P(X_B \in F \mid N(B) = k)$ ved at anvende definitionen af betinget sandsynlighed

$$P(X_B \in F \mid N(B) = k) = \frac{P(X_B \in F, N(B) = k)}{P(N(B) = k)}$$

og får

$$P(X_B \in F \mid N(B) = k) = \frac{\frac{\exp(-\mu(B))}{k!} \int_{B^k} \mathbf{1}\{(x_1, \ldots, x_k) \in F\} \prod_{i=1}^{k} \rho(x_i)\, dx_1 \cdots dx_k}{\frac{\mu(B)^k}{k!} \exp(-\mu(B))}$$

som kan forenkles til

$$P(X_B \in F \mid N(B) = k) = \int_{B^k} \mathbf{1}\{(x_1, \ldots, x_k) \in F\} \prod_{i=1}^{k} \frac{\rho(x_i)}{\mu(B)}\, dx_1 \cdots dx_k$$

Resultatet viser, at betinget på at der er præcis k punkter i B, så er punkterne $(x_1, \ldots, x_k)$ uafhængitgt og identisk fordelt med sandsynlighedstæthed $\frac{p(x_i)}{\mu(B)}$. Det betyder, at fordelingen af punkter i området B følger en binomialfordeling med intensitet $\frac{p(x_i)}{\mu(B)}$.

Beviset for Punkt (ii) i Sætning 3.6 gælder fra Punkt (i) og fra Standardbeviset, som ikke er taget med i dette projekt. □

Sætning 3.7.
Lad $X \sim \mathrm{Po}(S, \rho)$. Da eksisterer X og er entydigt bestemt af sine void sandsynligheder:

$$v(B) = \exp(-\mu(B)), \quad \text{for begrænset } B \subseteq S.$$

Bevis. Vi viser eksistensen af Poisson processen ved at konstruere den. Lad $\xi_0 \in S$ være et vilkårligt punkt i vores rum S, og definér mængderne B_i som

$$B_i = \{\eta \in S : i - 1 \leq \|\eta - \xi_0\| < i\} \quad \text{for} \quad i \in \mathbb{N}.$$

Disse B_i-mængder udgør en disjunkt forening af S, hvilket betyder, at hvert punkt i S tilhører præcis én af B_i-mængderne. Formålet med denne konstruktion er at opdele rummet S i afgrænsede regioner, som hver især kan behandles separat.

For hvert B_i defineres en stokastisk variabel N_i, som repræsenterer antallet af punkter i B_i. Denne følger en Poisson fordeling med parameter $\mu(B_i)$, altså $N_i \sim \text{Po}(\mu(B_i))$, hvor $\mu(B_i)$ angiver målet af B_i under intensitetsmålet μ. Her er N_i'erne uafhængige af hinanden, hvilket betyder, at antallet af punkter i ét område B_i ikke påvirker antallet af punkter i et andet område B_j.

Betinget på $N_i = n$ konstrueres en binomial proces X_i i B_i med n punkter. Målet μ_i og tætheden $f(\xi)$ på B_i defineres som

$$f(\xi) = \frac{\rho_i(\xi)}{\mu(B_i)},$$

hvor ρ_i er restriktionen af henholdsvis μ og ρ til B_i. Denne konstruktion sikrer, at punkterne fordeles jævnt i B_i i forhold til intensiteten ρ.

Ved at konstruere X_i processerne uafhængigt af hinanden, kan vi vise, at foreningen

$$X = \bigcup_{i=1}^{\infty} X_i$$

udgør en Poisson proces på S.

For at verificere dette, anvender vi loven om total sandsynlighed for en begrænset mængde $B \subseteq S$, hvilket giver:

$$P(n(X_B) = 0) = \prod_{i=1}^{\infty} P(n(X_i \cap B) = 0) = \prod_{i=1}^{\infty} \sum_{n=0}^{\infty} P(n(X_i \cap B) = 0 | N_i = n) P(N_i = n).$$

Da N_i er Poisson fordelt, har vi, at $P(N_i = n) = \frac{\exp(-\mu(B_i))\mu(B_i)^n}{n!}$.

Betinget sandsynligheden for, at $X_i \cap B$ er tom, givet $N_i = n$, kan skrives som

$$P(n(X_i \cap B) = 0 | N_i = n) = \left(1 - \int_{B \cap B_i} \frac{\rho_i(\xi)}{\mu(B_i)} d\xi\right)^n = \left(1 - \frac{\mu(B \cap B_i)}{\mu(B_i)}\right)^n.$$

Vi indsætter nu denne sandsynlighed tilbage i summen og produktet og forenkler trin for trin som følger:

$$P(n(X_i \cap B)) = \prod_{i=1}^{\infty} \sum_{n=0}^{\infty} \left(1 - \frac{\mu(B \cap B_i)}{\mu(B_i)}\right)^n \frac{\exp(-\mu(B_i))\mu(B_i)^n}{n!}$$

$$= \prod_{i=1}^{\infty} \sum_{n=0}^{\infty} \frac{(\mu(B_i) - \mu(B \cap B_i))^n}{n!} \exp(-\mu(B_i))$$

$$= \prod_{i=1}^{\infty} \exp(\mu(B_i) - \mu(B \cap B_i)) \exp(-\mu(B_i))$$

$$= \prod_{i=1}^{\infty} \exp(-\mu(B \cap B_i) = \exp\left(-\sum_{i=1}^{\infty} \mu(B \cap B_i)\right)$$

$$= \exp(-\mu(B)).$$

Udtrykket $\exp(-\mu(B))$ er void-sandsynligheden for en Poisson proces med intensitetsmål μ fra Afsnit 2.1, og dermed har vi entydigt konstrueret en Poisson proces på S med intensitet μ. Dette viser både eksistensen og entydigheden af Poisson processen, hvor entydigheden kommer af Sætning 2.5 □

Nu introduceres yderligere en sætning.

Sætning 3.8.
Hvis X er en Poisson proces på S, så er de stokastiske variable $X_{B_1}, X_{B_2}, \ldots$ uafhængige for disjunkte mængder $B_1, B_2, \ldots \subseteq S$.

Bevis. Det ønskes at påvise, at $X_{B_1}, X_{B_2}, \ldots, X_{B_n}$ er uafhængige for disjunkte mængder $B_1, \ldots, B_n$, hvor $n \geq 2$.
For $n = 2$ antager vi, at B_1 og B_2 er delmængder af S, disjunkte og begrænsede. Vi definerer da $B = B_1 \cup B_2$.

I beviset skal der vises, at følgende lighed gælder

$$P(X_{B_1} \in F_1, X_{B_2} \in F_2) = P(X_{B_1} \in F_1)P(X_{B_2} \in F_2).$$

Da X er en Poisson proces på $B \subseteq S$ så gælder det fra Sætning 3.6 at

$$
\begin{aligned}
P(X_{B_1} \in F_1, X_{B_2} \in F_2) &= \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B \cdots \int_B \mathbf{1}\left[\{x_1, \ldots, x_n\} \cap B_1 \in F_1, \{x_1, \ldots, x_n\} \cap B_2 \in F_2\right] \\
&\quad \times \prod_{i=1}^{n} \rho(x_i)\, dx_1 \ldots dx_n \\
&= \sum_{n=0}^{\infty} \frac{\exp(-\mu(B_1 \cup B_2))}{n!} \sum_{k=0}^{n} \binom{n}{k} \int_{B_1} \cdots \int_{B_1} \int_{B_2} \cdots \int_{B_2} \\
&\quad \times 1\left[\{x_1, \ldots, x_k\} \in F_1\right] 1\left[\{x_{k+1}, \ldots, x_n\} \in F_2\right] \prod_{i=1}^{n} \rho(x_i)\, dx_1 \ldots dx_n.
\end{aligned} \tag{3.6}
$$

Her følger denne omskrivning af, at der er n punkter samlet i B. Binomialkoefficienten $\binom{n}{k}$ angiver antallet af måder, hvorpå k punkter kan vælges ud af de n punkter til at ligge i B_1, hvilket samtidig resulterer i, at de resterende $n-k$ punkter må ligge i B_2. Her opdeles også i to summer, hvilket er nødvendigt for at tage højde for alle mulige konfigurationer af, hvordan punkterne fordeles mellem B_1 og B_2. Den indre sum opdeler efter, hvor mange af de n punkter der ligger i B_1, og hver term i denne sum vægtes af binomialkoefficienten $\binom{n}{k}$, som angiver antallet af måder, k punkter kan placeres i B_1, mens de resterende $n-k$ punkter placeres i B_2.

$$
\begin{aligned}
P(X_{B_1} \in F_1, X_{B_2} \in F_2) &= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{\exp(-\mu(B_1) - \mu(B_2))\, n!}{n!\, k!(n-k)!} \int_{B_1^k} \int_{B_2^{n-k}} \\
&\quad \times 1\left[\{x_1, \ldots, x_k\} \in F_1\right] 1\left[\{x_{k+1}, \ldots, x_n\} \in F_2\right] \prod_{i=1}^{n} \rho(x_i)\, dx_1 \ldots dx_n \\
&= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{\exp(-\mu(B_1) - \mu(B_2))}{k!(n-k)!} \int_{B_1^k} 1\left[\{x_1, \ldots, x_k\} \in F_1\right] \prod_{i=1}^{k} \rho(x_i)\, dx_1 \ldots dx_k \\
&\quad \times \int_{B_2^{n-k}} 1\left[\{x_{k+1}, \ldots, x_n\} \in F_2\right] \prod_{i=k+1}^{n} \rho(x_i)\, dx_{k+1} \ldots dx_n.
\end{aligned} \tag{3.7}
$$

Her gælder det, at de to summer kan samles til en dobbelt sum, og at binomialkoefficienten udskrives og ganges ind i brøken. Da B_1 og B_2 er disjunkte så kan eksponentialudtrykket omskrives til $\exp \mu(B) = \exp(\mu(B_1) + \mu(B_2))$. Her separeres integralerne og indikatorfunktionen også og vi reparametrisere $m = n - k$ så følgende lighed gælder

$$P(X_{B_1} \in F_1, X_{B_2} \in F_2) = \sum_{k=0}^{\infty} \frac{\exp(-\mu(B_1))}{k!} \int_{B_1^k} 1\left[\{x_1, \ldots, x_k\} \in F_1\right] \prod_{i=1}^{k} \rho(x_i)\, dx_1 \ldots dx_k$$
$$\times \sum_{m=0}^{\infty} \frac{\exp(-\mu(B_2))}{m!} \int_{B_2^m} 1\left[\{x_1, \ldots, x_m\} \in F_2\right] \prod_{i=1}^{m} \rho(x_i)\, dx_1 \ldots dx_m$$
$$= P(X_{B_1} \in F_1) P(X_{B_2} \in F_2). \tag{3.8}$$

Her gælder den sidste lighed igen fra Sætning 3.6.

Nu ses på induktionstrin. I basistrinnet er det blivet vist for $n = 2$
Antag, at påstanden gælder for n disjunkte mængder. Det vil sige, at for disjunkte $B_1, \ldots, B_n$ har vi:

$$P\big(X_{B_1} \in F_1, \ldots, X_{B_n} \in F_n\big) \;=\; \prod_{i=1}^{n} P\big(X_{B_i} \in F_i\big).$$

Vi ønsker at vise, at det også gælder for $n+1$ disjunkte mængder $B_1, \ldots, B_n, B_{n+1}$. Betragt

$$P\big(X_{B_1} \in F_1, \ldots, X_{B_n} \in F_n,\; X_{B_{n+1}} \in F_{n+1}\big).$$

Lad $B = B_1 \cup \cdots \cup B_n$. Da B og B_{n+1} er disjunkte, kan vi betragte begivenhederne $\{X_B \in F\}$ og $\{X_{B_{n+1}} \in F_{n+1}\}$.

Lad

$$\{X_B \in F\} \quad \longleftrightarrow \quad \{X_{B_1} \in F_1, \ldots, X_{B_n} \in F_n\}.$$

Ved basistrinnet for to disjunkte områder, B og B_{n+1}, fås

$$P\big(X_B \in F,\; X_{B_{n+1}} \in F_{n+1}\big) \;=\; P\big(X_B \in F\big)\, P\big(X_{B_{n+1}} \in F_{n+1}\big).$$

Da

$$\{X_B \in F\} \;=\; \{X_{B_1} \in F_1, \ldots, X_{B_n} \in F_n\},$$

kan vi skrive

$$P\big(X_{B_1} \in F_1, \ldots, X_{B_n} \in F_n,\; X_{B_{n+1}} \in F_{n+1}\big) \;=\; P\big(X_B \in F,\; X_{B_{n+1}} \in F_{n+1}\big).$$

Ved induktionsantagelsen er $P(X_B \in F) = \prod_{i=1}^{n} P(X_{B_i} \in F_i)$. Indsætter vi det, fås

$$P\big(X_{B_1} \in F_1, \ldots, X_{B_n} \in F_n,\; X_{B_{n+1}} \in F_{n+1}\big) \;=\; \Big(\prod_{i=1}^{n} P(X_{B_i} \in F_i)\Big)\, P(X_{B_{n+1}} \in F_{n+1}).$$

Samlet giver det

$$P\big(X_{B_1} \in F_1, \ldots, X_{B_{n+1}} \in F_{n+1}\big) \;=\; \prod_{i=1}^{n+1} P\big(X_{B_i} \in F_i\big).$$

Vi har dermed vist, at hvis påstanden gælder for n disjunkte mængder, gælder den også for $n+1$.

□

Som det fremgår af beviset, har Poisson processen den særlige egenskab, at de stokastiske variable for disjunkte mængder er indbyrdes uafhængige. Denne egenskab kaldes *uafhængig scattering*, hvilket betyder, at punkterne i forskellige disjunkte områder fordeles uafhængigt af hinanden. Egenskaben sikrer, at der ikke er nogen afhængighed mellem antallet eller placeringen af punkter i disjunkte områder.

Denne egenskab er vigtig for en Poisson proces og gør det muligt at vurdere, om en punktproces opfører sig fuldstændig tilfældigt, eller om der er tegn på afhængighed, i form af *klustering* eller *frastødelse*.

En nyttig karakterisering af en Poisson punktproces er givet ved følgende sætning. Den siger, at middelværdien over en sum af funktioner på mængden S, kan omskrives til et integrale over middelværdien ganget med intensiteten.

Sætning 3.9. Slivnyak-Meckes Sætning
Hvis $X \sim \text{Po}(S, \rho)$, så gælder det for funktioner $h : S \times N_{lf} \to [0, \infty)$, at $\mathbb{E}\left[\sum_{u \in X} h(u, X \setminus u)\right] = \int_S \mathbb{E}[h(u, X)]\rho(u)\, du$.

Bevis. Der vil kun blive set på det specielle tilfældet hvor

$$\sum_{u \in X} h(u, X \setminus \{u\})$$

kun afhænger af X gennem X_B for et $B \subseteq S$ med $\int_B \rho(u)\, du < \infty$. Til dette kan vi anvende Sætning 3.6 til at finde den forventede værdi hvor vi får følgende omskrivning

$$\mathbb{E}\left[\sum_{u \in X} h(u, X \setminus u)\right] = \sum_{n=1}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B \cdots \int_B \sum_{i=1}^{n} h(x_i, \{x_1, \ldots, x_n\} \setminus x_i) \prod_{i=1}^{n} \rho(x_i)\, dx_1 \cdots dx_n. \tag{3.9}$$

Det kan igen omskrives til

$$= \sum_{n=1}^{\infty} \frac{\exp(-\mu(B))}{(n-1)!} \int_B \cdots \int_B h(x_n, \{x_1, \ldots, x_{n-1}\}) \prod_{i=1}^{n} \rho(x_i)\, dx_1 \cdots dx_n \tag{3.10}$$

$$= \int_B \sum_{k=0}^{\infty} \frac{\exp(-\mu(B))}{k!} \int_B \cdots \int_B h(x_{k+1}, \{x_1, \ldots, x_k\}) \prod_{i=1}^{k+1} \rho(x_i)\, dx_1 \cdots dx_{k+1} \tag{3.11}$$

$$= \int_B \sum_{k=0}^{\infty} \frac{\exp(-\mu(B))}{k!} \mu(B)^k \int_B \cdots \int_B h(u, \{x_1, \ldots, x_k\}) \prod_{i=1}^{k} \frac{\rho(x_i)}{\mu(B)}\, dx_1 \cdots dx_k\, \rho(u)\, du. \tag{3.12}$$

Her gælder den første lighed, da integralet af hvert enkel led i h-funktionen giver det samme, og kan omskrives ved at gange med n, samt at der erstatte med $n = k + 1$.

Herefter gælder det, at tæthedsfunktionen for $\{x_1, \ldots, x_k\}$ er givet ved $\prod_{i=1}^{k} \frac{\rho(x_i)}{\mu(B)}$, hvor den forventede værdi er givet ved $E[g(x)] = \int g(x) f(x)\, dx$. Her er $g(x)$ en funktion og $f(x)$ er tæthedsfunktion. I Ligning (3.12) er $g(x) = h(u, \{x_1, \ldots, x_k\})$, hvor følgende omskrivning gælder

$$= \int_B \sum_{k=0}^{\infty} \frac{e^{-\mu(B)}}{k!}\, \mu(B)^k\, E[h(u, \{x_1, \ldots, x_k\}) | N(B) = k] \rho(u)\, du.$$

Her ser vi, at summen kan skrives som en forventningsværdi, da udtrykket $\frac{e^{-\mu(B)}}{k!}\, \mu(B)^k$ svarer til sandsynligheden $P(N(B) = k)$ for en Poisson fordeling med parameter $\mu(B)$. Her angiver $N(B)$ antallet af punkter i området B. Summen repræsenterer derfor en forventning, hvor antallet af punkter k vægtes af sandsynlighederne for, at $N(B) = k$, og derfor fås denne omskrivning

$$= \int_B E\left[E[h(u, \{x_1, \ldots, x_k\}) | N(B) = k]\right] \rho(u)\, du.$$

Nu kan formlen for loven om total forventning anvendes $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$ og vi får

$$\int_S \mathbb{E}[h(u, X)] \rho(u)\, du.$$

□

Beviset for det generelle tilfælde vil ikke blive bevist i projektet.

4 | Summary statistik

Følgende afsnit tager udgangspunkt i [1] og [2].

I dette afsnit vil Summary statistik blive introduceret. En summary statistik er en størrelse, enten et tal eller en funktion, der beregnes på baggrund af data for at beskrive eller opsummere vigtige egenskaber ved datasættet.

4.0.1 Første - og andenordens egenskaber

Første - og andenordens egenskaberne for de stokastiske tællevariable, $N(B)$ for $B \subseteq S$, kan beskrives ved intensitetsmål og *andenordens faktorielle momentmål*. Intensitetsmålet $\mu(B) = \mathbb{E}[N(B)]$ som blev introduceret i Kapitel 3, kaldes også førsteordens moment.
Nu kan definitionen for andenordens moment introduceres.

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

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

Her skal det forstås, at førsteordens faktorielle momentmål beskriver den overordnet tæthed af punkter, mens andenordens beskriver den afhængighed eller interaktion mellem par af punkter. $\alpha^{(2)}(C)$ er en mål funktion, der måler den forventet værdi for at finde par af punkter inden for et bestemt område $C \subseteq \mathbb{R}^d \times \mathbb{R}^d$. Her hvor andenordens faktorielle momentmål er blevet introduceret, så kan *andenordens produkttæthed* introduceres.

Definition 4.2. Andenordens produkttæthed
Hvis det andenordens faktorielle momentmål $\alpha^{(2)}$ kan skrives som

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

hvor $\rho^{(2)}$ er en ikke-negativ funktion, så kaldes $\rho^{(2)}$ anden ordens produkttætheden.

Her skal det intuitivt forståes, at $\rho^{(2)}(\xi,\eta)\,d\xi\,d\eta$ er sandsynligheden for at observere et par af punkter fra X, der optræder samtidigt i to uendeligt små kugler med centrum i ξ og η og volumener $d\xi$ og $d\eta$. Vi bruger altså disse to punkter som referencepunkter for at undersøge, hvordan andre punkter fra processen X placerer sig i forhold til disse.

For at undersøge, om processen afviger fra en Poisson proces, er det nyttigt at normalisere andenordens produkttætheden $\rho^{(2)}(\xi,\eta)$ ved at dividere med $\rho(\xi)\rho(\eta)$.

Definition 4.3. Parkorrelationsfunktionen
Hvis både ρ og $\rho^{(2)}$ eksisterer, defineres *parvis korrelationsfunktionen* som

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

hvor $g(\xi,\eta) = 0$ hvis $\rho(\xi)\rho(\eta) = 0$

Definition 4.3 giver en intuitiv forståelse af, hvordan punkter korrelerer og interagerer. Her gælder det så, at værdien kan fortolkes på følgende måde:

- $g(\xi,\eta) = 1$: Punktprocessen opfører sig som en Poisson proces, og der er ingen interaktion mellem punkterne ξ og η.
- $g(\xi,\eta) > 1$: Der er en *positiv korrelation* mellem punkterne, hvilket indikerer en tendens til klustering, altså at punkterne samler sig.
- $g(\xi,\eta) < 1$: Der er en *negativ korrelation* mellem punkterne, hvilket indikerer en tendens til frastødning.
- $g(\xi,\eta) = 0$: Punkterne ξ og η kan ikke forekomme samtidig, fordi mindst ét af punkterne ikke eksisterer i processen ($\rho(\xi)\rho(\eta) = 0$).

4.0.2 Mål $\mathcal{K}$ samt $K-$ og $L-funktion$

I denne del bliver målet ($\mathcal{K}$) introduceret samt *K- og L-funktionerne.*

Det andenordens reduceret momentmål kan udtrykkes ved hjælp af intensitetsfunktionen og den følgende målfunktion $\mathcal{K}$ på $\mathbb{R}^d$.

Definition 4.4. Andenordens reduceret momentsmål
Antag, at X har intensitetsfunktion ρ og den tilhørende målfunktion, hvor

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

for $B \subseteq \mathbb{R}^d$, ikke afhænger af valget af $A \subseteq \mathbb{R}^d$ med $0 < |A| < \infty$. Hvis $\rho(\xi)\rho(\eta) = 0$ så sættes leddet i summen lig 0 . Så siges X at være andenordens intensitetsvægtet stationær, og K kaldes andenordens reduceret momentmål.

Hvis parkorrelationsfunktionenen g eksister og er invariant under forskydning, så har vi en andenordens intensitetsvægtet staionæritet givet ved

$$\mathcal{K}(B) = \int_B g(\xi)d\xi, B \subseteq \mathbb{R}^d. \tag{4.1}$$

Nu kan K- og L-funktionen introduceres.

Definition 4.5. K- og L-funktion
K- og L-funktionerne for en stationær punktproces af anden orden med vægtning er defineret som

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

Her i Definition 4.5 skal $b(0,r)$ ses som en kugle med r angivet som radius og 0 angivet som centrum for kuglen. For L-funktionen gælder det at $\omega_d = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}$ angiver volumen af den d-dimensionelle enhedskugle.
Ud fra Definitionen kan det også ses, at K- og L- funktionerne afhænger af hinanden. Her kan det ses, at K-funktionen indgår i formlen for L-funktionen.

I praksis, så gælder det, at L-funktionen oftere bliver anvendt end K-funktionen. Årsagen til dette skal findes i, at L-funktionen er mere billedlig intuitiv og kan anses som en identitet for en Poisson proces, hvilket vil blive vist. Man kan altså bruge den til at vurdere om hvorvidt en punktproces er tilfældig eller ej. Her vil det gælde, i hvert fald for små r, at $L(r) - r > 0$ vil indikere en klustering, mens $L(r) - r < 0$ vil indikere en frastødning.
Hvis $\mathcal{K}$ er invariant under rotation, kan $\mathcal{K}$ bestemmes med K. Hvis parkorrelationsfunktionen g, fra Definition 4.3, er invariant under rotation, så kan det ved Ligning 4.1 vises, at

$$K(r) = \sigma_d \int_0^r t^{d-1} g(t)\, dt. \tag{4.2}$$

ifølge [1].

Ligning 4.2 viser forholdet mellem K og g. Her angiver $\sigma_d = 2\pi^{d/2}/\Gamma(d/2)$ overfladearealet af den d dimensionelle enhedskugle. Her gælder det dog i praksis, at man foretrækker at

bruge K- og L-funktioner end g.Det skyldes at de er nemmere at estimere.

Proposition 4.6.
Hvis X er en homogen Poisson proces og andenordens intensitetsvægtet stationær, så gælder for K- og L-funktionerne, at

$$K(r) = \sigma_d \frac{1}{d} r^d, \quad L(r) = r.$$

Bevis. Hvis X er en Poisson proces og andenordens intensitetsvægtet stationær, så gælder det fra Definition 4.3 at følgende lighed gælder $g(\xi, \eta) = 1$. Ved nu at anvende Ligning 4.2 får man følgende lighed

$$\begin{aligned} K(r) &= \sigma_d \int_0^r t^{d-1} g(t)\, dt \\ &= \sigma_d \int_0^r t^{d-1}\, dt \\ &= \sigma_d \left[\frac{1}{d} t^d \right]_0^r \\ &= \sigma_d \frac{1}{d} r^d. \end{aligned}$$

hvor $K(r)$ nu er vist. Nu kan $L(r)$ vises hvor det for $d \in \mathbb{N}$ gælder at

$$\Gamma(d/2) = \sqrt{\pi} \frac{(d-2)!!}{2^{\frac{d-1}{2}}}.$$

Heraf fås det at

$$\Gamma(1 + d/2) = \Gamma\left(\frac{d+2}{2} \right) = \sqrt{\pi} \frac{d!!}{2^{\frac{d+1}{2}}}.$$

Nu kan det indsættes i formlen fra Definition 4.5 hvor det bliver

$$\begin{aligned}
\frac{K(r)}{\omega_d} &= \frac{\sigma_d \frac{1}{d} r^d}{\pi^{d/2}/\Gamma(1+d/2)} \\
&= \frac{2\pi^{d/2}}{\sqrt{\pi}(d-2)!!\left(2^{\frac{d-1}{2}}\right)^{-1}\frac{1}{d}r^d} \div \frac{\pi^{d/2}}{\sqrt{\pi}d!!\left(2^{\frac{d+1}{2}}\right)^{-1}} \\
&= \frac{2\pi^{d/2}\sqrt{\pi}d!!\left(2^{\frac{d+1}{2}}\right)^{-1}\frac{1}{d}r^d}{\pi^{d/2}\sqrt{\pi}(d-2)!!\left(2^{\frac{d-1}{2}}\right)^{-1}} \\
&= \frac{2d!!\left(2^{\frac{d+1}{2}}\right)^{-1}}{(d-2)!!\left(2^{\frac{d-1}{2}}\right)^{-1}\frac{1}{d}r^d} \\
&= \frac{d!!}{(d-2)!!}\frac{1}{d}r^d \\
&= r^d.
\end{aligned}$$

Her følger den sidste lighed fra regneregler for dobbelt fakultet. Slutligt fås at

$$L(r) = (r^d)^{1/d} = r$$

som også var det der skulle vises. □

Nu introduceres tre definitioner for Summary statistikker. Disse baseret på afstand mellem punkter.

Først vil *tomrumsfunktionen* introduceres.

Definition 4.7. Tomrumsfunktion
Antag, at X er stationær. Den tomrumsfunktion F er fordelingsfunktionen for afstanden fra Origo (eller et andet fast punkt i $\mathbb{R}^d$) til det nærmeste punkt i X, dvs.

$$F(r) = P(X \cap b(0,r) \neq \emptyset), \quad r > 0,$$

hvor $b(0,r)$ er kuglen med centrum i Origo og radius r.

Tomrumsfunktionen beskriver, hvor tomt rummet er i forhold til punktprocessen X. Det angiver sandsynligheden for, at en kugle med en bestemt radius, centreret i Origo, indeholder mindst ét punkt fra punktprocessen X.
Hvis $F(r)$ er lille, så indikerer det store tomme områder uden punkter, mens et højt F(r) tyder på at der er mange punkter samtlet tæt.

Den næste funktion der kan introduceres er *nærmeste nabo funktionen.*

Definition 4.8. Nærmest nabo funktion
Nærmeste nabo funktionen $G(r)$ er givet ved:

$$G(r) = \frac{1}{\rho|A|}\mathbb{E}\sum_{\xi\in X\cap A}\mathbf{1}\big(X\setminus\{\xi\}\cap b(\xi,r)\neq\emptyset\big), \quad r>0,$$

for en vilkårlig mængde $A\subset\mathbb{R}^d$ med $0<|A|<\infty$. Her repræsenterer ρ intensiteten, $|A|$ er volumenet af området A, og $b(\xi,r)$ er kuglen med centrum i ξ og radius r.

Intuitivt kan nærmeste nabo funktionen fortolkes som en fordelingsfunktion, der beskriver afstanden fra et tilfældigt punkt til det nærmeste nabopunkt i X. Den måler, hvor tæt punkterne er lokalt, og viser, om punkterne danner klustering, eller om de frastøder.

Den sidste summary statistik der vil blive introduceret er *J funktionen.*

Definition 4.9. J funktionen
Antag at X er stationær. J funktionen er givet ved

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

for $F(r)<1$.

J funktionen sammenligner sandsynlighederne fra tomrumsfunktionen $F(r)$ og nærmeste nabo funktionen $G(r)$. Den kan bruges til at vurdere, om punkterne i en proces kluster sig sammen, frastøder hinanden eller er tilfældigt fordelt.

Hvis $J(r)$ er større end 1, betyder det, at punkterne frastøder hinanden og holder afstand. Derimod hvis $J(r)$ er mindre end 1, indikerer det, at punkterne kluster. Når $J(r)$ er lig med 1, opfører punkterne sig som i en Poisson proces, hvor de er tilfældigt fordelt.

J-funktionen giver derfor en simpel måde at analysere, hvordan punkterne fordeler sig i rummet, og om der er tegn på klustering eller frastødning.

5 | Markov punktproces

Følgende afsnit tager udgangspunkt i [1].

Markov punktprocesser beskriver punktprocesser, hvor punkterne påvirker hinanden. Modellerne laves ved at definere en tæthedsfunktion, der relaterer punktprocessen til en Poisson proces. Man sikrer, at processen har *Markov egenskaber*.

5.0.1 Endelige punktprocesser

Antag at en punktproces X på $S \subseteq \mathbb{R}^d$ har tæthedsfunktionen f med hensyn til en standard Poisson proces $Y \sim Po(S, 1)$ så er f kun veldefineret for $|S| < \infty$. Derfor vil der i hele dette kapitel antages, at $|S| < \infty$ medmindre andet er angivet.
Nu kan tæthedsfunktionen f defineres på mængden

$$N_f = \{x \subset S : n(x) < \infty\}.$$

Lad $F \subseteq N_f$ så er

$$P(X \in F) = \sum_{n=0}^{\infty} \frac{\exp(-|S|)}{n!} \int_S \cdots \int_S \mathbf{1}[\{x_1, \ldots, x_n\} \in F] f(\{x_1, \ldots, x_n\})\, dx_1 \cdots dx_n \quad (5.1)$$

hvor $n = 0$ skal ses som $\exp(-|S|)1[\emptyset \in F]f(\emptyset)$. Hvis $|S| = 0$ så gælder det at $P(X = \emptyset) = 1$. Derfor vil der fremadrettet antages at $|S| > 0$. I de fleste tilfælde, så er tæthedsfunktionen f oftest kun kendt op til en normeringskonstant, altså $f \propto h$ hvor $h : N_f \to [0, \infty)$ er en kendt funktion. Her er normaliseringskonstanten c ukendt hvor tæthedsfunktionen h kan angives som $h(x) = cf(x)$

Definition 5.1. Papangelou-intensitet
Papangelou-intensiteten for en punktproces X med tæthed f er defineret som

$$\lambda^*(x, \xi) = \frac{f(x \cup \xi)}{f(x)}, \quad x \in N_f, \xi \in S \setminus x,$$

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

Det gælder yderligere, at $\lambda^*(x,\xi)$ er uafhængig af normeringskonstanten c, fordi:

$$\frac{f(x\cup\xi)}{f(x)} = \frac{\frac{1}{c}h(x\cup\xi)}{\frac{1}{c}h(x)} = \frac{h(x\cup\xi)}{h(x)}.$$

Her ophæver normeringskonstanten sig selv, hvilket gør $\lambda^*(x,\xi)$ uafhængig af c.
Papangelou-intensiteten $\lambda^*(x,\xi)d\xi$ kan fortolkes som den betingede sandsynlighed for, at punktprocessen X har et punkt inden for det infinitesimale område $d\xi$, som indeholder ξ, givet resten af processen x.

Proposition 5.2.
Hvis X er en Poisson proces med intensitetsfunktion ρ, så er

$$\lambda^*(x,\xi) = \rho(\xi).$$

Bevis. Fra Definition 5.1 gælder det, at

$$\lambda^*(x,\xi) = \frac{f(x\cup\xi)}{f(x)} = \frac{\exp(|S|-\mu(S))\prod_{\eta\in x\cup\xi}\rho(\eta)}{\exp(|S|-\mu(S))\prod_{\eta\in x}\rho(\eta)} = \frac{\prod_{\eta\in x}\rho(\eta)\rho(\xi)}{\prod_{\eta\in x}\rho(\eta)} = \rho(\xi),$$

Her gælder den miderste omskrivning fra [1] s.82 hvor tæller og nævner går ud med hinanden, mens $\prod_{\eta\in x\cup\{\xi\}}\rho(\eta)$ kan omskrives til $\prod_{\eta\in x}\rho(\eta)\rho(\xi)$. Dette skyldes, at produktet $\prod_{\eta\in x\cup\{\xi\}}\rho(\eta)$ består af alle elementerne i x samt det ekstra punkt ξ, hvilket kan opdeles i:

$$\prod_{\eta\in x\cup\{\xi\}}\rho(\eta) = \prod_{\eta\in x}\rho(\eta)\cdot\rho(\xi).$$

□

Yderligere så gælder det, at en punktproces kaldes *tiltrækkende* hvis følgende ulighed er opfyldt

$$\lambda^*(x,\xi) \le \lambda^*(y,\xi), \quad \text{når } x\subset y,$$

ellers kaldens den frastødende hvis følgende ulighed er opfyldt

$$\lambda^*(x,\xi) \ge \lambda^*(y,\xi), \quad \text{når } x\subset y.$$

En punktproces siges altså at være tiltrækkende, hvis sandsynligheden for at finde et punkt ved positionen ξ stiger, når der tilføjes flere punkter til den eksisterende konfiguration. Dette afspejles i første ulighed.

På samme måde siges processen at være *frastødende*, hvis sandsynligheden for at finde et punkt ved ξ falder, når flere punkter tilføjes. Dette beskrives med den anden uligheden.

Definition 5.3. Arvelig funktion
En funktion $h : \mathcal{N} \to [0, \infty)$ er *arvelige*, hvilket betyder, at

$$h(x) > 0 \implies h(y) > 0 \quad \text{for alle } y \subset x.$$

Her skal arveligheden af en funktion intuitivt forstås som, at hvis en egenskab gælder for en større mængde, så vil den samme egenskab også gælde for enhver mindre delmængde.

Nu hvor arvelighed er introduceret, så kan følgende Proposition introduceres.

Proposition 5.4.
Hvis f er arvelig, så er der en en-til-en korrespondance mellem f and λ^*.

Bevis. Fra Definitionen 5.1 ser vi, at når $f(x \cup \{\xi\}) = 0$, så er $\lambda^*(x, \xi)$ defineret som 0. I tilfælde hvor $\lambda^*(x, \xi) = 0$, så må det gælde, at $f(x) = 0$ eller $f(x \cup \{\xi\}) = 0$.

Hvis $f(x \cup \{\xi\}) = 0$, er vi færdige. Hvis derimod $f(x) = 0$, følger det af, at f er arvelig, og dermed er $f(x \cup \{\xi\}) = 0$. Fra definitionen 5.1 har vi også at følgende ligheder gælder

$$\lambda^*(x \setminus \{\xi_1\}, \xi_1) = \frac{f(x \setminus \{\xi_1\} \cup \{\xi_1\})}{f(x \setminus \{\xi_1\})} = \frac{f(x)}{f(x \setminus \{\xi_1\})}. \tag{5.2}$$

Hvis punkt konfigurationen er givet som $x = \{\xi_1, \ldots, \xi_n\}$, som har endelig antal elementer, da S er begrænset, og $f(x) > 0$. Da f er arvelig, kan vi multiplicere begge sider med nævneren og få

$$\begin{aligned} f(x) &= \lambda^*(x \setminus \{\xi_1\}, \xi_1) f(x \setminus \{\xi_1\}) \\ &= \cdots \\ &= \lambda^*(x \setminus \{\xi_1\}, \xi_1) \lambda^*(x \setminus \{\xi_1, \xi_2\}, \xi_2) \cdots \lambda^*(x \setminus \{\xi_1, \ldots, \xi_n\}, \xi_n) f(\emptyset) \\ &\propto \prod_{i=1}^{n} \lambda^*(x \setminus \{\xi_1, \ldots, \xi_i\}, \xi_i). \end{aligned} \tag{5.3}$$

Her gælder den sidste lighed, da $f(\emptyset)$ bare er en konstant og indgår i proportionalitetskonstanten.
Dette viser at der er en en-til-en korrespondance mellem f og λ^*.

□

Hvis f ikke var arvelig, så vil det være muligt at $f(x \setminus \{\xi_1\}) = 0$ for $f(x) > 0$ hvilket vil medføre, at der ikke er en en-til-en korrespondance mellem f og λ^* .

5.0.2 Parvis interaktions punktproces

Nu kan vi undersøge et vigtigt aspekt inden for punktprocesser. I tidligere afsnit blev Poisson punktprocesser introduceret, hvor punkterne er tilfældigt placeret og uafhængige af hinanden. Dog er dette sjældent tilfældet i den virkelige verden eller i observerede datasæt. Her påvirker punkterne ofte hinanden. Dette kan intuitivt illustreres gennem en skov: Træer i en skov er sjældent tilfældigt placeret, da de konkurrerer om ressourcer som næring og sollys. Som resultat vokser træerne ikke tæt på hinanden, hvilket fører til en frastødning mellem punkterne.
Den klasse af punktprocesser som nu vil blive introduceret fokuserer på hvordan punkter interagere med hinanden.

Definition 5.5. Parvis interaktionsproces
I mange anvendelser har vi en *parvis interaktionspunktproces*, givet ved

$$f(x) \propto \prod_{\xi \in x} \phi(\xi) \prod_{\{\xi,\eta\} \subseteq x} \phi(\{\xi, \eta\}),$$

hvor ϕ er en *interaktionsfunktion*, dvs. en ikke-negativ funktion, hvor højresiden i ovenstående udtryk er integrerbar med hensyn til en Poisson proces $\mathrm{Po}(S, 1)$.

Vi kan også introducere *interaktionsrækkevide*. Her refererer denne til den maksimale afstand inden for hvilken punkterne i processen påvirker hinanden. Den er angivet med R og er givet ved

$$R = \inf\{r > 0 : \text{for all } \{\xi, \eta\} \subset S,\ \phi(\{\xi, \eta\}) = 1 \text{ if } \|\xi - \eta\| > r\}.$$

Her vil det for en Poisson punktproces gælde, at $R = 0$ da $\phi(\{\xi, \eta\}) = 1$. Omvendt så vil det ligeledes gælde, at $\phi(\{\xi, \eta\}) \leq 1$ medfører frastødning, mens $\phi(\{\xi, \eta\}) \geq 1$ vil medføre klustering.

Proposition 5.6.
Hvis X er en parvis interaktionsproces, så er tæthedsfunktionen f arvelig hvor $f(x) > 0$ og $\xi \notin x$ hvor

$$\lambda^*(x, \xi) = \phi(\xi) \prod_{\eta \in x} \phi(\{\xi, \eta\}).$$

Bevis. Lad $x \subseteq y$, så følger det af Definition 5.5, at tæthedsfunktionen f er arvelig og at

$$\lambda^*(x, \xi) = \frac{f(x \cup \xi)}{f(x)} = \frac{\prod_{\eta \in x \cup \xi} \phi(\eta) \prod_{\{\eta,\omega\} \subseteq x \cup \xi} \phi(\{\eta, \omega\})}{\prod_{\eta \in x} \phi(\eta) \prod_{\{\eta,\omega\} \subseteq x} \phi(\{\eta, \omega\})}$$

$$= \frac{\prod_{\eta\in x}\phi(\eta)\phi(\xi)\prod_{\{\eta,\omega\}\subseteq x}\phi(\{\eta,\omega\})\prod_{\eta\in x}\phi(\{\eta,\xi\})}{\prod_{\eta\in x}\phi(\eta)\prod_{\{\eta,\omega\}\subseteq x}\phi(\{\eta,\omega\})}$$
$$= \phi(\xi)\prod_{\eta\in x}\phi(\{\eta,\xi\}).$$

□

Yderligere så siges en parvis interaktionsproces at være homogen hvis det førsteordens led $\phi(\xi)$ er konstant og at det andenordens led er invariant under translationer og rotation. Intuitivt kan det forstås som, at vi kan forskyde eller rotere punkterne uden at ændre deres sandsynligheden for konfigurationen.

Før en Markov punktproces kan defineres så skal et *nabolag* introduceres.

Definition 5.7. Nabolag
Lad $\sim$ være en relation på mængden S, der er refleksiv og symmetrisk altså

- $\xi \sim \xi$ for alle $\xi \in S$,
- hvis $\xi \sim \eta$, så gælder også $\eta \sim \xi$.

To punkter ξ og η siges at være naboer, hvis $\xi \sim \eta$. Nabolaget for et punkt $\xi \in S$ defineres som mængden:

$$N_\xi = \{\eta \in S : \eta \sim \xi\}.$$

Intuitivt kan nabolaget forstås som de punkter, der ligger tæt nok på til at have en relation med et punkt. Hvert punkt er altid i sit eget nabolag(refleksivitet), og hvis et punkt er nabo til et andet, gælder det også omvendt (symmetri).
Et eksempel på et nabolag eller nabopunkter vil være med et *R-lukket nabolag*, hvor man på baggrund af afstand

$$N_\xi = \{\eta \in S : \|\xi - \eta\| \leq R\}$$

kan angive om hvorvidt punkter er i nabolag eller ej. Her vil punkter være i samme R-lukket nabolag hvis de højst har R afstand fra hinanden.
Nu kan *Markov funktionen* introduceres.

Definition 5.8. Markov funktion
Lad $h : N_f \to [0, \infty)$ være en arvelig funktion. Funktionen h kaldes en Markov funktion (med hensyn til relationen $\sim$), hvis følgende gælder

- For enhver konfiguration $x \in N_f$ med $h(x) > 0$ og ethvert punkt $\xi \in S \setminus x$, afhænger forholdet
$$\frac{h(x \cup \{\xi\})}{h(x)}$$

udelukkende af de punkter i x, der er relateret til ξ, givet som mængden $x \cap N_\xi = \{\eta \in x : \eta \sim \xi\}$.

Yderligere så gælder det at en tæthedsfunktion h, defineret med hensyn til en Poisson proces $Po(S, 1)$ kaldes en *Markov tæthedsfunktion* hvis h er en Markov funktion. En punktproces med en Markov tæthedsfunktion betegnes som en Markov punktproces.

Vi siger, at funktionen $\phi : N_f \to [0, \infty)$ er en interaktionsfunktion, hvis $\phi(x) = 1$, hver gang der findes $\xi, \eta \in x$ med $\xi \not\sim \eta$.

Sætning 5.9. Hammersley-Clifford-Ripley-Kelly
En funktion $h : N_f \to [0, \infty)$ er en Markov funktion, hvis og kun hvis der findes en interaktionsfunktion ϕ, sådan at

$$h(x) = \prod_{y \subseteq x} \phi(y), \quad x \in N_f. \tag{5.4}$$

For $h(x) > 0$ gælder da

$$\lambda^*(x, \xi) = \prod_{y \subseteq x} \phi(y \cup \xi), \quad x \in N_f, \, \xi \in S \setminus x. \tag{5.5}$$

Bevis. Vi starter med at vise, at tæthedsfunktionen $h(x)$ fra ligning 5.4 medfører Papangelou-intensiteten fra ligning 5.5. Lad $x \in N_f$ med $h(x) > 0$, og lad $\xi \in S \setminus x$. Papangelou-intensiteten er da givet som:

$$\lambda^*(x, \xi) = \frac{h(x \cup \{\xi\})}{h(x)} = \prod_{y \subseteq x} \phi(y \cup \{\xi\}).$$

Antag, at $h(x) = \prod_{y \subseteq x} \phi(y)$, hvor $\phi(y) > 0$ for alle $y \subseteq x$. Vi ønsker at vise, at $h(x) > 0$ medfører, at $h(x') > 0$ for $x' \subseteq x$. For $x' \subseteq x$ gælder

$$h(x') = \prod_{y' \subseteq x'} \phi(y').$$

Da $y' \subseteq x' \implies y' \subseteq x$, og $\phi(y') > 0$ for $y' \subseteq x$, følger

$$h(x') = \prod_{y' \subseteq x'} \phi(y') > 0.$$

Hermed er h arvelig, da $h(x) > 0 \implies h(x') > 0$ for $x' \subseteq x$.

Nu bruger vi definitionen af Papangelou-intensiteten og indsætter $h(x)$:

$$\lambda^*(x,\xi) = \frac{h(x \cup \{\xi\})}{h(x)} = \frac{\prod_{y \subseteq x \cup \{\xi\}} \phi(y)}{\prod_{y \subseteq x} \phi(y)} = \prod_{y \subseteq x} \phi(y \cup \{\xi\}).$$

Hvis $\phi(y \cup \{\xi\}) = 1$ for $y \subseteq x$, hvor $y \cap N_\xi = \emptyset$, bidrager disse y ikke til produktet. Derfor afhænger

$$\lambda^*(x,\xi) = \prod_{y \subseteq x \cap N_\xi} \phi(y \cup \{\xi\}),$$

kun af punkterne i $x \cap N_\xi$, hvilket viser, at $h(x)$ opfylder den lokale Markov egenskab. Da h dermed er en arvelig tæthedsfunktion der også overholder den lokale Markov egenskab, så er den også en Markov tæthedsfunktion
Nu skal vi vise den anden vej. Vi antager nu at h er en Markov tæthedsfunktion. Vi definer interaktionsfunktionen ϕ som en gaffelfunktion

$$\phi(x) = \begin{cases} h(\emptyset), & \text{hvis } x = \emptyset, \\ 1, & \text{hvis der findes } \eta, \xi \in x \text{ med } \eta \nsim \xi, \\ \frac{h(x)}{\prod_{y \subset x} \phi(y)}, & \text{ellers.} \end{cases}$$

Nu ses det på de tre muligheder, hvor det vises at ligning 5.4 er opfyldt.

For den første mulighed så antager vi, at $h(x) = 0$ og

$$\prod_{y \subset x} \phi(y) = 0,$$

hvilket betyder, at værdien for vores tæthed er 0.

Hvis blot én $\phi(y) = 0$, vil hele produktet

$$\prod_{y \subseteq x} \phi(y) = 0,$$

hvilket sikrer, at

$$h(x) = 0 = \prod_{y \subseteq x} \phi(y).$$

For den anden mulighed antag nu, at $h(x) = 0$ og $\prod_{y \subset x} \phi(y) > 0$. Dette indebærer, at alle delmængder $y \subseteq x$ bidrager positivt til produktet, hvilket vil sige, at $\phi(y) > 0$ for alle $y \subseteq x$. Ved definitionen af Papangelou-intensiteten og ϕ gælder det dog, at

$$\lambda^*(x \setminus \{\kappa\}, \kappa) = 0, \quad \text{for alle } \kappa \in x,$$

da $h(x) = 0$. Samtidig gælder det, at $h(y) > 0$ for alle $y \subset x$, hvilket giver

$$\lambda^*(x \setminus \{\kappa, \xi\}, \xi) > 0, \quad \text{for alle } \{\kappa, \xi\} \subseteq x.$$

Hvis der eksisterer $\xi, \eta \in x$, hvor $\xi \not\sim \eta$, gælder det, at

$$0 = \lambda^*(x \setminus \{\eta\}, \eta) = \lambda^*(x \setminus \{\xi, \eta\}, \eta) > 0.$$

hvilket er en direkte modsigelse.

Af denne modsigelse konkluderer vi, at $\phi(x) = 0$, og dermed er $h(x) = 0$ givet ved

$$h(x) = \prod_{y \subseteq x} \phi(y) = 0.$$

For det tredje tilfælde ses på $h(x) > 0$: Da h er arvelig, gælder $h(y) > 0$ for alle $y \subseteq x$. For at vise, at $h(x)$ har formen givet i Ligning (5.4), anvender vi induktion over $n = n(x)$, hvor $n(x)$ er antallet af punkter i x.

Hvis $n = 0$ eller $\xi \sim \eta$ for alle $\xi, \eta \in x$ så følger det af definitionen af ϕ og induktionsantagelsen at $\phi(y) > 0$ for alle $y \subset x$ og $h(x) = \prod_{y \subseteq x} \phi(y)$.

Ved definitionen af λ^* gælder:

$$h(x) = \lambda^*(z \cup \{\xi\}, \eta) h(z \cup \{\xi\}) = \lambda^*(z, \eta) h(z \cup \{\xi\}) = \frac{h(z \cup \{\eta\})}{h(z)} h(z \cup \{\xi\}).$$

hvor $x = z \cup \xi, \eta$
Ved at kombinere dette med induktionshypotesen, har vi:

$$h(x) = \frac{\prod_{y \subseteq z \cup \{\eta\}} \phi(y)}{\prod_{y \subseteq z} \phi(y)} \prod_{y \subseteq z \cup \{\xi\}} \phi(y).$$

Dette kan omskrives til:

$$h(x) = \prod_{y \subseteq z} \phi(y \cup \eta) \prod_{y \subseteq z \cup \{\xi\}} \phi(y \cup \xi).$$

Eftersom $\phi(y) = 1$ for $\{\xi, \eta\} \subseteq y$, har vi:

$$h(x) = \prod_{y \subseteq x} \phi(y),$$

hvilket viser, at $h(x)$ har den ønskede form.

□

6 Anvendelse af punktprocesmodeller på redwoodfull datasæt

I følgende afsnit er disse hjemmesider blandt andet brugt [3] og [4].

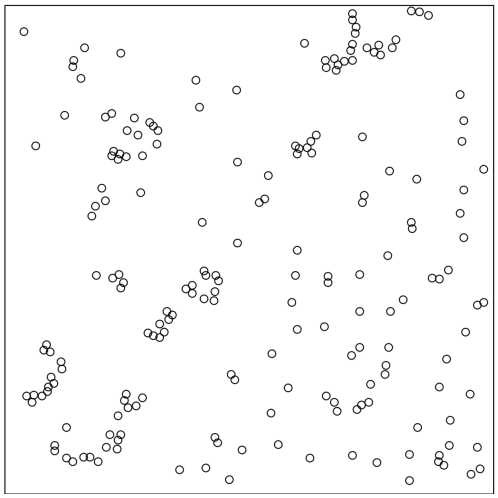
I denne del af projektet vil noget af den teori, der præsenteres tidligere, blive anvendt på rigtige data.

Det data som der bliver kigget på i denne del, er datasættet *redwoodfull* som angiver positionerne af 195 træer af typen California Giant Redwood.
Her vil der ses på K- og L - funktioner og se hvordan disse passer på det introduceret data. Tidligere i projektet er Poisson punktprocessen blevet introduceret, hvor den angiver en fuld tilfældig punktproces. I denne del af projektet vil to andre punktprocesser, som ikke er blevet nævnt tidligere i projektet, bruges. Der er tale om en Matérn cluster proces, som er et specialtilfælde af en Cox punktproces og Strauss punktproces. Disse vil blive simuleret, og deres K- og L-funktioner vil blive sammenlignet med datasættet. Til sidst sammenlignes resultaterne af de simulerede processer med hinanden. Disse punktprocesser er valgt, fordi de er gode eksempler på processer, der udviser henholdsvis klustering og frastødning.

Vi starter med at kigge på K- og L- funktionen for vores data.

```
1 install.packages("spatstat.data")
2 library(spatstat.data)
3 data(redwoodfull)
```

Listing 6.1: Indlæsning af det fulde Redwood Datasæt

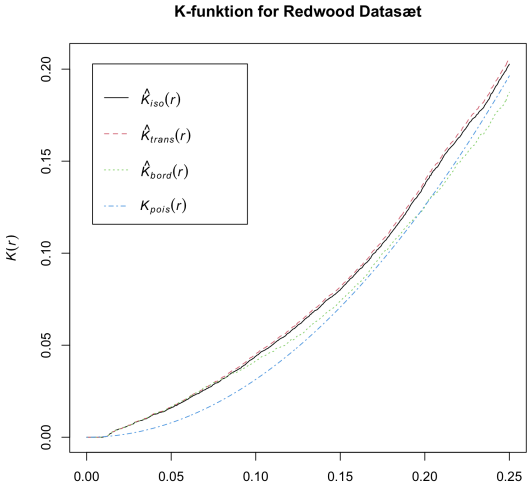


Figur 6.1: Det fulde datasæt redwoodfull

Som det kan ses på Figur 6.1 så er hvert træ erstattet med et punkt, men ud fra illustrationen er det svært at afgøre, om punkerne er tilfældige fordelt eller om de kluster eller frastøder hinanden. Dataen kan virke opdelt med en blanding af klustering og frastødning. For at kunne få en større indsigt i dette, kan vi se på K- og L-funktionerne og datasættet.

Vi ser først på K-funktionen.

```
1 k_funktion <- Kest(redwoodfull)
```



Figur 6.2: K-funktionen opdelt for datasættet

På Figur 6.2 vises forskellige metoder til estimering af K-funktionen, som bruges til at undersøge, hvor mange naboer et punkt har inden for en radius r. Til dette bruges *Kest*-

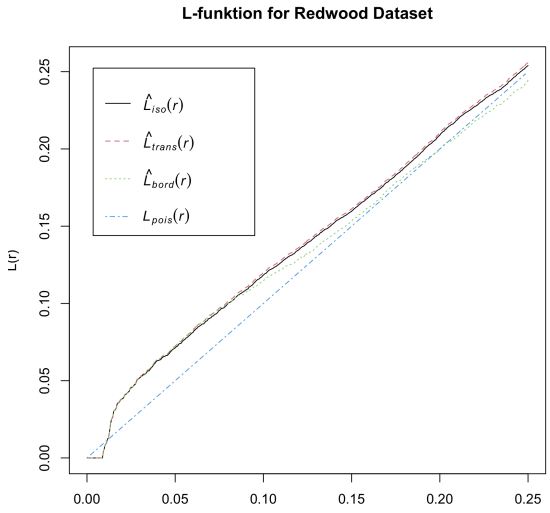
funktionen fra Rstudio.Her ses den isotropiske, translationskorrektionelle og ekskluderende metode på grafen, angivet med forskellige farver.

Den ekskluderende metode, angivet som $\hat{K}_{bord}$, ser på punkterne og fjerner dem, der ligger tæt på kanten af observationsvinduet, således at kuglen, hvori punkterne tælles, er helt indeholdt i vinduet. Dette indebærer, at kun punkter, der er fuldt inde i vinduet, og hvor hele deres radius r ligger inden for vinduets grænser, bidrager til estimeringen.

Den translationskorrektionelle metode og den isotropiske metode, angivet som $\hat{K}_{trans}$ og $\hat{K}_{iso}$, tager også højde for punkter tæt på kanten af observationsvinduet. For disse punkter bliver det mere kompliceret, fordi en del af cirklen med radius r vil være udenfor observationsvinduet og på den måde kan man ikke tælle præcis hvor mange naboer der er inden for cirklen.Hvis man gjorde observationsvinduet større, vil der muligvis være flere punker. Ved at anvende disse to metoder, prøver man altså kompencere for den manglende information ved at tilføje vægtning.
Den sidste er angivet som K_{pois} og er en K-funktion for en homogen Poisson proces, som fungerer som et referencepunkt som man kan sammenligne K-funktionen for vores datasæt. Ved at sammenligne disse to, kan det vurderes hvorvidt punkterne kluster eller frastøder hinanden sammenlignet med en Poisson proces.

På Figur 6.2 ses, at alle tilfælde at K-funktioner for det meste ligger en smule over referencepunktet, hvilket indikerer, at der er tegn på at punkterne har en tendens til at kluster sig sammen. Dette gælder især for små værdier af r. Jo større r, der betragtes, desto tættere kommer $\hat{K}_{Bord}$ på referencekurven, og det bemærkes, at $\hat{K}_{Bord}$ til sidst falder under referencekurven. Dette kan indikere en tendens til frastødning mellem punkterne på større skalaer. Analysen indikerer på, at træerne ikke er tilfældigt fordelt for små r, hvor klustering kan forekomme. Klustering kan for eksempel skyldes, at træerne er plantet tæt på hinanden med vilje, eller at næringsindholdet er særligt højt i visse områder og derfor vokser der kun træer i bestemte områder.
Det samme kan man gøre for L-funktionen. Her anvendes funktionen *Lest* som også allerede er en del af Rstudio.

```
1 l_funktion <- Lest(redwoodfull)
```

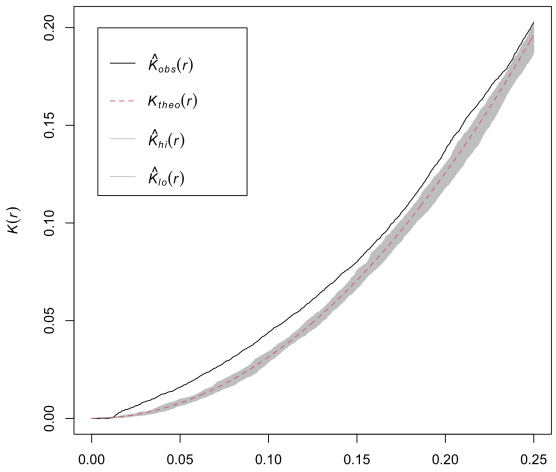


Figur 6.3: L-funktionen opdelt for datasættet

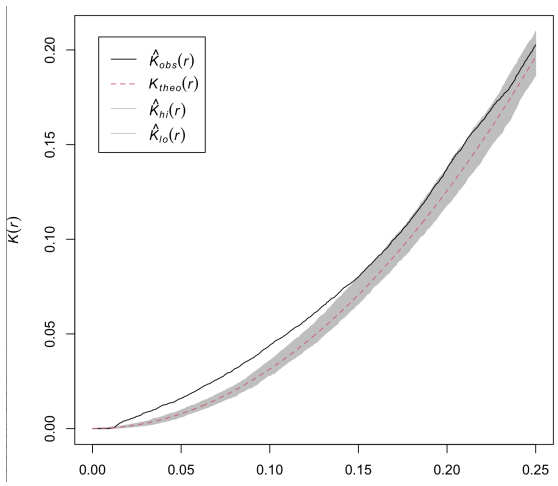
Fra Figur 6.3 ses noglelunde samme mønster ved små r-værdier. Her indikerer det også, at der er tale om en klustering, men jo større værdi af r, jo tættere kommes der på referencepunktet og igen ses det at $\hat{L}_{bord}$ faktisk ender med at ligge under referencepunktet og på den måde indikere en frastødning. Årsagen til dette kan skyldes, at den ekskluderende metode netop fjerner nogle punkter når man kommer langt ud, og reducerer den mængde af tilgængelig data tæt på kanten og på den måde kan det påvirke resultatet anderledes. Her ses også, at referencepunktet ser andereldes ud end for K-funktionen, og det skyldes, som nævnt tidligere i projektet, at $L(r) = r$ for en Poisson punktproces.

Som det ses på billederne, bruges den teoretiske homogene Poisson proces som referencepunkt. For at undersøge, om det observerede datasæt afviger signifikant fra tilfældighed, kan vi simulere en række homogene Poisson processer med samme intensitet som det observerede datasæt. Efter hver simulering beregnes en K-funktion, og for hver værdi af radius r findes en minimum- og maksimumværdi af K-funktionen. Disse værdier danner en *tilfældighedskonvolut*, som repræsenterer det forventede interval for K-funktionen under en tilfældig Poisson proces. Denne tilfældighedskonvolut bliver angivet som et grå område.

```
1 k_konvelut <- envelope(redwoodfull, Kest, nsim = 20, rank = 1)
```



Figur 6.4: Tilfældighedskonvelutten efter 20 simuleringer



Figur 6.5: Tilfældighedskonvelutten efter 200 simuleringer

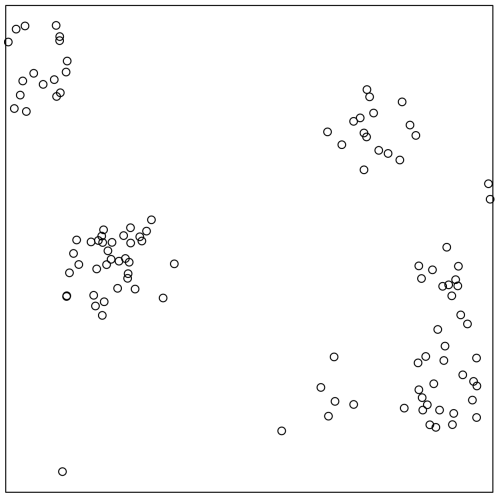
På Figur 6.4 og Figur 6.5 ses de simulerede tilfældighedskonvolutter for henholdsvis 20 og 200 simuleringer. For 20 simuleringer fremstår konvolutten smal og omslutter sig tæt omkring den teoretiske K-funktion. Dette medfører, at den observerede data for de fleste værdier af r afviger fra tilfældighed og indikerer en svag tendens til klustering. Omvendt viser konvolutten med 200 simuleringer en bredere struktur, hvilket gør den mere inkluderende for den observerede data, især for større r.

Ved hjælp af tilfældighedskonvolutten vurderes, om den observerede fordeling af punkter afviger signifikant fra en Poisson proces. For små værdier af r ligger $\hat{K}(r)$ udenfor konvolutten, hvilket understøtter en klusteringstendens. For større r falder $\hat{K}(r)$ indenfor konvolutten, hvilket indikerer, at punkterne her ikke skiller sig ud fra en tilfældig fordeling.

Nu kan der ses på de to andre punktprocesser og undersøges, hvordan de passer til datasættet. Vi starter med en *Matérn Cluster proces*, som anvender tre parametre: κ, radius og μ. Moderpunkterne fordeles som en homogen Poisson proces med intensitet κ, og omkring hvert moderpunkt placeres et stokastisk antal efterkommer punkter. Antallet af efterkommer punkter følger en Poisson fordeling med gennemsnit μ, mens deres positioner placeres inden for en given radius omkring moderpunktet. Denne proces skaber en kluster struktur, hvor parametrene κ, radius og μ styrer hvor tæt klusterne er, hvor stor frastødningen er og hvor mange punkter der forekommer i hver kluster.

Hvis vi simulerer en Matérn Cluster proces, kan vi se, at punkterne kluster en del mere sammen end for datasættet redwoodfull fra Figur 6.1.

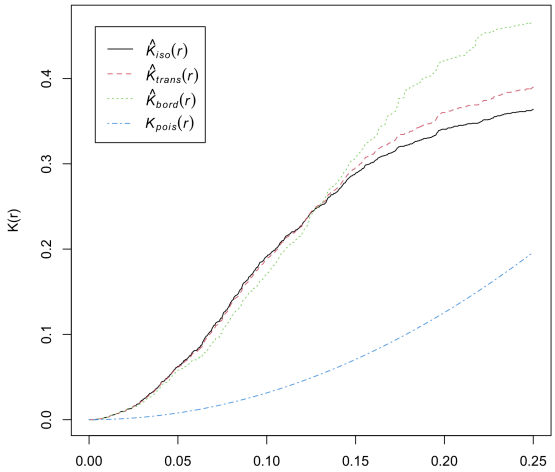
```
kappa <- 9
radius <- 0.1
mu <- 15
simulering <- rMatClust(kappa = kappa, r = radius, mu = mu)
k_funktion <- Kest(simulering)
```



Figur 6.6: Fordeling af punkter hvis vi simulerer en punktproces. Her er $\kappa = 9$, radius$= 0.1$ og $\mu = 15$

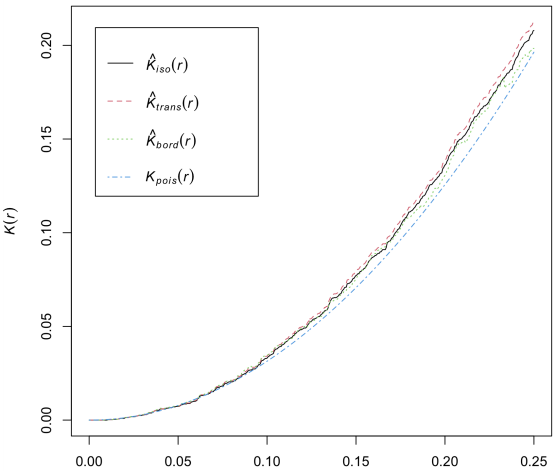
Nu kan vi se hvordan K-funktionen og L-funktionen vil se ud. Her vælger vi de samme parameterværdier.

```
k_funktion <- Kest(simulering)
```



Figur 6.7: K-funktion for en Matérn Cluster proces

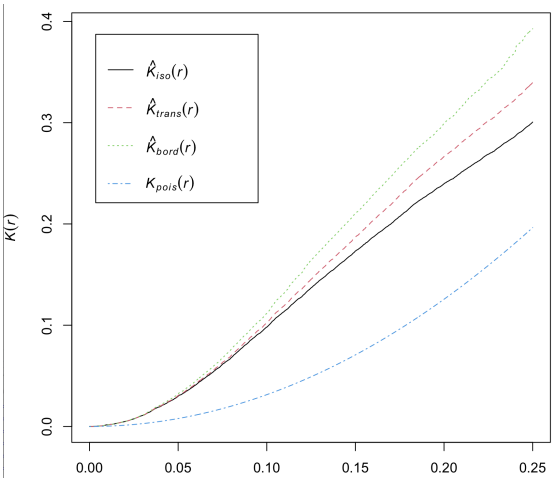
Her kan der tydeligt ses, hvordan K-funktionerne ligger langt over referencepunktet, og på den måde indikerer det en klustering. Dette afhænger stærk af hvilke parametre man vælger at anvende. Hvis man ændrer på radius, og øger den, så vil man kunne se at en Matérn Cluster proces begynder at ligne en Poisson proces, fordi punkterne spredes så meget, at klusterne næsten ikke er tydelige. Hvis radius sættes til 0.6 i stedet for 0.1 så ses netop dette.



Figur 6.8: K-funktion for en Matérn Cluster proces med radius = 0.6

Det samme gør sig også gældende hvis man øger κ. Punkterne bliver mere spredt på tværs af observationsvinduet, fordi flere moderpunkter betyder, at der er flere klusters at fordele punkterne over og derfor bliver klusterne mindre synlige. Omvendt hvis man øger μ så får man faktisk en mindre modsat effekt. Da μ regulerer hvor mange punkter der gennemsnitligt er omkrig et moderpunkt, og hvis man holder κ konstant, så øges ikke antal

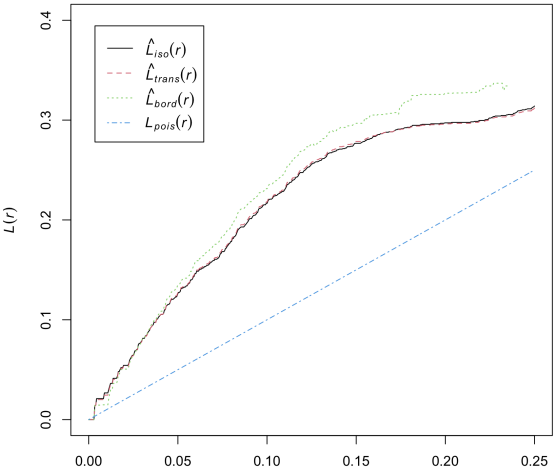
moderpunkter, men tætheden omkring enhver af de forskellige moderpunkter. Jo tættere de eksisterende moderpunkter bliver jo mere klustering forekommer der også.



Figur 6.9: K-funktion for en Matérn Cluster proces med parametrene $\kappa = 9$, $r = 0.1$ og $\mu = 30$

På Figur 6.9 ses netop denne øgening af $\mu = 30$ og ud fra grafen er det fortsat klart at der er tale om en klustering.

Det samme vil kunne ses hvis man ser på L-funktionen med de samme parameter værdier.

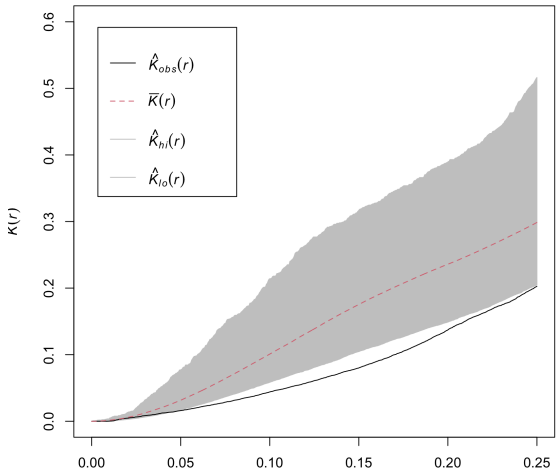


Figur 6.10: L-funktion for en Matérn Cluster proces med parametrene $\kappa = 9$, $r = 0.1$ og $\mu = 30$

Her kan det tydeligt ses, når man sammenligner med en Poisson proces, at der er tale om en klustering. Dog, ligesom før, bliver denne klustering mindre tydelig jo større radius bliver.

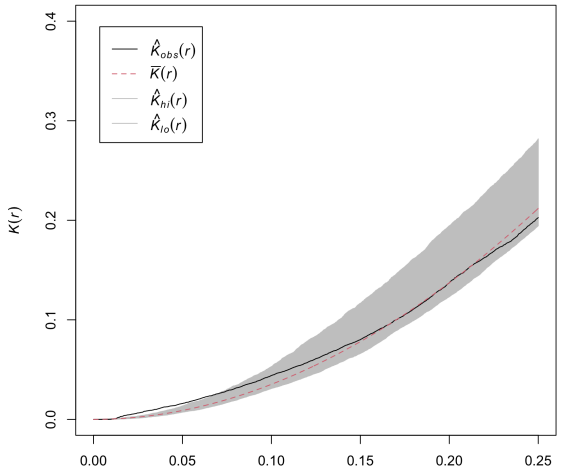
Nu kan Matérn Cluster processen sammenlignes med vores datasæt redwoodfull for at se, hvor god denne proces er til at forklare vores data. Her gør vi igen brug af tilfældighed-konvolut som i Figur 6.4 og 6.5 hvor der bliver brugt 200 simuleringer.

Vi starter med at se hvordan K-funktionen vil se ud fra forskellige valg af parametre og bagefter vil funktionen *kppm* fra R bruges til at finde de mest optimale parametre. Vi starter med de selvvalgte parametre $\kappa = 9$, $r = 0.1$ og $\mu = 15$.



Figur 6.11: K-funktion for en Matérn Cluster proces med parametrene $\kappa = 9$, $r = 0.1$ og $\mu = 15$ sammenlignet med den obsereveret data fra redwoodfull ved hjælp af tilfældighedskonvelut.

Fra Figur 6.11 kan det ses, at den observerede data ligger under tilfældighedskonvelutten for de fleste r og på den måde, så virker det ikke til, at de valgte parametre er så gode til at forklare dataen fra datasættet. Ligeledes kan det prøves med nyt udvalg af parametre.

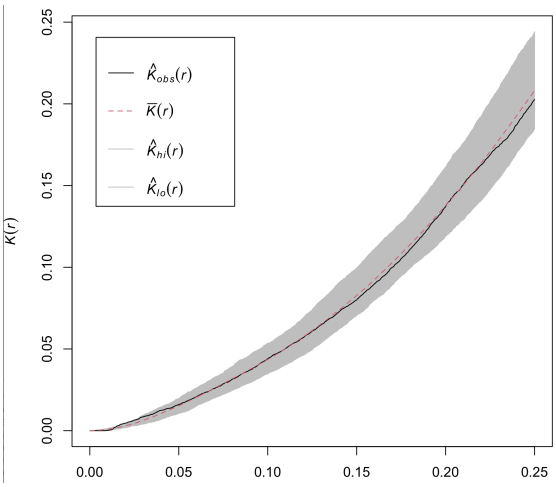


Figur 6.12: K-funktion for en Matérn Cluster proces med parametrene $\kappa = 20$, $r = 0.3$ og $\mu = 15$ sammenlignet med den obsereveret data fra Redwoodfull.

På Figur 6.12 kan det ses, at denne valg af parametre giver en bedre model, idet den observerede K-funktion for vores data for det meste ligger inden for tilfældighedskonvolutten. Ved at øge værdierne for κ og r skaber vi flere klynger, hvor punkterne i hver klynge spredes mere omkring moderpunktet. Dette resulterer i en model, der viser mindre markant klustring. Som det fremgår af grafen, passer denne model bedre til det observerede datasæt.

Nu kan vi bruge `kppm` til at finde mulige parametre til det data, vi har. Funktionen `kppm` bruger her *Maximum Likelihood Estimation* (MLE) og maksimerer sandsynligheden for at observere dataene givet Matérn Cluster modellen. MLE gør dette ved at opstille en sandsynlighedsfunktion, der beskriver, hvor sandsynligt det er at observere de givne data for forskellige parameterværdier. Derefter finder den de parametre, der maksimerer denne sandsynlighed.

```
tilpas_matclust <- kppm(redwoodfull, ~1, "MatClust")

summary(tilpas_matclust)
konvelut_tilpas <- envelope(tilpas_matclust, Kest, nsim = 200,
   global = FALSE)
```



Figur 6.13: K-funktion for en Matérn Cluster proces ved anvendelse af kppm.

Når `summary(tilpas_matclust)` bliver kørt, kan vi se, at de mest optimale parametre bliver sat til at være: $\kappa = 84.12$, $r = 0.0483$ og $\mu = 2.3182$. Det bliver repræsenteret i Figur 6.13. Sammenlignes det med Figur 6.12 og Figur 6.11 kan det ses, at de nye parametre kan beskrive den observeret data en del bedre end de selvvalgte parametre. Det gør sig gældende da vi for alle værdier af r, stadig ligger inde i tilfældighedskonveulutten.

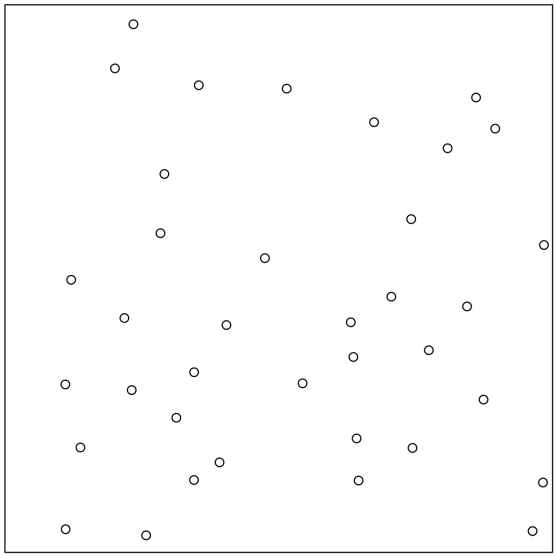
Nu kan der ses på en punktproces der er frastødende og hvordan den beskriver vores datasæt. Vi ser nu på Strauss punktprocess som er et eksempel på en Markov punktproces, da den opfylder Markov egenskaben.

Tætheden for en Strauss-proces er givet ved følgende formel

$$f(x) \propto \beta^{n(x)}\gamma^{s_R(x)},$$

hvor $n(x)$ er antallet af punkter, $\beta > 0$ er intensitetsparameteren, $\gamma \in [0, 1]$ er interaktionsparameteren, og $s_R(x)$ er antallet af punktpar inden for en given interaktionsradius R.

I Strauss-processen betyder parameteren $\gamma < 1$, at sandsynligheden for konfigurationen reduceres, jo flere punktpar der findes inden for interaktionsradiusen R. Dette resulterer i et mønster, hvor punkterne "holder afstand"fra hinanden, da sandsynligheden for hele konfigurationen bliver lavere, når punkter samles tættere.



Figur 6.14: En Strauss punktproces med $\beta = 100$, $\gamma = 0.3$ og r=0.1

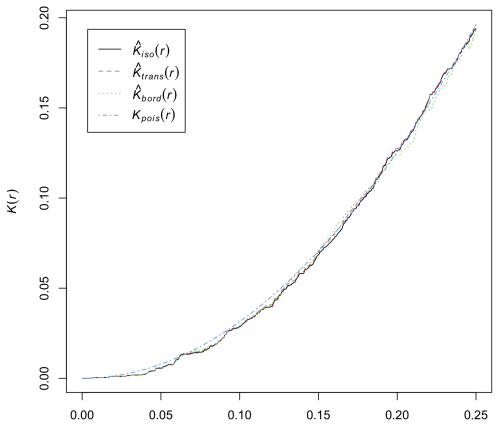
Det er vigtigt at bemærke, at Strauss processen ikke er velegnet til at modellere klustring. Selvom en værdi af $\gamma > 1$ teoretisk kunne indikere en form for klustering mellem punkter, fører dette til, at sandsynlighedstætheden bliver ikke-integrerbar og dermed matematisk ugyldig. Det skyldes at højre siden kan stige og divegere. Strauss processen er designet til at beskrive frastødning, hvor $\gamma \leq 1$, og kan derfor ikke håndtere klustringsmønstre.

Nu kan vi opskrive det i R og se hvordan K- og L-funktionen vil se ud for en Strauss punktproces. Vi starter med at simulere en Strauss proces ved hjælp af Metropolis-Hastings algoritmen.

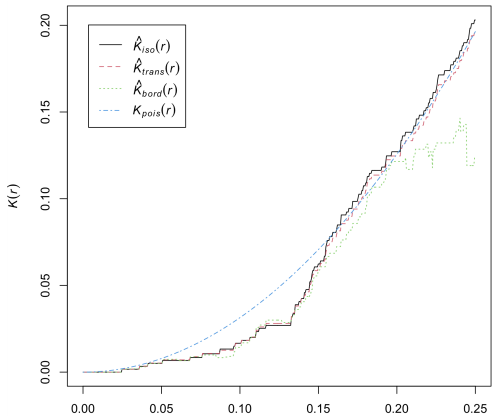
```
r <- 0.1
gamma <- 0.9
beta <- 100

strauss_model <- rmhmodel(cif = "strauss", par = list(beta = beta,
    gamma = gamma, r = r), w = obs_window)


sim_strauss <- rmh(model = strauss_model, control = list(nrep =
    10000))
```

Nu findes K-funktionen.

```
K_sim <- Kest(sim_strauss)
```



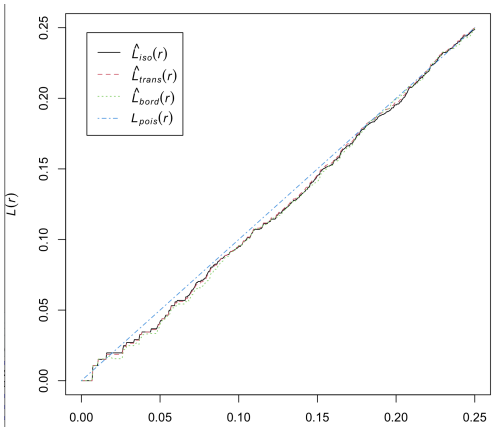
Figur 6.15: K-funktion for en Strauss proces med parametrene $r = 0.1$, $\gamma = 0.9$ og $\beta = 100$.



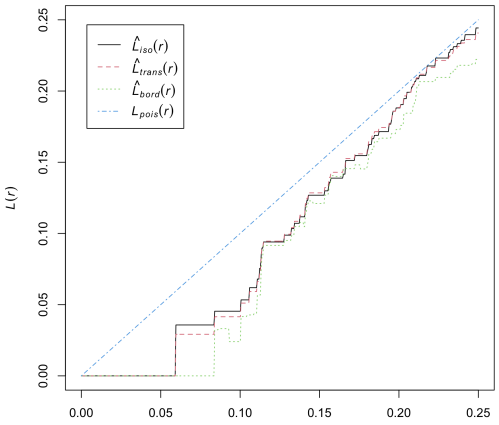
Figur 6.16: K-funktion for en Strauss proces med parametrene $r = 0.1$, $\gamma = 0.1$ og $\beta = 100$.

Ud fra Figur 6.15 og Figur 6.16 ses en tydelig forskel mellem graferne. Som nævnt tidligere, vil en mindre værdi af γ føre til en større indikation på frastødning, især for små værdier af r. For Figur 6.15, hvor $\gamma = 0.9$, ligger værdien tæt på 1, hvilket afspejler, at K-funktionen følger referencekurven for en Poisson punktproces. Når γ reduceres betydeligt, som i Figur 6.16 med $\gamma = 0.1$, bliver frastødningen tydelig for små r, hvilket illustreres ved, at K-funktionen ligger under Poisson-referencekurven.
Det samme kan ses for en L-funktion med de samme parametre.

```
L_sim <- Lest(sim_strauss)
```



Figur 6.17: L-funktion for en Strauss proces med parametrene $r = 0.1$, $\gamma = 0.9$ og $\beta = 100$.



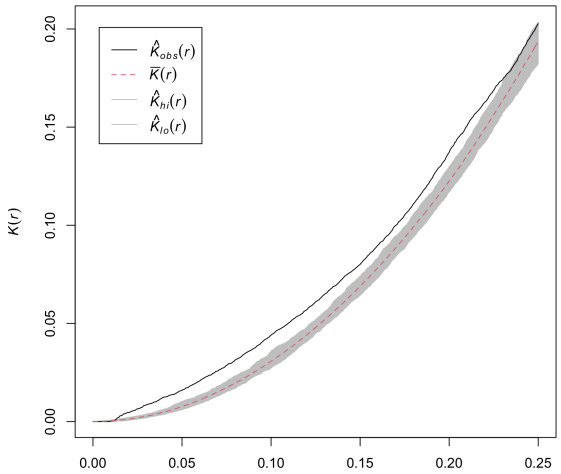
Figur 6.18: L-funktion for en Strauss proces med parametrene $r = 0.1$, $\gamma = 0.1$ og $\beta = 100$.

Nu kan vi anvende de mest opimale parametre. Her kan vi anvende funktionen *ppm* fra Rstudio til at estimere disse parametre.

```
1
2 R <- 0.2
3 strauss_fit <- ppm(redwoodfull, ~1, Strauss(R))
4 summary(strauss_fit)
```

Her bruges ppm til at vælge de bedst mulige parametre. Ved at køre `summary(strauss_fit)` kan det ses, at det bedste parametre vil være givet ved $\beta = 553.345$ og $\gamma = 0.9610556$. Hvis man anvender en tilfældighedskonvelut, kan vi nu vurdere hvor godt Strauss punktprocessen passer til vores datasæt.

```
1 K_redwood <- Kest(redwoodfull)
2 sim_strauss <- simulate(strauss_fit, nsim = 1)
3 K_sim_strauss <- Kest(sim_strauss[[1]])
4
5 envelope_strauss <- envelope(strauss_fit, Kest, nsim = 200)
```



Figur 6.19: K-funktion for en Strauss punktproces ved anvendelse af parametrene valgt af ppm funktionen.

Det kan ses, at den observerede data ligger over tilfældighedskonvolutten for Strauss processen. Dette indikerer, at Strauss modellen ikke er velegnet til at beskrive det observerede datasæt. Årsagen er, at modellen udviser mindre klustring, end hvad der observeres i datasættet. Strauss processen er designet til at modellere frastødning mellem punkter og kan derfor ikke håndtere den klustring der ligger i datasættet.

Ved sammenligning af Figur 6.5, Figur 6.13, og Figur 6.19 fremgår det også, at Poisson modellen ikke er en god beskrivelse af dataene. Selvom Poisson modellen opfanger mønsteret bedre end Strauss processen for større r, passer den stadig ikke til datasættet, da den ikke fanger klustringseffekterne ved små r.

Baseret på denne sammenligning kan det konkluderes, at Matérn Cluster modellen fra Figur 6.13 giver den bedste beskrivelse af dataene. Dette ses ved, at den observerede data i denne model ligger inden for tilfældighedskonvolutten. Matérn Cluster modellen er bedre i stand til at opfange de klustringsmønstre, der er karakteristiske for datasættet redwoodfull, hvilket gør den mere passende end både Strauss og Poisson modellen.

7 | Konklusion

I dette projekt blev først grundlæggende begreber om punktprocesser gennemgået, og det blev vist, hvordan de kan bruges til at beskrive fordelingen af punkter i tid og rum. Der blev også set på forskellen mellem endimensionale og flerdimensionale punktprocesser, samt hvordan man kan tilføje ekstra information til punkterne i mærkede punktprocesser.

Derefter blev Poisson processen introduceret. Dette er en simpel model for tilfældigt fordelte punkter. Noget vigtigt, som også blev anvendt til sidst i projektet, var introduktionen af K- og L-funktionerne. Disse hjælper med at afgøre, om punkterne kluster sig sammen eller frastøder hinanden. Ved at anvende disse funktioner i samspil med Poisson processen kan man vurdere, hvorvidt de data, man arbejder med, viser klustering, frastødning eller er tilfældigt fordelt.

Til sidst blev teorien afprøvet på datasættet redwoodfull, som indeholder positioner af 195 træer. Det kunne i projektet ses, at Poisson modellen ikke kunne forklare de mønstre, der blev observeret i datasættet, da der var tegn på klustering. Det samme gjorde sig gældende for Strauss processen. I stedet kunne det ses, at Matérn Cluster processen gav en langt bedre beskrivelse af dataene. Der kan derfor konkluderes, at det er vigtigt at vælge en korrekt model, der kan håndtere det datasæt, man arbejder med.

8 References

[1] Waagepetersen, Rasmus & Møller, Jesper. *Spatial Point Processes and their Applications*. A CRC Press Company, 2004.

[2] Baddeley, Adrian. *Spatial Point Processes and their Applications*. Springer, Berlin, Heidelberg, School og Matematics and Statistics, University of Western Australia, 2007.

[3] H. Paul Keeler. Simulating a matern cluster point process, 2018. URL `https://hpaulkeeler.com/simulating-a-matern-cluster-point-process/`.

[4] R-Project. Fit point process model to data. URL `https://search.r-project.org/CRAN/refmans/spatstat.model/html/ppm.html`.