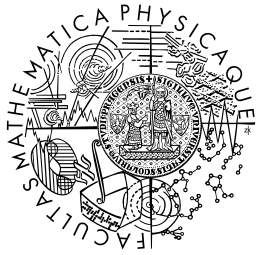


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Eliška Hájková

Shlukové bodové procesy s rodičovskými body

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Jiří Dvořák, Ph.D.
Studijní program: Matematika
Studijní obor: Obecná matematika

Praha 2022

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Ráda bych poděkovala RNDr. Jiřímu Dvořákovi, Ph.D. za uvedení do teorie bodových procesů, za poskytnuté cenné rady, odborný dohled, laskavost a především za čas, který mi v průběhu celé bakalářské práce velmi ochotně věnoval. Také bych ráda poděkovala prof. Aila Särkkä z Chalmers University of Technology and University of Gothenburg ve Švédsku za poskytnutí reálných dat o epidermálních nervových vláknech pro naší práci.

Název práce: Shlukové bodové procesy s rodičovskými body

Autor: Eliška Hájková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Jiří Dvořák, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V této práci představujeme základní pojmy pro bodové procesy a dvě metody odhadu parametrů Thomasové procesu. Zaprvé metodu minimálního kontrastu, která se používá pro situaci, kdy neznáme polohu rodičovských bodů, a zadruhé námi odvozený postup využívající informace o poloze rodičovských bodů. Pomocí simulací v programu R zjišťujeme, že naše metoda odhaduje zmíněné parametry lépe vzhledem k relativnímu vychýlení a relativní střední čtvercové chybě. Následně oběma metodami odhadujeme parametry reálných dat. Na závěr obálkovým testem testujeme, zda naše data odpovídají Thomasové procesu s parametry odhadnutými z dat zmíněnými metodami. Pro žádné získané odhady parametrů dat hypotézu nezamítáme.

Klíčová slova: bodový proces, shlukování, Thomasové proces, odhady parametrů, rodičovské body

Title: Non-orphan cluster point processes

Author: Eliška Hájková

Katedra: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Jiří Dvořák, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this work we introduce some basic concepts from the theory of spatial point processes and two methods of estimation of the parametres for Thomas process. Firstly the method of minimum contrast, that is used for situations when we do not know the position of parent points and secondly our proposed method using information about the parent points location. Using simulations in program R, we find out that our method estimated mentioned parameters better considering the relative bias and relative mean squared error. Subsequently, we estimate the parameters of the real data using both methods. Finally we test by Global Envelope Test, whether our data match Thomas process with parameters estimated from data with mentioned methods. For the combinations of parameters obtained by discussed methods we do not reject the hypothesis.

Keywords: point process, clustering, Thomas process, parameter estimation, parent points

Obsah

Seznam použitého značení

$\mathbb{R}^d$	d-rozměrná množina všech reálných čísel
$\mathbb{N}$	množina všech přirozených čísel
E	euklidovský prostor
(E, ρ)	úplný separabilní metrický prostor, v němž je každá omezená uzavřená podmnožina kompaktní
$\mathcal{B}(E)$	systém borelovských podmnožin na prostoru E
$\mathcal{B}_0(E)$	systém omezených borelovských podmnožin na prostoru E
$\mathcal{M}$	množina všech lokálně konečných měr na $(E, \mathcal{B})$
$\mathcal{N}$	množina všech lokálně konečných měr nabývajících pouze celočíselných hodnot na $(E, \mathcal{B})$
$\mathfrak{M}, \mathfrak{N}$	σ-algebra na $\mathcal{M}$ a $\mathcal{N}$
$\mathcal{N}^*$	množina všech měr z $\mathcal{N}$ nabývajících hodnoty maximálně 1 pro jednobodové množiny
$(\Omega, \mathcal{A}, \mathbb{P})$	pravděpodobnostní prostor
X	bodový proces
$X(B)$	počet prvků v B vzhledem k míře X
S	prostor kót
W	pozorovací okno
W_{-r}	zmenšené pozorovací okno o vzdálenost r
∂W	hranice množiny W
$\lambda(x)$	funkce intenzity bodového procesu
λ	intenzita stacionárního bodového procesu
κ, μ, σ	parametry Thomasové procesu
K, F	K-funkce a F-funkce
$b(x, r)$	uzavřená koule se středem x a poloměrem $r > 0$
o	počátek v $\mathbb{R}^2$
$dist(x, X)$	nejkratší vzdálenost mezi bodem x a procesem X, neboli $\inf\{\|\|x - y\|\|, y \in X\}$, inf je infimum
$\|A\|$	Lebesgueova míra množiny A
$\mathbf{1}$	indikátor
$\mathbb{E}$	střední hodnota
$\overline{X}_n$	výběrový průměr z X_i
$N(0, \sigma^2)$	normální rozdělení se střední hodnotou 0 a rozptylem σ^2
$L(0, \sigma)$, $\ell(0, \sigma)$	věrohodnostní funkce a logaritmus věrohodnostní funkce
$b(\theta)$, $rb(\theta)$	vychýlení a relativní vychýlení odhadu θ
MSE, $rMSE$	střední čtvercová chyba a relativní střední čtvercová chyba odhadu
MK	metoda minimálního kontrastu používající K-funkci
NO_1	odhady námi odvozeným postupem bez mínusové korekce
NO_2	odhady námi odvozeným postupem s mínusovou korekcí

Úvod

Bodové procesy mají širokou škálu uplatnění, neboť dobře znázorňují realitu. Teorii bodových procesů můžeme uplatnit například v zemědělství k zaznamenávaní polohy vzácného druhu rostlin v parku, ve zdravotnictví ke zkoumání nervových vláken nebo ve službách k zaznamenávání časů příchodů hostů do restaurace. Tuto reálnou situaci prezentujeme jako realizaci bodového procesu a dále můžeme zkoumat vlastnosti této situace pomocí různých metod pro bodové procesy.

Cílem této práce je analyzovat reálná data a odhadnout parametry modelu, ze kterého data mohou pocházet, pomocí známé metody minimálního kontrastu a námi odvozeným postupem.

Na úvod naší práce se seznámíme se základními pojmy z oblasti teorie bodových procesů. Mezi nejdůležitější pojmy naší práce budou patřit non-orphan procesy, Thomasové proces a K-funkce, s kterými budeme dále pracovat. Non-orphan proces je proces se dvěma typy bodů, prvním typem jsou rodičovské body a druhým typem dceřiné body, které náleží některému z rodičovských bodů.

V druhé kapitole si ukážeme, jak se odhadují parametry κ, μ a σ Thomasové procesu. Jedna z metod je metoda minimálního kontrastu, která odhaduje zmíněné parametry pomocí K-funkce, pokud máme informace pouze o dceřiných bodech. Avšak někdy můžeme mít informace i o rodičovských bodech, díky čemuž můžeme zkusit odhady parametrů zlepšit. K této situaci si odvodíme vlastní postup, jak parametry odhadnout.

Dále chceme porovnat odhady pomocí minimálního kontrastu a pomocí našeho postupu vzhledem k vychýlení a střední čtvercové chybě. To uskutečníme pomocí simulací v programu `R`. Ukázku našeho kódu pro tyto simulace přikládám v elektronické příloze.

Následně se seznámíme s reálným vzorkem dat, který odpovídá struktuře epidermálních nervových vlákem jednoho jedince. Na tato data aplikujeme metody z 2. kapitoly a odhadneme zmíněné parametry.

Ještě je potřeba ověřit, zda odhadnutý model dobře popisuje pozorovaná data. Toto provedeme pomocí obálkového testu s nulovou hypotézou, že naše data pocházejí z Thomasové procesu s parametry reálných dat, které jsme odhadli ve 4. kapitole.

1. Základní definice a pojmy

Pro intuitivní pochopení pojmu **bodový proces** si můžeme představit prostor, například $\mathbb{R}$ či $\mathbb{R}^2$, ve kterém jsou množiny s body. Pokud pro tyto body platí, že se v každé uzavřené a omezené podmnožině tohoto prostoru nachází pouze konečný počet bodů, pak můžeme hovořit o bodovém procesu. Konkrétně o jednoduchém bodovém procesu (viz definice 2), což nám ale stačí, neboť se v této práci budeme zabývat pouze těmito jednoduchými bodovými procesy. Nás by zajímalo, jaké má takový proces vlastnosti, například zda body tvoří shluky, zda se body vyskytují spíše na jedné straně prostoru, zda je nějaký bod ve vzdálenosti $r > 0$ od konkrétního bodu nebo zda mají jednotlivé body nějaké specifické vlastnosti. Avšak v praxi obvykle nepozorujeme celý prostor E, ale pouze část prostoru, kterou budeme nazývat **pozorovacím oknem** a v celé práci ho budeme značit jako W. Takovýto bodový proces s pozorovacím oknem W může dobře znázorňovat realitu.

Příklad. Mezi užitečné bodové procesy v $\mathbb{R}$ patří bodový proces zaznamenávající časy, kdy dojde k určité události, například čas příchozího hovoru v hotelu, či čas kdy, vkročí zákazník do obchodu. Pozorovacím oknem je zde typicky úsečka (viz obrázek 1.1 vlevo).

Příklad. V $\mathbb{R}^2$ a $\mathbb{R}^3$ se bodové procesy často využívají k zaznamenání lokace. Například ke znázornění epicenter zemětřesení, lokací galaxií, míst výskytu určitého typu koronaviru, či umístění domů, ze kterých se uskutečnil hovor. Pozorovacím oknem je zde nějaká kompaktní množina (viz obrázek 1.1 vpravo).

Příklad. Další situace, kterou můžeme modelovat bodovým procesem, je prostorové znázornění polohy vzniku požárů, které můžeme vidět na obrázku 1.2. Data jsou pojmenovaná jako `nbfires` a jsou obsažena v knihovně `spatstat` pro program `R`. Vznikla ze záznamů poskytnutých New Brunswick Department of Natural Resources o požárech spadajících pod jejich správu. Na obrázku 1.2 můžeme vidět, že máme více druhů bodů, takovéto bodové procesy se nazývají **kótované bodové procesy** a budeme se jim věnovat níže (viz definice 5). Dále stojí za povšimnutí, že pozorovací okno může mít jakýkoliv tvar, pokud je to kompaktní množina.

1.1 Bodový proces

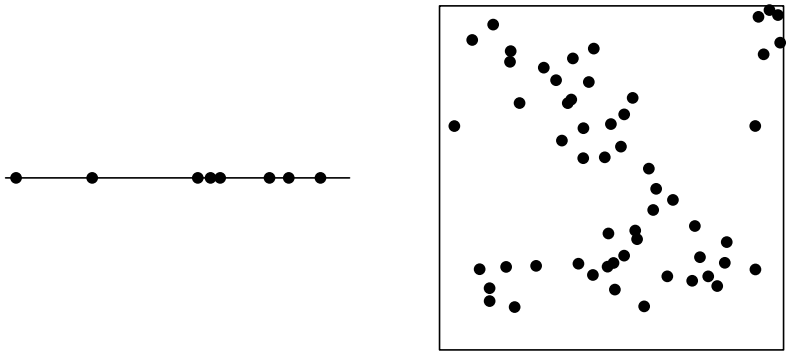
Následně si uvedeme formální definici bodového procesu a souvisejících pojmů. Definice a pojmy níže jsme čerpali převážně ze zdrojů Baddeley (2007), Rataj (2006), Daley a Vere-Jones (1988) a Andersson (2016).

Buď (E, ρ) úplný separabilní metrický prostor, v němž je každá omezená uzavřená podmnožina kompaktní. V našem textu budeme uvažovat jako E euklidovský prostor, konkrétně $\mathbb{R}^2$ pokud neuvedeme jinak. V euklidovských prostorech jsou množiny kompaktní právě tehdy, když jsou omezené a uzavřené, tedy druhá podmínka je automaticky splněna. Dále si uvedeme základní značení a definice:

$$\mathcal{B}(E) \text{ ... systém borelovských podmnožin na } E,$$

$$\mathcal{B}_0(E) \text{ ... systém omezených borelovských podmnožin na } E.$$

Definice 1 (lokálně konečná míra)**.** *Míra μ (tj. nezáporná σ-aditivní množinová funkce) na $(E, \mathcal{B})$ je* **lokálně konečná***, jestliže je konečná na $\mathcal{B}_0$.*



Obrázek 1.1: Na levém obrázku vidíme realizaci bodového procesu v jedné dimenzi. Na pravém obrázku ve dvou dimenzích.

Dále

$$\mathcal{M} \equiv \mathcal{M}(E) = \{\mu : \mu \text{ je lokálně konečná na } (E, \mathcal{B})\}, \tag{1.1}$$

$$\mathcal{N} \equiv \mathcal{N}(E) = \{\mu \in \mathcal{M} : \mu(B) \in \mathbb{N} \cup \{0, \infty\} \text{ pro každou } B \in \mathcal{B}\}. \tag{1.2}$$

Tímto jsme zavedli množinu všech lokálně konečných měr na $(E, \mathcal{B})$ a množinu všech lokálně konečných měr nabývajících pouze celočíselných hodnot. Ještě uvedeme značení pro nejmenší σ-algebru na $\mathcal{M}$ pomocí vzorů měřitelných množin a značení pro dvě σ-algebry na $\mathcal{M}$ a $\mathcal{N}$:

$$\begin{aligned} &\sigma\{\{\mu \in \mathcal{M} : \mu(B) \le r\}, B \in \mathcal{B}, r \ge 0\}, \\ &\mathfrak{M} = \sigma\Big\{\mu \to \mu(B) \text{ měř., } B \in \mathcal{B}\Big\}, \\ &\mathfrak{N} = \Big\{M \cap \mathcal{N} : M \in \mathfrak{M}\Big\}. \end{aligned} \tag{1.3}$$

Definice 2 (bodový proces)**. Bodový proces** *na E je měřitelné zobrazení*

$$X : (\Omega, \Sigma, \mathbb{P}) \to (\mathcal{N}, \mathfrak{N}),$$

kde Ω je stavový prostor, Σ je σ-algebra na Ω, $\mathbb{P}$ je pravděpodobnost na Ω, $(\mathcal{N}, \mathfrak{N})$ je měřitelný prostor všech lokálně konečných celočíselných měr.

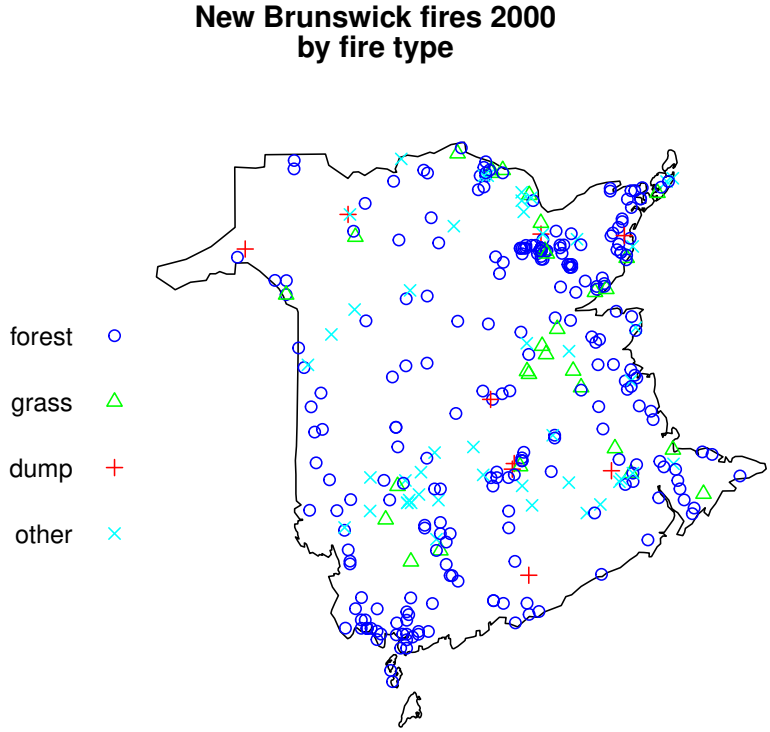
Bodový proces je **jednoduchý**, *jestliže $\mathbb{P}[X \in \mathcal{N}^*] = 1$, kde*

$$\mathcal{N}^* = \{v \in \mathcal{N} : v(\{x\}) \le 1 \text{ pro každé } x \in E\}.$$

Pro jednoduchý bodový proces platí, že míra z $\mathcal{N}$ každé jednobodové množiny je rovna maximálně 1. Pokud vezmeme náhodnou lokálně konečnou množinu, tak každý její bod bude mít míru 1, nebo 0, neboli tento bod v naší množině buď bude ležet (když bude mít míru 1), nebo nebude. Tedy můžeme ztotožnit míru (čítací) s jejím nosičem. Tím pádem můžeme psát, jak je zvykem v prostorové statistice, $x \in X$, což znamená, že x je prvkem nosiče náhodné míry.

Dále budeme psát

$$X(B) \text{ ... počet prvků v } B \text{ vzhledem k míře } X,\ B \in \mathcal{B}, \tag{1.4}$$



Obrázek 1.2: Realizace bodového procesu znázorňující typ požáru a místo jeho vzniku. Tmavě modrá kolečka jsou lesní požáry, zelené trojúhelníky travnaté požáry, červené plusy požáry skládek a světle modré křížky ostatní požáry.

tedy že míra X na množině B změří počet prvků v B. Díky ztotožnění atomické míry (čítací) a jejího nosiče můžeme realizaci bodového procesu chápat jako neuspořádanou množinu bodů $x = \{x_1, \ldots, x_n\}$, kde $x_i \in W$ značí polohu pozorovaného bodu, W je uzavřená omezená podmnožina E, kterou budeme nazývat **pozorovacím oknem**, $n \in \mathbb{N}$.

Poznámka. V celé práci budeme uvažovat jednoduché bodové procesy, tedy pravděpodobnost, že dva různé body budou ležet na stejném místě bude rovna nule.

Nyní ještě doplníme dvě definice k bodovým procesům, které se nám budou později v práci hodit.

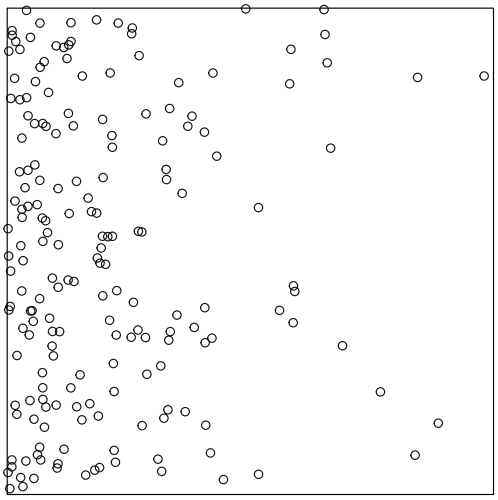
Definice 3 (funkce intenzity bodového procesu)**.** *Nechť X je bodový proces na $\mathbb{R}^2$.* **Funkce intenzity** *je měřitelná nezáporná funkce $\lambda(x)$ splňující (pokud taková funkce existuje)*

$$\mathbb{E}X(B) = \int_B \lambda(x)\mathrm{d}x$$

pro všechny borelovské množiny B.

Definice 4 (stacionární bodový proces)**.** *Bodový proces X v $\mathbb{R}^2$ je* **stacionární***, pokud pro každý fixní vektor $v \in \mathbb{R}^2$ je rozdělení posunutého bodového procesu $X + v$ (získaný posunutím každého bodu $x \in X$ o v, tedy $X + v = \{x + v : x \in X\}$) identické s rozdělením bodového procesu X.*

Poznámka. V celé práci budeme uvažovat stacionární bodové procesy.



Obrázek 1.3: Nestacionární bodový proces s funkcí intenzity rostoucí doleva.

Poznámka. Pro stacionární bodový proces v $\mathbb{R}^2$ platí, že existuje nezáporné reálné číslo λ splňující

$$\mathbb{E}X(B) = \lambda |B|,$$

kde $|B|$ je Lebesgueova míra množiny B. Tedy intenzita stacionárního bodového procesu je konstantní a odpovídá očekávanému počtu bodů na množině o velikosti 1. Také z toho plyne, že ať si vezmeme jakoukoliv podmnožinu pozorovacího okna W, tak by se bodový proces na těchto podmnožinách měl chovat obdobně. Příklad realizace nestacionárního bodového procesu můžeme vidět na obrázku 1.3.

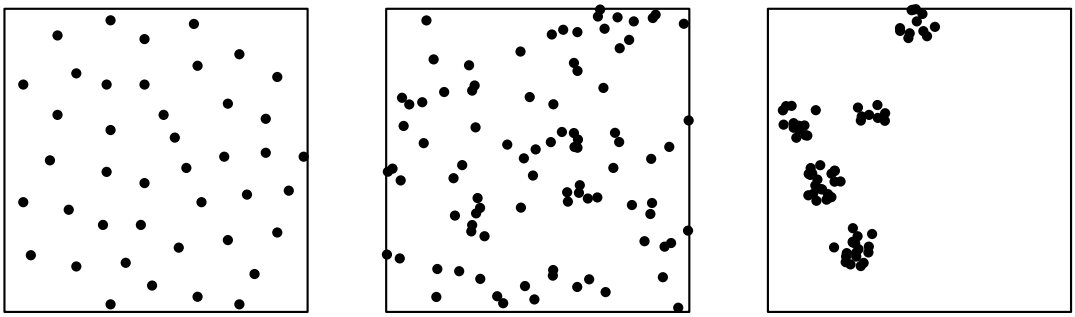
1.1.1 Typy interakcí

Při práci s bodovými procesy obvykle rozlišujeme tři hlavní skupiny bodových procesů dle typu bodových interakcí uvnitř pozorovacího okna W (viz obrázek 1.4):

- proces, jehož body mezi sebou nemají žádné interakce,
- *regulární proces*, který vykazuje odpudivé interakce mezi body. Body typicky vykazují pravidelnost a jsou od sebe dále než u prvního případu,
- *shlukový proces*, který vykazuje přitažlivé interakce mezi body. Body mají tendenci být více ve skupinkách (neboli opak regulárního procesu). Jeden z postupů, jak vytvořit shlukový proces je takový, že vezmeme bodový proces X a každý bod $x_i \in X$ nahradíme konečnou množinou bodů Z_{x_i} nazývanou shluk náležící bodu x_i, $i \in \mathbb{N}$. Obvykle se předpokládá, že shluky Z_{x_i} jsou pro různé rodičovské body x_i navzájem nezávislé.

1.2 Shlukový proces s rodičovskými body

Už dříve jsme zmínili, že body bodového procesu mohou mít nějakou přidanou informaci, takovouto vlastnost budeme nazývat **kóta**. Například u znázornění lokace stromů



Obrázek 1.4: Vlevo vidíme realizaci regulárního procesu, jehož body mezi sebou mají odpudivé interakce. Uprostřed je realizace bodového procesu, jehož body mezi sebou nemají žádné interakce a úplně vpravo vidíme realizaci shlukového bodového procesu, jehož body mají přitažlivé interakce.

v lese může každý bod obsahovat přidanou informaci o výšce daného stromu. **Kótovaným bodovým procesem** můžeme rozumět dvojici (x_i, m_i), kde x_i je lokace daného bodu a m_i je jeho kóta. Nyní si uvedeme formální definici.

Definice 5 (kótovaný bodový proces)**.** Kótovaný bodový proces *na prostoru E a s prostorem kót S je bodový proces Y na $E \times S$ takový, že počet bodů Y v $K \times S$ pro každou kompaktní množinu $K \subset E$ je konečný. Neboli bodový proces na součinovém prostoru*

$$Y : (\Omega, \Sigma, Pr) \to (\mathcal{N}(E \times S), \mathfrak{N}(E \times S)),$$

splňující, že tzv. podkladový proces $Y(\cdot \times S)$ na E je náhodná lokálně konečná míra.

Příklad kótovaného procesu jsme mohli vidět na obrázku 1.2. Prostor kót může být velmi obecný. Může to být konečná množina, spojitý interval reálných čísel, nebo více komplikovaný prostor množiny všech konvexních mnohoúhelníků, či prostor bodových procesů.

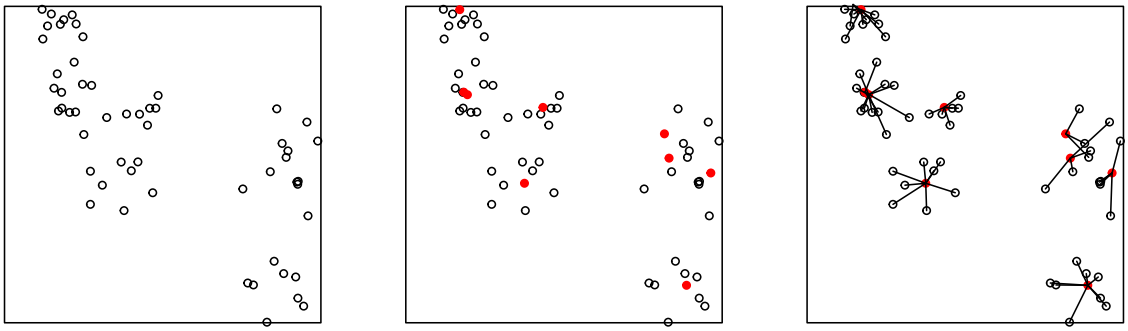
Právě posledním příkladem se teď budeme zabývat a zavedeme si pojem **shlukový proces s rodičovskými body** (dále jen **non-orphan proces**). Význam tohoto procesu je, že máme tzv. **rodičovské body** (rodiče) a **dceřiné body** (dcery) náležící některému z rodičů.

Příklad. V reálné situaci může rodičovské body znázorňovat například pozice jabloní v sadě a dceřiné body místa dopadů jablek. V praxi můžeme mít data obsahující informace pouze o místech dopadu jablek, nebo i o pozicích stromů, nebo navíc i o tom, z jakého stromu jablko spadlo a další.

Formálně řečeno, naším podkladovým procesem bude proces rodičovských bodů a prostorem kót prostor bodových procesů dcer, neboli $S = \mathcal{N}$ (viz (1.2)) a $(x_i, m_i) = (x_i, Z_{x_i})$, kde x_i je lokace rodiče a Z_{x_i} je bodový proces dcer náležící danému rodiči.

Obvykle pozorujeme pouze dceřiné body, ale můžou nastat situace, kdy máme data i o rodičovských bodech a vazbách mezi rodičem a dcerou. Realizaci těchto bodových procesů můžeme vidět na obrázku 1.5. Těmito situacemi se budeme zabývat v dalších kapitolách.

Poznámka. Pro definici 2 bodového procesu se předpokládá, že (E, ρ) je úplný separabilní metrický prostor. Aby naše definice byla korektní, tak $(E \times S)$ toto musí také splňovat,



Obrázek 1.5: Realizace non-orphan bodového procesu. Na levém obrázku můžeme vidět situaci, kdy máme data pouze o dceřiných bodech. Na prostředním obrázku máme navíc informaci o rodičovských bodech (červené body) a vpravo navíc víme, který dceřiný bod patří k jakému rodiči.

tedy že i prostor kót bodových procesů s dcerami a kartézský součin úplných separabilních metrických prostorů je opět úplný separabilní metrický prostor.

- Prostor S bodových procesů dcer je prostor lokálně konečných měr $\mathcal{M}_Y$ (viz (1.1)). Tento prostor je úplný separabilní metrický prostor a příslušná borelovská σ-algebra je přesně $\mathfrak{M}$, což nám říká např. kniha Daley a Vere-Jones (1988, věta A2.6.III.).
- Kartézský součin konečně mnoha separabilních prostorů je separabilní, což se můžeme dočíst např. v knize Willard (1970, věta 16.4c)).
- Platí, že kartézský součin dvou úplných metrických prostorů je úplný metrický prostor, avšak kvůli dodržení přiměřeného rozsahu práce necháváme bez důkazu.

1.3 Poissonův proces

Speciálním příkladem bodového procesu je Poissonův proces, který může být charakterizován pomocí dvou vlastností. Zaprvé, že počet bodů uvnitř uzavřené množiny se řídí Poissonovým rozdělením a zadruhé, že pro všechny navzájem disjunktní uzavřené množiny je počet bodů uvnitř daných množin nezávislý. Následně si uvedeme formální definici, kde budeme uvažovat opět E jako euklidovský prostor a $X(B)$ jako náhodnou veličinu (viz (1.4)).

Definice 6 (Poissonův proces)**.** Poissonův proces*, s konstantní intenzitou $\lambda > 0$, je bodový proces v E splňující:*

- *pro každou omezenou uzavřenou množinu B v E má $X(B)$ Poissonovo rozdělení s parametrem $\lambda|B|$, kde $|B|$ je Lebesgueova míra B na E,*
- *pokud $B_1, B_2, \ldots, B_{n-1}, B_n$, $n \in \mathbb{N}$, jsou navzájem disjunktní uzavřené množiny v E, pak $X(B_1), X(B_2), \ldots, X(B_n)$ jsou nezávislé.*

1.4 Thomasové proces

Thomasové proces je příklad shlukového bodového procesu (Thomas, 1949). Postup konstrukce takového procesu je následující:

1. vezmeme Poissonův proces s konstantní intenzitou $\kappa > 0$, jehož body nazveme rodiče,
2. každý rodičovský bod nahradíme několika dceřinými body, kde se toto číslo řídí Poissonovým rozdělením s parametrem $\mu > 0$, příčemž počty dcer pro jednotlivé rodiče jsou nezávislé,
3. dceřiné body kolem rodičovského bodu x jsou i.i.d. náhodné body $y_i = x + e_i$, kde e_i je vektor posunutí se souřadnicemi, které jsou nezávislé a s normálním rozdělením $N(0, \sigma^2)$, $\sigma > 0$.

Tento proces je pro nás zásadní, neboť na tomto modelu budeme následně odhadovat zmíněné parametry κ, μ a σ.

1.5 K-funkce

Během naší práce budeme potřebovat nějakou funkcionální charakteristiku, která bude dobře vypovídat o našem bodovém procesu. K tomu se hodí K-funkce (Baddeley, 2007, kapitola 2.6), která je definovaná jako

$$K(r) = \frac{\mathbb{E} \sum_{x \in (X \cap B)} X\Big(b(x,r) \setminus \{x\}\Big)}{\lambda \mathbb{E} X(B)}, r > 0, \tag{1.5}$$

kde $\mathbb{E}$ značí střední hodnotu, B je omezená borelovská množina s kladnou Lebesgueovo mírou, na jejíž volbě hodnota K-funkce nezávisí, X je bodový proces, $b(x,r)$ je koule se středem x a poloměrem r, λ je intenzita procesu. Tato K-funkce je velmi užitečná, neboť $\lambda K(r)$ nám udává očekávaný počet bodů y, který splňuje $0 < ||y - x|| \leq r$ pro daný bod procesu x, neboli očekávaný počet bodů ve vzdálenosti $r > 0$ od typického bodu procesu.

1.5.1 Odhad K-funkce

Často se v praxi musíme vypořádat s problémem, kdy máme informace pouze z pozorovacího okna W, ale ne z okolí za tímto oknem. Například u (1.5) máme uvedenou praktickou interpretaci K-funkce, která pracuje s body y splňujícími $0 < ||y - x|| \leq r$, my však pozorujeme pouze $X \cap W$. Tedy chceme informaci o očekávaném počtu bodů ve vzdálenosti r od jiného bodu. Avšak některé body, které jsou blízko hranice W, mohou mít okolní body mimo pozorovací okno W. Jinak řečeno $\text{dist}(x, X) \leq r$ právě tehdy, když počet bodů uvnitř množiny $\Big(X \cap b(x,r)\Big)$ je kladný, kde X je bodový proces, x je bod a $\text{dist}(x, X)$ je nejkratší vzdálenost mezi bodem x a zbylými body bodového procesu X, neboli $\text{dist}(x, X) = \inf\{||x - y||, y \in X\}$. Avšak my můžeme pozorovat pouze $\Big(X \cap W \cap b(x,r)\Big)$, čímž nebereme v potaz body za hranicí, a tedy K-funkce odhadnutá z našich dat bude typicky menší než by byla ve skutečnosti, kdybychom pozorovali celý prostor. Říkáme, že odhad K-funkce je záporně vychýlený. Metodám odstraňujícím tento problém se obecně říká **okrajová korekce**.

Mínusová korekce

Jedna ze základních metod, jak se s tímto vypořádat, je tzv. **mínusová korekce**, jejíž znázornění můžeme vidět na obrázku 1.6. Tato metoda spočívá v tom, že naše pozorovací okno zmenšíme, abychom měli informace i o okolních bodech za hranicí nového zmenšeného pozorovacího okna. A jak zvolit velikost, o kterou okno zmenšit? Chtěli bychom informace o všech bodech v okolí $r > 0$ od typického bodu procesu, kde r je pevně dané. Pokud okno zmenšíme o r, pak naší podmínku splníme a naše zmenšené okno bude

$$W_{-r} = \Big\{u \in W : \mathrm{dist}(u, \partial W) \geq r\Big\}, \tag{1.6}$$

kde ∂W značí hranici pozorovacího okna, $\mathrm{dist}(x, \partial W)$ je nejkratší vzdálenost mezi bodem x a množinou ∂W, neboli $\mathrm{dist}(x, \partial W) = \inf\{||x - y||, y \in \partial W\}$.

Tedy tuto K-funkci můžeme pomocí mínusové korekce neparametricky odhanout jako

$$\begin{aligned} \hat{K}(r) &= \frac{\sum_{x \in X \cap W_{-r}} X(b(x,r) \setminus \{x\})}{\hat{\lambda} X(W_{-r})} \\ &= \frac{\sum^{\neq}_{x,y \in X \cap W} \mathbf{1}\Big\{x \in W_{-r}\Big\}\mathbf{1}\Big\{||x - y|| \leq r\Big\}}{\hat{\lambda} X(W_{-r})}, \; r > 0, \end{aligned} \tag{1.7}$$

kde $0 < r$ je pevná hodnota menší než nějaké R stanovené tak, aby naše zmenšené pozorovací okno bylo neprázdné, X je bodový proces, $\mathbf{1}$ je indikátor, x jsou napozorované hodnoty bodového procesu X, $\neq$ v sumě značí, že bereme pouze dvojice různých bodů a $\hat{\lambda} = X(x)/|W|$ je odhad λ vzniklý pomocí metody momentů (podrobněji například v knize Illian a kol. (2008, kapitoly 2.6 a 7.2.2)). Tedy odhad počtu bodů kolem typického bodu procesu ve vzdálenosti $r > 0$ dostaneme tak, že pro každý bod uvnitř zmenšeného pozorovacího okna W_{-r} spočteme body ve vzdálenosti r od něj, tyto hodnoty pak sečteme a následně vydělíme celkovým počtem bodů uvnitř pozorovacího okna W_{-r} a odhadem intenzity.

Translační metoda

Další metoda, která nám pomůže se s tímto problémem vypořádat, je **translační metoda**, která pracuje s celým pozorovacím oknem W, ale jednotlivým dvojicím bodů přiřazuje různé váhy. Body blízko hranice budou mít okolo sebe menší počet bodů, tedy těmto bodům se budeme snažit dát větší váhu. Vezmeme tedy dvojice bodů uvnitř W, spočteme jejich vzdálenost, a pokud je jejich vzdálenost velká, pak tyto body dostanou větší váhu.

Konkrétně tyto váhy budeme určovat takto:

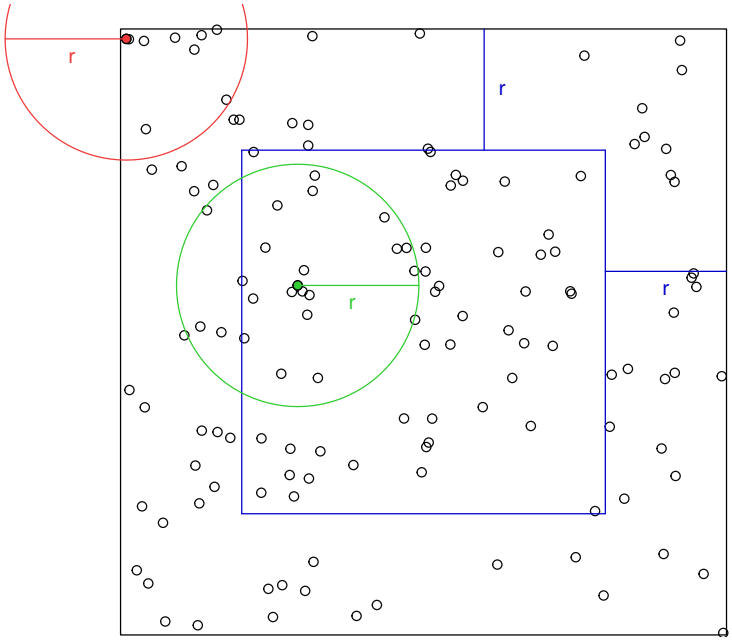
$$\mathrm{váha}(x_1, x_2) = \frac{1}{|(W_{x_1} \cap W_{x_2})|},$$

kde $|(W_{x_1} \cap W_{x_2})| = |(W \cap W_{x_2 - x_1})|$ je obsah průniku W_{x_1} a W_{x_2}, kde W_{x_1} je posunuté pozorovací okno $W_{x_1} = \{z + x_1 : z \in W\}$.

Nyní budeme chtít pomocí této metody odhadnout K-funkci:

$$\tilde{K}(r) = \sum^{\neq}_{x_1, x_2 \in W} \frac{|W|^2 \, \mathbf{1}\Big\{||x_2 - x_1|| \leq r\Big\}}{n(n-1)\,|(W_{x_1} \cap W_{x_2})|}, \text{ pro } 0 < r < r_{st}, \tag{1.8}$$

kde $n(n-1)/|W|^2$ je odhad čtverce intenzity $\tilde{\lambda}^2$, n je počet bodů uvnitř W, a pokud W je obdélník, pak r_{st} je délka kratší strany W. Podrobněji o této metodě v Illian a kol. (2008, kapitoly 4.2.2 a 4.3.3).



Obrázek 1.6: Mínusová korekce. Modrý čtverec je zmenšené pozorovací okno pro danou hodnotu r. Červený bod značí bod ve vnějším pozorovacím okně, pro který nejsme schopni určit kompletní informace o bodech do vzdálenosti r od něj, zelený bod značí bod ve zmenšeném pozorovacím okně, pro který jsme schopni určit počet bodů v jeho okolí do vzdálenosti r.

1.6 F-funkce

Další funkcionální charakteristika, která dobře vypovídá o procesu, je F-funkce, známá také jako contact distribution funkce nebo empty space funkce. Tato F-funkce je definovaná jako

$$F(r) = \mathbb{P}(\text{dist}(o, X) \leq r) = \mathbb{P}(X(b(o, r)) > 0), r > 0, \tag{1.9}$$

kde o je počátek, X je bodový proces, $X(b(u, r))$ je počet bodů do vzdálenosti r od polohy u. Pro stacionární bodové procesy hodnota F-funkce nezávisí na volbě u.

F-funkce odpovídá pravděpodobnosti, že v okolí r náhodně zvoleného bodu uvnitř W bude ležet nějaký bod procesu.

1.6.1 Odhad F-funkce

Jako u K-funkce, tak i zde budeme chtít naší funkci odhadnout z dat, což můžeme udělat následovně:

$$\hat{F}(r) = \frac{1}{X(M)} \sum_{u \in M} \mathbf{1}\{\text{dist}(u, X) \leq r\}, r > 0, \tag{1.10}$$

kde M je mřížka bodů uvnitř W.

Avšak i zde máme problém s okraji pozorovacího okna W, a tedy opět aplikujeme mínusovou korekci (viz (1.6)).

Tím dostaneme odhad

$$\hat{F}_{W_{-r}}(r) = \frac{1}{X(M \cap W_{-r})} \sum_{u \in M \cap W_{-r}} \mathbf{1}\{\mathrm{dist}(u, X) \leq r\}, r > 0, \tag{1.11}$$

kde W_{-r} je jako v (1.6) a r zvoleno tak, aby zmenšené okno bylo neprázdné.

1.7 Chyby odhadů

Pokud se odhadují parametry v nějakém modelu, typicky nás zajímá, jak přesné jsou získané odhady. Zde si uvedeme dva základní pojmy, které nám takovou otázku pomohou zodpovědět. Následující definice přebíráme zjednodušené z Nagy (2022).

Definice 7 (vychýlení)**.** *Nechť $\hat{\theta}$ je odhad parametru $\theta \in \Theta \subset \mathbb{R}$ v modelu $\mathcal{F}$. Pokud střední hodnota $\hat{\theta}$ existuje, pak* **vychýlením** *odhadu $\hat{\theta}$ rozumíme funkci $b : \Theta \to \mathbb{R}$, splňující*

$$\mathbb{E}_\theta(\hat{\theta}) = \theta + b(\theta), \text{ pro } \theta \in \Theta,$$

kde $\mathbb{E}_\theta(\hat{\theta})$ je střední hodnota $\hat{\theta}$, která závisí na parametru θ.

Definice 8 (střední čtvercová chyba)**.** *Nechť $\hat{\theta}$ je odhad parametru $\theta \in \Theta \subset \mathbb{R}$ v modelu $\mathcal{F}$. Pokud střední hodnota veličiny $(\hat{\theta} - \theta)^2$ existuje, potom* **střední čtvercovou chybou***, neboli MSE (mean squared error), rozumíme*

$$MSE_\theta(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta} - \theta)^2, \text{ pro } \theta \in \Theta,$$

kde $\mathbb{E}_\theta(\hat{\theta} - \theta)^2$ je střední hodnota $(\hat{\theta} - \theta)^2$, která závisí na parametru θ.

Zde budeme mluvit o **relativním vychýlení** a **relativní střední čtvercové chybě**, budeme značit $rb(\theta)$ a $rMSE(\theta)$, čímž rozumíme, že odhad podělíme skutečnou hodnotou parametru, u případu s MSE podělíme kvadrátem skutečné hodnoty parametru. O relativním vychýlení a relativním MSE budeme mluvit pouze pokud bude hodnota skutečného parametru nenulová.

2. Odhady parametrů

V následujících kapitolách budeme odhadovat parametry Thomasové procesu a následně odhady zkoumat. Vždy budeme uvažovat jeden bodový vzorek, u kterého chceme odhadnout ony parametry. Bodovým vzorkem zde myslíme simulaci jednoho bodového procesu našeho modelu. Již poměrně dlouhou dobu jsou zavedené metody pro odhady parametrů shlukových procesů založené na jejich momentových vlastnostech. Na jednu ze základních metod pro toto odhadování se nyní podíváme.

Budeme chtít odhadnout parametry κ, μ a σ v Thomasové procesu, které pro připomenutí jsou:

- κ – střední počet rodičovských bodů na jednotkové pozorovací okno,
- μ – průměrný počet dceřiných bodů v jednom shluku,
- σ^2 – směrodatná odchylka posunutí dceřiných bodů od rodičovského bodu.

Pro tyto odhady budeme předpokládat, že máme informace o dceřiných bodech, ale o rodičích ne.

2.1 Metoda minimálního kontrastu

Metoda minimálního kontrastu je obecná technika pro nalezení parametrů modelu bodového procesu pomocí dat z bodového vzorku. Nejprve se z dat vypočítá funkcionální charakteristika $\hat{T}(x)$, velice často se jako tato funkce používá K-funkce (1.5). Dále se vyjádří teoretická hodnota $T(\theta)$ této souhrnné statistiky v modelu (pokud možno jako algebraický výraz zahrnující parametry modelu) nebo se odhadne ze simulací modelu. Poté hledáme optimální hodnoty parametrů pro model, aby se dosáhlo co nejtěsnější shody mezi teoretickými a získanými výsledky. Matematicky řečeno, náš odhad je

$$\hat{\theta} = \operatorname{argmin}_{\theta\in\Theta} D(\hat{T}(x), [T(\theta)]),$$

kde T je vybraná funkcionální statistika, x napozorovaná data a D je míra nepodobnosti.

Naším modelem tedy bude Thomasové proces a vybranou statistikou K-funkce, která zde má známý analytický tvar parametru θ (Baddeley, 2007, kapitola 4.2), kde θ je v tomto případě vektor se složkami κ, μ, σ:

$$K_\theta(r) = \pi r^2 + \frac{1}{\kappa}\Big(1 - \exp\{-\frac{r^2}{4\sigma^2}\}\Big). \tag{2.1}$$

Intenzita tohoto procesu je $\lambda = \kappa\mu$. Tato K-funkce nezávisí na parametru μ, proto μ nejde odhadnout pomocí metody minimálního kontrastu. Tedy μ odhadneme pomocí intenzity λ Thomasové procesu, pro kterou platí $\lambda = \kappa\mu$. Tedy odhad $\hat{\mu}$ dostaneme jako $\hat{\mu} = \frac{\hat{\lambda}}{\hat{\kappa}}$, kde $\hat{\lambda}$ je počet dceřiných bodů uvnitř W dělený plochou W a $\hat{\kappa}$ odhad počtu rodičů uvnitř W spočítaný z analytického tvaru (2.1).

Následně musíme odhadnout K-funkci z dat, označme jako neparametrický odhad $\hat{K}$. K tomuto odhadu lze využít například odhad (1.7). Poté už nám stačí nalézt $\hat{\theta}$, které minimalizuje

$$\int_a^b |(K_\theta(r))^q - (\hat{K}(r))^q|^p \ \mathrm{d}r,\ r > 0, \tag{2.2}$$

kde $0 \leq a < b,\ p, q > 0$ jsou zvolené hodnoty. Ve výpočtech používáme hodnoty $q = 1/4$ a $p = 2$ (Diggle, 2003, kapitola 6.1.1). Toto je metoda minimálního kontrastu založená na statistice $T(x) = (\hat{K}(r), a < r < b)$ a s mírou nepodobnosti definovanou jako integrál kvadrátu rozdílu čtvrtých odmocnin dvou funkcí.

2.2 Naše odhady

Tentokrát budeme v situaci, kdy budeme mít i dodatečné informace o rodičovských bodech. V našem konkrétním případě máme informace o dceřiných bodech v pozorovacím okně W a o všech rodičovských bodech uvnitř W i v okolí $t > 0$ od hranice W, kde $t > 0$ je zvolené tak, aby každý dceřiný bod uvnitř W měl informaci o svém rodiči. Pak by mohly být vhodnější jiné metody, které odhadnou neznámé parametry lépe než metoda minimálního kontrastu. Na to se teď podíváme opět v případě s Thomasové procesem.

2.2.1 Odhad κ

Parametr κ v Thomasové procesu odpovídá střednímu počtu rodičů na jednotkové pozorovací okno v modelu, neboli Poissonovo rozdělení s parametrem $\kappa\,|W|$, kde $|W|$ je Lebesgueova míra W. Avšak tyto počty rodičů známe, tedy náš odhad parametru κ bude

$$\hat{\kappa} = \frac{n}{|W|},$$

kde W je pozorovací okno, $|W|$ Lebesgueova míra W a $n \in \mathbb{N}$ je počet rodičovských bodů uvnitř W.

2.2.2 Odhad μ

Parametr μ nám říká, kolik dceřiných bodů má průměrně jeden rodičovský bod, neboli kolik bodů je průměrně v jednom shluku. Z definice se toto číslo řídí Poissonovým rozdělením s parametrem μ. V situaci, kdy pozorujeme náhodný výběr o rozsahu n z Poissonova rozdělení s parametrem μ, pak střední hodnota tohoto rozdělní je μ, přičemž výběrový průměr je nestranný a konzistentní odhad střední hodnoty. Protože ke každému pozorovanému rodiči známe počet odpovídajících dceřiných bodů, náš odhad parametru μ bude již zmíněný výběrový průměr

$$\hat{\mu} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

kde $n \in \mathbb{N}$ je celkový počet rodičů uvnitř W, $X_i \in \mathbb{N}$ je počet dceřiných bodů i-tého rodiče, který je uvnitř W.

Avšak jako u odhadu K-funkce (1.7), i zde je problém s pozorovacím oknem. Rodiče blízko hranice budou mít své dcery i mimo pozorovací okno, a tedy náš odhad vyjde o něco menší než skutečná hodnota parametru, neboli záporně vychýlený. Proto použijeme mínusovou korekci (1.6), avšak zde nechceme body ve vzdálenosti $r > 0$ od typického bodu, ale chceme všechny dceřiné body náležící typickému rodičovskému bodu. Tedy pokud okno zmenšíme o největší vzdálenost mezi rodičem a dcerou, pak budeme mít informace o většině dcer náležících rodičovským bodům uvnitř zmenšeného okna. Nebudeme mít informaci pouze o dcerách mimo W, které mají vzdálenost ke svému rodiči větší, než

je největší vzdálenost rodiče a příslušné dcery uvnitř W. Náš odhad tedy bude

$$\hat{\mu} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

kde $n \in \mathbb{N}$ je celkový počet rodičů uvnitř W_{-r} a $X_i \in \mathbb{N}$ je počet dceřiných bodů i-tého rodiče, který je uvnitř W_{-r}.

2.2.3 Odhad σ

Zbývá už jen parametr σ, který odpovídá směrodatné odchylce posunutí dceřiného bodu od rodičovského (vzhledem ke každé souřadnicové ose). Pokud vezmeme jeden shluk kolem rodičovského bodu, tak má ono posunutí e_i rozdělení $N(0, \sigma^2)$ na ose x i na ose y, přičemž jednotlivá posunutí jsou nezávislá. Centrované normální rozdělení $N(0, \sigma^2)$ má hustotu

$$f(y_i;\, 0, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{{y_i}^2}{2\sigma^2}},\ y_i \in \mathbb{R}. \tag{2.3}$$

Pro nalezení parametru normálního rozdělení je poměrně snadné použít metodu maximální věrohodnosti, která nám dá analytický odhad hledaného parametru, který je maximálně věrohodný. Více o této metodě např. v Anděl (2011, kapitola 7.6).

Tedy logaritmus věrohodnostní funkce bude mít tvar

$$L(0, \sigma) = \prod_{i=1}^{n} f(y_i;\, 0, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i)^2}{2\sigma^2}},$$

$$\ell(0, \sigma) = \log\Big\{L(0, \sigma)\Big\} = \sum_{i=1}^{n} \log\Big\{f(y_i;\, 0, \sigma)\Big\} = -\frac{n}{2}\,\log(2\pi\sigma^2) \,-\, \frac{1}{2\sigma^2}\sum_{i=1}^{n} (y_i)^2,$$

kde n je počet pozorování, neboli dvojnásobek počtu dceřiných bodů (odpovídající souřadnicím vzhledem k ose x a y).

Následně zbývá už jen zderivovat podle proměnné σ, položit rovné nule a vyjádřit σ. Tím dostaneme maximum funkce (protože po dosazení našeho odhadu za parametr σ vyjde druhá derivace záporná), a tedy odhad parametru σ. Výsledná derivace bude

$$\frac{\partial \ell(0, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n} (y_i)^2 \stackrel{!}{=} 0, \tag{2.4}$$

kde ∂ značí derivaci a $\stackrel{!}{=} 0$ znamená, že rovnici položíme rovnu 0. Vyřešením poslední rovnosti (2.4) dostáváme odhad parametru σ

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} {y_i}^2}{n}}. \tag{2.5}$$

Ještě nás zajímá, zda je odhad opravdu maximálně věrohodný. K tomu je potřeba, aby po dosazení našeho odhadu za parametr σ vyšla druhá derivace záporná.

$$\frac{\partial^2 \ell(0, \hat{\sigma})}{\partial^2 \sigma} = \frac{n^2}{\sum_{i=1}^{n} {y_i}^2} - \frac{3n^2 \sum_{i=1}^{n} {y_i}^2}{(\sum_{i=1}^{n} {y_i}^2)^2} = -\frac{2n^2}{\sum_{i=1}^{n} {y_i}^2} < 0. \tag{2.6}$$

Tedy náš odhad je opravdu maximálně věrohodný.

3. Simulace

V této kapitole se podíváme, jak přesné odhady získáme pomocí metod popsaných výše na simulovaných datech. Tyto metody si označíme pomocí MK pro metodu minimálního kontrastu, NO_1 pro naše odhady bez mínusové korekce a NO_2 pro naše odhady s mínusovou korekcí. Zejména nás bude zajímat relativní vychýlení a relativní MSE.

3.1 Program

V programu `R` pomocí knihovny `spatstat` jsme zvolili dvě různé hodnoty pro každý z parametrů κ, μ, σ v Thomasové procesu. Konkrétně:

- $\kappa \in \{15, 30\}$,
- $\mu \in \{5, 10\}$,
- $\sigma \in \{0.02, 0.05\}$.

Vzali jsme všechny kombinace parametrů, tím jsme získali 8 různých modelů Thomasové procesu K_1 až K_8. Pro každý z těchto modelů jsme vygenerovali 500 realizací, z každé realizace odhadli zmíněné tři parametry pomocí metod MK, NO_1, NO_2 popsaných výše a pro tyto odhady odhadli relativní vychýlení a relativní MSE. Avšak střední hodnotu odhadu ani rozdílu odhadu a skutečného parametru neznáme. Proto tuto střední hodnotu odhadneme výběrovým průměrem. Relativní vychýlení jsme spočítali jako rozdíl výběrového průměru z jednotlivých odhadů a skutečné hodnoty parametru, to celé dělené skutečnou hodnotou parametru. Relativní MSE jsme spočítali obdobně, jen celé umocněné na druhou

$$rb(\theta) = \frac{\overline{\theta_n} - \theta}{\theta}, \tag{3.1}$$

$$rMSE(\hat{\theta}) = \frac{\sum_i^n (\hat{\theta}_i - \theta)^2}{n\theta^2}, \tag{3.2}$$

kde $\hat{\theta}_i$ je odhad parametru odhadnutý z i-té realizace, $\overline{\theta_n}$ je výběrový průměr z $\hat{\theta}_i$, n je celkový počet odhadů daného parametru (tedy 500) a θ je skutečná hodnota parametru. V tabulkách 3.1 a 3.2 můžeme vidět relativní vychýlení a relativní MSE pro všechny kombinace parametrů K_1 až K_8.

3.1.1 Výsledky

Relativní vychýlení

V tabulce 3.1 můžeme vidět kladné i záporné hodnoty. Kladná hodnota znamená, že odhady parametru vycházely spíše vyšší než skutečná hodnota parametru. Záporná hodnota nám říká, že odhady parametru vycházely spíše menší než skutečný parametr. Z tabulky hned vyčnívají odhady μ pomocí metody minimálního kontrastu, neboť ostatní odhady mají chybu menší než 1, pouze tyto odhady vycházejí výrazně větší. Je to způsobeno tím, že μ odhadujeme pomocí intenzity jako $\hat{\mu} = \frac{\hat{\lambda}}{\hat{\kappa}}$. Avšak odhad $\hat{\kappa}$ může být někdy dost blízký nule, pak dělíme velmi malým číslem a náš odhad $\hat{\mu}$ vyjde příliš vysoký. Dalším pozorováním je, že všechny naše odhady mají ve všech případech v absolutní hodnotě

menší relativní vychýlení než pomocí minimálního kontrastu. Tedy naše odhady by měly být blíže skutečné hodnotě parametru než odhady minimálním kontrastem. Toto pozorování se dalo očekávat, neboť v našich odhadech využíváme více informací z dat než metoda minimálního kontrastu.

Relativní MSE

Relativní MSE nám více ukáže rozdíly v jednotlivých chybách, neboť rozdíl mezi skutečnou hodnotou parametru a jeho odhadem je umocněn na druhou. Proto v tabulce 3.2 u metody minimálního kontrastu můžeme opět vidět velmi vysoké hodnoty u odhadu μ, poměrné vysoké hodnoty i u odhadu σ a všechny ostatní hodnoty malé. Opět si potvrzujeme, že všechny naše odhady mají menší relativní MSE než odhady pomocí metody minimálního kontrastu.

3.1.2 Výběr nejlepších 95 % odhadů

Někdy se může stát, že se v našich datech objeví velmi odlehlé hodnoty, které nám podstatně ovlivní výsledné relativní vychýlení nebo relativní MSE. Například výběrový průměr, s kterým také pracujeme, je velmi citlivý na odlehlá pozorování. Proto vez-

Relativní vychýlení

			MK			**NO$_2$**		
κ	μ	σ	$\hat{\kappa}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\kappa}$	$\hat{\mu}$	$\hat{\sigma}$
15	5	0.02	0.0957	2.1321	0.6132	-0.0058	-0.0020	-0.0042
		0.05	0.2800	2.0986	0.3134	-0.0050	0.0101	-0.0066
	10	0.02	0.0841	3.3287	0.8547	-0.0009	-0.0053	-0.0008
		0.05	0.2433	2.9585	0.3994	0.0014	0.0076	-0.0027
30	5	0.02	0.0991	1.9746	0.4674	-0.0052	0.0075	-0.0001
		0.05	0.3780	0.5923	0.1154	-0.0043	0.0031	0.0008
	10	0.02	0.0933	1.4392	0.3383	0.0012	0.0024	0.0011
		0.05	0.3073	1.1247	0.1776	0.0030	-0.0019	0.0003

Tabulka 3.1: Shrnutí výsledků simulace Thomasové procesu – relativní vychýlení.

Relativní MSE

			MK			**NO$_2$**		
κ	μ	σ	$\hat{\kappa}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\kappa}$	$\hat{\mu}$	$\hat{\sigma}$
15	5	0.02	0.2112	418.6139	23.5934	0.0509	0.0181	0.0050
		0.05	0.6149	282.5742	5.0791	0.0350	0.0289	0.0089
	10	0.02	0.1948	708.4691	37.3086	0.0518	0.0101	0.0023
		0.05	0.4730	477.4194	6.8412	0.0350	0.0165	0.0046
30	5	0.02	0.1530	446.3837	19.3677	0.0252	0.0084	0.0025
		0.05	0.8572	54.0600	2.1760	0.0172	0.0136	0.0044
	10	0.02	0.1281	351.9434	15.0162	0.0253	0.0044	0.0012
		0.05	0.5278	133.8034	3.2856	0.0174	0.0080	0.0022

Tabulka 3.2: Shrnutí výsledků simulace Thomasové procesu – relativní MSE.

meme pouze 95 % nejlepších odhadů pro každý parametr zvlášť, vypočítáme tabulky znovu a uvidíme, jestli se hodnoty nějak změní. Nejlepší odhady zde myslíme ve smyslu co nejmenší hodnoty $(\hat{\theta}_i - \theta)^2$, tedy hodnoty kvadrátu rozdílu odhadu a skutečné hodnoty parametru. Relativní vychýlení a relativní MSE z 95 % nejlepších výsledků můžeme vidět v tabulkách 3.3 a 3.4.

V tabulkách 3.1 a 3.2 nám vycházely velké hodnoty pro odhady μ. Zde můžeme vidět, že se naše chyby v odhadech podstatně zmenšily, neboť jsme odstranily těch 5 % nejhorších pozorování, která byla v tomto případě opravdu velmi odlehlá.

Nyní se nabízí otázka, zda se naše odhady obecně zlepšily. Vezmeme rozdíl absolutních hodnot relativního vychýlení a relativního vychýlení z 95 % nejlepších dat a následně zopakujeme to stejné s relativním MSE. Pokud vyjde výsledný rozdíl kladný, tak to znamená, že chyba odhadu spočítaná z 95 % nejlepších dat je menší než ze 100 % dat, tedy že jsme odhad zlepšili. U relativních vychýlení nám rozdíly vychází převážně kladné, záporné hodnoty jsme dostali ve 13 případech z 48. Tedy při použití 95 % nejlepších dat se vychýlení odhadů obecně zlepšilo. U relativního MSE vyšly všechny hodnoty kladné, tedy u všech odhadů se naše chyby zmenšily. Avšak z principu zahození 5 % nejhorších odhadů tato situace měla nastat. Zahazujeme odhady, pro které byl výraz $(\hat{\theta}_i - \theta)^2$ nej-

Relativní vychýlení z 95 % dat

			MK			**NO$_2$**		
κ	μ	σ	$\hat{\kappa}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\kappa}$	$\hat{\mu}$	$\hat{\sigma}$
15	5	0.02	0.0537	-0.0508	-0.0175	-0.0077	-0.0003	-0.0029
		0.05	0.1719	-0.0959	-0.0266	-0.0082	0.0045	-0.0043
	10	0.02	0.0593	-0.0541	-0.0358	-0.0047	-0.0011	-0.0012
		0.05	0.1565	-0.0996	-0.0489	-0.0047	0.0064	-0.0036
30	5	0.02	0.0806	-0.0550	-0.0294	-0.0104	0.0039	-0.0002
		0.05	0.2454	-0.1216	-0.0527	-0.0064	-0.0016	0.0011
	10	0.02	0.0719	-0.0602	-0.0377	-0.0052	0.0017	0.0007
		0.05	0.2185	-0.1163	-0.0592	-0.0024	-0.0036	-0.0014

Tabulka 3.3: Shrnutí výsledků simulace Thomasové procesu – relativní vychýlení z 95 % nejlepších dat.

Relativní MSE z 95 % dat

			MK			**NO$_2$**		
κ	μ	σ	$\hat{\kappa}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\kappa}$	$\hat{\mu}$	$\hat{\sigma}$
15	5	0.02	0.1426	0.0865	0.0398	0.0400	0.0133	0.0037
		0.05	0.3412	0.1713	0.0552	0.0274	0.0209	0.0061
	10	0.02	0.1273	0.0729	0.0315	0.0406	0.0072	0.0018
		0.05	0.3024	0.1548	0.0376	0.0273	0.0117	0.0035
30	5	0.02	0.1084	0.0673	0.0266	0.0200	0.0060	0.0019
		0.05	0.4058	0.1826	0.0495	0.0136	0.0098	0.0033
	10	0.02	0.0922	0.0532	0.0220	0.0199	0.0032	0.0009
		0.05	0.3365	0.1575	0.0381	0.0136	0.0055	0.0016

Tabulka 3.4: Shrnutí výsledků simulace Thomasové procesu – relativní MSE z 95 % nejlepších dat.

větší, tedy nezapočítáváme odhady nejvíce zvyšující MSE, tím pádem se MSE muselo zmenšit. U vychýlení toto zlepšení nastat nemusí. Například v situaci s mírně záporně vychýlenými odhady, které obsahují několik velmi vysokých odlehlých hodnot, se může vychýlení zvětšit.

4. Reálná data

V předchozích kapitolách jsme se seznámili se základními pojmy, uvedli si, jak budeme hledat odhady κ, μ a σ pokud máme i nemáme informace o rodičovských bodech a následně na simulovaných datech testovali, jak přesné odhady jsou.

Nyní provedeme odhady na reálných datech, která nám poskytla paní profesorka Aila Särkkä z Chalmers University of Technology and University of Gothenburg ve Švédsku. Pro popis těchto dat čerpáme informace z diplomové práce jejího studenta Andersson (2016).

Jedná se o problematiku onemocnění diabetickou neuropatií, což je nejčastější pozdní komplikace u lidí nemocných s cukrovkou. Při diabetické neuropatii dochází k postižení funkce i struktury nervu v důsledku vysoké hladiny cukru v krvi. Nejčastěji poškozuje nervy v nohou a chodidlech (Vitalion.cz., 2022).

Hlavním cílem v diplomové práci Andersson (2016) bylo zjistit více o vlivech diabetické neuropatie na epidermální nervová vlákna (dále jen ENF) a tím nalézt způsoby, jak tuto poruchu detekovat v časnějším stádiu. Dřívější diagnostika může pomoci zpomalit postup nemoci a oddálit příznaky nemoci. Epidermální nervová vlákna jsou malá senzorická nervová vlákna v epidermis (pokožka), nejsvrchnější vrstvě kůže, která snímají teplo a bolest. Jednotlivá vlákna procházejí hranicí mezi epidermis a dermis (vrstva pod pokožkou) a vstupují do epidermis, kde se často rozvětví.

Tyto ENF můžeme modelovat jako realizaci bodového procesu, kde rodičovské body jsou místa vstupů ENF do epidermis (pokožky) a dceřiné body místa jejich zakončení.

4.1 Popis dat

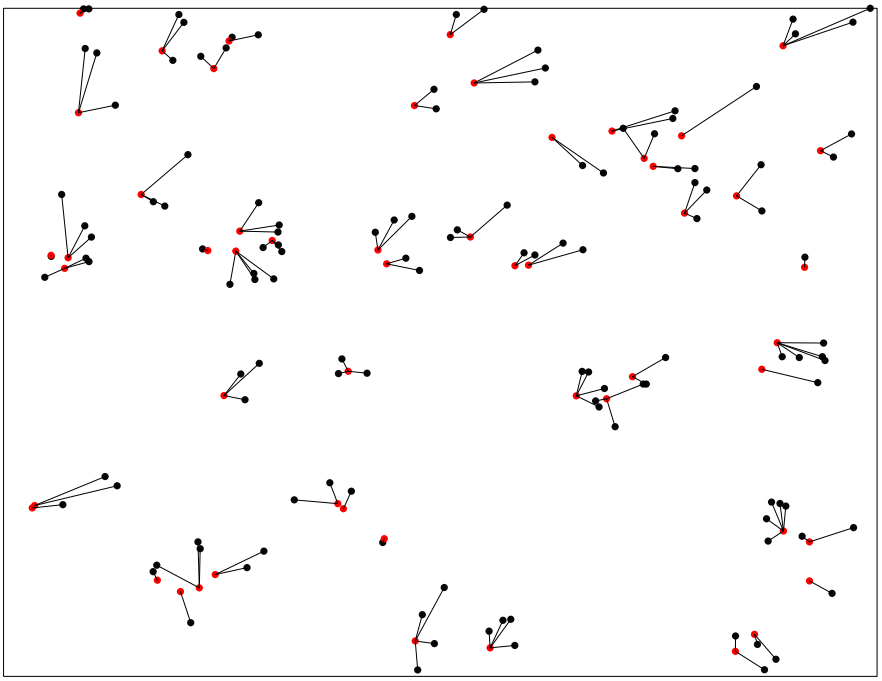
Náš datový vzorek byl odebrán z pravé nohy zdravého jedince, myšleno bez cukrovky. Konkrétně máme data v určitém pozorovacím okně o místech vstupu ENF z dermis do epidermis (pokožky) a o místech, kde vlákna končí. Navíc máme informace o tom, ke kterému rodiči každý dceřiný bod patří. Celkem máme 182 bodů, z čehož je 54 bodů rodičovských a 128 bodů dceřiných. Rozměry pozorovacího okna W jsou $332.3\,\mu m \times 434.5\,\mu m$.

Na obrázku 4.1 můžeme vidět znázornění těchto bodů pomocí bodového procesu. Červené body značí vstupy vláken do epidermis (rodiče) a černé body jsou jejich konce v epidermis (dcery). Úsečky znázorňují, jaký dceřiný bod náleží jakému rodiči.

Záznam dat je ovlivněn okrajovými efekty. Podle dostupných informací byly okrajové efekty zohledněny následujícím způsobem:

- pokud se vstup vlákna do epidermis nachází mimo okno W, pak se nezaznamenají ani konce příslušných vláken uvnitř W,
- pokud se v okně W nachází pouze vstup vlákna do epidermis a konce vlákna leží mimo W, pak se vstup nezaznamená,
- pokud je vstup vlákna uvnitř W a některé konce vlákna se nachází uvnitř W a některé mimo W, pak se konce mimo W nezaznamenají.

Nyní odhadneme κ, μ a σ těchto dat pomocí metody minimálního kontrastu i pomocí námi odvozených postupů, neboť máme dodatečné informace o rodičích.



Obrázek 4.1: Znázornění reálných dat jako realizace bodového procesu. Červené body jsou rodičovské body, černé dceřiné body a úsečky znázorňují, ke kterému rodičovskému bodu dcery patří.

4.2 Výsledky

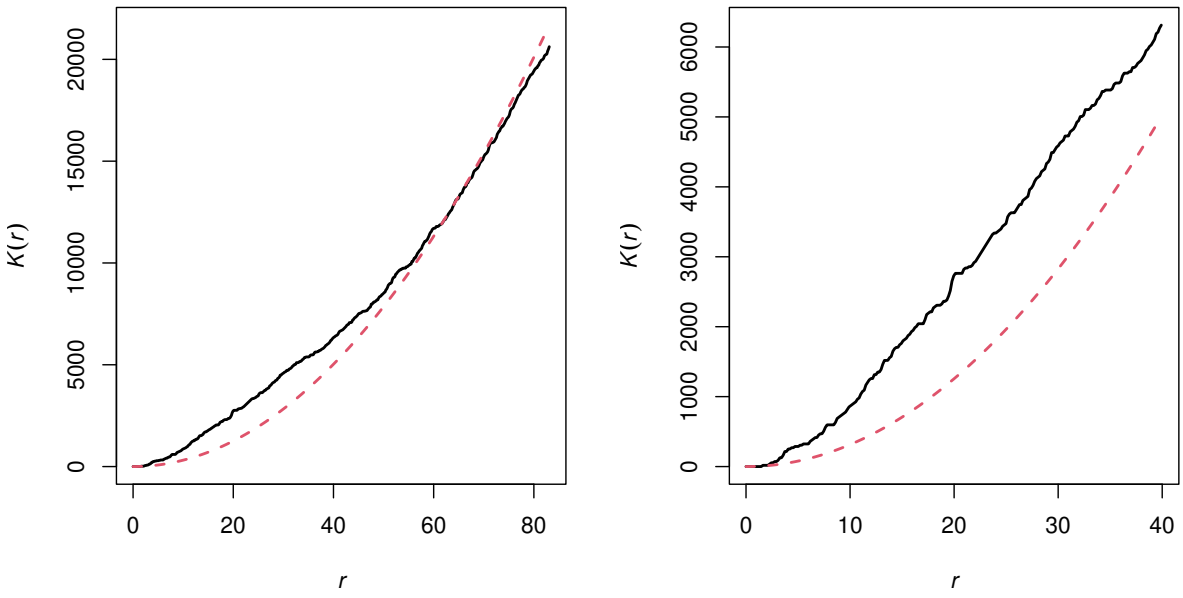
Na datový vzorek jsme spustili náš kód v programu `R` pro odhad pomocí metod MK (s translační metodou (1.8)), NO_1 a NO_2. Při použití metody NO_2 nám vyšla hodnota, o kterou okno zmenšujeme, 47.21 μm. V následující tabulce 4.1 můžeme vidět odhadnuté hodnoty parametrů.

Zajímavé je, že metoda minimálního kontrastu nám odhadla κ dvakrát větší než naše odhady, následně tedy μ dvakrát menší než naše odhady a σ také dvakrát menší. Může to být tím, že pro metodu MK je potřeba zvolit vhodnou maximální hodnotu pro horní mez integrálu (2.2), dále jen $rmax$, aby tato hodnota odpovídala dosahu interakcí. Defaultně se tato hodnota bere jako 1/4 délky kratší strany pozorovacího okna, tedy v našem případě se $rmax$ rovná 83.08. Můžeme se tedy podívat na graf 4.2 empirické K-funkce odhadnuté neparametricky z dat a teoretické K-funkce pro Poissonův proces $K(r) = \pi r^2$.

Odhady reálných dat

	MK	NO_1	NO_2
$\hat{\kappa}$	0.0008	0.0004	0.0004
$\hat{\mu}$	1.1420	2.3704	2.3333
$\hat{\sigma}$	6.6712	13.3511	12.6171

Tabulka 4.1: Odhady parametrů reálných dat pomocí minimálního kontrastu a námi odvozených metod. Maximální horní mez pro integrál v minimálním kontrastu je stanovena defaultně na 1/4 délky menší strany pozorovacího okna.



Obrázek 4.2: Graf K-funkcí dle rostoucího r. Černá plná křivka je empirická K-funkce dat a červená čárkovaná křivka je teoretická K-funkce Poissonova procesu.

Vidíme, že se empirická K-funkce začíná přibližovat k teoretické K-funkci Poissonova procesu kolem hodnoty $r = 40$. To znamená, že vysoká r už nepřináší informace o interakčních parametrech, ale vnáší pouze nežádoucí šum. Pro kontrolu můžeme přidat tabulku citlivosti odhadů na volbu $rmax$.

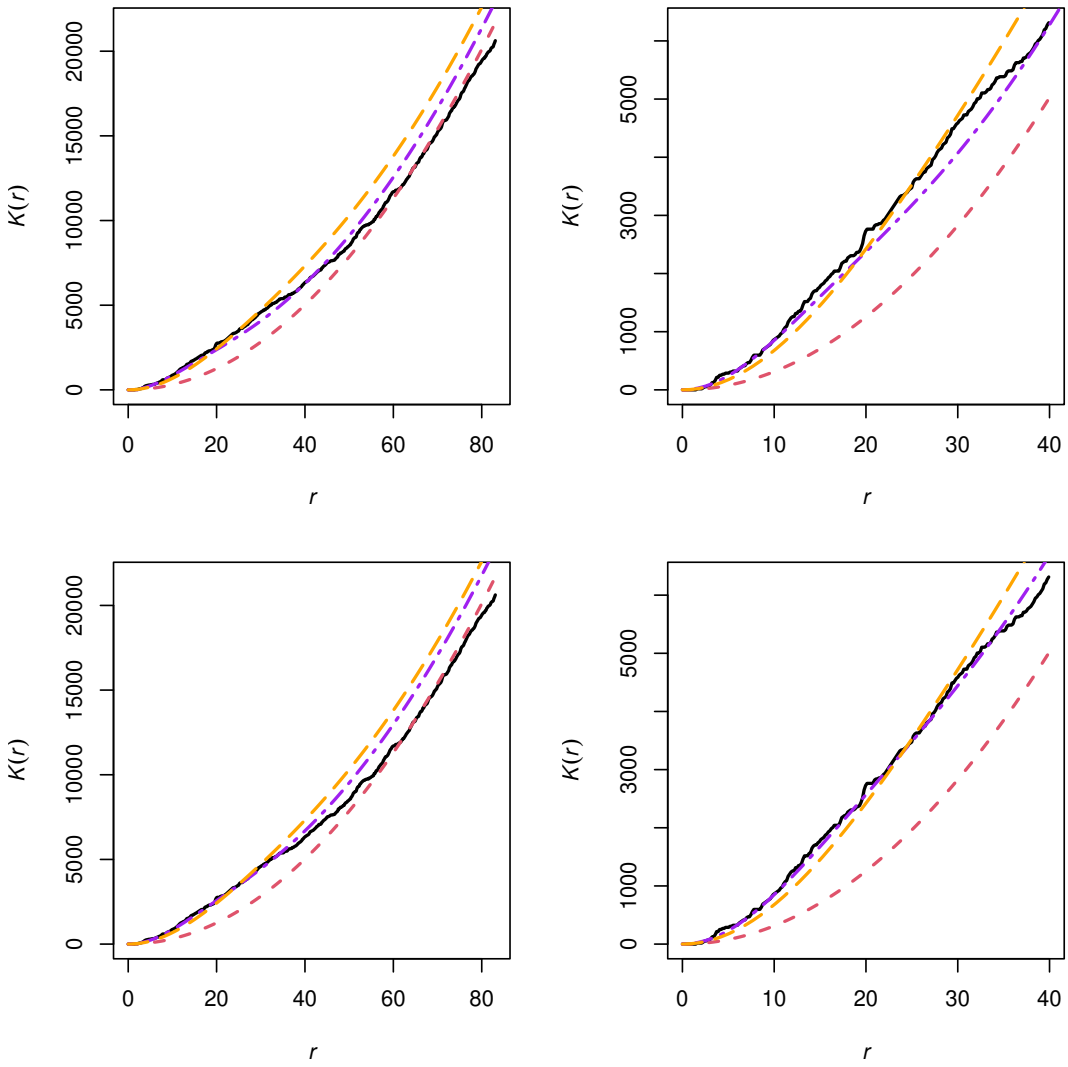
Z tabulky 4.2 vidíme, že se od $r = 40$ odhady parametrů opravdu moc nemění. Vezmeme tedy nové odhady pomocí MK s hodnotou $rmax = 40$. Této hodnotě odpovídají odhady $\hat{\kappa} = 0.0006, \hat{\mu} = 1.4965, \hat{\sigma} = 8.0332$. Všechny tyto hodnoty jsou blíže našim odhadům. Na závěr si ke grafu 4.2 vykreslíme grafy parametrických K-funkcí (spočtené z odhadů parametrů) pomocí MK a NO_2. Graf parametrické K-funkce spočítané z odhadů NO_1 neuvádíme, neboť tato křivka je téměř identická s K-funkcí dle NO_2.

Odhady MK dle $rmax$ pro reálná data

	5	10	15	20	25	30	35	40	
$\hat{\kappa}$	0.0000	0.0000	0.0001	0.0004	0.0005	0.0006	0.0006	0.0006	
$\hat{\mu}$	51.0670	23.7485	13.2679	1.9929	1.6741	1.6107	1.5866	1.4965	
$\hat{\sigma}$	59.0368	34.6998	26.1527	9.6577	8.6875	8.4702	8.3809	8.0332	
	45	50	55	60	65	70	75	80	max
$\hat{\kappa}$	0.0006	0.0006	0.0007	0.0007	0.0007	0.0007	0.0008	0.0008	0.0008
$\hat{\mu}$	1.4351	1.3714	1.3251	1.2801	1.2399	1.2078	1.1806	1.1572	1.1420
$\hat{\sigma}$	7.7944	7.5475	7.3706	7.1970	7.0434	6.9205	6.8180	6.7292	6.6712

Tabulka 4.2: Odhady parametrů reálných dat pomocí minimálního kontrastu v závislosti na hodnotě maximální meze integrálu (2.2).

Na pravém obrázku 4.3 můžeme vidět, že pro r do hodnoty 40 odhaduje parametrická K-funkce s odhady parametrů pomocí MK s $rmax = 40$ empirickou K-funkci lépe než s defaultním $rmax$ a navíc je blíže K-funkci z našeho odhadu.



Obrázek 4.3: Graf K-funkcí dle rostoucího r. Černá plná křivka je empirická K-funkce spočítaná z dat pomocí translační metody (1.8), červená křivka s kratšími čárkami je teoretická křivka pro Poissonův proces, fialová křivka s tečkami a čárkami je parametrická K-funkce s odhady parametrů spočítanými pomocí MK a oranžová křivka s dlouhými čárkami je parametrická K-funkce s parametry odhadnutými NO_2. V levém horním obrázku je K-funkce z MK spočítaná s defaultní hodnotou $rmax$ a v levém dolním obrázku se stanoveným $rmax = 40$. Obrázky na pravé straně ukazují r pouze do hodnoty 40 pro lepší přehlednost.

5. Test

Výše jsme odhadli parametry reálných dat pomocí metod pro Thomasové procesy. Nyní se ale nabízí otázka, zda jsou tyto metody vhodné pro náš datový vzorek. Nevíme, zda tato data odpovídají Thomasové procesu. Proto ještě provedeme test hypotézy, zda jsou naše data z modelu Thomasové procesů s konkrétními parametry proti alternativě, že nejsou.

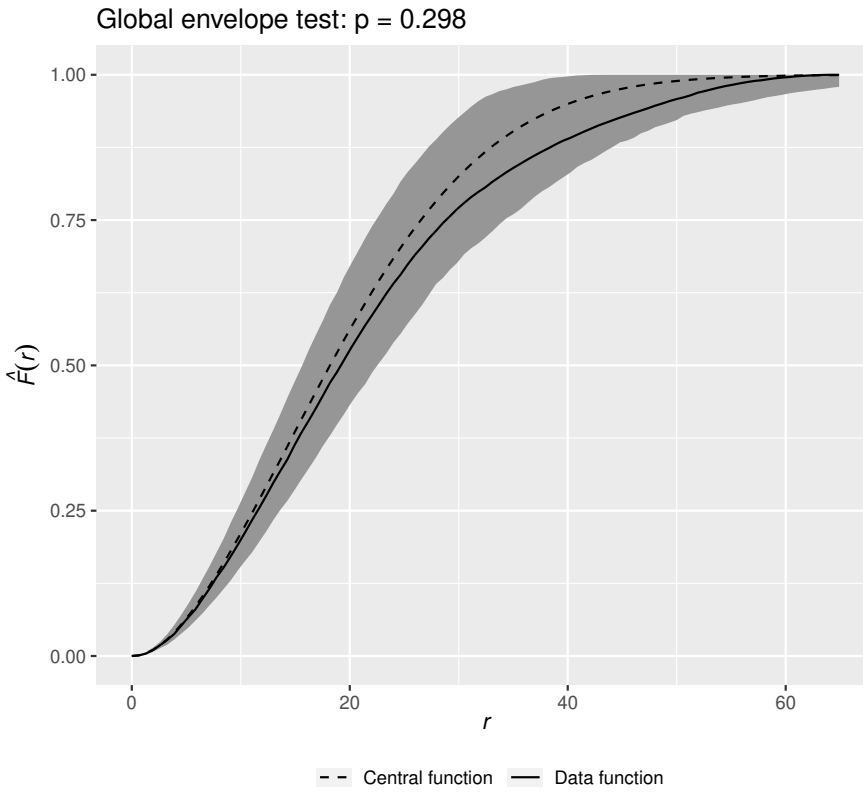
5.1 Obálkový test

Pro testování našich dat použijeme obálkový test. Myšlenka tohoto obálkového testu je taková, že pokud jsou naše data z testovaného modelu, pak by se i funkcionální charakteristiky našich dat a dat z testovaného modelu měly chovat stejně. Tedy zvolíme nějakou funkcionální charakteristiku dat (dále jen datovou křivku), která vypovídá o struktuře dat. Chceme zjistit, zda je tato datová křivka typickou datovou křivkou v testovaném modelu. Provede se určité množství simulací z našeho modelu, z těchto simulací se odhadnou datové křivky a seřadí se. V našem případě budeme řadit pomocí kritéria ERL, neboli extreme rank lengths. Více o této metodě například v článku Myllymäki a kol. (2017, kapitola 4). Pomocí tohoto řazení se vybere 95 % nejtypičtějších datových křivek a vytvoří se tzv. obálka z těchto křivek. Datovým křivkám, které nejsou typické, říkáme extrémní. Nás pak zajímá, zda je naše datová křivka v této obálce typických křivek. Pokud naše datová křivka kdekoliv vystoupí z obálky, pak je křivka extrémní a zamítáme hypotézu na hladině 5 %, že jsou naše data z testovaného modelu, ve prospěch alternativy.

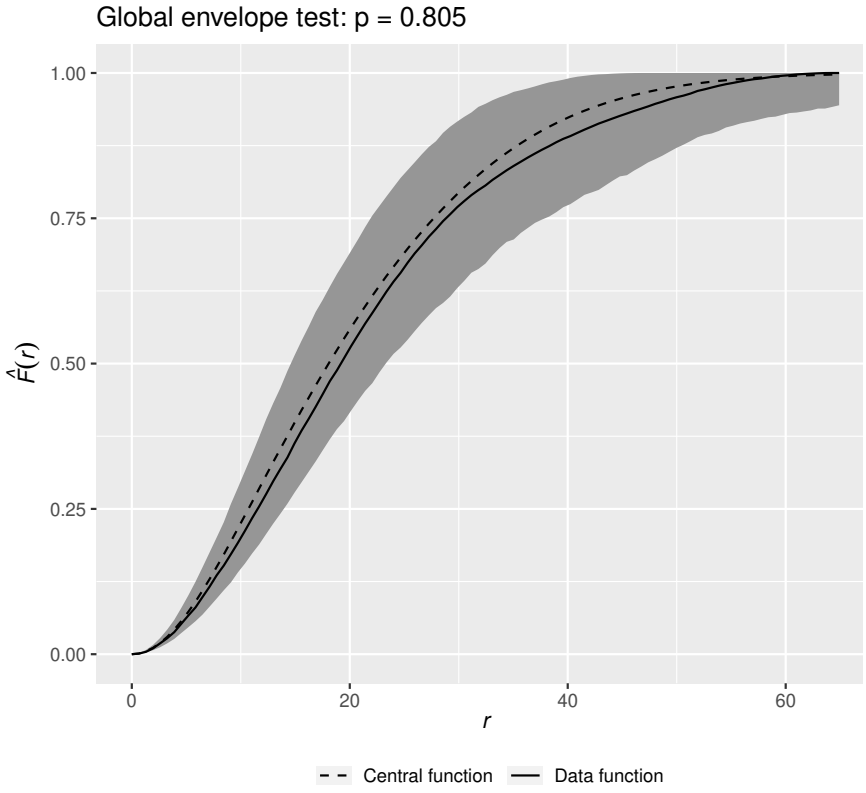
Jako funkcionální charakteristiku jsme zvolili F-funkci (1.9) s mínusovou korekcí. Nebylo by vhodné volit K-funkci, neboť pomocí této datové křivky jsme počítali odhad metodou minimálního kontrastu. Takový test by tedy neměl žádnou sílu. Dále jsme pro náš obálkový test zvolili hladinu 5 %, hodnoty argumentu r, při kterých má být F-funkce v bodě r vyhodnocena, nechali defaultní a počet simulací z testovaného modelu nastavili na 999, abychom měli celkem 1000 datových křivek včetně té z našich dat.

Na obrázku 5.1 můžeme vidět výsledek obálkového testu hypotézy, že naše data jsou z Thomasové modelu s parametry získanými metodou MK. Šedá plocha značí obálku typických datových křivek, čárkovaná křivka nejtypičtější datovou křivku a plná křivka datovou křivku našich dat. Vidíme, že celá naše datová křivka leží uvnitř obálky, a tedy hypotézu nezamítáme. Zároveň nahoře u obrázku vidíme uvedenou p-hodnotu, která je rovna 0.298. Tato p-hodnota vyjadřuje pořadí naší datové křivky seřazené od nejextrémnější po nejtypičtější dělené počtem všech datových křivek (včetně té naší). Podrobněji o výpočtu této p-hodnoty v Myllymäki a kol. (2017, kapitola 4.2). Tedy naše křivka je v tomto modelu 298. nejextrémnější.

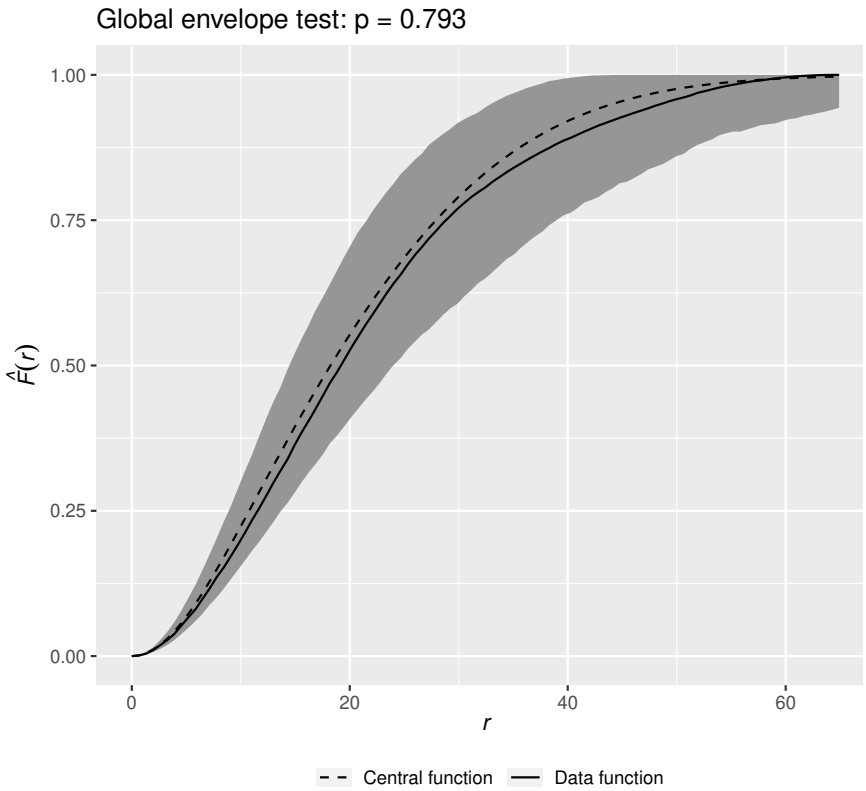
Obdobně na obrázcích 5.2 a 5.3 vidíme, že naše datové křivky leží uvnitř obálky. Tím pádem ani jednu hypotézu, že naše datová křivka odpovídá datovým křivkám z Thomasové modelu s parametry odhadnutými pomocí MK, NO_1 a NO_2, nebudeme zamítat.



Obrázek 5.1: Obálkový test hypotézy, že naše data odpovídají Thomasové modelu s odhadnutými parametry metodou MK.



Obrázek 5.2: Obálkový test hypotézy, že naše data odpovídají Thomasové modelu s odhadnutými parametry metodou NO_1.



Obrázek 5.3: Obálkový test hypotézy, že naše data odpovídají Thomasové modelu s odhadnutými parametry metodou NO_2.

Závěr

V této práci jsme se na úvod seznámili se základními koncepty teorie bodových procesů a s metodou minimálního kontrastu pro odhady parametrů Thomasové procesu, pokud nemáme informace o rodičovských bodech.

Přínosem této práce je odvození našeho postupu pro odhady parametrů Thomasové procesu pracující s informacemi nejen o dcerách, ale i o rodičích. Dále následné provedení simulací v programu `R`, díky kterým jsme byli schopni porovnat odhady získané metodou minimálního kontrastu a tímto naším postupem. Odhady jsme porovnávali pomocí relativního vychýlení a relativní střední čtvercové chyby. Také jsme tyto metody pro odhady parametrů aplikovali na reálná data o poloze epidermálních nervových vláknech u jednoho zdravého jedince a testovali, zda odhadnutý model dobře popisuje pozorovaná data.

Ze simulací jsme zjistili, že naše odhady měly ve všech zkoumaných případech menší relativní vychýlení i relativní střední čtvercovou chybu oproti odhadům metodou minimálního kontrastu a tedy lépe odpovídají skutečné hodnotě parametru.

Pro testování, zda jsou naše metody pro Thomasové proces vhodné i pro naše data, jsme použili obálkový test na hladině 5 %. Nulovou hypotézu, že náš datový vzorek odpovídá Thomasové procesu s odhadnutými parametry, jsme nezamítli ani v jednom případě s metodou minimálního kontrastu ani s námi odvezeným postupem. A tedy odhadnuté modely dobře popisují reálná data.

Seznam použité literatury

Andersson, C. (2016). Statistical methods for early discovery of diabetic neuropathy using epidermal nerve fiber data. Diplomová práce. Chalmers University of Technology and University of Gothenburg.

Anděl, J. (2011). *Základy matematické statistiky*. MatFyzPress. ISBN 978-80-7378-162-0.

Baddeley, A. (2007). *Spatial Point Processes and their Applications*. Springer-Verlag Berlin Heidelberg. ISBN 978-3-540-38174-7.

Daley, D. a Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes*. Second edition. ISBN 0-387-95541-0.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. Hodder Education Publishers, second edition. ISBN 978-03-4074-070-5.

Illian, J., Penttinen, A., Stoyan, H. a Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. ISBN 978-0-470-01491-2.

Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H. a Hahn, U. (2017). Global envelope tests for spatial processes: Series b (statistical methodology). *Royal Statistical Society*, **79**(2), 381–404.

Nagy, S. (2022). NMSA332: Mathematical Statistics II. URL `https://www2.karlin.mff.cuni.cz/~nagy/NMSA332/NMSA332.pdf`. [cit. 25.5.2022].

Rataj, J. (2006). *Bodové procesy*. Druhé opravené vydání. Karolinum, Praha. ISBN 80-246-1182-1.

Thomas, M. (1949). A generalization of poisson's binomial limit for use in ecology. **36** (1/2), 18–25.

Vitalion.cz. (2022). Diabetická neuropatie. URL `https://nemoci.vitalion.cz/diabeticka-neuropatie/`. [cit. 6.6.2022].

Willard, S. (1970). *General Topology*. Addison-Wesley. ISBN 978-0-201-08707-9.