ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

FAKULTA ŽIVOTNÍHO PROSTŘEDÍ

Katedra vodního hospodářství a environmentálního modelování

Využití jádrových odhadů hustoty pro popis rozdělení pravděpodobnosti srážkových úhrnů

DIPLOMOVÁ PRÁCE

Vedoucí práce: Ing. Jan Hnilica, Ph.D.

Diplomant: Bc. Anna Maksaeva

2023

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Fakulta životního prostředí

ZADÁNÍ DIPLOMOVÉ PRÁCE

Autorka práce:	Anna Maksaeva
Studijní program:	Krajinné inženýrství
Obor:	Environmentální modelování
Vedoucí práce:	Ing. Jan Hnilica, Ph.D.
Garantující pracoviště:	Katedra vodního hospodářství a environmentálního modelování
Jazyk práce:	Čeština
Název práce:	**Využití jádrových odhadů hustoty pro popis rozdělení pravděpodobnosti srážkových úhrnů**
Název anglicky:	**Use of the kernel density estimates for the description of probability distribution of daily precipitation sums**
Cíle práce:	Cílem práce je analýza jádrových odhadů hustoty při popisu rozdělení denních srážkových úhrnů. Pozornost bude věnována problematice shlazení odhadu a zejména pak aplikaci jádrových odhadů na přirozeně ne-negativní náhodné veličiny, s čímž je spjat tzv. „boundary problem“. S využitím syntetických i reálných dat budou testovány možné postupy konstrukce jádrových odhadů, výsledné odhady budou porovnány s parametrickými odhady hustoty.
Metodika:	Práce bude omezena na odhady jednorozměrných rozdělení. Rešerše: - odhady hustoty – obecně, neparametrické odhady, jádrové odhady - přehled metod odhadu hustoty srážkových úhrnů Praktická část: - výběr datových podkladů (reálná i syntetická data) - aplikace jádrových odhadů, porovnání s ostatními metodami - zhodnocení: posouzení úspěšnosti metod, citlivosti na nastavení, robustnosti, dostupnosti software (případně náročnosti naprogramování)
Doporučený rozsah práce:	40 stran
Klíčová slova:	hustota pravděpodobnosti, srážky, jádrové odhady

Doporučené zdroje informací:

1. Rajagopalan, B., Lall, U., Tarboton, D. G. (1997): Evaluation of kernel density estimation methods for daily precipitation resampling. Stochastic hydrology and hydraulics, 11(6): 523 – 547.
2. Silverman, B. W. (1986): Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York.

Předběžný termín obhajoby: 2022/23 LS - FŽP

Elektronicky schváleno: 29. 3. 2020
doc. Ing. Martin Hanel, Ph.D.
Vedoucí katedry

Elektronicky schváleno: 30. 3. 2020
prof. RNDr. Vladimír Bejček, CSc.
Děkan

ČESTNÉ PROHLÁŠENÍ

Prohlašuji, že jsem diplomovou práci na téma „Využití jádrových odhadů hustoty pro popis rozdělení pravděpodobnosti srážkových úhrnů“ vypracovala samostatně a citovala jsem všechny informační zdroje, které jsem v práci použila a které jsem rovněž uvedla na konci práce v seznamu použitých informačních zdrojů.

Jsem si vědoma, že na moji diplomovou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů, ve znění pozdějších předpisů, především ustanovení § 35 odst. 3 tohoto zákona, tj. o užití tohoto díla.

Jsem si vědoma, že odevzdáním diplomové práce souhlasím s jejím zveřejněním podle zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů, ve znění pozdějších předpisů, a to i bez ohledu na výsledek její obhajoby.

Svým podpisem rovněž prohlašuji, že elektronická verze práce je totožná s verzi tištěnou a že s údaji uvedenými v práci bylo nakládáno v souvislosti s GDPR.

v Praze dne 20.03.2023 …….

PODĚKOVÁNÍ

Upřímně děkuji Ing. Janu Hnilicovi, Ph.D., za trpělivost, za odborné vedení při zpracování diplomové práce a věnovaný čas.

Srdečně děkuji svému tatínkovi, bez jehož podpory by se nic z toho nestalo.

Abstrakt

Cílem diplomové práce bylo analyzovat účinnost jádrových odhadů hustoty při popisu rozložení pravděpodobnosti denních srážkových úhrnů. Byla provedena literární rešerše, která se zabývala jednak problematikou jádrových odhadů obecně, dále potom specifickými problémy spjatými se srážkovými daty. Na základě rešerše bylo vybráno osm metod jádrového odhadu, které byly naprogramovány. Metody byly nejprve testovány na syntetických datech pro porovnání jejich účinnosti. Na základě tohoto testování byly vybrány dvě metody („cut and normalize“ a logaritmická transformace), které byly aplikovány na měřená data a porovnány s parametrickým odhadem. Bylo zjištěno, že jádrový odhad hustoty je při správném výběru metody efektivnější a poskytuje přesnější odhad rozdělení pravděpodobnosti než parametrický odhad.

Klíčová slova: jádrový odhad, kernelový odhad, šířka okna, srážky, hustota rozdělení, pravděpodobnost

Abstract

The aim of the diploma thesis was to analyse the efficiency of kernel density estimates for daily precipitation data. Eight methods of the kernel density estimation were selected on the basis of a literature research. Synthetic datasets (samples from various gamma distributions) were used to test the methods to compare their efficiency. On the basis of the test, two methods (“cut and normalize” and logarithmic transformation) were selected and applied on real measured data. Their efficiency was compared with the efficiency of parametric density estimates. Finally, it was found, that the kernel density estimates provide more accurate density estimation than their parametric counterparts.

Keywords: kernel density estimate, bandwidth, precipitation, probability density function, likelihood

Obsah

1. Úvod

Údaje o srážkových úhrnech jsou nezbytným podkladem v mnoha odvětvích vědecké a inženýrské činnosti, od hydrologických a klimatologických studií až po posuzování a navrhování hydrotechnických opatření (Mosthaf a Bardossy, 2017).

Při mnoha praktických aplikacích, jako např. při konstrukci srážkových generátorů (Wilks a Wilby, 1999) nebo při posuzování klimatických změn (Teutschbein a Seibert, 2012), hraje klíčovou úlohu rozdělení pravděpodobnosti srážkových úhrnů. Rozdělení pravděpodobnosti denních srážek se vyznačuje specifickým tvarem – je výrazně kladně šikmé, s velkou převahou malých úhrnů a řídkým výskytem vysokých srážek (Chen a Brissette, 2014).

Při odhadování rozdělení na dané lokalitě v praxi převládají parametrické metody odhadu. Zřejmě nejčastěji používaným parametrickým modelem je gama rozdělení, jehož použití pro popis denních srážek lze nalézt ve velkém množství prací (např. Block et al., 2009, a mnoho dalších). Nicméně existují studie, např. Vlček a Huth (2009), poukazující na to, že automatické používání gama rozdělení nemusí v případě srážek přinášet přesné výsledky.

Vedle parametrických odhadů existuje množství neparametrických metod odhadu rozdělení, jejich obsáhlý a podrobný výčet představuje např. Silverman (1986). Předmětem této práce bylo posoudit možnosti využití jedné z těchto metod, jádrových odhadů hustoty, při popisu rozdělení pravděpodobnosti denních srážkových úhrnů.

2. Cíl práce

Cílem diplomové práce bylo analyzovat možnost využití jádrových odhadů hustoty při popisu rozložení pravděpodobnosti denních srážkových úhrnů a porovnat jejich účinnost s běžně používanými parametrickými odhady.

Diplomová práce je rozdělena do dvou základních částí. V první části jsou shrnuty výsledky literární rešerše, která byla zaměřena jednak na problematiku jádrových odhadů obecně, tak také na specifické problémy, které nastávají při aplikaci jádrových odhadů na srážková data. Tyto problémy pocházejí z charakteristického tvaru rozdělení srážek a z toho, že srážky jsou náhodná veličina přirozeně nezáporná. Velká pozornost je v rešerši věnována právě metodám, které pomáhají adaptovat jádrové odhady hustoty na uvedené okolnosti.

Druhá část práce se věnuje praktickému testování metod popsaných v první části. Nejprve jsou metody testovány na syntetických datech, což umožňuje jejich vzájemné porovnání. Na jeho základě jsou pak vybrány nejúčinnější postupy, které jsou testovány na reálných měřených datech. Spolu s nimi jsou testovány také parametrické metody odhadu a výsledky obou těchto přístupů jsou nakonec porovnány.

Veškeré zpracování dat v této práci bylo provedeno v programovacím jazyce R (R Core Team, 2015).

3. Literární rešerše

3.1 Základní pojmy

Před rozborem metod použitých v této práci je vhodné uvést stručný přehled základních pojmů, se kterými bude v dalším textu často operováno. Práce se týká pravděpodobnostního rozložení srážkových úhrnů a některých metod jeho odhadu. Na srážkové úhrny je tedy pohlíženo jako na náhodnou veličinu, proto jsou v následujících odstavcích definovány některé základní pojmy z teorie pravděpodobnosti.

Náhodná veličina je číselný popis výsledku náhodného pokusu. Před realizací pokusu je výsledek náhodný, tzn. není známý a z vnějších okolností jej nelze předem určit. Náhodné veličiny lze rozdělit na diskrétní a spojité.

Diskrétní náhodná veličina může nabývat pouze izolovaných hodnot. Množina těchto hodnot pak může být konečná i nekonečná. Pravděpodobnostní rozdělení diskrétní náhodné veličiny je popsáno pravděpodobnostní a distribuční funkcí.

Pravděpodobnostní funkce každému x přiřazuje pravděpodobnost, se kterou náhodná veličina X nabývá této dané hodnoty a je definována jako $P(X = x) = p$. Pravděpodobnost p musí splňovat následující:

$$0 \leq p_i \leq 1 \; pro \; každé \; i$$

$$\sum_i p_i = 1.$$

Distribuční funkce $F(x)$ diskrétní náhodné veličiny X každému x přiřazuje pravděpodobnost, s jakou je hodnota veličiny X menší nebo rovna hodnotě x:

$$F(x) = P(X \leq x) = \sum_i P(X = x_i) \text{ pro všechna } x_i \leq x.$$

Spojitá náhodná veličina může nabývat jakékoli hodnoty z určitého intervalu na reálné číselné ose. Pravděpodobnost, že náhodná veličina nabude jakékoli jednotlivé hodnoty je rovna 0, protože počet hodnot, kterých může náhodná veličina nabývat, je nekonečný. Pravděpodobnostní rozdělení spojité náhodné veličiny je popsáno hustotou pravděpodobnosti, distribuční a kvantilovou funkcí.

Hustota pravděpodobnosti spojité náhodné veličiny $f(x)$ se používá k vyjádření pravděpodobnosti s jakou náhodná proměnná spadá do určitého intervalu hodnot. Tato pravděpodobnost je dána integrálem hustoty na daném intervalu, tzn. je dána

obsahem plochy pod funkcí hustoty pravděpodobnosti. Vlastnosti funkce hustoty pravděpodobnosti jsou

$$f(x) \geq 0,$$

$$\int f(x)\, dx = 1.$$

Dalším způsobem popisu pravděpodobnostního rozdělení spojité náhodné veličiny je distribuční funkce. *Distribuční funkce* spojité náhodné veličiny X má následující tvar:

$$F(x) = \int_{-\infty}^{x} f(u)du.$$

Tato funkce každému reálnému číslu x přiřazuje pravděpodobnost, že náhodná veličina nabude hodnoty, která je rovna nebo menší než toto číslo, tzn. bude spadat do polouzavřeného intervalu $< -\infty, x)$. Hustotu pravděpodobnosti získáme derivací distribuční funkce. Vlastnosti distribuční funkce jsou následující:

- $0 \leq F(x) \leq 1$, *pro* $\forall x$ *(funkce je nezáporná),*
- $F(x) \leq F(y)$, *pro* $\forall x < y$ *(funkce je neklesající),*
- $\lim_{h \to 0} F(x + h) = F(x)$, *(funkce je zprava spojitá),*
- $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$,
- $F(\infty) = \lim_{x \to \infty} F(x) = 1$.

Pro popis pravděpodobnostního rozdělení spojité náhodné veličiny se používá také *kvantilová funkce*, která je inverzní k funkci distribuční (někdy proto nazývána také jako inverzní distribuční funkce). Pokud uvažujeme rozdělení spojité náhodné veličiny s distribuční funkcí $F(x)$, potom je kvantil Q_p je číslo, pro něž platí následující:

$$P(X \leq Q_p) = F(Q_p) = p$$

Distribuční funkce $F(Q_p) = p$ náhodné veličiny X udává s jakou pravděpodobností bude hodnota náhodného pokusu menší nebo rovna Q_p a nabývá hodnot z intervalu $< 0, 1 >$, inverzní distribuční funkce $F^{-1}(p) = Q_p$ potom udává, pro jaká x bude výsledek náhodného pokusu s požadovanou pravděpodobností p menší nebo roven Q_p. V případě, že distribuční funkce není prostá může jedné hodnotě p odpovídat několik hodnot Q_p.

3.2 Odhady hustoty pravděpodobnosti

Přesné určení (resp. odhad) rozdělení pravděpodobnosti dané konkrétní náhodné veličiny je jednou ze základních úloh statistické praxe a o jedné z metod odhadu pojednává i tato práce. Odhad rozdělení vždy vychází z tzv. *náhodného výběru*, který můžeme definovat jako konečný soubor naměřených hodnot (realizací náhodného pokusu) dané náhodné veličiny. Při odhadu hustoty pak mohou nastat dvě rozdílné výchozí situace, o kterých pojednávají další odstavce.

3.2.1 Parametrické odhady

První situace je ta, kdy máme znalost o typu rozdělení dané veličiny. Tato znalost může být dána teoretickými důvody, nebo konvencí, kdy u mnoha veličin je některé rozdělení všeobecně akceptováno jako vyhovující. Náhodný výběr pak použijeme pro odhad neznámých parametrů daného rozdělení. To může být provedeno několika způsoby, obvykle se používá metoda momentů nebo metoda maximální věrohodnosti (Anděl, 1998).

Při výběru metody se berou v úvahu střední hodnota a rozptyl výsledného odhadu, které se obvykle vyjadřují pomocí následujících pojmů:

- *nestrannost* – odhad je nestranný, pokud je jeho střední hodnota rovna hledanému parametru (tzn. odhad není zatížen systematickou chybou)
- *konzistence* – odhad je konzistentní, pokud se jeho rozptyl zmenšuje s rostoucí velikostí výběru (tzn. dostatečně velkým výběrem dosáhneme dostatečně přesného odhadu).

3.2.1.1 Metoda momentů

Momentová metoda využívá toho, že momenty náhodné veličiny jsou funkcemi parametrů rozdělení, které nejsou známé. Momentový odhad obecně najdeme tak, že teoretické momenty nahradíme jejich výběrovými momenty, které získáme z náhodného výběru. Následně potom řešíme rovnici (nebo soustavu rovnic) vzhledem k neznámým parametrům.

Obecný a centrální moment *k*-tého řádu náhodné veličiny se vypočte jako

$$\mu_k' = E\,(X^k)$$

$$\mu_k = E(X-\mu)^k.$$

Odpovídající výběrové momenty se vypočtou jako

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

$$m_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^k.$$

Metoda momentů spočívá tedy v tom, že odhady parametrů se stanoví jako řešení rovnic $\mu'_k = m'_k$ (resp. $\mu_k = m_k$), pro *k = 1, 2,...,n*, kde *n* je počet parametrů, které mají být odhadnuty. Výhodou použití metody momentů je především její snadnost, nicméně odhady získané touto metodou obvykle nejsou nestranné. Metoda ale poskytuje konzistentní odhady.

Výrazně lepších výsledků je dosahováno tzv. metodou L-momentů (Hosking, 1990), která k odhadu využívá lineární kombinace běžných momentů.

3.2.1.2 Metoda maximální věrohodnosti

Další metodou, která se používá pro odhad parametrů pravděpodobnostního rozdělení je metoda maximální věrohodnosti (Maximum Likelihood), která předpokládá, že nejlepší odhad parametru rozdělení by měl maximalizovat pravděpodobnost, že pozorované hodnoty pocházejí z předpokládaného rozdělení. Tato metoda je nejčastěji používanou. Uvažujme, že $x = (x_1, \ldots, x_n)$ je náhodný výběr a $f(x, \theta)$ je funkce rozdělení pravděpodobnosti s parametrem θ. Sdruženou hustotu pravděpodobnosti (*věrohodnostní funkci*) lze zapsat následovně:

$$L = \prod_{i=1}^{n} f(x_i, \theta)$$

Maximálně věrohodný odhad je takový odhad, který maximalizuje věrohodnostní funkci. Při praktické aplikaci je vhodné pracovat s logaritmem věrohodnostní funkce, tj.

$$\ln L = \sum_{i=1}^{n} ln\,[f(x_i, \theta)]$$

Maximálně věrohodný odhad $\hat{\theta}$ hledáme jako bod, kde se parciální derivace věrohodnostní funkce podle jednotlivých parametrů rovnají nule (řešíme tzv. soustavu věrohodnostních rovnic). Nezkreslený maximálně věrohodný odhad je rovněž nejlepším nezkresleným odhadem, tzn. jeho rozptyl mezi všemi nezkreslenými

odhady je nejmenší. Zároveň jsou tyto odhady konzistentní. Problémem je, že soustava věrohodnostních rovnic často vyžaduje numerické řešení.

3.2.2 Neparametrické odhady

Při použití neparametrických metod předem neděláme žádné předpoklady o funkční formě pravděpodobnostního rozdělení. Neparametrické metody obecně hledají takový odhad, který bude co nejbližší k datovým bodům, aniž by byl příliš skokovitý nebo nerovný. Tento přístup má velkou výhodu oproti parametrickým metodám v tom, že výsledné odhady se mohou přizpůsobit větší škále možných tvarů pravděpodobnostních rozdělení. Použití jakéhokoli parametrického přístupu s sebou přináší možnost, že zvolené pravděpodobnostní rozdělení je výrazně odlišné od skutečnosti a v takovém případě nebude výsledný model dobře odpovídat datům. Oproti tomu neparametrické přístupy se tomuto nebezpečí zcela vyhýbají, protože žádný předpoklad o tvaru rozdělení nečiníme. To je obzvlášť výhodné pro smíšená a vícemodální rozdělení, kde by volba parametrického modelu byla velmi komplikovaná. Nevýhoda neparametrických přístupů však spočívá v tom, že tyto metody neredukují problém odhadu na malý počet parametrů, a proto je pro použití těchto metod vyžadován velmi velký počet pozorování (mnohem více, než je obvykle potřeba pro parametrický přístup), aby byl získaný odhad dostatečně přesný.

3.2.2.1 Histogram

Histogram je nejjednodušším případem neparametrického odhadu hustoty pravděpodobnostního rozdělení. Prostor jevů je rozdělen do nepřekrývajících se přihrádek (počet těchto přihrádek je vyjádřen přirozeným číslem) a hustota je aproximována kvantifikací počtu pozorování, která spadají do jednotlivých přihrádek. Vzhledem k počátku x_0 a šířce přihrádky, kterou označíme jako h, definujeme přihrádky jako intervaly $[x_0 + mh, x_0 + (m + 1)\, h]$ pro kladná a záporná celá čísla m. Pro jednoznačnost jsou intervaly obvykle voleny zleva uzavřené a zprava otevřené. Histogram je pak definován jako

$$\hat{f}(x) = \frac{n_x}{nh} \tag{1}$$

kde n_x je počet hodnot ve stejné přihrádce jako x. Pro konstrukci histogramu musí být zvolen jak počátek, tak šířka přihrádky, kterou lze ovlivnit míru vyhlazení

výsledného histogramu. Histogram lze zobecnit tak, že se šířky přihrádek mohou měnit, odhad je pak definován následovně:

$$\hat{f}(x) = \frac{1}{n} \times \frac{počet\ X_i\ ve\ stejné\ přihrádce\ jako\ x}{šířka\ přihrádky\ obsahující\ x}.$$

Jak je možné vidět, pro histogram je nutné definovat dva parametry: šířka přihrádky h a počátek přihrádky x_0. Histogram je velmi jednoduchou formou neparametrického odhadu hustoty a jeho použití je spojeno s určitými nedostatky. Za prvé, konečný tvar odhadu hustoty velmi závisí na výchozí poloze první přihrádky, která je dána zcela subjektivně. Za druhé, přirozenou vlastností histogramu je nespojitost odhadu hustoty na hranicích přihrádek, která však nesouvisí s původní hustotou, ale je pouze následkem volby polohy jednotlivých přihrádek. Třetí nevýhodou je skutečnost, že počet přihrádek roste exponenciálně s počtem dimenzí. Vyšší dimenze vyžadují velký rozsah náhodného výběru, v opačném případě by většina přihrádek byla prázdná (tento problém je společným problémem všech neparametrických metod odhadu hustoty). Všechny tyto nevýhody činí histogram nevhodným pro většinu praktických aplikací kromě rychlé vizualizace výsledků v jednom nebo dvou rozměrech. Histogramu se také běžně používá pro vytvoření prvotní představy pro výběr parametrického modelu rozdělení.

3.2.2.2 Naivní odhad

Z definice hustoty pravděpodobnosti plyne, že pokud má náhodná veličina X hustotu f, pak:

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h).$$

Pro libovolné dané h můžeme samozřejmě odhadnout $P(x - h < X < x + h)$ podílem výběru spadajícího do intervalu $(x - h, x + h)$. Tedy přirozený odhad hustoty je dán výběrem malého čísla h a výpočtem:

$$\hat{f}(x) = \frac{1}{2nh}\ počet\ x_i\ spadající\ do\ (x - h, x + h).$$

Tato metoda se nazývá naivní odhad (naive estimator). Pro transparentnější vyjádření odhadu definujeme váhovou funkci w takto:

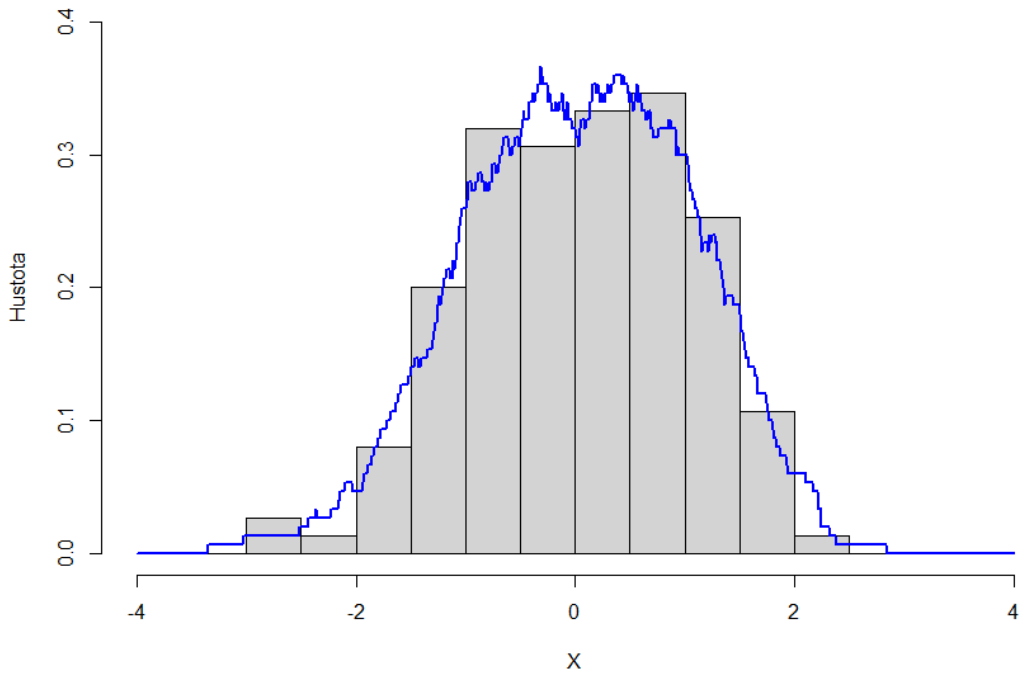
$$w(x) = \begin{cases} 0.5 & jestli\ |x| < 1 \\ 0 & v\ ostatních\ případech. \end{cases} \quad (2)$$

Pak lze snadno vidět, že naivní odhad lze napsat jako

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} w\left(\frac{x - x_i}{h}\right) \tag{3}$$

Z toho vyplývá, že odhad je konstruován umístěním pole o šířce $2h$ a výšce $(2nh)^{-1}$ na každé pozorování a následným sečtením pro získání odhadu. Konstrukce naivního odhadu je podobná konstrukci histogramu, ale v případě naivního odhadu je každý bod středem svého intervalu, čímž se histogram osvobodí od subjektivního výběru pozic přihrádek. Volba šířky přihrádky ale stále existuje a řídí se parametrem h, který určuje, jak bude výsledný odhad vyhlazen.

Naivní odhad není pro odhad hustoty ideální z hlediska interpretace a prezentace. Z definice vyplývá, že $\hat{f}$ není spojitá funkce, ale má skoky v bodech $x_i \pm h$ a všude jinde má nulovou derivaci. To způsobuje, že odhady nejsou hladké, což jednak není žádoucí esteticky, ale také to ztěžuje interpretaci výsledného odhadu.



Obr. 1 Histogram a naivní odhad (modrá čára) ze náhodného samplu 150 hodnot z normálního rozdělení N(0, 1). Šířka okna v obou případech je 0.5.

3.2.2.3 Kernelový (jádrový) odhad

Jádrový odhad (KDE – kernel density estimate) získáme tak, že v definici (3) nahradíme váhovou funkci w jádrovou funkcí (kernelem) K. Odhad je tedy ve tvaru

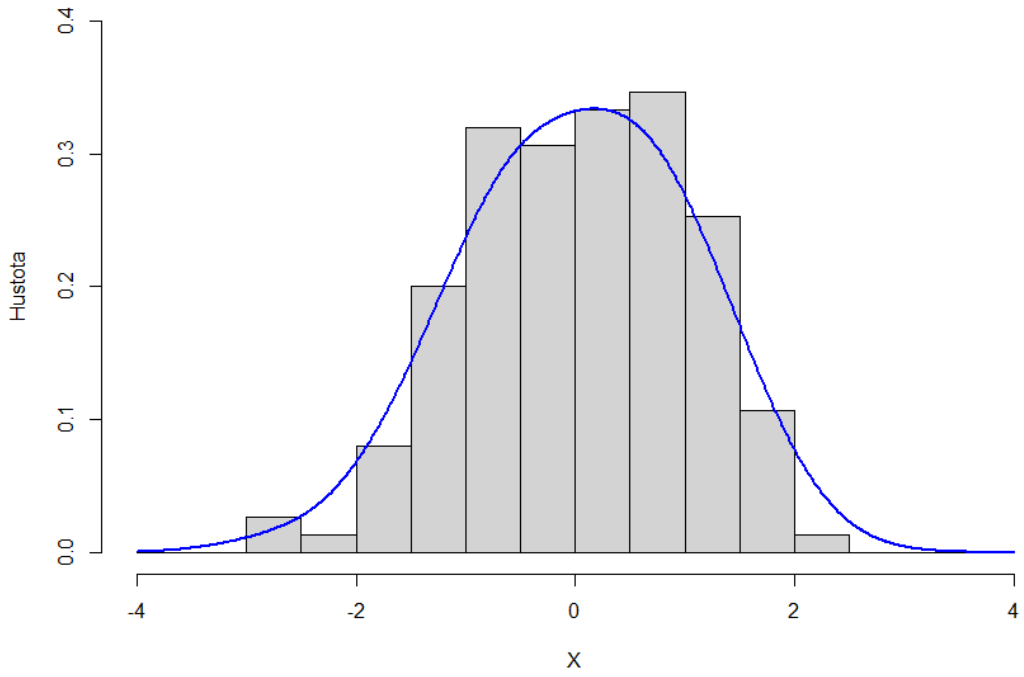
$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{4}$$

Jádrová funkce (jádro) musí splňovat následující požadavky:

- $\int K(x)dx = 1,$
- $\int x\,K(x)dx = 0,$
- $\int x^2\ K(x)dx\ <\ \infty,$
- $K(x) \geq 0\ pro\ všechny\ x,$
- $K(x)\ =\ K(-x),$

což znamená, že jádro je symetrická hustota rozdělení pravděpodobnosti se střední hodnotou rovnou nule a s konečným rozptylem. Proměnná h v definici (4) je reálné kladné číslo, které se obvykle nazývá šířka pásma nebo šířka okna (v anglicky psané literatuře „bandwidth“) a kontroluje stupeň vyhlazení odhadu.

Z definice plyne, že pokud je K hustota pravděpodobnosti, pak také $\hat{f}$ podle (4) je hustota a po funkci K přebírá spojitost a diferencovatelnost. Jádrový odhad tedy řeší problém se spojitostí, který byl u histogramu i naivního odhadu, a také nevyžaduje žádnou subjektivní volbu analogickou k volbě rozmístění přihrádek. Odhad hustoty je konstruován centralizací škálovaného jádra na každý bod náhodného výběru. Hodnota odhadu v bodě x je pak jednoduše součtem hodnot všech jader v tomto bodě. Kombinace příspěvků z každého datového bodu znamená, že v oblastech, kde je mnoho pozorování a očekává se, že skutečná hustota má relativně velkou hodnotu, by měl jádrový odhad také nabývat relativně velké hodnoty. Opak by měl nastat v oblastech, kde je relativně málo pozorování.



Obr. 2 Histogram a jádrový odhad (modrá čára) ze náhodného samplu 150 hodnot z normálního rozdělení N(0, 1).Použita byla jádrová funkce gaussova, šířka okna v obou případech je 0.5.

Kromě metod popsaných výše existuje několik dalších metod neparametrického odhadu hustoty, jako jsou například

- nearest neighbour estimator
- maximum penalized likelihood estimator
- variable kernel estimator.

Podrobnosti, včetně způsobu implementace, uvádí Silverman (1986). Nicméně tyto metody, jak ukazuje Silverman, jsou pouze speciálními případy jádrových odhadů, resp. jsou jádrovými odhady se speciálně volenými kernely, a obvykle vykazují některé problematické aspekty (výpočetní náročnost, citlivost na lokální šum v datech atp.). Z těchto důvodů se běžně nepoužívají a tato práce se jimi dále nezabývá.

3.3 Aplikace kernelových odhadů

Praktická aplikace kernelových odhadů znamená výběr vhodné kernelové funkce a nastavení ideální hodnoty šířky okna (h). Pro tyto volby neexistuje univerzální a všeobecně akceptované řešení, ale je nutné posuzovat každý případ (veličinu) individuálně. Při teoretických úvahách je volba kernelu a šířky okna motivována analýzou střední kvadratické chyby (MSE), které se dopustíme, pokud skutečnou hustotu $f(x)$ v bodě x odhadujeme pomocí $\hat{f}(x)$ podle (4). MSE je dána vztahem

$$MSE_x(\hat{f}) = E\left(\hat{f}(x) - f(x)\right)^2 \tag{5}$$

kde $E(.)$ označuje střední hodnotu. Je nutné vysvětlit, že obecně výsledek odhadu závisí na zvoleném kernelu, zvoleném h a na datech, které máme k dispozici. Při výpočtu chyby (podle (5) a i při některých dalších uvedených definicích) je míněna závislost na datech, tzn. kernel a šířku okna bereme jako fixní a počítáme střední hodnotu při použití různých datových výběrů.

Standardními elementárními vlastnostmi střední hodnoty a rozptylu je možné střední kvadratickou chybu rozepsat jako

$$MSE_x(\hat{f}) = \left(E\hat{f}(x) - f(x)\right)^2 + var\hat{f}(x) \tag{6}$$

kde var označuje rozptyl (varianci). Vztah (6) ukazuje, že při jádrovém odhadu se celková kvadratická chyba skládá ze dvou složek – z chyby odhadu a jeho rozptylu. Silverman (1986) dál ukazuje, že při rozvoji (6) podle Taylorova polynomu se dají obě komponenty (chyba a rozptyl) vyjádřit jako funkce šířky okna (h) tak, že chyba je přímo úměrná hodnotě h, zatímco rozptyl nepřímo. Z toho plyne, že volbou velkého h dosáhneme redukce rozptylu odhadu, ale zvýšíme jeho chybu a naopak. Toto přelévání celkové chyby (trade-off) mezi chybou a rozptylem je průvodním znakem všech neparametrických metod.

MSE není vždy dostačujícím měřítkem kvality odhadu, jelikož se zaměřuje pouze na lokální chování odhadu. Celková (globální) přesnost odhadu se potom získá integrováním MSE jako střední integrovaná kvadratická chyba (zkráceně $MISE$) definovaná jako

$$MISE(\hat{f}) = E\int \left(\hat{f}(x) - f(x)\right)^2 \mathrm{d}x. \tag{7}$$

3.3.1 Výběr kernelu

Pomocí minimalizace Taylorova rozvoje střední kvadratické chyby odhadu (6) lze vyjádřit teoretickou optimální hodnotu šířky okna jako

$$h_{opt} = k_2^{-2/5} \left(\int K(x)^2 \mathrm{d}x \right)^{1/5} \left(\int f^{\prime\prime}(x)^2 \mathrm{d}x \right)^{1/5} n^{-1/5} \tag{8}$$

kde

$$k_2 = \int x^2 K(x)\, \mathrm{d}x$$

a f je skutečná hustota, kterou se snažíme odhadnout. Pokud dosadíme tuto optimální hodnotu do vyjádření $MISE$ podle (7), dostaneme

$$MISE(\hat{f}\,) = \frac{5}{4} C(K) \left(\int f^{\prime\prime}(x)^2 \mathrm{d}x \right)^{1/5} n^{-4/5}$$

kde $C(K)$ je konstanta (vzhledem k datům) závisející pouze na použitém kernelu K a její hodnota je

$$C(K) = {k_2}^{2/5} \left(\int K(x^2) \mathrm{d}x \right)^{4/5}$$

Optimální kernel tedy teoreticky dostaneme minimalizací $C(K)$ vzhledem k použité kernelové funkci. Nejmenší hodnotu $C(K)$ má tzv. *Epanechnikovův* kernel

$$K(x) = \begin{cases} 0.75\,(1 - x^2) & pro\ |x| \leq 1 \\ 0 & pro\ |x|\ > 1 \end{cases} \tag{9}$$

Nicméně ostatní kernely mají hodnotu $C(K)$ velmi podobnou a výsledky odhadu dosažené s různými kernely mohou být ekvivalentní prostřednictvím vhodného výběru šířky vyhlazovacího okna. V důsledku toho se obecně má za to, že volba kernelové funkce není pro odhad hustoty příliš důležitá a je zcela legitimní vybrat kernel podle dalších vlastností (diferencovatelnost, definiční obor, výpočetní náročnost).

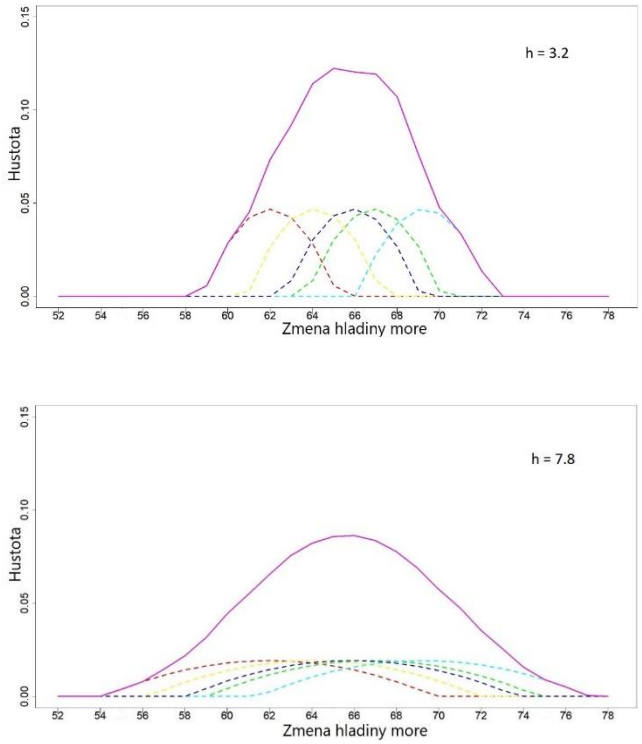
Obvykle se používá těchto několik kernelových funkcí:

Název kernelu	Funkce	Definiční obor
Normální (Gaussovský)	$\mathrm{K(x)} = (2\pi)^{-1/2} e^{-x^2/2}$	$(-\infty, \infty)$
Epanechnikov	$\mathrm{K(x)} = 0.75\,(1 - x^2)$ 0	$\lvert \mathrm{x} \rvert \leq 1$ v opačném případě
Bisquare	$\mathrm{K(x)} = 0.9375(1 - x^2)^2$ 0	$\lvert \mathrm{x} \rvert \leq 1$ v opačném případě

Tab. 1 Nejčastěji používané kernelové funkce.

3.3.2 Volba úrovně shlazení

Problém volby parametru šířky vyhlazovacího okna (h) je oproti volbě kernelové funkce zásadní a silně ovlivňuje přesnost kernelových odhadů. Obecně malé hodnoty h vedou k zeštíhlení jader, zatímco velké hodnoty h vedou k jejich zploštění.



Obr. 3 Jádrové odhady s různou šířkou okna, aplikované na data globální průměrné změny absolutní hladiny moře, pozorované od 15.04.2012 do 15.08.2012. Šířka okna = 3.2 a 7.8, bylo použito jádro Epanechnikova.

Pro volbu ideální hodnoty šířky okna neexistuje univerzální řešení, nicméně bylo vyvinuto několik metod, které jsou popsány v následujících odstavcích.

3.3.2.1 Reference na parametrické rozdělení

Při této metodě předpokládáme, že skutečná hustota f zhruba odpovídá určitému parametrickému rozdělení. Potom můžeme tuto předpokládanou hustotu dosadit do vztahu (8) pro výpočet optimální hodnoty h. Hodnoty optimálního h pro některé typy rozdělení jsou k dispozici v literatuře. Např. Silverman (1986) ukazuje, že pro normální rozdělení je za použití Gaussovského kernelu dosaženo hodnoty

$$h = 1.06\,\sigma\,n^{-1/5}$$

kde σ je směrodatná odchylka vypočítaná z náhodného výběru. Rajagopalan et al. (1997) pro normálně rozdělená data s použitím Epanechnikovova kernelu dochází k hodnotě

$$h = 2.13\,\sigma\,n^{-1/5}$$

a pro data pocházející z exponenciálního rozdělení (opět s Epanechnikovovým kernelem) uvádí hodnotu

$$h = 1.97\,\sigma\,n^{-1/5} \tag{10}$$

Je potřeba poznamenat, že uvedené vzorce poskytují jen hrubý odhad optimální hodnoty, který málokdy vede k uspokojujícímu výsledku. Celá metoda je také nepraktická kvůli potřebě předem uvažovat nějaké parametrické rozdělení, což je v rozporu s celkovou filosofií neparametrických odhadů. Na druhou stranu ale odhady šířky okna touto metodou poskytují kvalitní výchozí bod pro mnohé další metody.

3.3.2.2 Cross-validace metodou nejmenších čtverců

Cross-validace metodou nejmenších čtverců (least-squares cross-validation) je zcela automatická metoda pro výběr parametru vyhlazování. Při daném libovolném odhadu $\hat{f}$ skutečné hustoty f lze zapsat integrovanou čtvercovou chybu jako

$$\int (\hat{f} - f)^2 = \int \hat{f} - 2\int \hat{f} f - \int f^2 \tag{11}$$

Poslední člen rovnice (11) nezávisí na h, a tak ideální volba šířky okna (ve smyslu minimalizace integrované čtvercové chyby) bude odpovídat hodnotě, která minimalizuje veličinu R definovanou jako

$$R(\hat{f}) = \int \hat{f} - 2\int \hat{f}f \tag{12}$$

Základním principem metody je sestrojit odhad $R(\hat{f})$ ze samotných dat a poté tento odhad minimalizovat přes h.

Silverman (1986) ukazuje, že odhadem R je veličina

$$M_0(h) = \int \hat{f} - 2n^{-1}\sum_i \hat{f}_{-i}(x_i) \tag{13}$$

kde

$$\hat{f}_{-i}(x) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{x - x_j}{h}\right). \tag{14}$$

Skóre M_0 závisí pouze na datech (i když pro snadný výpočet není v příliš vhodné formě). Lze ukázat (viz Silverman, 1986, kapitola 3.4.3), že minimalizace M_0 odpovídá minimalizaci $MISE$.

Je nutné uvést, že celá procedura může přinášet problematické výsledky v případě, že vstupní data jsou nějakým způsobem zaokrouhlená (diskretizovaná), což je přesně případ srážek. V takovém případě může výpočet degenerovat do $M_0 \to -\infty$ když $h \to 0$. V takovém případě je lepší provádět minimalizaci pouze v určitém intervalu kolem výchozí hodnoty h, dané např. výpočtem podle reference na některé parametrické rozdělení (viz předchozí odstavec).

3.3.2.3 Likelihood cross-validace

Tato metoda je založena na podobném principu jako metoda předešlá. Vychází z toho, že na odhady hustoty $\hat{f}$ s danou kernelovou funkcí je možné pohlížet jako na parametrickou rodinu rozdělení s parametrem h (data tentokrát bereme jako fixní). Náhodný výběr, který máme pro odhad k dispozici, můžeme použít ke cross-validaci věrohodnostní funkce

$$CV(h) = \frac{1}{n}\sum_i log\left(\hat{f}_{-i}(x_i)\right) \tag{15}$$

kde symbol $\hat{f}_{-i}$ má stejný význam jako u předchozí metody, viz vztah (14).

Metoda je velmi citlivá na odlehlé hodnoty (outliery), při jejich výskytu vede volba kernelu s omezeným definičním oborem k přílišnému shlazení (příliš velké h). Rovněž

zde mohou nastat problémy v případě diskretizovaných data, stejně jako u předchozí metody.

3.3.2.4 Přímá minimalizace kvadratické chyby

Přímá minimalizace kvadratické chyby zahrnuje celý komplex metod odhadu h, založený na přímé minimalizaci aproximativního vyjádření $MISE$ podle Taylorova rozvoje. Pokud v rovnici (8) vyjádříme

$$\alpha(K) = k_2^{-2/5} \left(\int K(x)^2 \mathrm{d}x \right)^{1/5}$$

a dále

$$\beta(f) = \left(\int f''(x)^2 \mathrm{d}x \right)^{1/5},$$

optimální hodnotu h lze pak vyjádřit jako

$$h_{opt} = \alpha(K)\, \beta(f) n^{-1/5}, \tag{16}$$

kde člen $\beta(f)$ je neznámý (závisí na neznámé odhadované hustotě). Princip přímé minimalizace spočívá v tom, že provedeme odhad $\hat{\beta}(h_0)$ členu β, při kterém místo neznámé hustoty využijeme kernelový odhad hustoty s nějakou prvotní hodnotu h_0. Tento postup je pak opakován podle schématu

$$h_{i+1} = \alpha(K)\, \beta(h_i) n^{-1/5}, \tag{17}$$

dokud není dosaženo konvergence k optimální hodnotě. První koncept metody navrhnul Woodroofe (1970) a později byl rozpracován v pracích Scott et al. (1977), Sheater (1983, 1986), Park a Marron (1990) a dalších. V této práci je použita metoda z Sheater a Jones (1991). Přesný postup je obsáhlý a je rozepsán v uvedené práci a dále sumarizován v Rajagopalan et al. (1997). Pro praktickou aplikaci byla použita implementace metody v R balíku KernSmooth (Wand a Jones, 1995).

3.4 Aplikace kernelových odhadů na srážková data

Srážková data mají specifický charakter, který při aplikaci kernelových odhadů přináší některé problémy. Zaprvé se ve srážkových datech vyskytuje velké množství nul, které jsou důležitou součástí popisu časové řady srážek. Zde vzniká určitý rozpor v charakteru srážek, které jako spojitá náhodná veličina mají nulu jako významnou diskrétní hodnotu. Z tohoto důvodu se obvykle suché dny (nuly) analyzují separátně od nenulových dat. V této práci je tento postup dodržen a kernelové odhady jsou aplikovány výhradně na nenulová data.

Druhým charakteristickým rysem je to, že nenulové údaje jsou silně nahloučeny kolem počátku, kde je koncentrována hlavní část pravděpodobnosti. To způsobuje charakteristický tvar křivky hustoty pravděpodobnosti, která se v případě srážek vyznačuje vysokými hodnotami kolem počátku s následným exponenciálním poklesem.

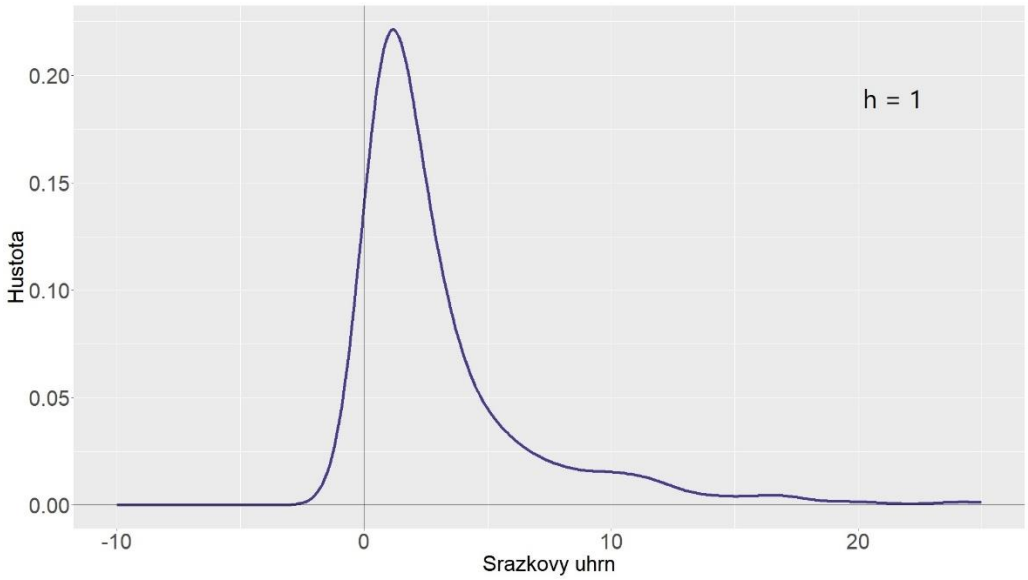
Co se týče parametrických modelů srážek, používají se obvykle ty, které mají schopnost vystihnout uvedený pokles hustoty. Typicky jsou používána tato rozdělení:

- exponenciální rozdělení (např. v Monjo et al. (2012))
- weibullovo rozdělení (např. v Castellvi et al. (2004))
- gama rozdělení (např. v Teutschbein a Seibert (2012))
- log-logistické rozdělení (např. v Monjo et al. (2014))
- mix dvou exponenciálních rozdělení (např. Wilks, 2011)

Nejběžněji je používáno gama rozdělení, které je považováno za standardní model k popisu pravděpodobnostního rozložení denních srážek.

Uvedený tvar rozdělení působí při kernelových odhadech problémy, protože výrazně ztěžuje nastavení optimální šířky okna. V levé části rozdělení kolem počátku, kde je koncentrováno velké množství dat, je zapotřebí použití spíše úzkých kernelů, které ale způsobují šum v pravé části rozdělení, kde je relativně málo hodnot. Pravá část rozdělení naopak vyžaduje spíše ploché kernely, které ale nedovolují dobře vystihnout vysoké hodnoty hustot kolem počátku. Dalším problémem je fakt, že srážky jsou ze své podstaty nezáporné, takže pravděpodobnostní funkce nesmí (neměla by) být nenulová v záporných hodnotách. Tento fakt (tzv. „boundary problem“) působí při aplikaci kernelových odhadů problémy, protože symetrické kernely umístěné v těsné blízkosti počátku mají tendenci vytvářet nenulovou pravděpodobnost i v záporných hodnotách. Pokud při jádrovém odhadu hustoty srážek použijeme kernely s omezeným definičním oborem (Epanechnikov, Bisquare),

dojde v oblasti $\langle 0, h\rangle$ ke ztrátě pravděpodobnosti do záporných hodnot a k deformaci hustoty. Pro kernely s neomezeným definičním oborem (Gaussovský kernel) to platí dvojnásob, tam k určité ztrátě pravděpodobnosti dojde vždy. V obou případech také integrál hustoty přes definiční obor srážek není rovný 1.



Obr. 4 Ilustrace „boundary problému" - jádrový odhad hustoty srážkových dat, zpracovaný pomocí dat ze stanici Dolní Dvořiště, pozorované od 01.01.1961 do 01.09.1963. Šířka okna = 1, bylo použito Gaussovo jádro.

Pro řešení uvedených problémů bylo vyvinuto několik postupů, které jsou popsány v dalších kapitolách.

3.4.1 Proměnlivá šířka kernelu

Jedna z metod pro řešení problémů, které byly popsány výše, spočívá v tom, že optimální šířku okna zvolíme různou pro různé části rozdělení. Místo jedné globální hodnoty h dostaneme specifickou hodnotu h_i pro každé jádro zvlášť. Princip metody je obecně v tom, že h_i volíme nepřímo úměrné druhé odmocnině odhadu hustoty na daném bodě $\hat{f}(x_i)$ (Abramson, 1982). To obecně zaručuje, že pro místa s vysokou hustotou je dosaženo úzkých kernelů a naopak. Procedura je iterační a postupuje se podle následujícího schémata:

1. zvolíme úvodní globální h a vytvoříme prvotní odhady $\hat{f}(x_i)$
2. pro všechna i vypočteme individuální faktory $\lambda_i = \left(\hat{f}(x_i)/g\right)^{-1/2}$, kde g je geometrický průměr $\hat{f}(x_i)$, který v praxi počítáme jako $\ln(g) = n^{-1}\sum \ln(x_i)$

3. pro všechna i spočteme upravené šířky okna jako $h_i = \lambda_i\, h$ a na jejich základě nové hodnoty $\hat{f}(x_i)$

Postup je opakován, podle Silvermana (1986) ke konvergenci stačí cca 2-3 iterace. Pro určení první hodnoty globálního h se dá použít např. reference na vybrané parametrické rozdělení. Výsledné odhady s individuálně nastavenými šířkami jednotlivých kernelů počítáme podle

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - x_i}{h_i}\right) \tag{18}$$

V literatuře je tento postup většinou označován jako „adaptive kernel“ nebo „variable kernel“. Tato metoda tedy řeší hlavně přílišné shlazení píků rozdělení a přílišnou drsnost konců rozdělení, únik pravděpodobnosti do záporných hodnot („boundary problem“) je řešen jen částečně spíše jako vedlejší efekt.

3.4.2 Metoda cut and normalize

Metoda představuje velmi přímočarý postup k řešení „boundary“ problému. Spočívá v tom, že pro každý použitý kernel vypočítáme plochu, která unikne do záporné oblasti a zbytek kernelu se normalizuje tak, aby měl odpovídající integrál v povolené oblasti. Nejjednodušší příklad je ten, že vypočteme globální h_{opt} a potom omezíme hodnotu h_i pro každý kernel tak, že nastavíme

$$h_i = \min(h_{opt}, x_i) \tag{19}$$

kde x_i je daný bod náhodného výběru, na kterém je umístěn příslušný kernel s šířkou okna h_i. Metoda neřeší žádné problematické aspekty, které jejím použitím vzniknou, např. deformaci tvaru hustoty na definičním oboru srážek v okolí počátku.

3.4.3 Speciální kernely

Jiným přístupem je použití kernelů specificky vytvořených pro oblast v blízkosti počátku. Příklady takových kernelů uvádí např. Muller (1991), nicméně tyto kernely jsou asymetrické a v některých případech jsou v určité své části záporné, takže samy o sobě nepředstavují hustotu. Jejich použití pro řešení boundary problému není obvyklé a jsou zde uvedeny pouze pro úplnost.

Odlišné kernely navrhl ve své práci Chen (2000), který jako základ kernelu zvolil hustotu gama rozdělení. Parametr tvaru rozdělení je v rámci odhadu počítán pro každý bod náhodného výběru zvlášť a výsledný odhad tak poskytuje automatickou adaptaci tvaru rozdělení na nahloučení hodnot kolem počátku. Jádrový odhad hustoty s použitím tohoto tzv. gama kernelu má tvar

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f_{\gamma}\left(x, \frac{x_i}{h} + 1, 1/h\right) \tag{20}$$

kde $f_{\gamma}(x, \alpha, \beta)$ je hustota gama rozdělení s parametry tvaru (α) a měřítka (β). Tento způsob odhadu úspěšně aplikovali např. Peel a Wilson (2008).

3.4.4 Logaritmická transformace dat

Metoda je obecně použitelná pro jakákoliv data nahloučená v blízkosti počátku a spočívá v tom, že se ještě před odhadem hustoty provede logaritmická transformace náhodného výběru, výpočet h a odhad hustoty se provede nějakou z dříve uvedených metod pro logaritmovaná data a výsledná hustota je pak převedena zpět. Celá procedura se dá zapsat jako

$$\hat{f}(x) = \frac{1}{nh_L}\sum_{i=1}^{n} K\left(\frac{\log(x) - \log(x_i)}{h_L}\right) \tag{21}$$

kde h_L je šířka okna odhadnutá pro logaritmovaná data. Metoda z podstaty poskytuje jak automatickou adaptaci šířky okna jednotlivých kernelů, tak také zabraňuje úniku pravděpodobnosti do záporných hodnot.

4. Metodika praktické části

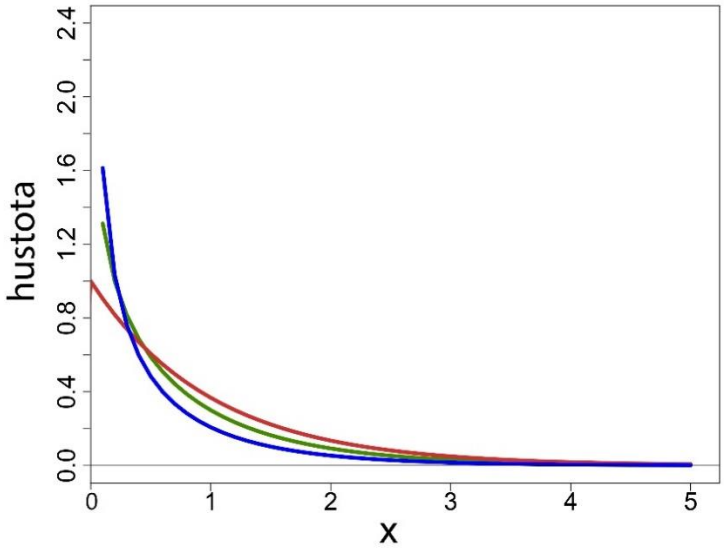
Praktická část měla dva cíle. Prvním cílem bylo nalezení optimálního způsobu jádrového odhadu hustoty (tedy volby kernelu a šířky okna) pro denní srážková data. Druhým cílem bylo porovnat účinnost vybraného způsobu s parametrickými odhady hustoty. Veškeré výpočty byly prováděny v jazyce R.

4.1 Výběr optimálního způsobu jádrového odhadu

Cílem této části bylo vzájemným porovnáním metod popsaných v předchozí části práce vybrat optimální způsob výběru kernelu a šířky okna pro odhad hustoty denních srážkových dat. Pro testování bylo vybráno těchto osm způsobů:

Metoda	**Postup**
Reference na parametrické rozdělení	rovnice (10)
Cross-validace metodou nejmenších čtverců	rovnice (13)
Likelihood cross-validace	rovnice (15)
Přímá minimalizace kvadratické chyby	rovnice (17)
Proměnlivá šířka kernelu	rovnice (18)
Metoda Cut and normalize	rovnice (19)
Gama kernel	rovnice (20)
Logaritmická transformace dat	rovnice (21)

Při porovnávání metod bylo nutné posoudit shodu rozdělení vzorových data a provedeného jádrového odhadu, proto jako vzorová data pro testování byly použity samply z gama rozdělení, které je běžně používaným parametrickým modelem. Dá se tedy předpokládat, že pokud bude metoda úspěšná při rekonstrukci gama rozdělení, bude také použitelná pro srážková data. Aby bylo možné posoudit schopnost jádrových odhadů vystihnout různý tvar rozdělení, byly používány samply z gama rozdělení s parametry tvaru 0.5, 0.75 a 1, parametr měřítka byl vždy 1.



Obr. 5 Hustoty gama-rozdělení s parametry tvaru 0.5 (modrá čára), 0.75 (zelená čára) a 1 (červená čára).

Pro posouzení shody rozdělení samplu a kernelového odhadu byly použity dvě statistiky:

1. jednovýběrový Kolmogorovův-Smirnovův (KS) test
2. integrál rozdílu hustot

KS test testuje hypotézu, že náhodný výběr pochází z rozdělení s distribuční funkcí $F(x)$. Používá testovací statistiku ve tvaru

$$D_n = \max_x |F_n(x) - F(x)|.$$

kde $F_n(x)$ je empirická distribuční funkce testovaného výběru. Při praktické aplikaci byla nalezena největší odchylka mezi distribuční funkcí vzorového gama rozdělení a distribuční funkcí vypočítanou z jádrového odhadu. Hladina významnosti testu byla zvolena 5%, ale spíše než výsledku testu (který poskytuje binární informaci) byla pozornost věnována konkrétním hodnotám D_n, které umožňovaly porovnání různých metod jádrového odhadu. Integrál rozdílu hustot byl zvolen jako doplňující kritérium, protože KS test zaznamenává pouze maximální rozdíl, tzn. hodnotu v jediném bodě. Integrál rozdílu hustot je dále značen jako I a je popsán následujícím způsobem:

$$I = \int_x |pdf_\Gamma(x) - pdf_{kernel}(x)|\, dx. \tag{22}$$

Každá metoda jádrového odhadu byla postupně testována pro odhad všech tří tvarů gama rozdělení, pro každý bylo použito 100 samplů o velikosti 1000 bodů. Výsledkem testování jedné metody tedy byl soubor obsahující 3-krát 100 hodnot testovací statistiky D_n (spolu s výsledky testů) a 3-krát 100 hodnot integrálu rozdílů hustot.

Pro přehlednost dál v praktické části jsou výsledky testů sumarizovány do tabulek, střední hodnotu značíme jako E a směrodatnou odchylku jako σ.

4.2 Porovnání jádrového a parametrického odhadu

Metoda jádrového odhadu, která byla v první části na základě provedených testů vyhodnocena jako nejlepší, byla následně použita pro odhad hustoty reálných měřených srážek a výsledky byly porovnány s parametrickým odhadem.

Jako reprezentant parametrického odhadu hustoty bylo vybráno gama rozdělení odhadované metodou maximální věrohodnosti. Uvedená kombinace představuje zřejmě nejčastější a všeobecně akceptovaný postup při práci se srážkovými úhrny, pro praktickou aplikaci byla použita implementace v R balíku univariateML (Moss, 2019).

Jako testovací data byly použity denní srážkové úhrny z 5 stanic na povodí řeky Malše: Benešov nad Černou, Besednice, České Budějovice, Dolní Dvořiště a Kaplice. K dispozici byly 38leté řady denních úhrnů v období od 1. 1. 1961 do 31. 12. 1998.

Protože skutečné rozdělení použitých srážkových dat je neznámé, bylo využito toho, že srážková data jsou měřena s přesností na 0.1 mm, takže měřené údaje jsou vlastně k dispozici ve formě podrobného histogramu. Při velkém množství dat, které byly k dispozici, je možné histogram považovat za dobrý odhad skutečné hustoty měřených dat. Obě odhadnuté hustoty (gama a jádrový odhad) byly tedy porovnávány s histogramem srážkových dat, jednak v jeho nejpodrobnějším nastavení (šířka sloupce 0.1) a dále i s širšími sloupci. Hodnotícím kritériem byl zvolen integrál rozdílu hustot (mezi histogramem a odhadem), pomocí kterého byla porovnávána úspěšnost jádrového a parametrického odhadu.

5. Výsledky

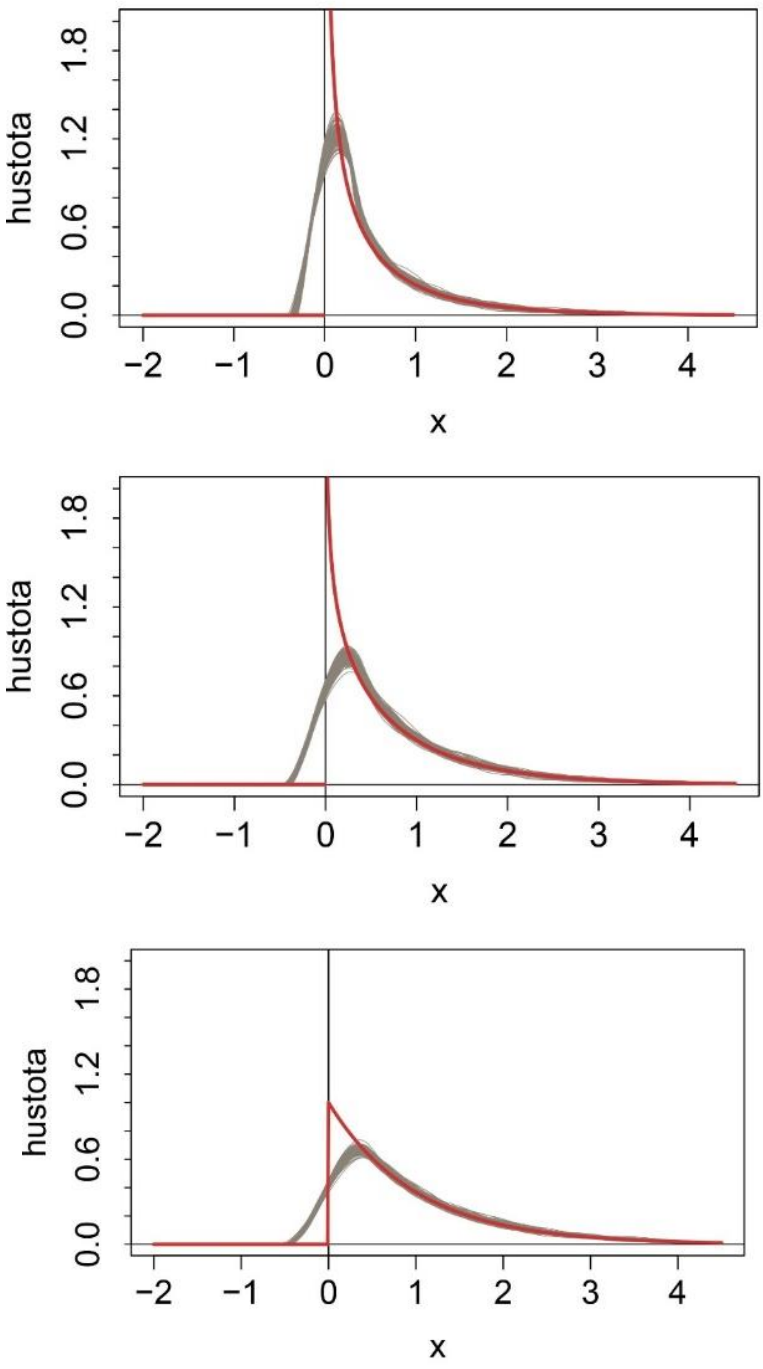
5.1 Porovnání metod jádrového odhadu

V následujících kapitolách je uveden stručný souhrn výsledků porovnání jednotlivých metod jádrového odhadu pomocí syntetických data z daných gama rozdělení.

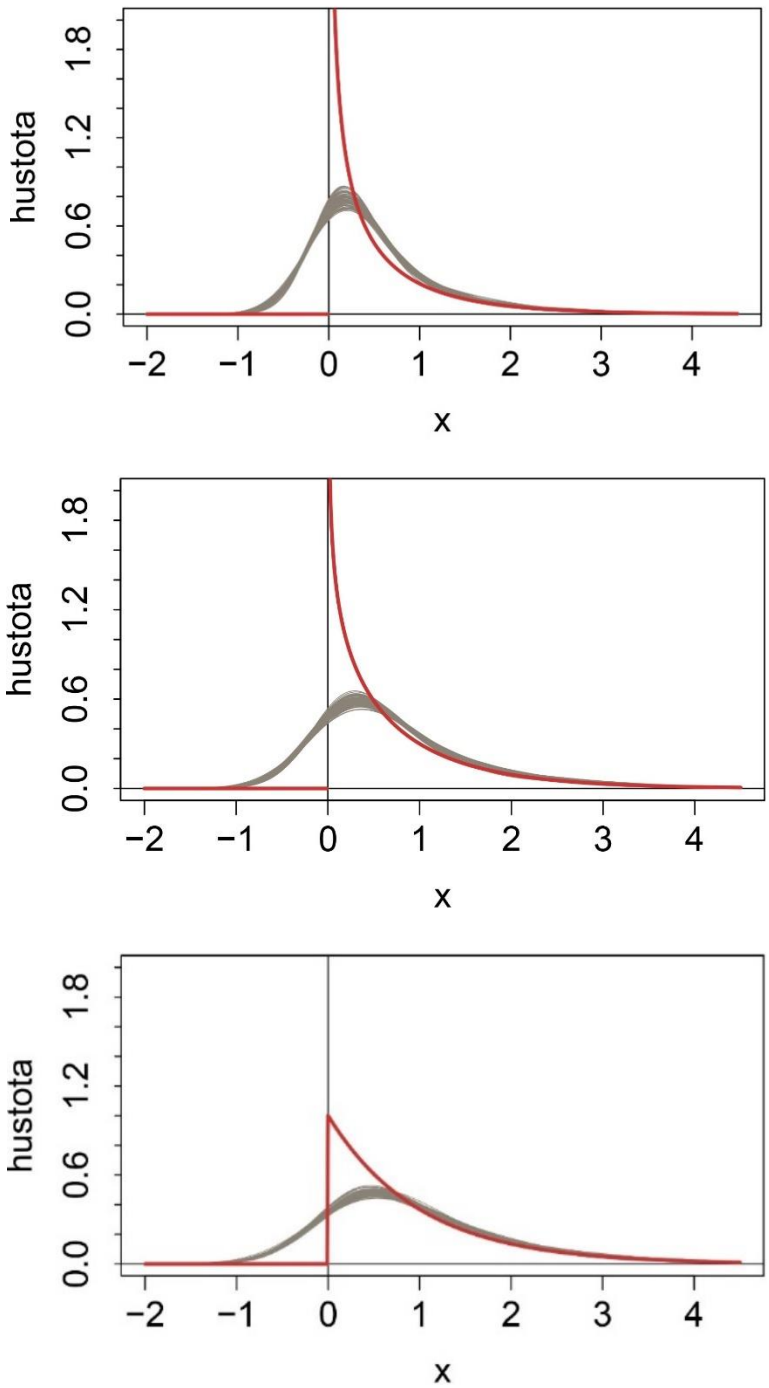
5.1.1 Reference na parametrické rozdělení

Šířka okna pro jádrový odhad byla počítána ze vztahu (10), který předpokládá, že data mají exponenciální rozdělení, což zhruba odpovídá srážkovým datům i gama rozdělením, použitým pro testovací účely.

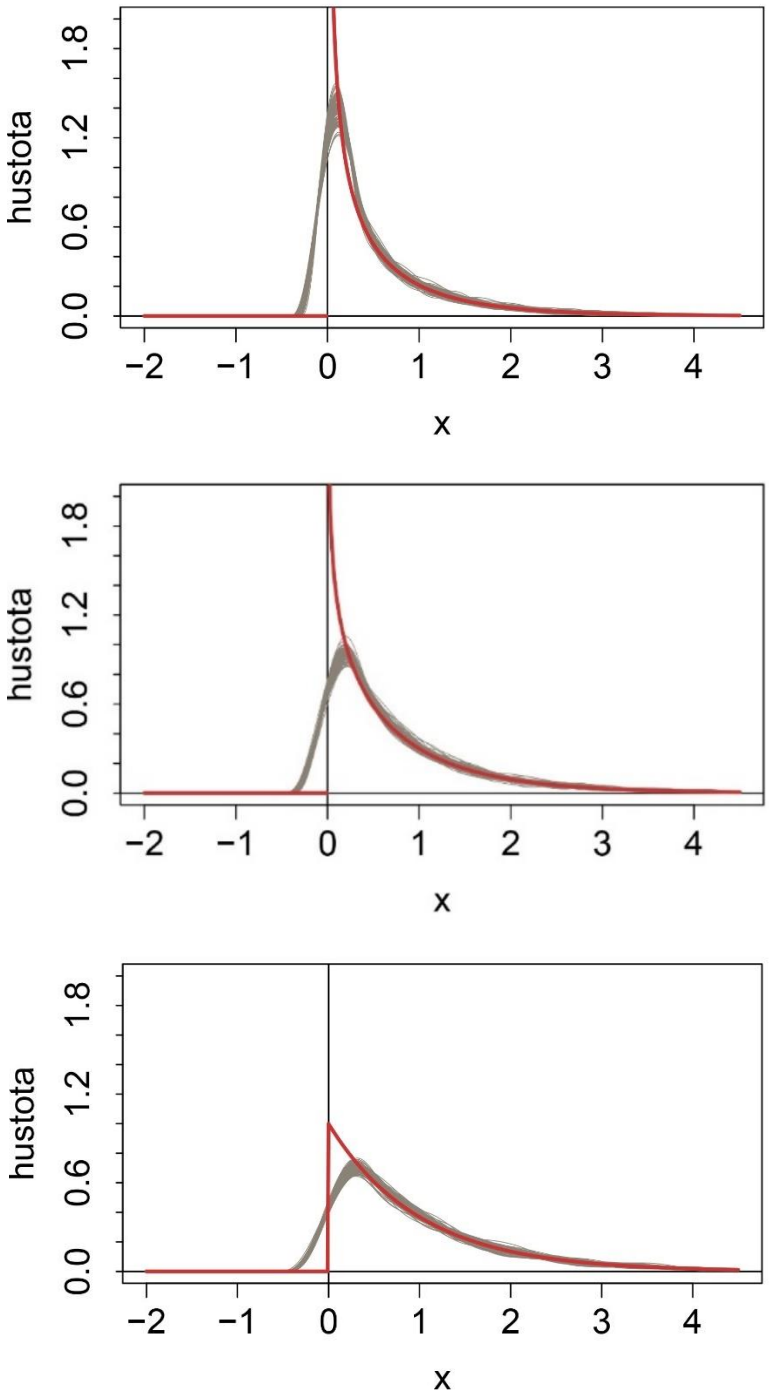
Při testování této metody bylo také provedeno porovnání účinnosti odhadu při použití tří různých kernelových funkcí: Epanechnikov, Gauss a Bisquare. Následující tři obrázky ilustrují porovnání výsledků dosažených různými kernely a pro různé tvary vzorového gama rozdělení. Tabulka pak obsahuje stručnou sumarizaci číselných výsledků.



Obr. 6 Vizuální porovnání výsledků při použití Epanechnikovova kernelu: červená čára představuje vzorovou hustotu gama rozdělení pro parametr tvaru 0.5 (nahoře), 0.75 (uprostřed) a 1 (dole). Šedé čáry představují jádrové odhady ze 100 různých samplů daného gama rozdělení.



Obr. 7 Vizuální porovnání výsledků při použití Gaussova kernelu: červená čára představuje vzorovou hustotu gama rozdělení pro parametr tvaru 0.5 (nahoře), 0.75 (uprostřed) a 1 (dole). Šedé čáry představují jádrové odhady ze 100 různých samplů daného gama rozdělení.



Obr. 8 Vizuální porovnání výsledků při použití Bisquare kernelu: červená čára představuje vzorovou hustotu gama rozdělení pro parametr tvaru 0.5 (nahoře), 0.75 (uprostřed) a 1 (dole). Šedé čáry představují jádrové odhady ze 100 různých samplů daného gama rozdělení.

Kernel	Poměr úspěšných testů	$E(D_n)$	$\sigma(D_n)$	$E(I)$	$\sigma(I)$
Parametr tvaru = 0.5					
Epanechnikov	0	0.18096	0.00747	0.39416	0.01356
Gaussov	0	0.24580	0.00714	0.56365	0.01766
Bisquare	0	0.16674	0.00639	0.35247	0.01488
Parametr tvaru = 0.75					
Epanechnikov	0	0.12039	0.00574	0.28918	0.01231
Gaussov	0	0.18842	0.00780	0.46731	0.01853
Bisquare	0	0.10641	0.00576	0.25370	0.01114
Parametr tvaru = 1					
Epanechnikov	0	0.08112	0.00481	0.21677	0.01471
Gaussov	0	0.15032	0.00705	0.39020	0.02176
Bisquare	0.27	0.06908	0.00440	0.19188	0.01404

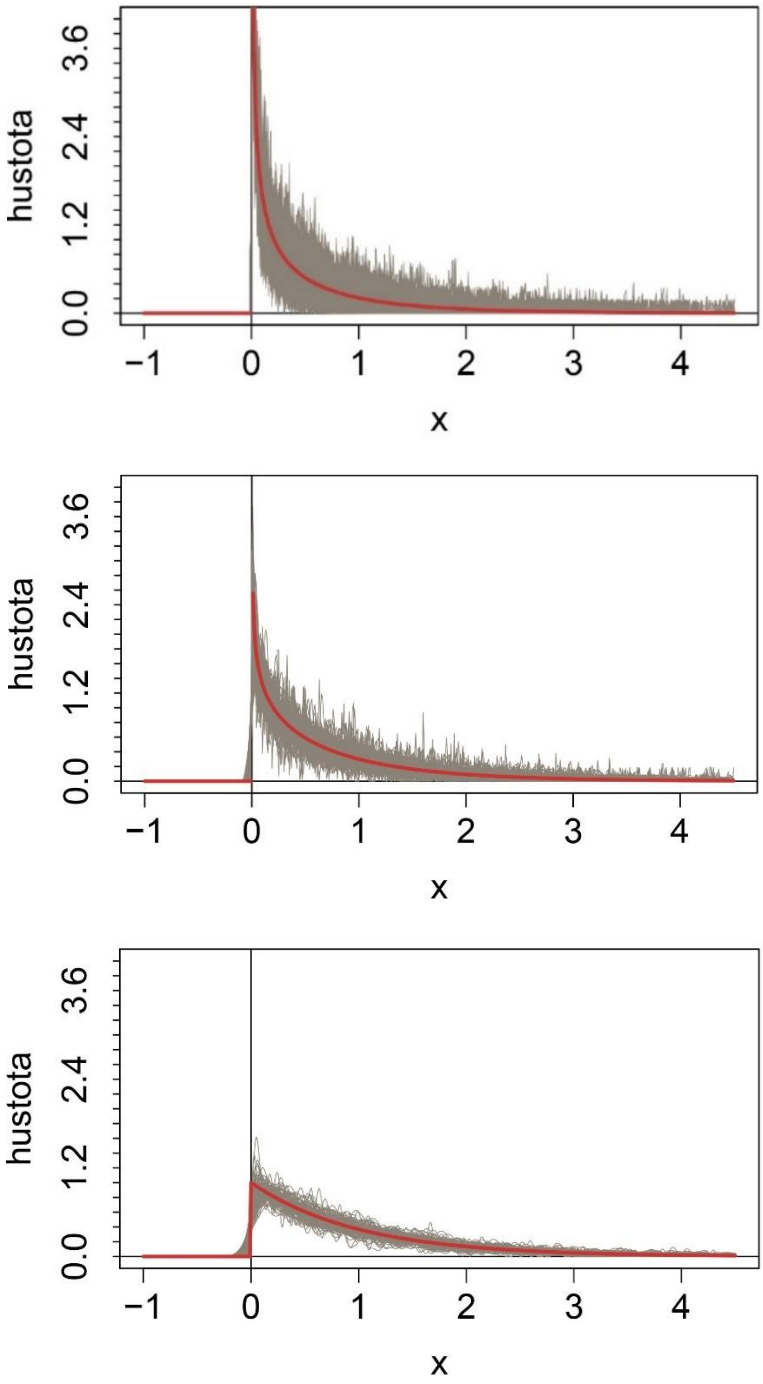
Tab. 2 Výsledky analýzy metody reference na parametrické rozdělení.

Z vizuálního porovnání i z číselných výsledků plyne, že použití Gaussova jádra je nevhodné pro data s exponenciálním tvarem rozdělení, únik pravděpodobnosti do záporných hodnot je ze všech tří jader největší. Tomu napomáhá i neomezený definiční obor jádra. Ostatní dvě jádra dosahují podobných výsledků, ale ve všech ukazatelích jsou trochu úspěšnější odhady provedené s jádrem Bisquare. Z těchto důvodů bude v dalších analýzách používáno pouze jádro Bisquare, protože po provedení několika zkušebních odhadů s dalšími metodami vyplynulo, že bychom s jádry Epanechnikov a Bisquare dosahovali velmi podobných výsledků. To je také ve shodě s poznatky z literatury (např. Silverman, 1986).

Z výsledků také plyne, že podle očekávání je nejvíce problematický odhad pro parametr tvaru 0.5, kde je v okolí počátku soustředěno nejvíce hustoty, křivka hustoty je nejvíce strmá a ztráty odhadu do záporných hodnot jsou nejvyšší.

5.1.2 Cross-validace metodou nejmenších čtverců

Šířka okna pro jádrový odhad byla počítána minimalizací výrazu (13). Používáno bylo již jen jádro Bisquare.



Obr. 9 Vizuální porovnání výsledků: červená čára představuje vzorovou hustotu gama rozdělení pro parametr tvaru 0.5 (nahoře), 0.75 (uprostřed) a 1 (dole). Šedé čáry představují jádrové odhady ze 100 různých samplů daného gama rozdělení.

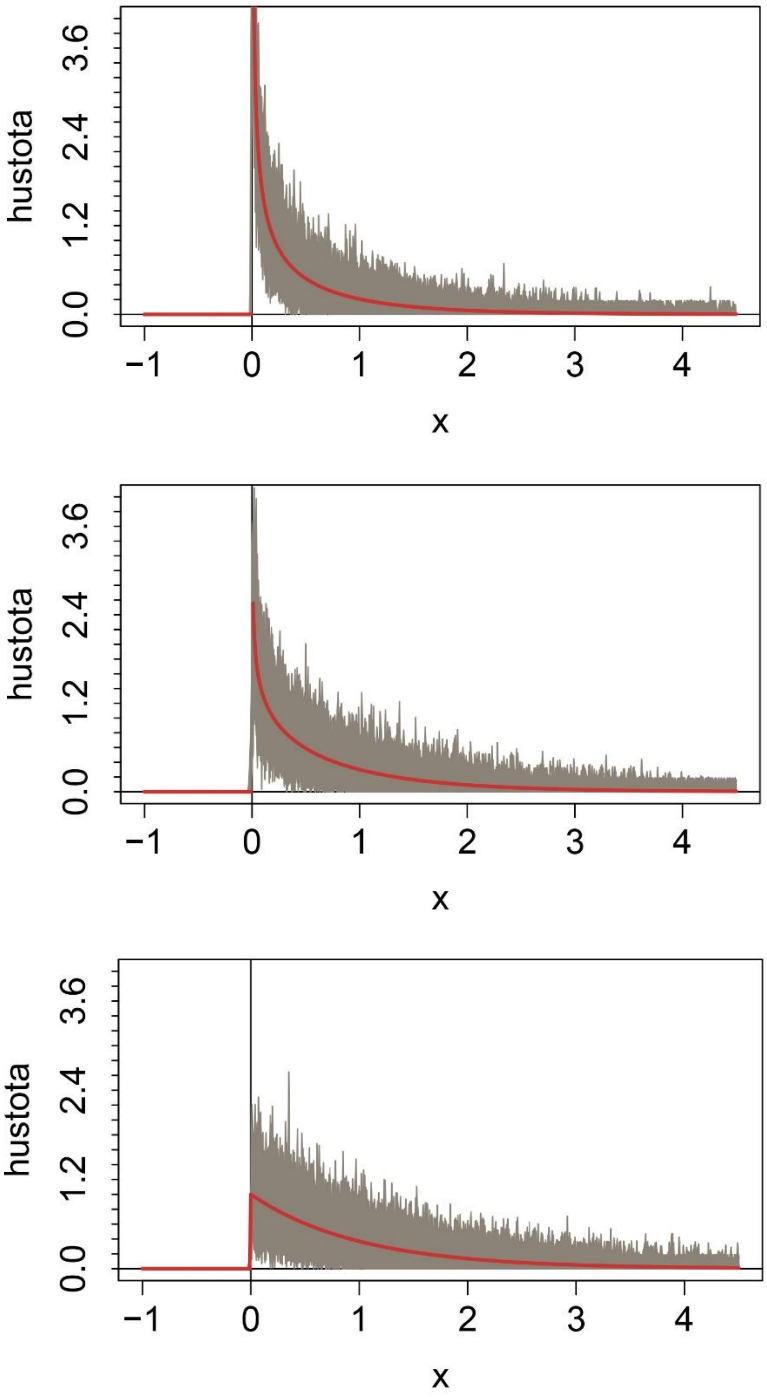
Parametr tvaru	Poměr úspěšných testů	$E(D_n)$	$\sigma(D_n)$	$E(I)$	$\sigma(I)$
0.5	1	0.04219	0.00873	0.31495	0.02209
0.75	1	0.03509	0.01008	0.24981	0.03825
1	1	0.03229	0.01105	0.19885	0.03749

Tab. 3 Výsledky analýzy cross-validace metodou nejmenších čtverců.

Z hlediska Kolmogorovova-Smirnovova testu metoda dosahuje velmi kvalitních výsledků při libovolném tvaru vzorového gama rozdělení. Grafy však ukazují velkou rozkolísanost výsledných hustot, která odpovídá tomu, že cross-validace má tendenci produkovat příliš malé hodnoty šířky okna. Tomu také odpovídají hodnoty integrálů rozdílů hustot, které jsou obdobné jako u reference na parametrické rozdělení (s jádrem Bisquare). Tyto výsledky také ukazují, že samotné testování (např. KS testem) nevypovídá o výsledcích vše a ukazuje se důležitost vizuální kontroly výsledků.

5.1.3 Likelihood cross-validace

Šířka okna pro jádrový odhad byla počítána pomocí rovnice (15).



Obr.10 Vizuální porovnání výsledků: červená čára představuje vzorovou hustotu gama rozdělení pro parametr tvaru 0.5 (nahoře), 0.75 (uprostřed) a 1 (dole). Šedé čáry představují jádrové odhady ze 100 různých samplů daného gama rozdělení.

Parametr tvaru	Poměr úspěšných testů	$E(D_n)$	$\sigma(D_n)$	$E(I)$	$\sigma(I)$
0.5	1	0.04271	0.00765	0.32123	0.02007
0.75	1	0.03408	0.01026	0.37751	0.03098
1	1	0.03561	0.01134	0.42507	0.03298

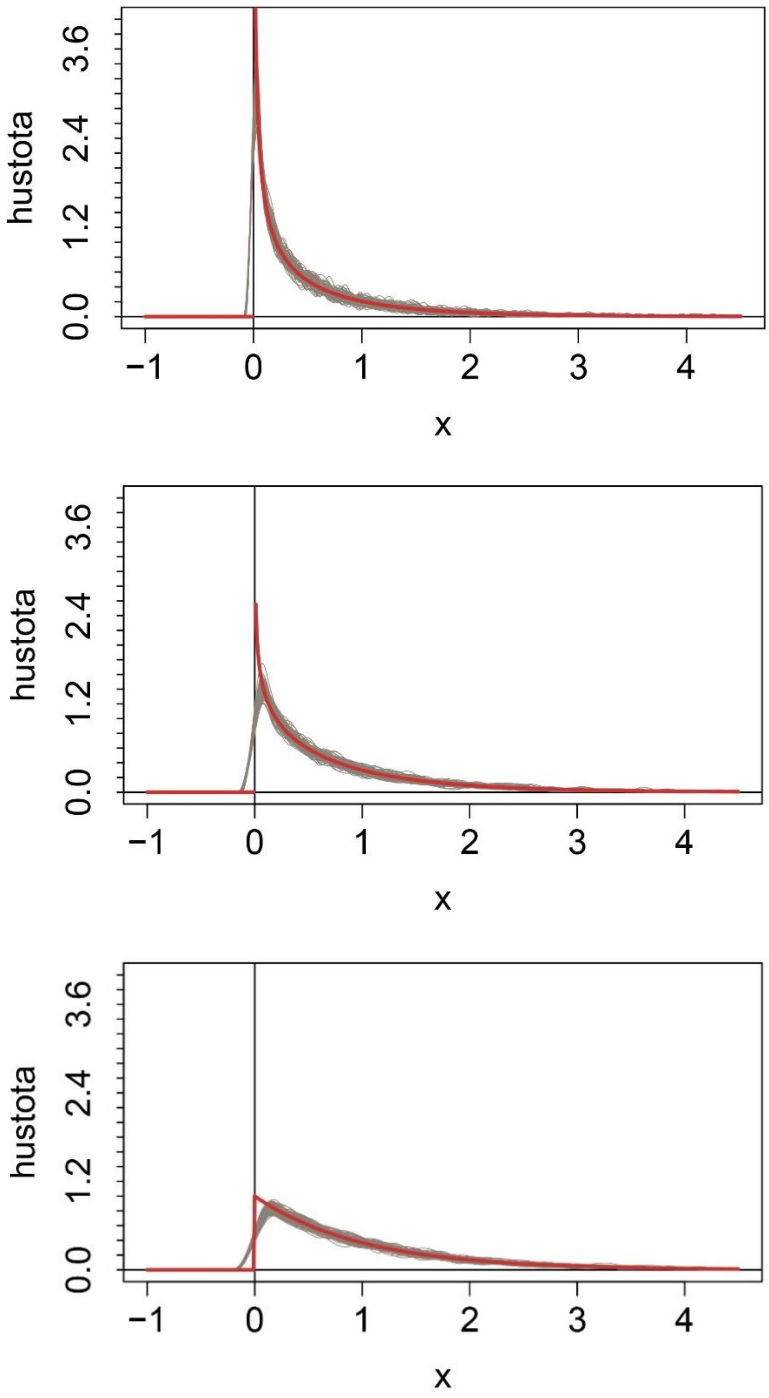
Tab. 4 Výsledky analýzy metody likelihood cross-validace.

Metoda likelihood-cross-validace ukazuje kvalitní výsledky podle Kolmogorovova-Smirnovova testu se všemi třemi parametry tvaru gama rozdělení. Nicméně při vizuálním zhodnocení se objevuje stejná situace jako u cross-validace metodou nejmenších čtverců – metoda produkuje příliš malé hodnoty šířky okna, což způsobuje vysokou rozkolísanost odhadu a je také odraženo ve vysokých hodnotách integrálu rozdílu hustot. Na základě vizuálního posouzení lze konstatovat, že metoda je prakticky nepoužitelná, přestože z hlediska KS testu vychází výsledky dobře.

Zajímavostí je, že z hlediska kritéria I bylo dosaženo nejhoršího výsledku pro gama rozdělení s parametrem tvaru 1, u kterého bylo u předchozích metod dosaženo naopak nejlepších výsledků.

5.1.4 Přímá minimalizace kvadratické chyby

Pro přímou minimalizace kvadratické chyby byla šířka okna počítána pomocí výrazu (17), využita byla implementace metody v R balíku KernSmooth (Wand a Jones, 1995). Jako v předchozích případech bylo používáno jádro Bisquare.



Obr. 11 Vizuální porovnání výsledků: červená čára představuje vzorovou hustotu gama rozdělení pro parametr tvaru 0.5 (nahoře), 0.75 (uprostřed) a 1 (dole). Šedé čáry představují jádrové odhady ze 100 různých samplů daného gama rozdělení.

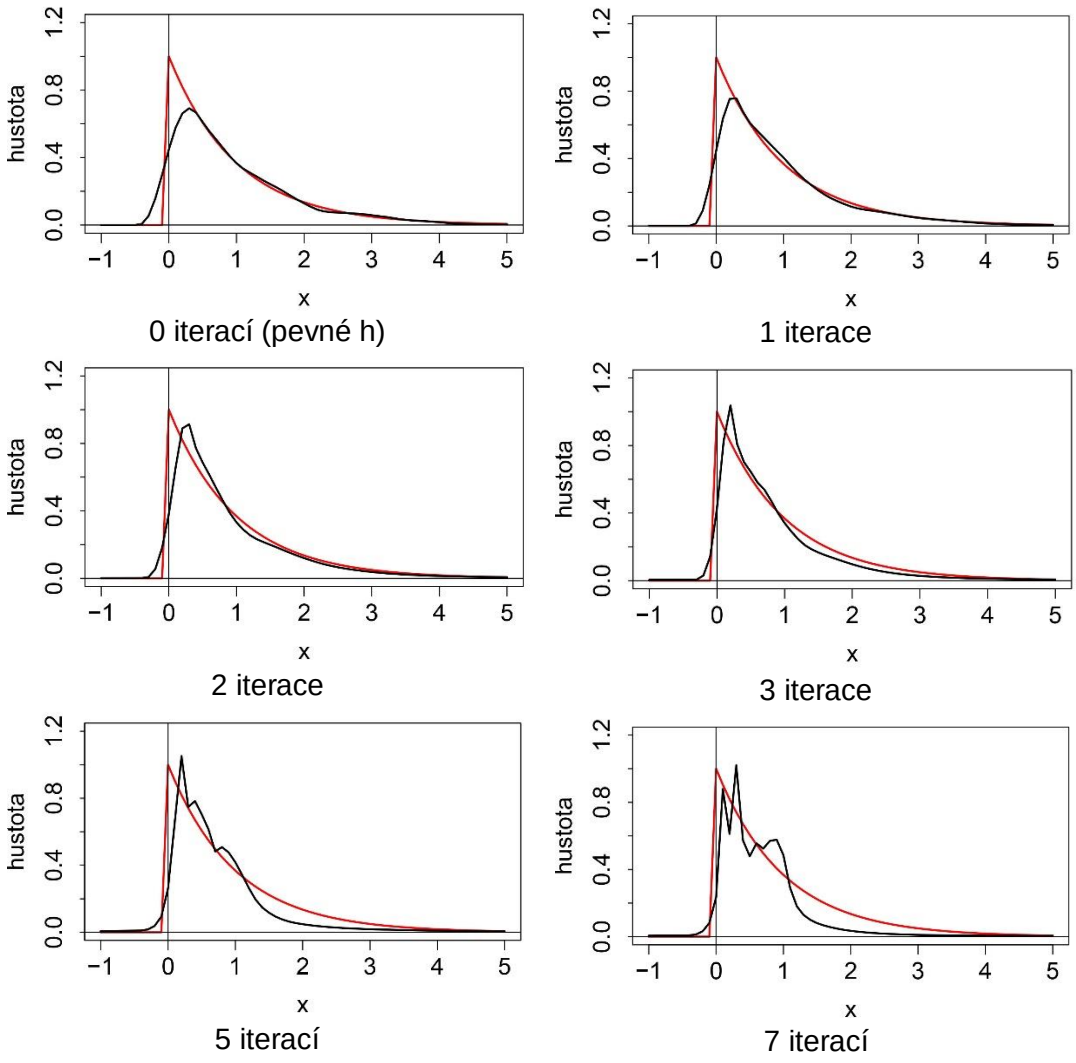
Parametr tvaru	Poměr úspěšných testů	$E(D_n)$	$\sigma(D_n)$	$E(I)$	$\sigma(I)$
0.5	0	0.08447	0.00447	0.19821	0.01307
0.75	1	0.05075	0.00354	0.16556	0.01194
1	1	0.03286	0.00326	0.13385	0.01362

Tab. 5 Výsledky analýzy přímé minimalizace kvadratické chyby.

Přímá minimalizace kvadratické chyby ukázala dobré výsledky pro parametry tvaru 0.75 a 1, pro gama hustotu s parametrem 0.5 nastávají výrazné chyby v okolí počátku, což vedlo k zamítnutí hypotézy o shodě rozdělení ve všech případech. Podle grafů je zřejmé, že odhad je rozkolísaný a taky se stále částečně nachází v oblasti záporných hodnot. Rozkolísanost je však výrazně zredukována oproti cross-validačním metodám, což se odráží v hodnotách integrálu rozdílů hustot, které jsou oproti předchozím testovaným metodám výrazně zredukovány.

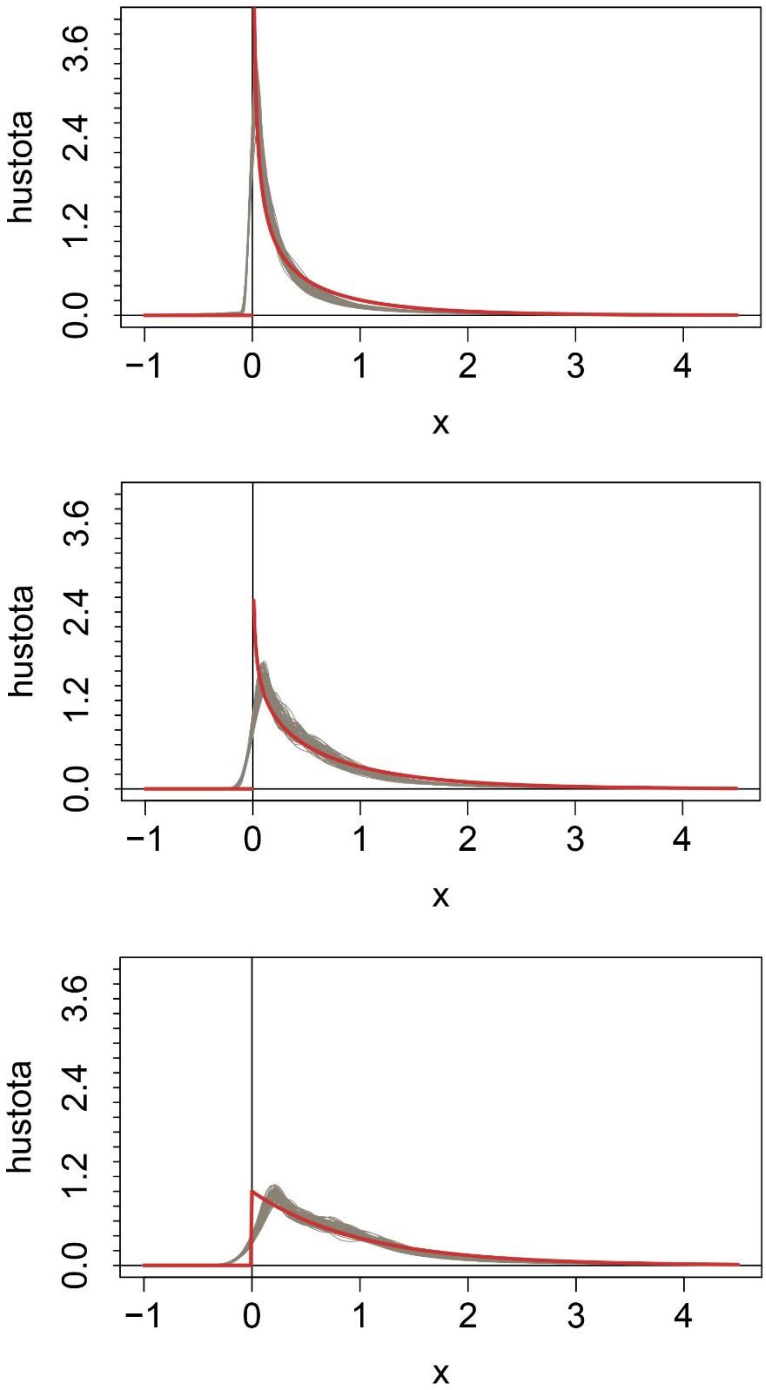
5.1.5 Proměnlivá šířka kernelu

Úvodní globální hodnota šířky okna byla spočtena metodou parametrické reference, individuální hodnoty h_i byly spočteny iteračním postupem popsaným v kapitole 3.4.1 a aplikovány podle rovnice (18). Před testem metody byl nejprve testován potřebný počet iterací. Na samplech z gama rozdělení byla provedena řada testů, ze kterých vyplynulo, že nejlepších výsledků je v průměru dosahováno při použití 3 iterací. Pro menší počet iterací nebyl efekt adaptivního kernelu dostatečný, pro vyšší počty má metoda tendenci produkovat příliš úzké kernely v oblasti počátku a dochází tak k velké rozkolísanosti odhadu. Toto zjištění je v souladu se závěry publikovanými v Silverman (1986). Následující obrázek ilustruje tuto skutečnost na syntetických datech.



Obr. 12 Ilustrace vlivu počtu iterací na výsledek odhadu s proměnlivou šířkou jádra. Červená čára představuje vzorovou hustotu gama rozdělení (parametr tvaru 1), černá čára kernelový odhad.

Nasledující obrázek a tabulka ilustrují výsledky pro 3 iterace.



Obr. 13 Vizuální porovnání výsledků: červená čára představuje vzorovou hustotu gama rozdělení pro parametr tvaru 0.5 (nahoře), 0.75 (uprostřed) a 1 (dole). Šedé čáry představují jádrové odhady ze 100 různých samplů daného gama rozdělení. Počet iterací 3.

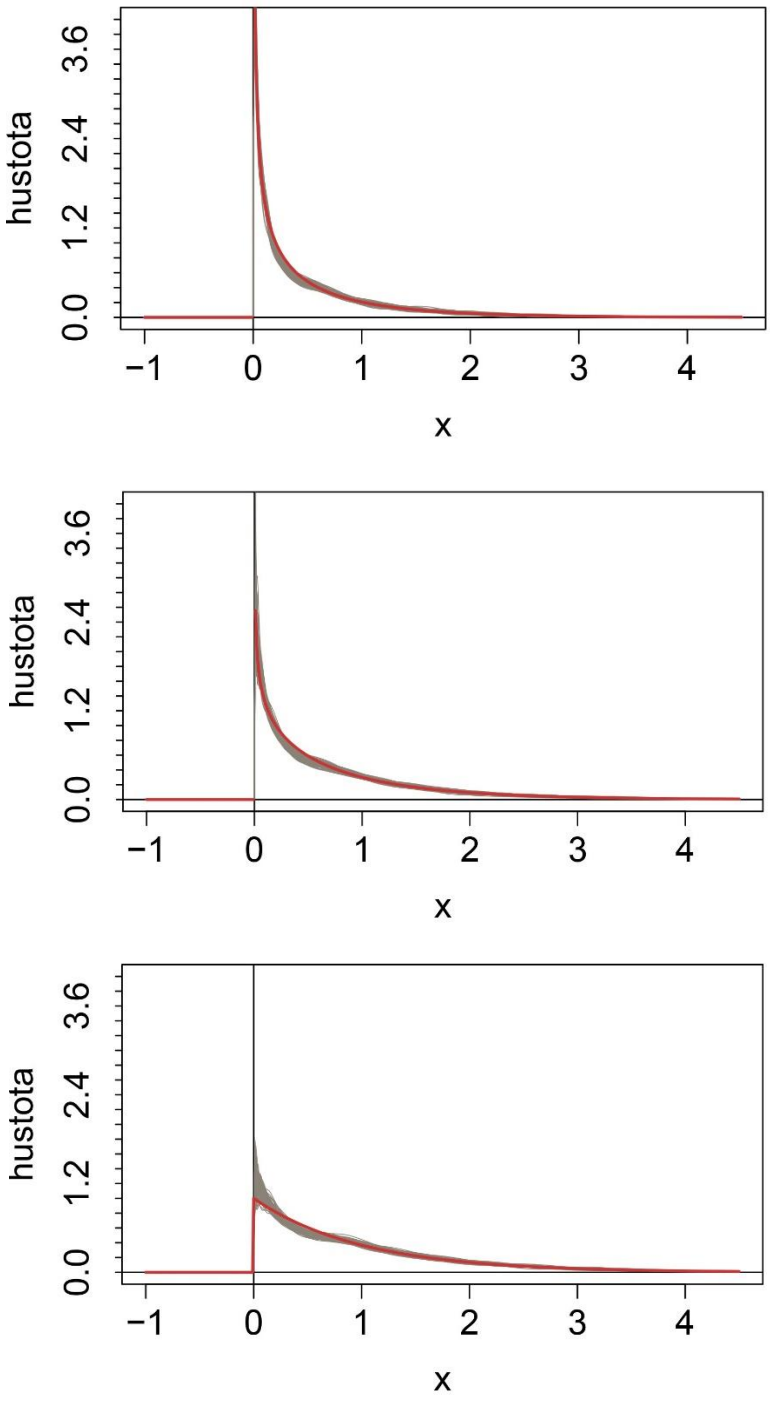
Parametr tvaru	Poměr úspěšných testů	$E(D_n)$	$\sigma(D_n)$	$E(I)$	$\sigma(I)$
0.5	0	0.10371	0.00606	0.31366	0.03253
0.75	0.81	0.05900	0.00769	0.25614	0.03333
1	0.93	0.04893	0.00941	0.23673	0.02983

Tab. 6 Výsledky analýzy metody proměnlivá šířka kernelu.

Oproti předchozí metodě došlo k citelnému zvýšení chybových kritérií, což je patrné hlavně u integrálu rozdílů hustot, přestože výsledky vypadají opticky velice dobře. Nulové procento úspěšných KS testů pro parametr tvaru 0.5 je dáno odklonem jádrového odhadu v oblasti kolem počátku, kdy k zamítnutí nulové hypotézy stačí jedna výrazná odchylka od vzorové hustoty, která je pro tento tvar rozdělení kolem počátku velmi strmá.

5.1.6 Cut and Normalize

Úvodní globální šířka okna byla spočtena metodou reference na parametrické rozdělení a pro úpravu jednotlivých kernelů byl posléze použit vzoreček (19). Jako v předchozích odstavcích bylo používáno už jen jádro Bisquare.



Obr. 14 Vizuální porovnání výsledků: červená čára představuje vzorovou hustotu gama rozdělení pro parametr tvaru 0.5 (nahoře), 0.75 (uprostřed) a 1 (dole). Šedé čáry představují jádrové odhady ze 100 různých samplů daného gama rozdělení.

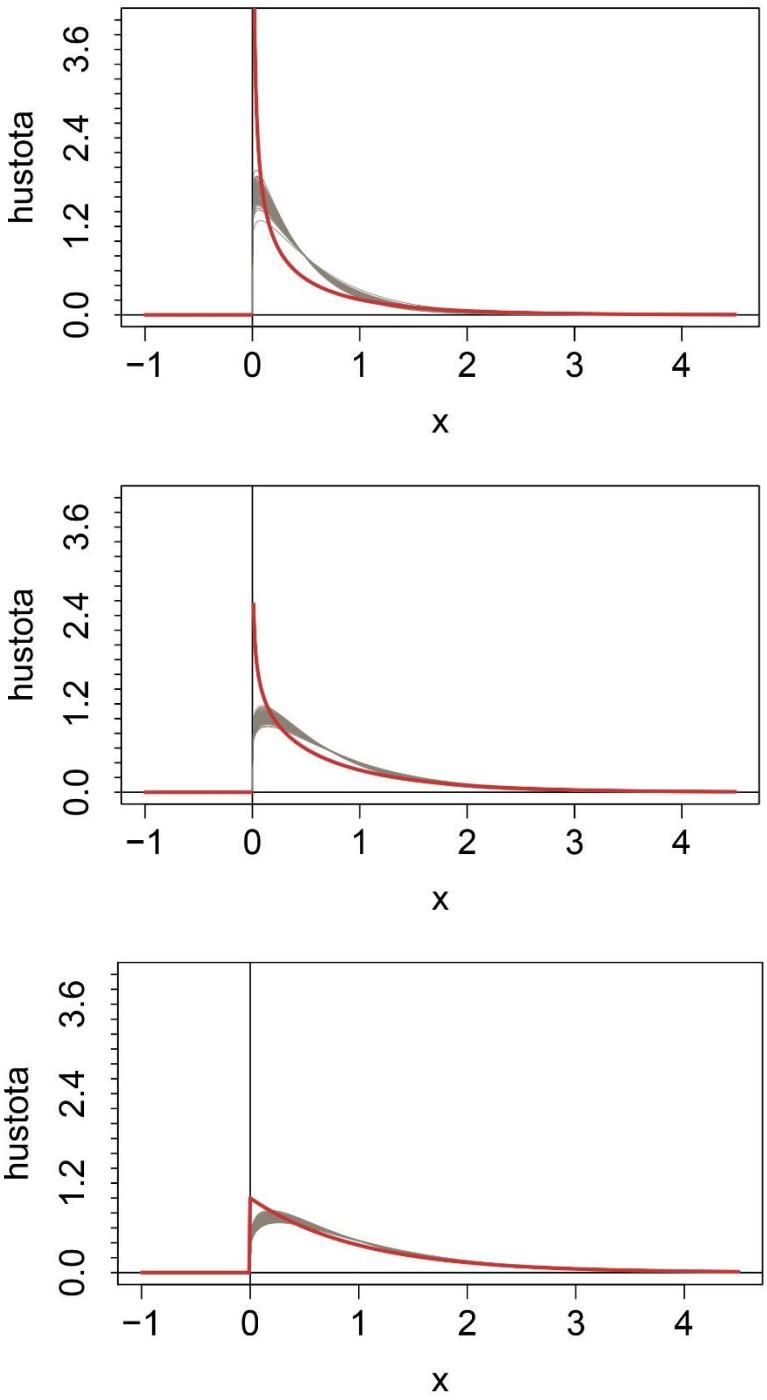
Parametr tvaru	Poměr úspěšných testů	$E(D_n)$	$\sigma(D_n)$	$E(I)$	$\sigma(I)$
0.5	0	0.09832	0.01020	0.07987	0.01559
0.75	1	0.03695	0.01047	0.08831	0.01866
1	1	0.02667	0.00779	0.10092	0.01652

Tab. 7 Výsledky analýzy metody Cut and Normalize.

Výsledky ukazují, že při použití kernelu s omezeným definičním oborem metoda řeší únik pravděpodobnosti do záporných hodnot, což je zřejmé už z definice ve vzorci (19). Hodnoty D_n pro tvar 0.5 jsou několikanásobně vyšší než pro ostatní tvary, což je způsobeno výraznými rozdíly mezi vzorovou a odhadnutou hustotou právě v oblasti kolem počátku. Pro ostatní parametry funguje metoda dobře a celkově lze shrnout, že touto metodou je dosaženo zatím nejlepších výsledků. Integrál rozdílů hustot je oproti předchozím metodám několikanásobně redukovaný.

5.1.7 Gama kernel

Odhad hustoty byl počítán s využitím gama kernelu podle výrazu (20). Pro výpočet optimální hodnoty globálního h bylo testováno několik metod, nakonec byla použita metoda likelihood cross-validace.



Obr. 15 Vizuální porovnání výsledků: červená čára představuje vzorovou hustotu gama rozdělení pro parametr tvaru 0.5 (nahoře), 0.75 (uprostřed) a 1 (dole). Šedé čáry představují jádrové odhady ze 100 různých samplů daného gama rozdělení.

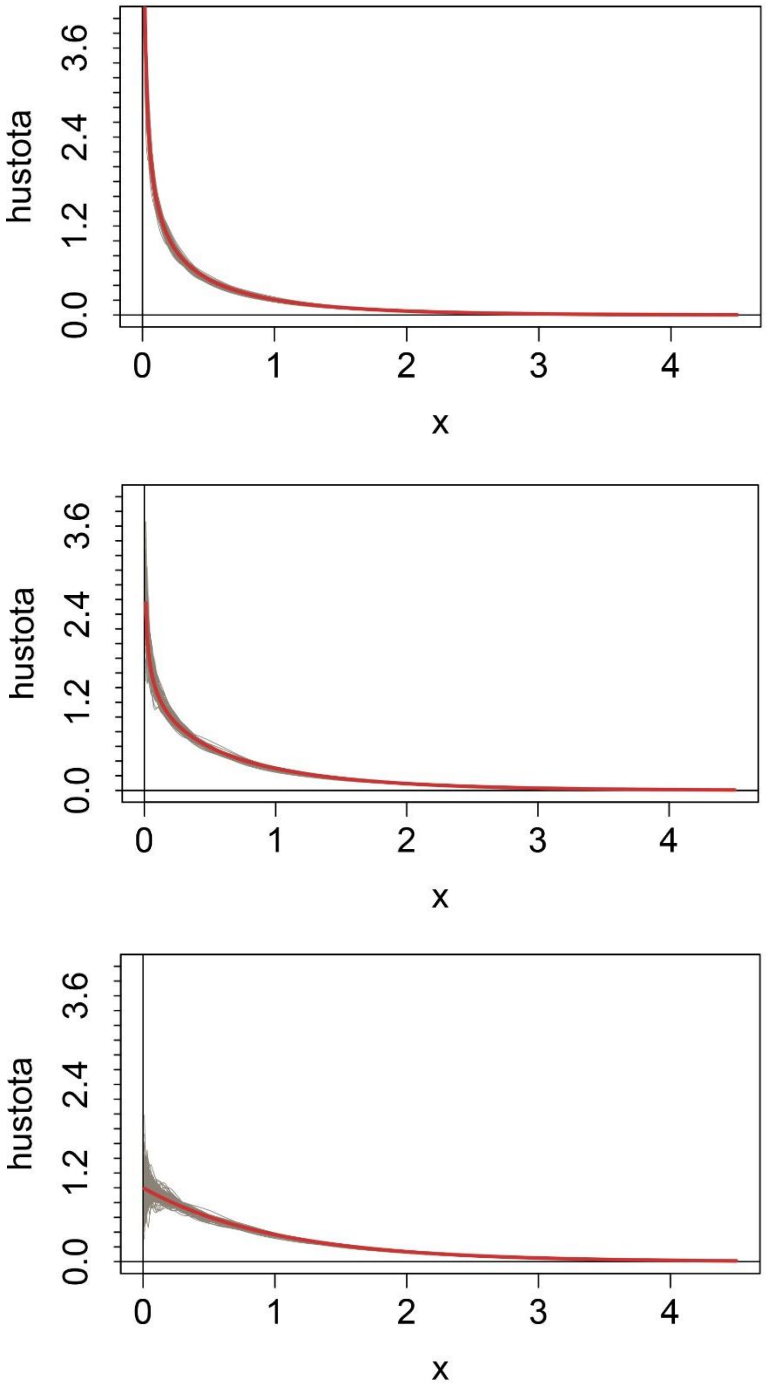
Parametr tvaru	Poměr úspěšných testů	$E(D_n)$	$\sigma(D_n)$	$E(I)$	$\sigma(I)$
0.5	0	0.19093	0.01197	0.35973	0.00352
0.75	0	0.10560	0.01162	0.23646	0.00646
1	0.96	0.05568	0.01187	0.14792	0.00971

Tab. 8 Výsledky analýzy metody Gama kernel.

Jak můžeme vidět z vizuálního porovnání výsledků, odhad se nenachází v oblasti záporných hodnot, protože metoda se pomocí změny parametrů gama kernelu automaticky adaptuje k nashromáždění dat kolem počátku v oblasti kladných hodnot. Odhad taky zjevně nemá žádnou rozkolísanost, výsledné křivky jsou hladké. Nicméně tvar rozdělení není gama kernely vystihnut přesně, jak je možné vidět z úspěšnosti KS testů a hodnot dalších chybových kritérií. Obzvlášť z hlediska integrálu rozdílů hustot došlo k citelnému zhoršení oproti předchozí metodě Cut and normalize.

5.1.8 Logaritmická transformace dat

Data náhodných výběrů byla před odhadem hustoty zlogaritmována, byl použit přirozený logaritmus. Odhad (globální) hodnoty šířky okna pro logaritmovaná data byl proveden metodou přímé minimalizace chyby (vztah 17) a finální odhad hustoty byl spočten podle vzorečku (21). Bylo používáno jádro Bisquare.



Obr. 16 Vizuální porovnání výsledků: červená čára představuje vzorovou hustotu gama rozdělení pro parametr tvaru 0.5 (nahoře), 0.75 (uprostřed) a 1 (dole). Šedé čáry představují jádrové odhady ze 100 různých samplů daného gama rozdělení.

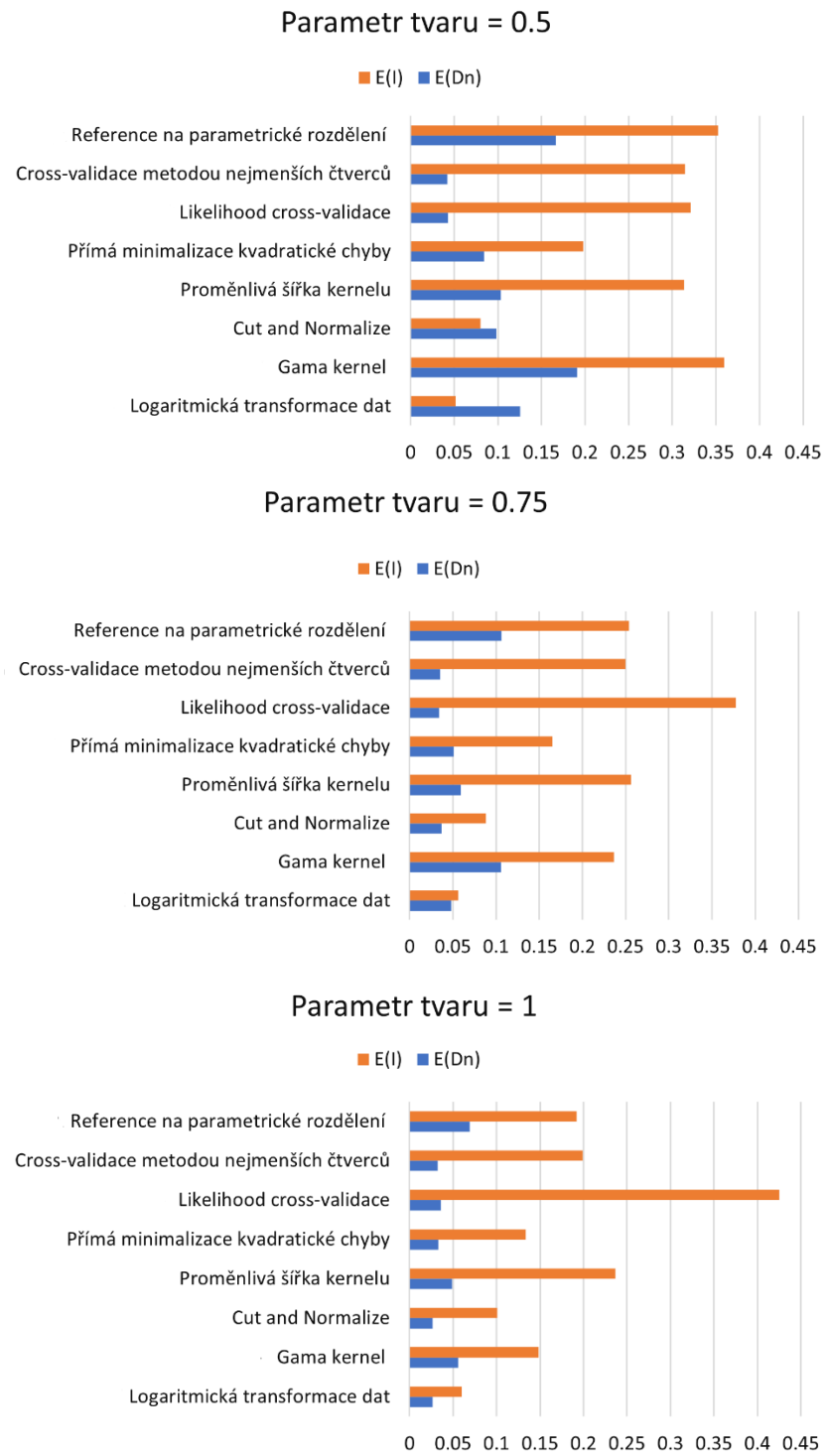
Parametr tvaru	Poměr úspěšných testů	$E(D_n)$	$\sigma(D_n)$	$E(I)$	$\sigma(I)$
0.5	0	0.12533	0.00846	0.05171	0.01523
0.75	1	0.04796	0.00750	0.05615	0.01578
1	1	0.02628	0.00920	0.05970	0.01887

Tab. 9 Výsledky analýzy metody logaritmická transformace dat.

Z vizuálního porovnání je patrné, že metoda velmi dobře vystihuje tvar rozdělení, obzvláště pro nízké hodnoty parametru tvaru vzorového gama rozdělení, a zároveň zabraňuje úniku pravděpodobnosti do záporných hodnot. Pro gama rozdělení s tvarem 1 můžeme vidět určitou rozkolísanost odhadů kolem počátku, které ale nemají vliv na úspěšnost KS testu. Z hlediska celkové úrovně chybových ukazatelů je tato metoda nejúspěšnější ze všech testovaných.

5.1.9 Shrnutí porovnání metod

Následující tabulka shrnuje celkové výsledky dosažené jednotlivými metodami pro jednotlivé rozdělení.



Obr. 17 Porovnávání výsledků testovaných metod s parametry tvaru 0.5 (nahoře), 0.75 (uprostřed) a 1 (dole). Oranžové čáry představují střední hodnotu integrálů rozdílu hustoty rozdělení pravděpodobnosti, modré čáry představují střední hodnotu D_n.

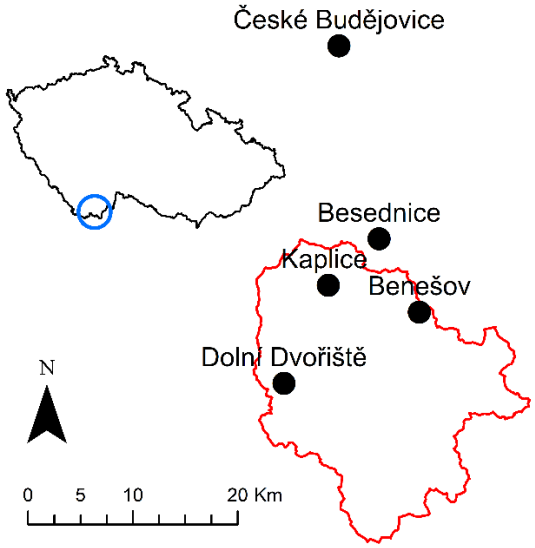
Z výsledů se dá konstatovat, že metoda logaritmické transformace dat se ukázala jako nejúspěšnější, především z hlediska kritéria integrálu rozdílu hustot I, kde několikanásobně předčila ostatní metody. Z pohledu tohoto kritéria se jako velice dobrá projevila také metoda Cut and normalize. Za těmito dvěma metodami je již v úspěšnosti velice značný skok a další metody výrazně zaostávají.

Z hlediska druhého kritéria – maxima rozdílů distribučních funkcí D_n, je situace vyrovnanější, obzvláště pro parametr gama rozdělení 0.5, kde se jako úspěšnější ukázaly cross-validační metody. Ty ale zcela selhaly z pohledu prvního kritéria a jejich problematické výsledky byly potvrzeny také vizuálním prozkoumáním výsledků. Obecně můžeme konstatovat, že kritérium D_n můžeme označit pouze za doplňkovou informaci (stejně jako výsledek KS testu), protože vysoká hodnota může být způsobena jediným velkým rozdílem, obzvláště v oblasti kolem počátku, kde hustota strmě klesá.

Celkově byly pro další část práce (testování na měřených datech) vybrány dvě metody – Cut and normalize a logaritmická transformace. Po úvaze byla provedena změna v metodě Cut and normalize, konkrétně ve stanovení úvodní globální hodnoty h. Místo reference na parametrické rozdělení, která byla použita při testování na gama rozděleních, byla pro reálná data použita přímá minimalizace kvadratické chyby. Důvodem jsou výsledky této kapitoly, které ukazují, že přímá minimalizace poskytuje kvalitnější odhady než metoda reference na parametrické rozdělení.

5.2 Porovnání jádrového a parametrického odhadu

Datové podklady tvořily denní srážkové úhrny z pěti stanic na povodí řeky Malše a jeho okolí: Benešov nad Černou, Besednice, České Budějovice, Dolní Dvořiště a Kaplice. Použity byly 38leté řady denních úhrnů v období od 1. 1. 1961 do 31. 12. 1998.



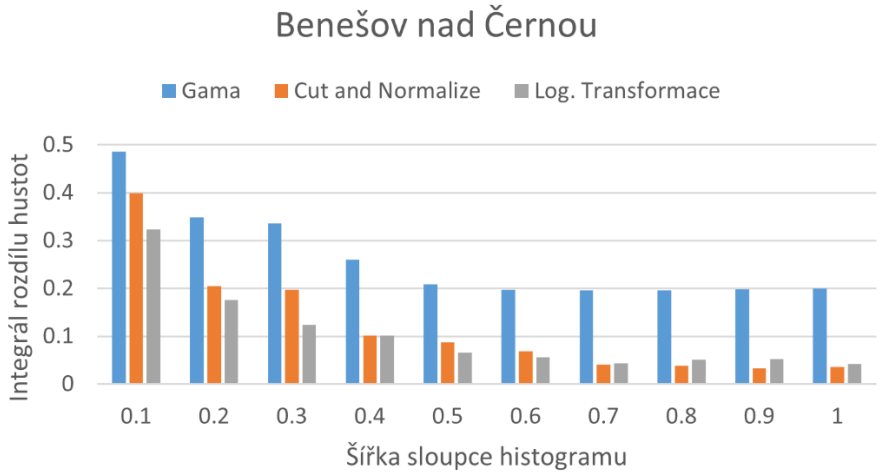
Obr. 18. Poloha stanic použitých k testování vybraných metod.

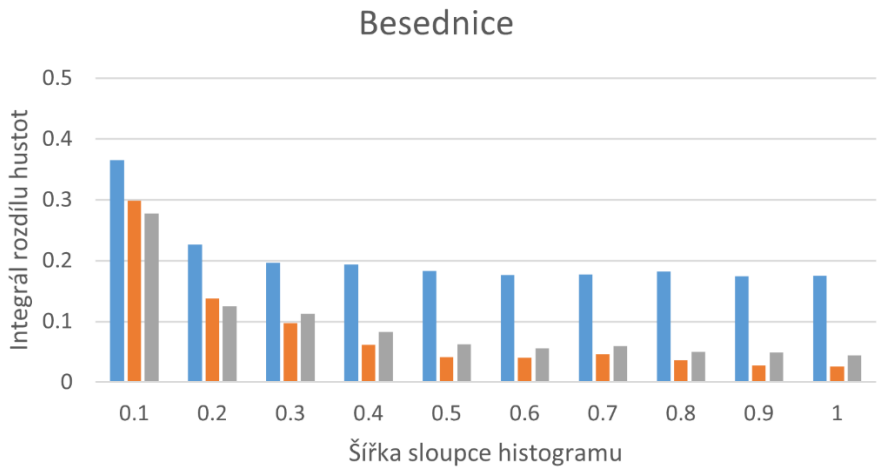
Ze změřených dat byly nejprve provedeny tři odhady hustoty: hustota gama rozdělení (metodou maximum likelihood) a oba vybrané jádrové odhady – cut and normalize a logaritmická transformace.

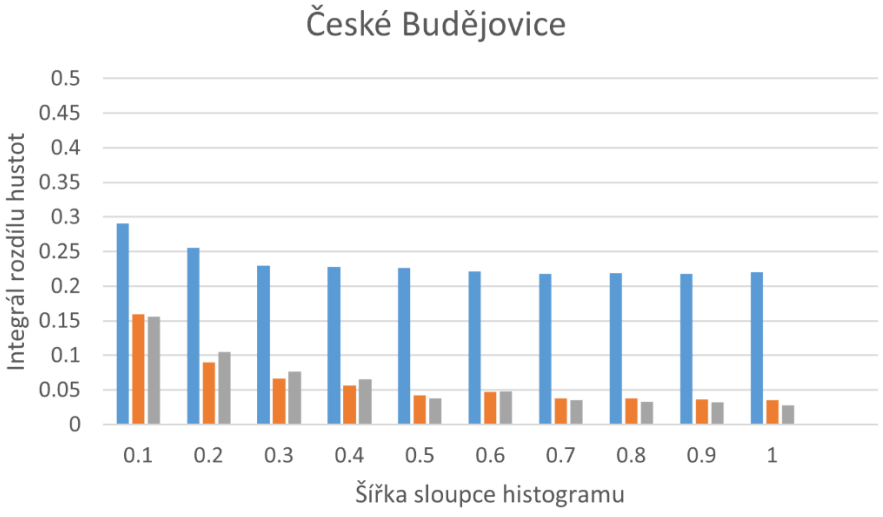
Z měřených dat pak byly postupně tvořeny histogramy s šířkou sloupce 0.1, 0.2, … 1 mm. Tyto histogramy byly uvažovány za „skutečnou“ hustotu rozdělení, vzhledem k délce použitých dat, nebo alespoň za nejlepší možnou informaci, kterou můžeme o skutečné hustotě mít, protože měřená data jsou zaokrouhlena na 0.1 mm a přesnější hodnoty nelze dostat. S uvedenými histogramy byly všechny odhady porovnány tak, že byl počítán integrál rozdílů I mezi histogramem a odhadem, analogicky jako v předchozí kapitole.

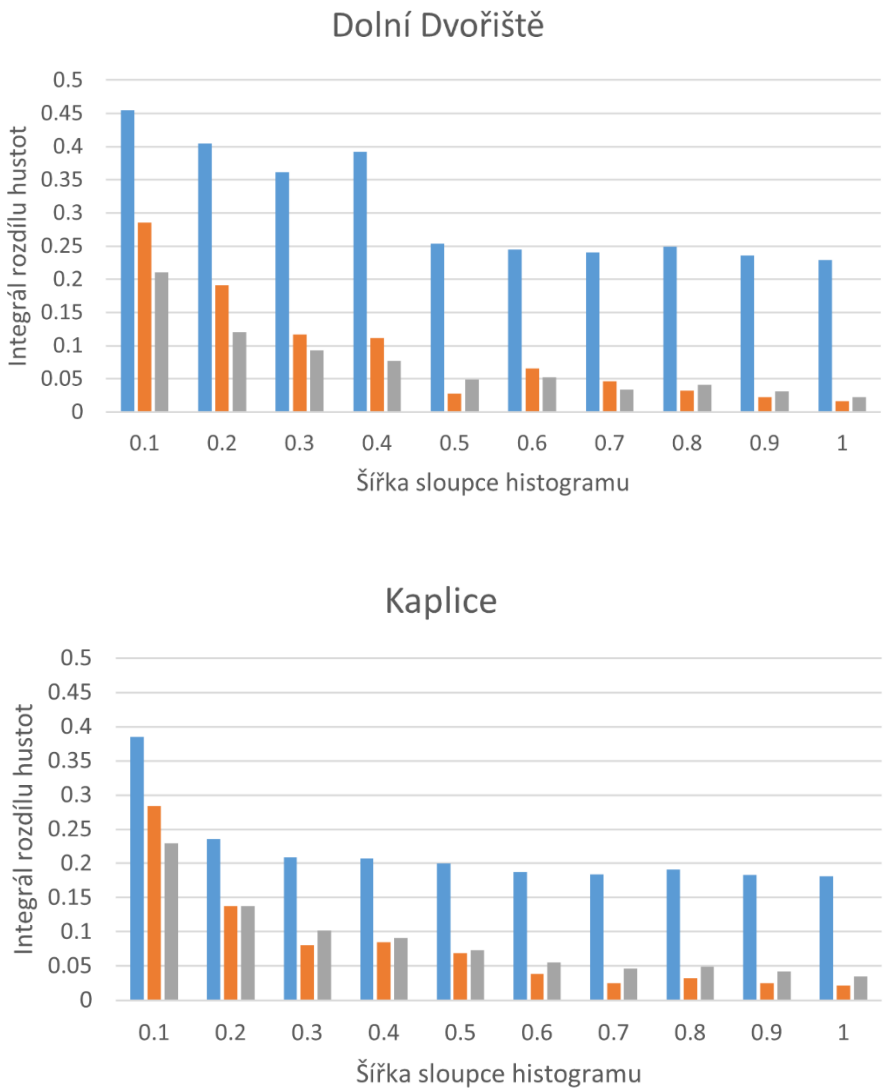
Histogram s šířkou sloupce 0.1 mm je nejpodrobnější možné vyjádření hustoty měřených dat, který by při velmi velkém množství dat představoval přesný odhad hustoty na dané lokalitě. Protože dat je přece jen omezený počet, byly použity i histogramy s širšími sloupci (násobky 0.1) až do 1 mm, které zanechávají detailní kolísání hustoty skryté. Zároveň bylo možné sledovat vliv shlazení vzorové hustoty na hodnoty chybového kritéria.

Následující obrázky ukazují kompletní výsledky pro všechny lokality. Modře jsou vyznačeny hodnoty chyby pro parametrickou gama hustotu, oranžově pro jádrový odhad metodou Cut and normalize a šedivě pro jádrový odhad metodou logaritmické transformace.









Obr. 19. Porovnání úspěšnosti parametrického a jádrového odhadu pro jednotlivé lokality. Chyby parametrických odhadů jsou vybarveny modře, chyby jádrového odhadu metodou Cut and normalize jsou vybarveny oranžově, chyby jádrového odhadu metodou logaritmické transformace jsou šedě.

Z výsledků plyne, že obě metody jádrového odhadu jasně překonávají parametrický odhad. S rostoucí šířkou sloupce histogramu dochází samozřejmě k redukci chyby, protože jsou potlačeny lokální kolísání hustoty, ale i tato redukce je pro jádrové odhady mnohem rychlejší. Například, v lokalitě Kaplice dosáhnou jádrové odhady chybu integrálu rozdílu hustot 0.05 při použití histogramu s šířkou sloupce 0.6 mm. Pokud bychom chtěli dosáhnout takové chyby s gama rozdělením, museli bychom použít histogram o šířce sloupce 9.6 mm. Přestože je tedy gama rozdělení všeobecně akceptovaným modelem pro denní srážkové úhrny, je zřejmé, že správným výběrem metody jádrového odhadu dosáhneme výrazně lepší přesnosti.

6. Diskuse a závěr

Cílem předložené diplomové práce bylo analyzovat využití jádrových odhadů hustoty pro popis rozdělení pravděpodobnosti denních srážkových úhrnů. Byla provedena literární rešerše, zaměřená jednak na problematiku jádrových odhadů obecně, ale hlavně na specifické problémy, týkající se jejich využití pro srážková data. Jednalo se především o vystihnutí specifického tvaru rozdělení srážek a řešení tzv. „boundary“ problému, který plyne z toho, že srážky jsou nezáporná náhodná veličina.

Na základě rešerše bylo vybráno osm metod volby šířky kernelového okna. Tyto metody byly naprogramovány v jazyce R, v jednom případně byla využito hotové řešení v podobě R balíku. Metody byly nejprve porovnány mezi sebou, když byla testována jejich schopnost vystihnout různá rozdělení na základě náhodných samplů.

Na základě testů byly vybrány dvě metody (Cut and normalize a logaritmická transformace dat), které byly použity pro odhad rozdělení reálných měřených srážek. Přitom byla jejich úspěšnost porovnána s parametrickým odhadem, kde jako reprezentant parametrických metod bylo použito gama rozdělení, odhadované metodou maximální věrohodnosti. Bylo zjištěno, že jádrové odhady z hlediska přesnosti zcela překonaly parametrický odhad.

Z výsledků plyne, že jádrové odhady v případě srážek představují velmi dobrou alternativu k parametrickým odhadům a že s pomocí jádrových odhadů lze dosáhnou výrazně vyšší přesnosti odhadu. Výhodou je také to, že pro obtížné části výpočtu lze použít R balík KernSmooth, který obsahuje funkce pro výpočet šířky okna metodou přímé minimalizace kvadratické chyby. Ta sama o sobě nebyla hodnocena jako nejlepší, ale výsledky ukazují, je velmi vhodné ji zařadit do metod Cut and normalize nebo logaritmické transformace, které se ukázaly jako nejúčinnější

Lze shrnout, že diplomová práce splnila vytčené cíle.

7. Přehled literatury a použitých zdrojů

Abramson I. S., 1982: On bandwith variation in kernel estimates - a square root law. The annals of statistics 10(4): 1217 – 1223. DOI: 10.1214/aos/1176345986.

Anděl, J., 1998: Statistické metody. MatfyzPress, 1998. ISBN: 80-85863-27-8. 274 stran.

Block, P.J., Filho, F.A.S., Sun, L., Kwon, H., 2009: A streamflow forecasting framework using multiple climate and hydrological models. Journal of the American Water Resources Association, 45(4):828-843. doi: 10.1111/j.1752-1688.2009.00327.x.

Castellvi F., Mormeneo I., Perez P.J., 2004: Generation of daily amounts of precipitation from standard climatic data: a case study for Argentina. Journal of Hydrology 289(1-4): 286-302. doi: 10.1016/j.jhydrol.2003.11.027.

Chen, J., Brissette, F., 2014: Stochastic generation of daily precipitation amounts: review and evaluation of different models, Climate Research 59(3): 189–206. DOI:10.3354/cr01214.

Chen S. X., 2000: Probability density function estimation using gamma kernels. Annals of the institute of statistical mathematics 52(3), 471-480, DOI: 10.1023/A:1004165218295.

Gramacki A., 2018: Nonparametric Kernel Density Estimation and Its Computational Aspects. Springer International Publishing, Poland, ISBN 978-3-319-71687-9.

Hosking J.R.M., 1990: L-moments: analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society B, 52 (1), 105–124.

James G., Witten D., Hastie T., Tibshirani R., 2013: An Introduction to Statistical Learning with Applications in R. Springer Science+Business Media, New York, ISBN 978-1-4614-7137-0.

Monjo R., Caselles V., Chust G., 2012: Alternative model for precipitation probability distribution: application to Spain. Climate Research 51(1): 23-33. doi: 10.3354/cr01055.

Monjo R., Caselles V., Chust G., 2014: Probabilistic correction of RCM precipitation in the Basque Country (Northern Spain). Theoretical and Applied Climatology 117:317-329. doi: 10.1007/s00704-013-1008-8.

Moss J (2019). “univariateML: An R package for maximum likelihood estimation of univariate densities.” Journal of Open Source Software, 4(44), 1863.

Mosthaf, T., Bardossy, A., 2017: Regionalizing nonparametric models of precipitation amounts on different temporal scales. Hydrology and Earth System Sciences 21(5): 2463-2481. DOI: 10.5194/hess-21-2463-2017.

Muller H.G., 1992: Smooth optimum kernel estimators near endpoints. Biometrika, 78, 521-530.

Park B.U., Macron J.S., 1991: Comparison of data-driven bandwidth selectors. Journal of the American Statistical Association 85, 66-72.

Peel S., Wilson L.J., 2008: Modelling the distribution of precipitation forecasts from the Canadian ensemble prediction system using kernel density estimation. Weather and forecasting 23(4), 575-595. DOI: 10.1175/2007WAF2007023.1.

R Core Team, 2015: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, availabale at: https://www.R-project.org.

Rajagopalan B., 1997: Evaluation of kernel density estimation methods for daily precipitation resampling. Springer-Verlag, Stochastic Hydrology and Hydraulics 11:523-547.

Rajib M., 2018: Statistical Methods in Hydrology and Hydroclimatology. Springer Nature, Singapore, ISBN 978-981-10-8778- 3.

Riley K.F., Hobson M.P., Bence S.J., 2006: Mathematical Methods For Physics And Engineering, 3rd edn. Cambridge University Press, UK, ISBN 978-0-511-16842-0.

Scott D.W., Tapia, R.A., Thompson, J.R., 1977: Kernel density estimation revisited. Nonlinear Analysis 1, 339-372.

Sheather S.J., 1983: A data-based algorithm for choosing the window width when estimating the density at a point. Computational Statistics and Data Analysis 1, 229-238 .

Sheather S.J., 1986: An improved data-based algorithm for choosing the window width when estimating the density at a point. Computational Statistics and Data Analysis 4, 61-65.

Sheather S.J., Jones, M.C., 1991: A reliable data-based bandwidth selection, method for kernel density estimation. Journal of the Royal Statistical Society B53,683-690.

Silverman B.W., 1986: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London, ISBN 978-0412246203.

Teutschbein, C, Seibert, J., 2012: Bias correction of regional climatemodel simulations for hydrological climate-change impact studies: Review and evaluation of different methods. Journal of Hydrology. 456–457: 12–29. DOI:10.1016/j.jhydrol.2012.05.052.

Teutschbein C., Seibert J., 2013: Is bias correction of regional climate model (RCM) simulations possible for non-stationary conditions? Hydrology and Earth systém Sciences 17(12): 5061-5077. doi: 10.5194/hess-17-5061-2013.

Vlček, O., Huth, R. 2009: Is daily precipitation Gamma-distributed? Adverse effects of an incorrect use of the Kolmogorov-Smirnov test. Atmospheric Research, 93:759-766. doi: 10.1016/j.atmosres.2009.03.005.

Wand, M.P. and Jones, M.C. (1995) "Kernel Smoothing". Chapman and Hall, New York. DOI: https://doi.org/10.1201/b14876.

Wackerly D., Mendehall W., Scheaffer R.L., 2008: Mathematical Statistics With Application. Thomson Learning, USA, ISBN-13: 978-0-495-38508-0.

Wilks D.S., 2011: Statistical methods in the atmospheric sciences, 3rd edn. Academic, Amsterdam, ISBN 978-0-12-385022-5.

Wilks, D.S., Wilby, R.L., 1999: The weather generation game: a review of stochastic weather models, Prog. Phys. Geog. 23: 329–357. DOI:10.1177/030913339902300302.

Woodroofe M., 1970: On choosing a delta-sequence. Annals of Mathematical Statistics 41, 1665-1671.