MEX - Zápočtová úloha

Miroslav Kubů

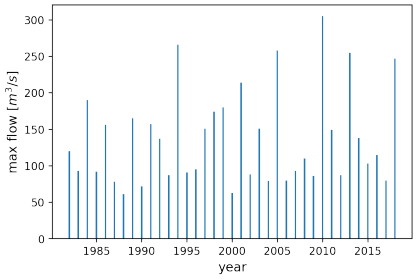
9. září 2019

Abstrakt

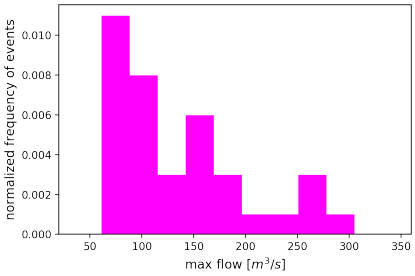
V této práci se budeme zabývat modelováním maximálních ročních průtoků řeky N. Analýzu dat provádíme s využitím technik modelování extrémních událostí. Prvotní odhady pro překročení hladiny průtoku získáváme z tzv. distribution-free nerovností. Poté prezentujeme modely využívající jádrové odhady či pro extrémní události specifická rozdělení, díky kterým můžeme předpovídat maximální hladiny průtoků. Pozorujeme přitom rozdíly mezi jednotlivými přístupy a komentujeme jejich možné výhody a nevýhody.

1 Analýza dat

K dispozici máme celkově 37 pozorování maximálního ročního průtoku mezi léty 1982 až 2018 udávaného v m^3/s. Hodnoty pro jednotlivé roky vykreslujeme na obrázku 1a. Právě z obrázku 1a je poté patrný široký rozptyl v jednotlivých hodnotách průtoků. To numericky demonstrujeme též v tabulce 1, odkud je zřejmé, že pracujeme s hodnotami od 61 do 305 m^3/s.



(a) hodnoty maximálního průtoku v m^3/s



(b) histogram maximálních průtoků

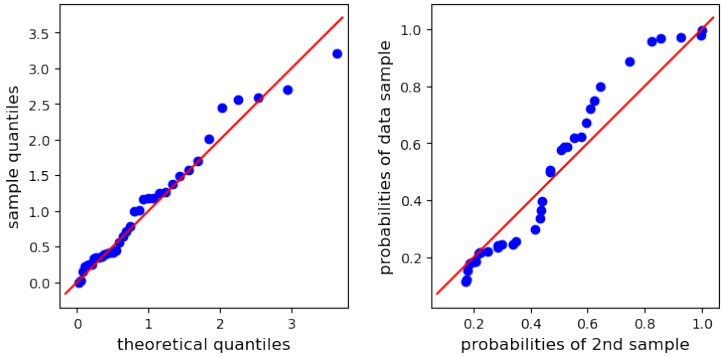
K vykreslení histogramu průtoku využíváme metodu Doane [1], která pro daná data doporučuje použít 9 binů. S tímto nastavením vykreslujeme histogram na obrázku 1b, čímž získáváme intuitivní představu o tvaru rozdělení náhodné veličiny. Patrná je zhruba exponenciálně klesající tendence se třemi význačnými peaky. Tato tendence nicméně vzhledem k nízkému počtu dat nemusí odpovídat skutečné podobě rozdělení průtoku. V další fázi analýzy porovnáváme rozdělení dat s normálním exponenciálním rozdělením.

min	max	průměr	výb. rozptyl
61	305	136.92	4189.58

Tabulka 1: Vybrané popisné statistiky datové množiny.

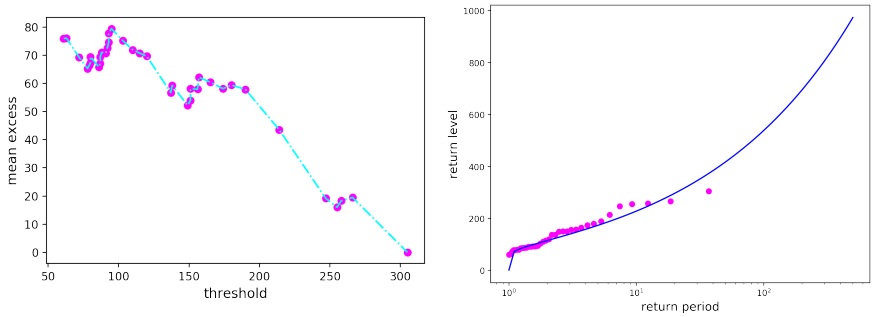
Na obrázku 2 lze pozorovat rozdíly v chování rozdělení našich dat v porovnání s nafitovaným exponenciálním rozdělením. Z QQ plotu na obrázku 2 můžeme pozorovat, že pravý chvost teoretického rozdělení se poměrně liší od rozdělení empirického. Na druhou stranu levý chvost empirického rozdělení je popsán poměrně přesně. V případě PP plotu poté pozorujeme dvě výrazné odchylky v distribučních funkcích. Vzhledem k tomu, že obě výchylky jsou pozorovány v opačných polorovinách tvořených vyznačenou přímkou, může být toto pozorování způsobené větším rozptylem empirického rozdělení oproti teoretickému rozptylu. Z obrázku 2 lze tedy usoudit, že model vzniklý fitováním exponenciálního rozdělení nemusí dostatečně přesně popisovat naměřené hodnoty.

Chování průtoku lze dále pozorovat na mean excess plotu vykresleném na obrázku 3a. Z obrázku je patrné, že s rostoucím průtokem hodnota mean excess klesá, a to téměř lineárně. Na obrázku 3b poté prezentujeme též return plot. Na něm je možno pozorovat očekávaná doba návratu pro jednotlivé hodnoty průtoků. Pro tři nejvodnatější roky je poté doba návratu vyšší než 10 let.



(a) QQ plot (b) PP plot

Obrázek 2: QQ plot a PP plot pro porovnání se standardním exponenciálním rozdělení.



(a) ME plot (b) Return plot

Obrázek 3: ME plot a Return plot pro daná data.

2 Modely předpovědí

V následující sekci budeme výši maximálního ročního průtoku předpovídat s využitím odlišných modelů. V první části odhadujeme pravděpodobnosti překročení určité hladiny skrze statistické nerovnosti. V dalších dvou fázích poté pro výpočet pravděpodobnosti překročení hladiny konstruujeme modely využívající rozdělení pro modelování extrémních událostí a jádrové odhady. Konkrétně je naším úkolem modelovat pravděpodobnosti překročení hladin 230 m^3/s a 310 m^3/s, nepřekročení hladin 50 m^3/s a 70 m^3/s, a na závěr překročení rekordní hladiny $X_m = 305$ m^3/s a zábrany $u = 230$ m^3/s o více než 50 m^3/s za podmínky jejího překročení.

2.1 Distribution-free nerovnosti

Před konstrukcí samotných modelů ukazujeme odhady pravděpodobností překročení hladiny skrze Markovovu (1), Chernoffovu (2), Čebyševovu (3) a Čebyšev-Cantelliho (4) nerovnost pro náhodnou veličinu průtoku $X \in \mathbb{R}^+$. Uvažujeme-li tedy nezápornou náhodnou veličinu $X \in \mathbb{R}^+$ a kladnou hladinu $\varepsilon > 0$, používáme pro odhad pravděpodobnosti překročení hladiny ε Markovovu nerovnost ve tvaru

$$\mathrm{P}(X \geq \varepsilon) \leq \frac{\mathrm{E}X}{\varepsilon}. \tag{1}$$

Z Markovovy nerovnosti následně vychází nerovnost Chernoffova ve tvaru

$$\mathrm{P}(X \geq \varepsilon) \leq \min_{t>0} \frac{\mathrm{E}[e^{tX}]}{e^{t\varepsilon}}. \tag{2}$$

Pro Čebyševovu a Čebyšev-Cantelliho nerovnost poté krom nezápornosti náhodné veličiny uvažujeme též její nulovou střední hodnotu. Z tohoto důvodu tedy využíváme náhodnou veličinu $Y = X - \mathrm{E}X$, jež využíváme v nerovnostech [2]

$$\mathrm{P}(Y \geq \varepsilon) \leq \frac{\mathrm{Var}(Y)}{\varepsilon^2}, \tag{3}$$

$$\mathrm{P}(Y \geq \varepsilon) \leq \frac{\mathrm{Var}(Y)}{\varepsilon^2 + \mathrm{Var}(Y)}, \tag{4}$$

kde $\mathrm{Var}(Y)$ je rozptyl Y. Díky tomu můžeme získat horní odhad pro pravděpodobnosti překročení dané hladiny ε. Dle zadání se snažíme získat odhady pro pravděpodobnost překročení hladn $\varepsilon = 230$ a $\varepsilon = 310$. Odhady získané z nerovností (1), (3) a (4) zobrazujeme v tabulce . Z tabulky je patrné, že horní hranice

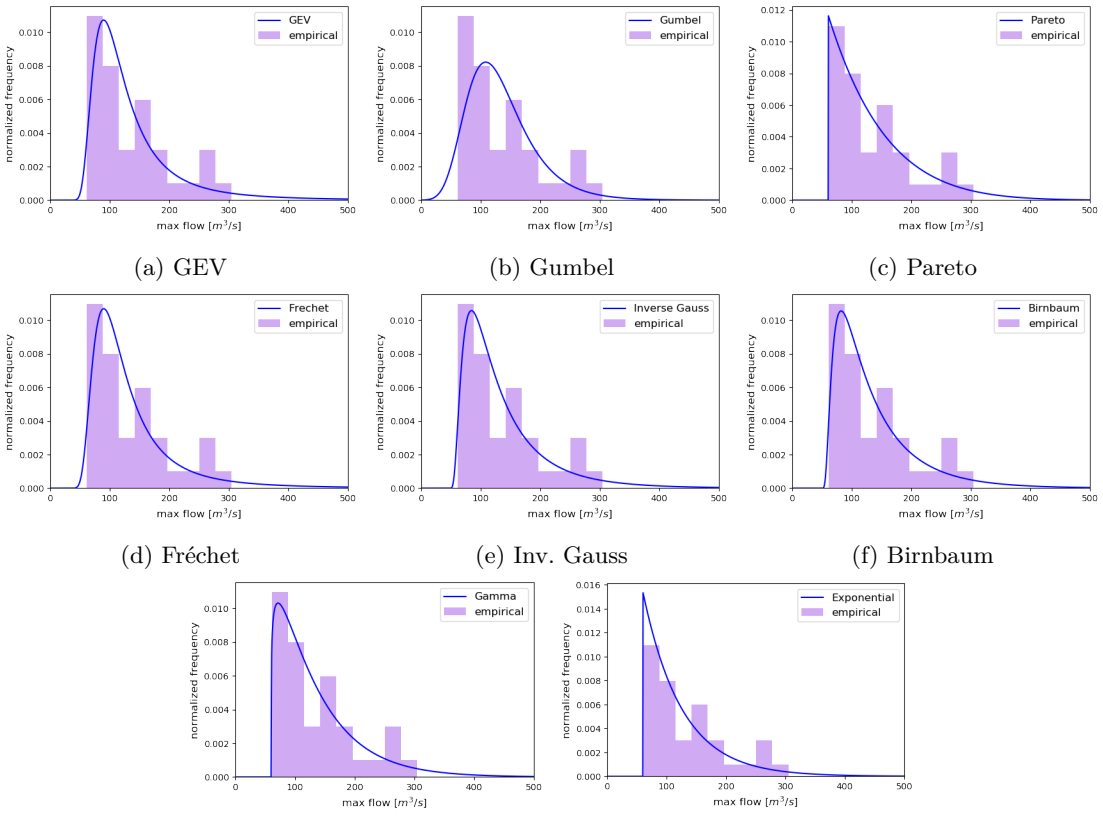
Nerovnost	$\mathrm{P}(X \geq 230)$	$\mathrm{P}(X \geq 310)$
Markov	≤ 0.5953	≤ 0.4417
Chernoff	≤ 0.4303	≤ 0.0000
Čebyšev	≤ 0.4705	≤ 0.1361
Cantelli	≤ 0.3199	≤ 0.1198

Tabulka 2: Odhady pravděpodobností překročení vybraných hladin získané skrze statistické nerovnosti.

odhadu se pro použité nerovnosti výrazně liší. Pro hodnoty vyšší než maximálně naměřených $X_{max} = 305$ poté Chernoffova mez s $t \to +\infty$ klesá k 0. Nejlepší přesnost poté získáváme u Čebyšev-Cantelliho nerovnosti. Přesnost odhadů lze nicméně lépe diskutovat až po srovnání s pokročilejšími modely v následujících sekcích.

2.2 Parametrické modely

V další části práce budeme modelovat rozdělení maximálních průtoků řeky skrze rozdělení pro statistiku extrémních událostí. K modelování použijeme Gumbelovo rozdělení, Frechetovo rozdělení a obecnější generalized extreme value (GEV) rozdělení. Tato rozdělení se vyznačují těžkými pravými chvosty, skrze něž se pokusíme lépe zachytit trend v rozdělení. Dále pro ilustraci rozdělení modelujeme též Paretovým rozdělení pro těžký pravý chvost. To má sice nevýhodu v omezeném supportu pro levý chvost, ale dokáže zachytit rozdělení v chvostu pravém. To znamená, že pro daný případ je Paretovo rozdělení nevhodné pro modelování extrémního sucha, ale lze jej použít pro modelování extrémně vysokých průtoků. Dále použijeme modely s inverzním gaussovským rozdělením, Gamma rozdělením a Birnbaumovým rozdělením s lehkými levými chvosty a těžkými pravými chvosty.



(a) GEV (b) Gumbel (c) Pareto

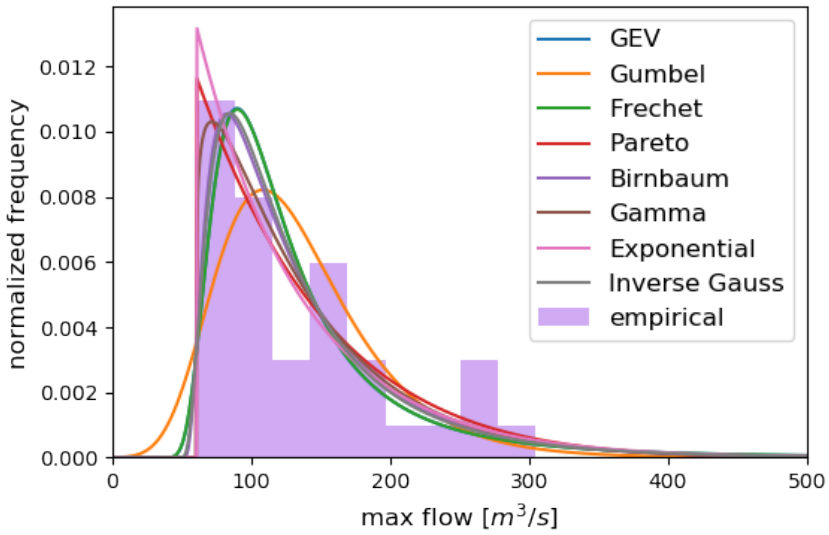
(d) Fréchet (e) Inv. Gauss (f) Birnbaum

(g) Gamma (h) Exponential

Obrázek 4: Přehled odhadů skrze vybraná statistická rozdělení.

Parametry rozdělení fitujeme pomocí statistického balíčku SciPy a pokud nebude uvedeno jinak, používáme výchozí nastavení balíčku SciPy v1.2.9. Jednotlivá nafitovaná rozdělení společně s empirickým

rozdělením dat zobrazujeme na obrázku 4. Z obrázku 4 je patrné, že rozdělení GEV v našem případě téměř přechází do tvaru Fréchetova rozdělení. Zřejmý překryv rozdělení poté demonstrujeme též na obrázku 5. Námi používaná rozdělení se prezentují těžšími pravými chvosty, což by bylo přínosné pro modelování pojistných událostí při povodních. Gumbelovo rozdělení nicméně na první pohled ne zcela přesně kopíruje hlavní peak histogramu a prezentuje se patrně těžším levým chvostem. Zmíněné vlastnosti použitých



Obrázek 5: Porovnání jednotlivých parametrických odhadů.

rozdělení lze též ukázat numericky v tabulce 3. Z té je vidět skutečnost, že vyšší pravděpodobnosti pro extrémní sucha predikuje pouze Gumbelovo, a méně výrazně pak též Fréchetovo a GEV rozdělení. Ostatní modely poté pro předpověď průtoku pod 50 m^3/s predikují buď jako přímo nulovou, nebo téměř nulovou. Dále pozorujeme jen velice malé odchylky hodnot v případě Fréchetova a GEV rozdělení. Zobecněné GEV rozdělení tedy zjevně přechází v rozdělení Fréchetovo. Co se předpovědí extrémních průtoků týče, největší pravděpodobnost rekordního průtoku predikuje Fréchetovo rozdělení.

Shodu rozdělení s daty lze kontrolovat na QQ plotech vykreslených na obrázku 6. Z QQ plotů je patrné, že GEV rozdělení přecházející ve Fréchetovo rozdělení má pro naše data příliš těžký pravý chvost. Naopak Gumbelovo rozdělení má patrně pravý chvost až příliš lehký. Naopak poměrně dobrá shoda je patrná u Paretova rozdělení, Birnbaumova rozdělení a Gamma rozdělení. Právě Birnbaumův a Gamma model, jež se i na základě hodnot v tabulce 3 silně podobají, tak z vybraných modelů nejlépe popisují zadaná data. Těžké chvosty Fréchetova rozdělení jsou patrné i na hodnotách překročení 230 m vysoké hráze o více než 50 m vypsaných v tabulce 3. Zatímco v případě Gumbelova rozdělení pravděpodobnost $\mathrm{P}(X > 280)$ klesne oproti $\mathrm{P}(X > 230)$ na třetinu hodnoty, v případě Fréchetova rozdělení není pokles z $\mathrm{P}(X > 230)$ na $\mathrm{P}(X > 280)$ ani na polovinu. V tabulce 4 dále uvádíme hodnoty hladin pro 10-letou, 100-letou a

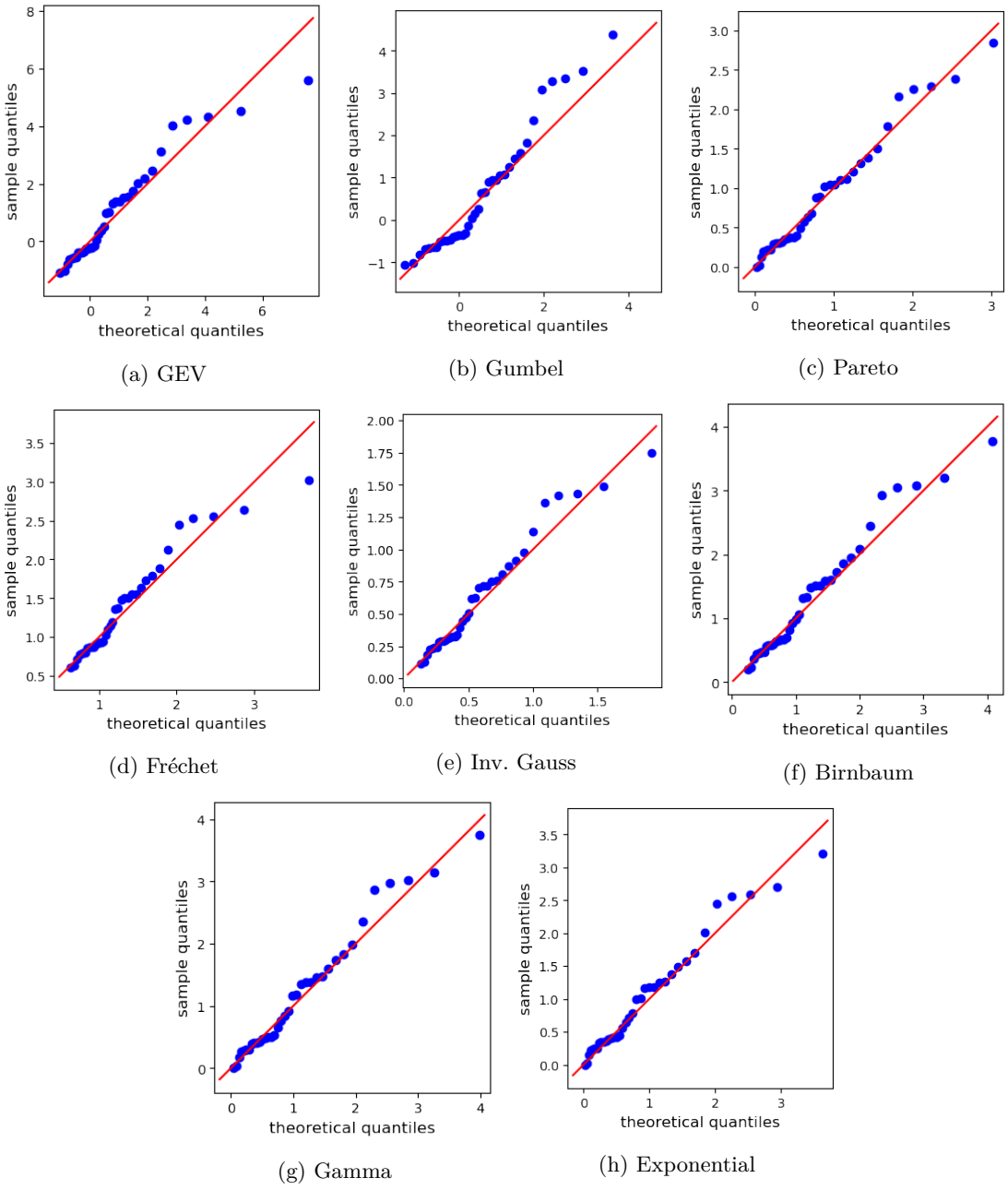
Rozdělení	$\mathrm{P}(X > 230)$	$\mathrm{P}(X > 310)$	$\mathrm{P}(X < 50)$	$\mathrm{P}(X < 70)$	$\mathrm{P}(X > X_m)$	$\mathrm{P}(X > 280 \mid X > u)$
GEV	0.0981	0.0446	0.0009	0.0651	0.0465	0.5953
Gumbel	0.0644	0.0111	0.0248	0.0939	0.0124	0.3354
Pareto	0.1106	0.0317	0.0000	0.1000	0.0344	0.4665
Fréchet	0.0975	0.0440	0.0010	0.0653	0.0459	0.5928
Birnbaum	0.0965	0.0325	0.0000	0.0691	0.0348	0.5047
Gamma	0.0995	0.0308	0.0000	0.0891	0.0331	0.4814
Expon.	0.1080	0.0376	0.0000	0.1118	0.0402	0.5176
In. Gauss	0.0951	0.0338	0.0001	0.0666	0.0359	0.5184

Tabulka 3: Výpočty příslušných pravděpodobností hladin pro vybrané modely.

1000-letou vodu s přesností na 0.5 m^3/s. Pro případ 10-leté vody se všechna rozdělení chovají poměrně podobně, pro 100-letou a pak převážně 1000-letou vodu je nicméně patrný vliv těžkých chvostů GEV a Fréchetova rozdělení. Ty totiž předpovídají výrazně vyšší maximální průtoky.

Rozdělení	$h_{0.1}$	$h_{0.01}$	$h_{0.001}$
GEV	228.5	539.5	1256.0
Gumbel	209.5	315.0	418.5
Pareto	237.0	375.5	484.0
Fréchet	228.0	533.5	1229.0
Birnbaum	227.5	399.5	579.0
Gamma	230.0	386.0	540.0
Exponential	236.0	411.0	585.5

Tabulka 4: Výpočty konkrétních hladin významnsti pro vybrané modely.



(a) GEV (b) Gumbel (c) Pareto

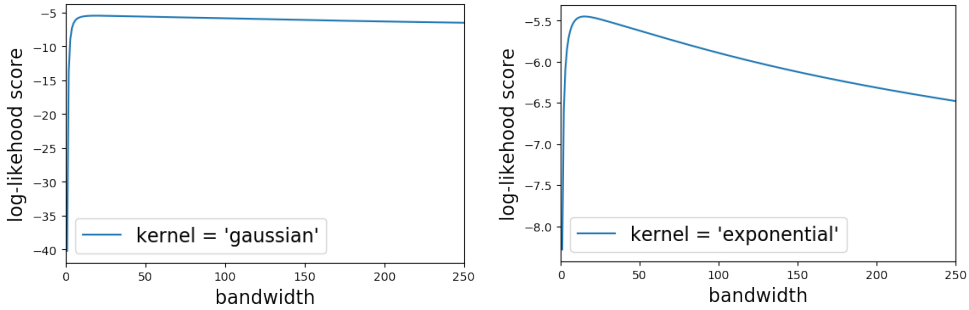
(d) Fréchet (e) Inv. Gauss (f) Birnbaum

(g) Gamma (h) Exponential

Obrázek 6: Přehled QQ plotů pro vybraná statistická rozdělení.

2.3 Jádrové odhady

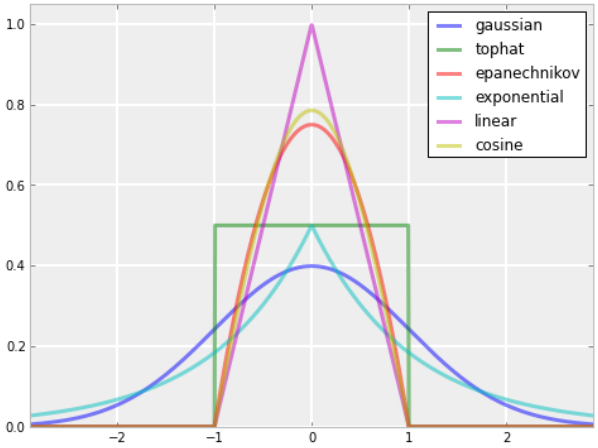
V další sekci se zaměříme na modelování rozdělení skrze jádrové odhady. Prvně je tedy třeba optimalizovat hodnoty parametru bandwidth pro dosažení optimální shody s daty. K tomu použijeme metodu grid search, která parametr bandwidth odhaduje za pomocí cross-validation. Konkrétně tak opakovaně



(a) gaussovské jádro

(b) exponenciální jádro

Obrázek 7: Ukázka vlivu bandwidth na hodnoty log-likehood skóre pro vybraná jádra.



Obrázek 8: Přehled použitých jader.

náhodně vybíráme 80% dat do fitovací množiny a na zbytku dat pozorujeme shodu s modelem. Skrze kombinaci grid search a cross-validation poté můžeme vybrat bandwidth zaručující největší shodu s testovacími daty.

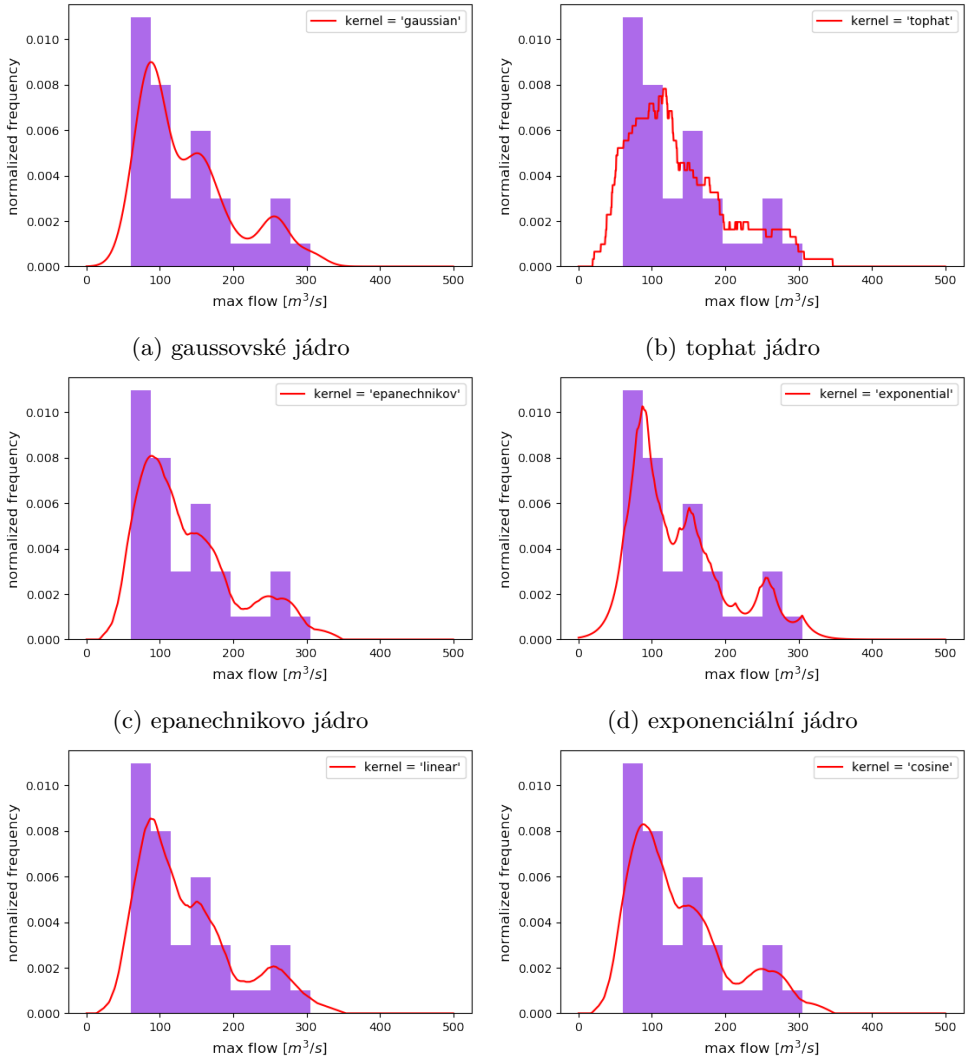
Jádro	opt. bandwidth
Gauss	19.0
Tophat	41.5
Epanechnikov	44.5
Exponential	15.0
Linear	48.5
Cosine	44.5

Tabulka 5: Porovnání optimálních hodnot bandwidth.

Jak ukazuje obrázek 7, zatímco shoda modelu pro gaussovské jádro s daty je velice podobná pro téměř všechny hodnoty z mřížky, v případě exponenciálního jádra je shoda velice závislá na konkrétní hodnotě bandwidth. V případě nízkého počtu dat nicméně může tato skutečnost vést k přefitování modelu. Na obrázku 9 poté vykreslujeme odhady pro vybraná jádra. Z obrázku je zřejmé, že tři výrazné peaky v histogramu nejlépe kopíruje exponenciální jádro. Vzhledem k velice nízkému počtu dat je nicméně touto volbou velice snadné docílit přefitování, a proto může být v praxi výhodnější použít modely s jádrem epanechnikovým či cosinovým, které jsou lépe vyhlazené.

Podíváme-li se na srovnání modelů na obrázku 11, zjišťujeme, že na chvostech se všechny modely chovají téměř totožně. Pro modelování extrémních událostí tedy všechny modely fungují podobně, zásadně se nicméně liší pro predikci na oblasti hojně pozorovaných průtoků.

V našem případě se pro modelování jeví vhodné Epanechnikovo jádro, zaměříme se tedy podrobněji na příslušný model. Provádíme porovnání modelu vzniklého použitím cross-validation s modelem určeným volbou bandwidth na základě D-metody. Dle D-metody je vhodné použít volbu bandwidth $h = 52$, v případě cross-validation je optimální hodnota vypočtena jako $h = 44.5$. Výsledky porovnání jsou ilustro-



(a) gaussovské jádro

(b) tophat jádro

(c) epanechnikovo jádro

(d) exponenciální jádro

(e) lineární jádro

(f) cosinové jádro

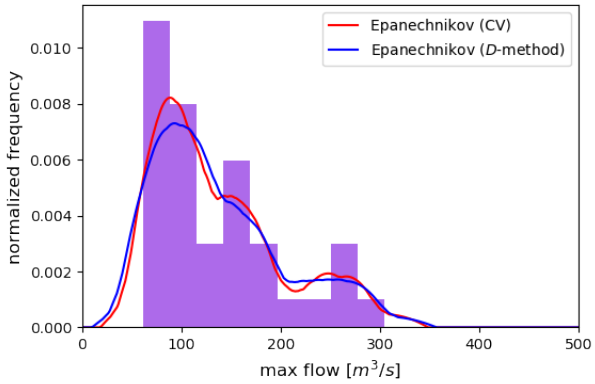
Obrázek 9: Přehled jádrových odhadů pro optimální hodnoty bandwidth.

vány na obrázku 10. Ač je model využívající bandwidth dle D-metody mírně vyhlazenější, oba modely se výrazně neliší. Výsledky hodnot predikcí pro zadané hladiny pro vybraná jádra jsou vypsané v tabulce 6. Vzhledem k tomu, že ze zadání predikujeme pravděpodobnosti extrémního sucha a extrémních záplav, numerické výsledky se pro jednotlivé modely velice podobají. Vypíšeme-li nicméně v tabulce 7 hodnoty

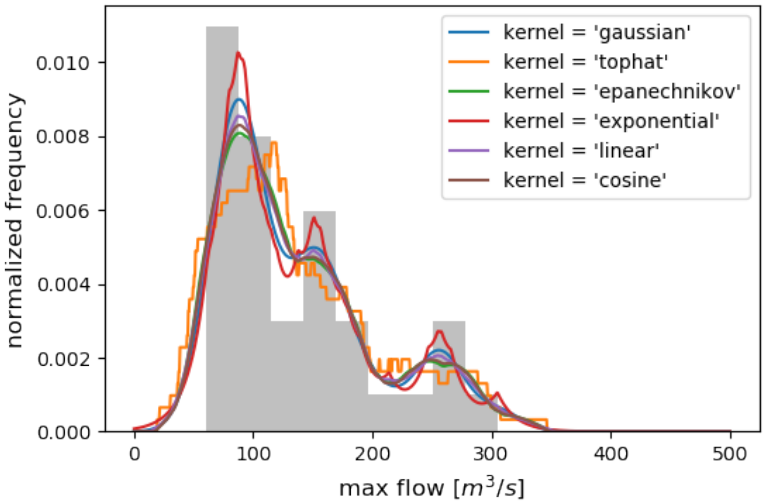
Jádro	$P(X > 230)$	$P(X > 310)$	$P(X < 50)$	$P(X < 70)$	$P(X > X_{max})$	$P(X > 280 \mid X > u)$
Gauss	0.1308	0.0111	0.0289	0.1199	0.0143	0.2875
Tophat	0.1241	0.0114	0.0476	0.1529	0.0149	0.3634
Epanechnikov	0.1291	0.0112	0.0316	0.1308	0.0137	0.3075
Exponential	0.1329	0.0112	0.0315	0.1147	0.0169	0.2874
Linear	0.1293	0.0109	0.0320	0.1270	0.0139	0.3016
Cosine	0.1296	0.0110	0.0298	0.1278	0.0137	0.3017

Tabulka 6: Výpočty příslušných pravděpodobností hladin pro vybrané modely.

hladin pro 10-letou, 100-letou a 1000-letou vodu, zjistíme, že exponenciální jádro má v dlouhodobém horizontu výrazně těžší pravý chvost, nežli ostatní modely. To je patrné především na hodnotě hladiny pro 1000-letou vodu. Ta je patrně nerealistická, což naznačuje, že model pro predikci s exponenciálním jádrem není pro modelování použitelný. Podobně nerealistické hodnoty byly pozorovány i při volbě vyšších hodnot bandwidth.



Obrázek 10: Porovnání modelů s epanechniovým jádrem pro různé metody volby bandwidth.



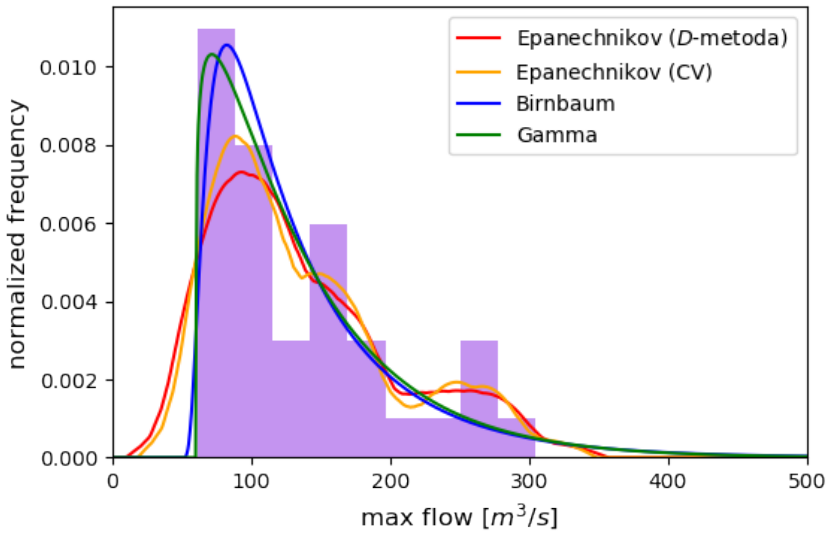
Obrázek 11: Porovnání jednotlivých jádrových odhadů.

Jádro	$h_{0.1}$	$h_{0.01}$	$h_{0.001}$
Gauss	248.0	312.0	339.5
Tophat	250.0	310.5	323.0
Epanechnikov	251.5	308.5	319.5
Exponential	250.5	324.5	7145.5
Linear	251.5	308.0	319.0
Cosine	251.5	308.5	319.5

Tabulka 7: Výpočty konkrétních hladin významnosti pro vybrané modely.

2.4 Srovnání modelů

V rámci srovnání modelů lze hned ze začátku pozorovat platnost odhadů získaných ze statistických nerovností. Přesnost těchto odhadů se nicméně pro praxi nejeví jako dostatečná. Porovnáme-li dále hodnoty predikcí z tabulek 3 a 6 s tvary rozdělení na obrázku 12, pozorujeme výrazné rozdíly v podobě jednotlivých modelů. Jádrové odhady se totiž prezentují výrazně těžšími levými chvosty, a naopak lehčími pravými chvosty. Hodnoty predikcí pro extrémní sucho jsou tak u jádrových odhadů výrazně vyšší. Naopak hodnoty predikcí pro extrémně vysoké průtoky jsou znatelně vyšší u modelů využívajících rozdělení pro statistiku extrémních událostí. Druhým zásadním rozdílem je popis modelů pro standardně pozorované průtoky. Zatímco jádrové odhady dokázaly zachytit tři výrazné peaky v histogramu, parametrické modely nikoliv. Pokud se tyto tři peaky v rozdělení náhodné veličiny mají skutečně objevovat, modely s jádrovými odhady získávají výraznou výhodu. Jedná-li se však pouze o zkreslení vzniklé nízkým počtem dat, dochází v případě jádrových odhadů k přefitování modelu.



Obrázek 12: Porovnání vybraných jádrových odhadů a parametrických modelů.

3 Závěr

V rámci této práce jsme modelovali maximální roční průtoky dané řeky. Po deskriptivní analýze dat jsme na základě QQ plotů a PP plotů prověřili rozdělení průtoků. Následně jsme skrze distribution-free nerovnosti získali odhady predikcí extrémních průtoků. Poté jsme skrze vybraná rozdělení pro statistiku extrémních událostí modelovali rozdělení průtoků. Na závěr jsme rozdělení průtoků modelovali též skrze jádrové odhady s šesti vybranými druhy jader.

V rámci analýzy modelů jsme pozorovali výrazné rozdíly v popisu chvostů pro různé typy modelů. Jádrové odhady se prezentovaly těžkými levými chvosty, zatímco modely s extremálními rozděleními se prezentovaly spíše těžkými pravými chvosty. Další zásadní rozdíl je v podobě chování modelu pro hojně pozorované průtoky. Jádrové modely výrazně přesněji popisují tři výrazné peaky v rozdělení. Pozorování těchto peaků nicméně mohlo být způsobeno málo početnou datovou množinou či zvoleným počtem binů histogramu. Z tohoto důvodu bychom pro modelování průtoků volili buď model s Epanechnikovým jádrem, který se oproti zbylým jádrovým modelům jevil jako nejvíce vyhlazený, nebo modely s Birnbaumovým či Gamma rozdělením. Ty na základě QQ plotů vykazovaly nejlepší shodu s empirickým rozdělením dat.

Reference

[1] Legg, P. - Rosin, P. Improving accuracy and efficiency of mutual information for multi-modal retinal image registration using adaptive probability density estimation. *Computerized Medical Imaging and Graphics*, 37 (7-8). str. 597-606. ISSN 0895-6111.

[2] Wikipedia Contributors Cantelli's inequality, `https://en.wikipedia.org/wiki/Cantelli%27s_inequality`, [online cit 04/09/2019].