

Použití modifikovaného LN5 rozdělení v hydrologii a klimatologii

Ladislav Budík
Český hydrometeorologický ústav
Pobočka Brno
Jan Holub
Ústav matematiky a statistiky
Přírodovědecká fakulta MU Brno

Statistické dny 2023, 19. 5. – 21. 5. 2023

Osnova

Empirická a teoretická křivka překročení, funkce přežití

Rozdělení LN5 a jeho modifikace mLN5

Metody odhadu parametrů

Příklady aplikace mLN5

Závěr

Empirická a teoretická křivka překročení, funkce přežití

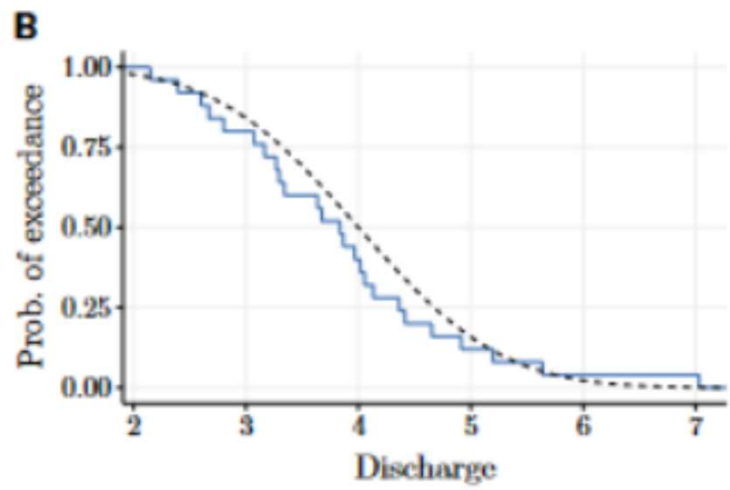
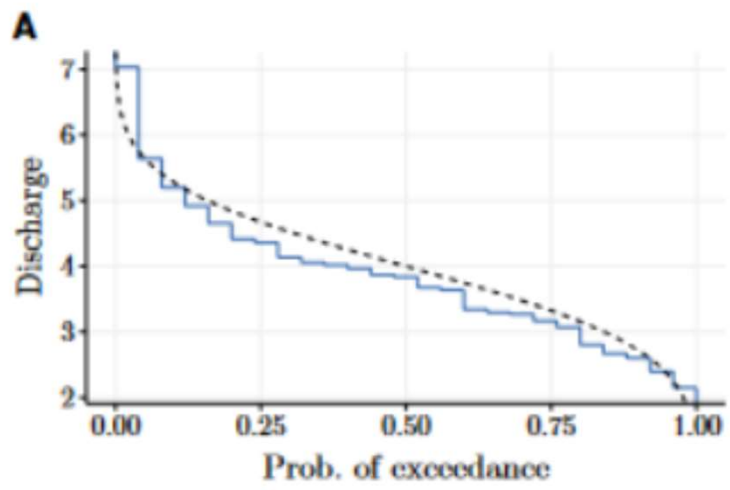
V hydrologii a klimatologii je jedním ze základních nástrojů při analýze dat křivka překročení. V jiných oblastech, např. v medicíně, se používá křivka přežití. Tyto termíny se často používají při studiu extrémních událostí, jako jsou povodně, zemětřesení nebo pády na burze.

Empirická křivka překročení ukazuje vztah mezi sestupnými hodnotami dané proměnné (na svislé ose) a odhadovanými pravděpodobnostmi překročení těchto hodnot (na vodorovné ose). Pravděpodobnost překročení daného prahu se odhaduje jako relativní četnost těch hodnot v souboru dat, které jsou nad tímto prahem.

Teoretická křivka překročení je založena na modelu rozdělení pravděpodobnosti sledované proměnné. Kvantily zvoleného rozdělení pravděpodobnosti jsou uspořádány sestupně na svislé ose a pravděpodobnosti překročení těchto kvantilů jsou vyneseny ve stejném pořadí na vodorovnou osu. Teoretickou křivku překročení lze použít k odhadu pravděpodobnosti odlehlých hodnot mimo rozsah dostupných údajů.

Funkce přežití je inverzní funkcí teoretické křivky překročení. Vyjadřuje pravděpodobnost, že realizace sledované proměnné bude větší než daná hodnota. Funkce přežití je nezbytná pro analýzu pravděpodobnosti přežití nebo doby přežití v různých oblastech.

Stručně řečeno, zatímco empirická křivka překročení je založena na naměřených datech, teoretická křivka překročení je založena na modelu základního rozdělení pravděpodobnosti a funkce přežití je inverzní křivkou překročení. Získání nejlepší teoretické křivky překročení je nezbytné pro odhad velikosti extrémní události s danou pravděpodobností překročení.



Pětiparametrické log-normální rozdělení LN5 a jeho modifikace mLN5

$$X \sim N(\mu, \sigma^2), a > 0, b > 0, y_0 \in \mathbb{R}$$

Definice rozdělení LN3:

$$Y = e^X + y_0, Y \sim \text{LN3}(\mu, \sigma^2, y_0)$$

Definice rozdělení LN5:

$$Y = a \cdot \exp(\text{sgn } X \cdot |X|^b + y_0), Y \sim \text{LN5}(a, b, \mu, \sigma^2, y_0), \boldsymbol{\theta} = (a, b, \mu, \sigma^2, y_0)^T$$

Hustota pravděpodobnosti:

$$f(y, \boldsymbol{\theta}) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{\ln^{\frac{1-b}{b}}\left(\frac{a}{y-y_0}\right)}{b(y-y_0)} \cdot \exp\left\{-\frac{\left(-\ln^{\frac{1}{b}}\left(\frac{a}{y-y_0}\right) - \mu\right)^2}{2\sigma^2}\right\}, & y \in (y_0, y_0 + a), \\ \frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{\ln^{\frac{1-b}{b}}\left(\frac{y-y_0}{a}\right)}{b(y-y_0)} \cdot \exp\left\{-\frac{\left(\ln^{\frac{1}{b}}\left(\frac{y-y_0}{a}\right) - \mu\right)^2}{2\sigma^2}\right\}, & y \in (y_0 + a, \infty), \\ 0, & \text{otherwise.} \end{cases}$$

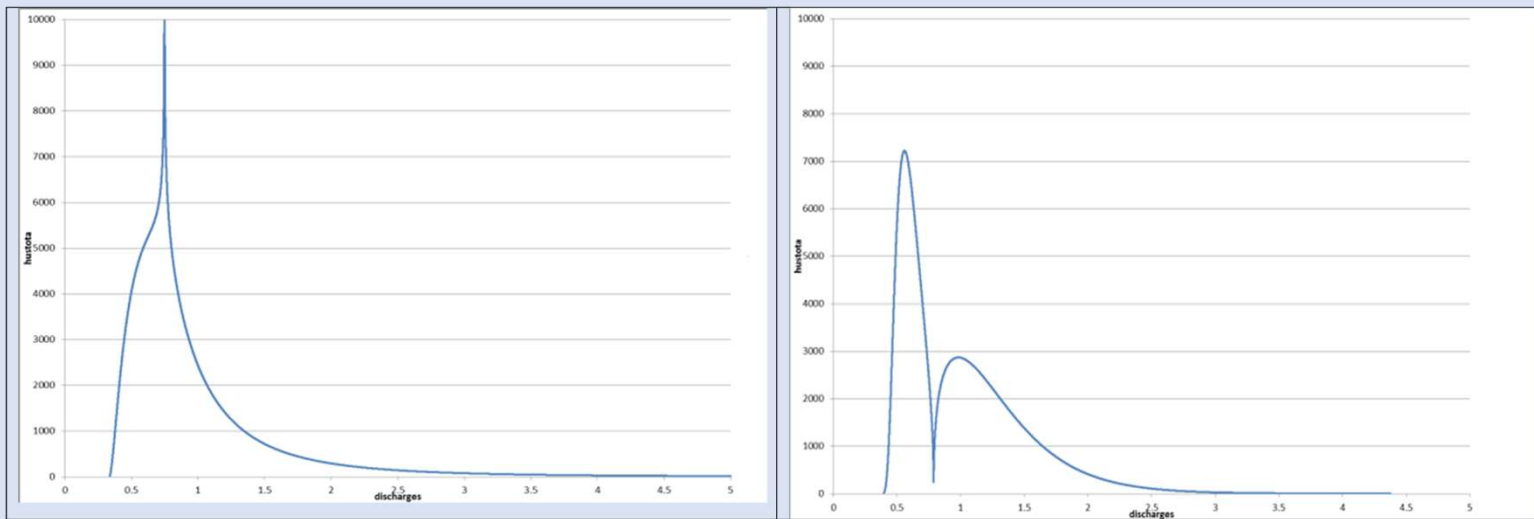
Distribuční funkce:

$$F(y, \boldsymbol{\theta}) = \begin{cases} 0, & y \in (-\infty, y_0), \\ \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{-\ln^{\frac{1}{b}} \left(\frac{a}{y-y_0} \right) - \mu}{\sqrt{2}\sigma} \right) \right], & y \in [y_0, a + y_0), \\ \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\ln^{\frac{1}{b}} \left(\frac{y-y_0}{a} \right) - \mu}{\sqrt{2}\sigma} \right) \right], & y \in [a + y_0, \infty), \end{cases}$$

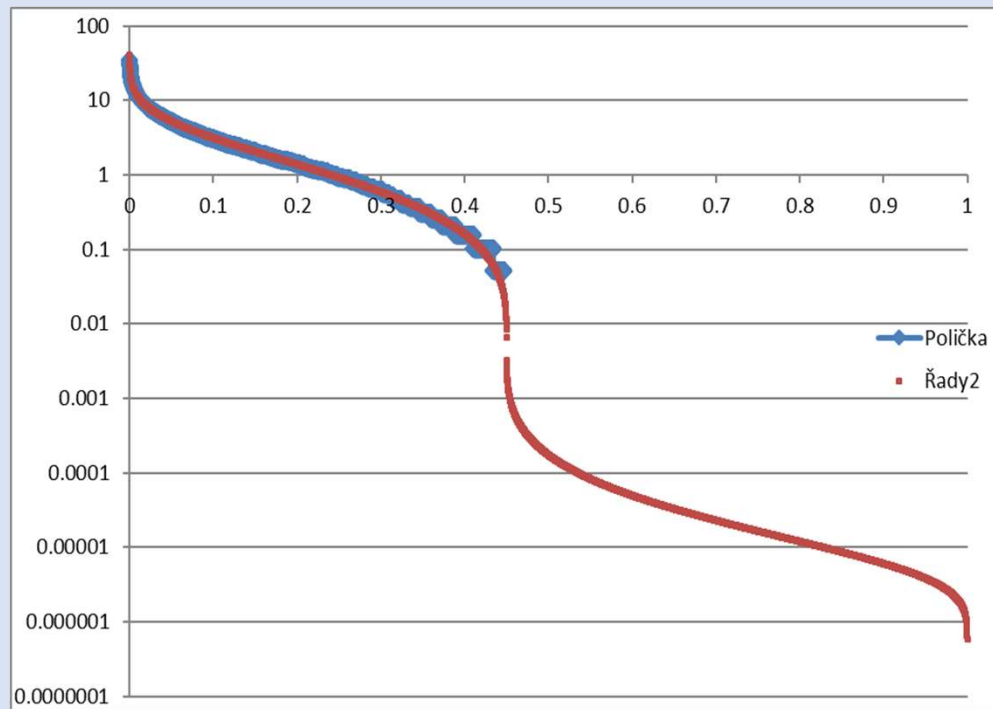
Kvantilová funkce:

$$F^{-1}(\alpha, \boldsymbol{\theta}) = \begin{cases} a \exp \left\{ - \left[-\mu - \sqrt{2}\sigma \operatorname{erf}^{-1}(2\alpha - 1) \right]^b \right\} + y_0, & 0 < \alpha < \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{-\mu}{\sqrt{2}\sigma} \right) \right], \\ a \exp \left\{ \left[\mu + \sqrt{2}\sigma \operatorname{erf}^{-1}(2\alpha - 1) \right]^b \right\} + y_0, & \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{-\mu}{\sqrt{2}\sigma} \right) \right] \leq \alpha < 1, \end{cases}$$

Příklady průběhu hustot LN5 pro $b < 1$ a $b > 1$



Někdy se takový průběh hustoty může i hodit. Např. křivku překročení denních srážkových úhrnů jsme dosud uměli vytvořit jen pro srážkové dny. Tento průběh umožní zobrazit vše, tedy i počty „suchých,“ dnů.



Definice rozdělení mLN5:

$$Y = \begin{cases} a \cdot \exp(\text{sign}X \cdot |X|^b) + y_0, & |X| \geq 1 \\ a \cdot \exp(\text{sign}X \cdot |X|^{b+(1-b)(1-|X|)}) + y_0, & |X| < 1 \end{cases}$$

Odvození hustoty pravděpodobnosti:

Zaveďme intervaly

$$G_1 = (-\infty, -1),$$

$$G_2 = (-1, 0),$$

$$G_3 = (0, 1),$$

$$G_4 = (1, \infty).$$

Na těchto intervalech definujme funkci $t(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ danou předpisem:

$$t(x) = \begin{cases} t_1(x) = a \exp\{-(-x)^b\} + y_0, & x \in G_1, \\ t_2(x) = a \exp\{-(-x)^{b+(1-b)(1+x)}\} + y_0, & x \in G_2, \\ t_3(x) = a \exp\{x^{b+(1-b)(1-x)}\} + y_0, & x \in G_3, \\ t_4(x) = a \exp\{x^b\} + y_0, & x \in G_4. \end{cases}$$

V tom případě obrazy intervalů G_1, G_2, G_3 a G_4 , které označíme H_1, H_2, H_3, H_4 , jsou

$$H_1 = t(G_1) = (y_0, ae^{-1} + y_0),$$

$$H_2 = t(G_2) = (ae^{-1} + y_0, a + y_0),$$

$$H_3 = t(G_3) = (a + y_0, ae + y_0),$$

$$H_4 = t(G_4) = (ae + y_0, \infty).$$

Dále nechť

$$\tau_1(\cdot) : H_1 \rightarrow G_1 = (y_0, ae^{-1} + y_0) \rightarrow (-\infty, -1) \text{ je inv. funkce k } t(x) \text{ na } H_1,$$

$$\tau_2(\cdot) : H_2 \rightarrow G_2 = (ae^{-1} + y_0, a + y_0) \rightarrow (-1, 0) \text{ je inv. funkce k } t(x) \text{ na } H_2,$$

$$\tau_3(\cdot) : H_3 \rightarrow G_3 = (a + y_0, ae + y_0) \rightarrow (0, 1) \text{ je inv. funkce k } t(x) \text{ na } H_3,$$

$$\tau_4(\cdot) : H_4 \rightarrow G_4 = (ae + y_0, \infty) \rightarrow (1, \infty) \text{ je inv. funkce k } t(x) \text{ na } H_4.$$

Pak hustota má tvar:

$$f(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2\sigma^2} \left[\ln^{\frac{1}{b}} \left(\frac{a}{y-y_0} \right) + \mu \right]^2 \right\} |\tau'_1(y)|, & y \in H_1, \\ \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2\sigma^2} [\tau_2(y) - \mu]^2 \right\} |\tau'_2(y)|, & y \in H_2, \\ \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2\sigma^2} [\tau_3(y) - \mu]^2 \right\} |\tau'_3(y)|, & y \in H_3, \\ \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2\sigma^2} \left[\ln^{\frac{1}{b}} \left(\frac{y-y_0}{a} \right) - \mu \right]^2 \right\} |\tau'_4(y)|, & y \in H_4, \\ 0, & \text{jinak,} \end{cases}$$

Distribuční funkce:

$$F(y) = \begin{cases} 0, & y \in (-\infty, y_0), \\ \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{-\ln^{\frac{1}{b}} \left(\frac{a}{y-y_0} \right) - \mu}{\sqrt{2}\sigma} \right) \right], & y \in [y_0, ae^{-1} + y_0), \\ \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\tau_2(y) - \mu}{\sqrt{2}\sigma} \right) \right], & y \in [ae^{-1} + y_0, a + y_0), \\ \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\tau_3(y) - \mu}{\sqrt{2}\sigma} \right) \right], & y \in [a + y_0, ae + y_0), \\ \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\ln^{\frac{1}{b}} \left(\frac{y-y_0}{a} \right) - \mu}{\sqrt{2}\sigma} \right) \right], & y \in [ae + y_0, \infty), \end{cases}$$

Kvantilová funkce:

$$F^{-1}(\alpha) = \begin{cases} a \exp \left\{ - \left(-\mu - \sqrt{2}\sigma \operatorname{erf}^{-1}(2\alpha - 1) \right)^b \right\} + y_0, & \alpha \in I_1, \\ t_2 \left(\mu + \sqrt{2}\sigma \operatorname{erf}^{-1}(2\alpha - 1) \right), & \alpha \in I_2, \\ t_3 \left(\mu + \sqrt{2}\sigma \operatorname{erf}^{-1}(2\alpha - 1) \right), & \alpha \in I_3, \\ a \exp \left\{ \left(\mu + \sqrt{2}\sigma \operatorname{erf}^{-1}(2\alpha - 1) \right)^b \right\} + y_0, & \alpha \in I_4, \end{cases}$$

kde

$$t_2(x) = a \exp \left\{ - (-x)^{b+(1-b)(1+x)} \right\} + y_0,$$

$$t_3(x) = a \exp \left\{ x^{b+(1-b)(1-x)} \right\} + y_0,$$

intervaly

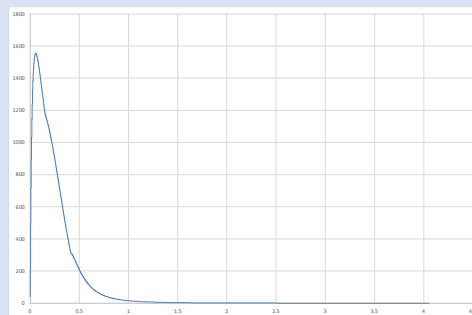
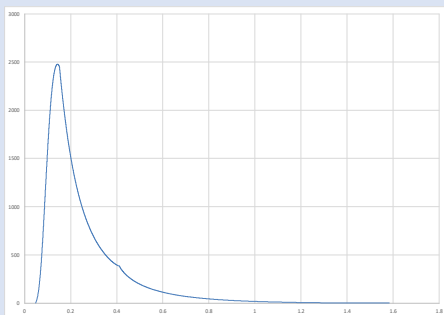
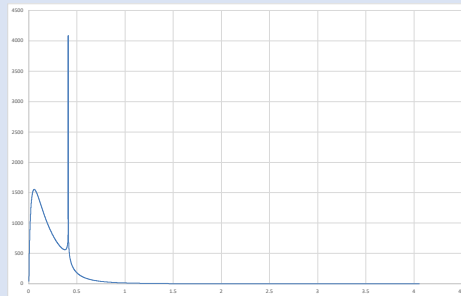
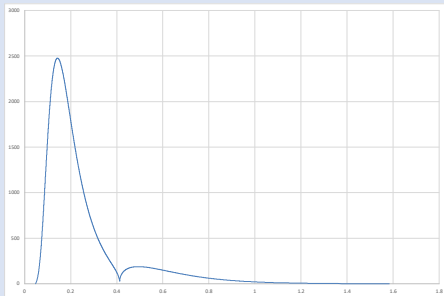
$$I_1 = \left(0, \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{-1-\mu}{\sqrt{2}\sigma} \right) \right] \right),$$

$$I_2 = \left[\frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{-1-\mu}{\sqrt{2}\sigma} \right) \right], \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{-\mu}{\sqrt{2}\sigma} \right) \right] \right),$$

$$I_3 = \left[\frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{-\mu}{\sqrt{2}\sigma} \right) \right], \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{1-\mu}{\sqrt{2}\sigma} \right) \right] \right),$$

$$I_4 = \left[\frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{1-\mu}{\sqrt{2}\sigma} \right) \right], 1 \right).$$

Příklady tvaru hustot pro $b < 1$ a $b > 1$, transformace LN5 a mLN5



Alternativní parametrizace LN5 a mLN5

Kvůli zlepšení numerických vlastností metod odhadu parametrů zavedeme následující parametrizaci:

$$\mu' = \frac{\mu}{\sigma}, \sigma' = \sigma^b$$

Pak rozdělení LN5 má tvar:

$$Y = a \exp \{ \operatorname{sgn}(U + \mu') \sigma' |U + \mu'|^b \} + y_0$$

a rozdělení mLN5 je tvaru:

$$Y = \begin{cases} a \exp \{ \operatorname{sgn}(U + \mu') \sigma' |U + \mu'|^b \} + y_0, & |X| \geq 1 \\ a \exp \{ \operatorname{sgn}(U + \mu') |U + \mu'|^{b+(1-b)(1-|X|)} (\sigma')^{(1+(\frac{1}{b}-1)(1-|X|))} \} + y_0, & |X| < 1 \end{cases}$$

přičemž $U \sim N(0, 1)$.

Poznámka k parametru y_0 : U teoretických křivek teplot vzduchu lépe funguje jako bod, kolem nějž křivku s pomocí parametru a otáčíme. U srážek a průtoků se ukazuje, že y_0 je nejspíše rovno 0 (nulový posun).

Metody odhadu parametrů

n ... celkový počet měření

y_i ... i -té měření

y_{ti} ... i -tá teoretická hodnota

$p_i = \frac{i}{n+1}$... pravděpodobnost překročení hodnoty y_i na empirické křivce překročení

$p_{ti} = F(y_i; \theta)$... pravděpodobnost překročení hodnoty y_{ti} na teoretické křivce překročení

$u_i = \frac{1}{\sigma}(\tau(y_i, a, b, y_0) - \mu)$... zpětná transformace hodnoty y_i na standardizované normální rozdělení, kde τ je inverzní funkce k funkci t použité při odvozování hustoty mLN5

$u_{ti} = \Phi^{-1}(p_i)$... transformace pravděpodobnosti p_i na standardizované normální rozdělení

Metoda relativních nejmenších čtverců

Tato metoda hledá parametry křivky překročení tak, aby výraz $\sum_{i=1}^n \left(\frac{y_i - y_{ti}}{y_i} \right)^2$ nabýval minimální hodnoty. Minimalizují se tedy čtverce relativních odchylek empirické a teoretické křivky překročení vzhledem k ose y .

Metoda nejmenších čtverců se zpětnou transformací

Při použití metody nejmenších čtverců minimalizujeme výraz $\sum_{i=1}^n (u_i - u_{ti})^2$. Jde tedy o minimalizaci odchylek empirické a teoretické křivky překročení vzhledem k ose y.

Metoda optimalizace pravděpodobnosti

Hledáme parametry křivky překročení tak, aby výraz $\sum_{i=1}^n (p_i - p_{ti})^2$ nabýval svého minima. Dochází tedy k minimalizaci čtverců odchylek pravděpodobností překročení empirických a teoretických hodnot vzhledem k ose x.

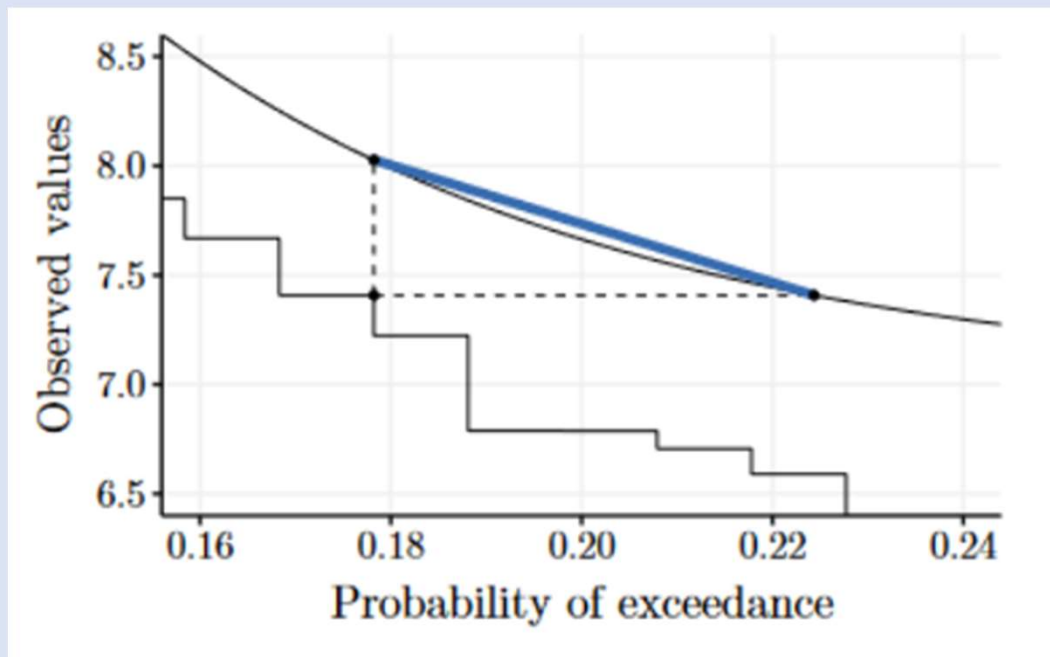
Trojúhelníková metoda

Tato metoda vychází z kombinace metody relativních nejmenších čtverců (resp. metody nejmenších čtverců se zpětnou transformací) a metody optimalizace

pravděpodobností. Minimalizuje se výraz: $\sum_{i=1}^n \left(\frac{y_i - y_{ti}}{y_i} \right)^2 + \sum_{i=1}^n (p_i - p_{ti})^2$ (resp. výraz

$$\sum_{i=1}^n (u_i - u_{ti})^2 + \sum_{i=1}^n (p_i - p_{ti})^2 .$$

Ilustrace trojúhelníkové metody:



Metoda maximální věrohodnosti – zde není použita.

Poznámka k parametru y_0 : U teoretických křivek teplot vzduchu lépe funguje jako bod, kolem nějž křivku s pomocí parametru a otáčíme. U srážek a průtoků se ukazuje, že y_0 je nejspíše rovno 0 (nulový posun).

V následujícím vztahu je zahrnuta i možnost rotace, která je potřebná pro prokládání teplot zároveň s úpravou pro extrémní σ a μ . Tato úprava umožňuje zvládat proložení různě asymetrických až symetrických dat. Počet parametrů je stále 5.

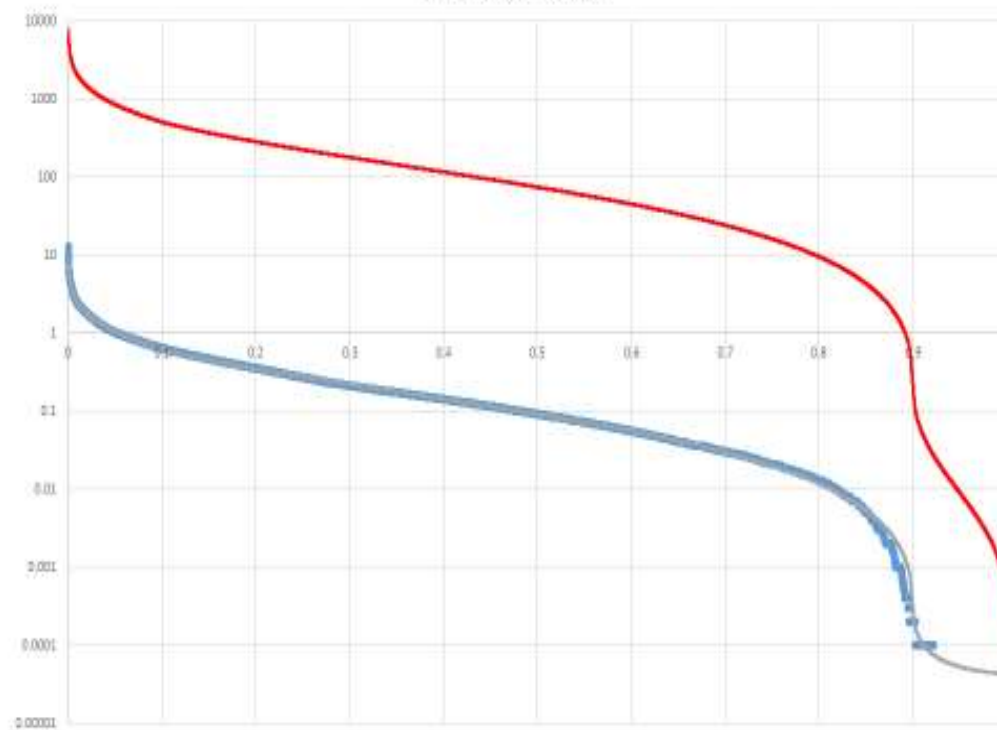
$$Y = \{a\bar{Y} \left[e^{\left(x + \frac{\mu}{\sigma}\right)^b \sigma^b} \right] - a \cdot y_0'\} , \text{ pro } |X| \geq 1$$

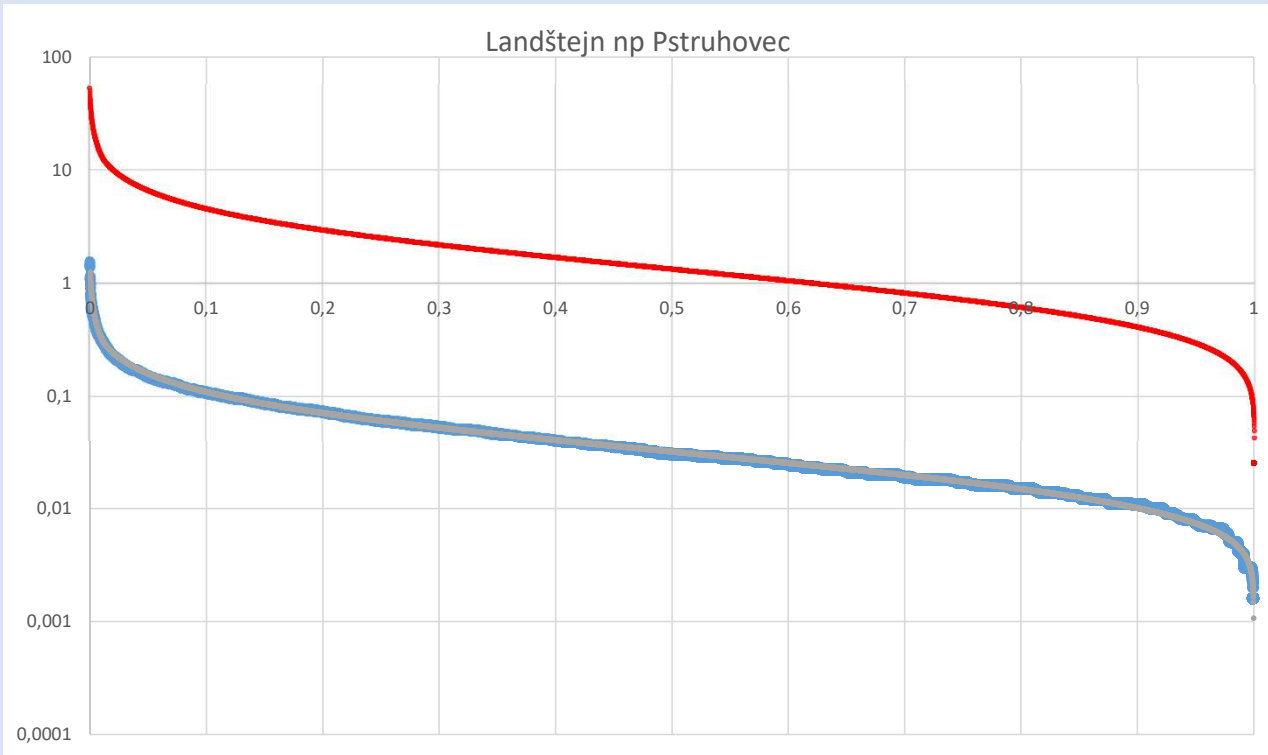
$$Y = \{a\bar{Y} \left[e^{\left(x + \frac{\mu}{\sigma}\right)^{b+(1-b)(1-|x\sigma+\mu|)} \sigma^{b+(1-b)(1-|x\sigma+\mu|)}} \right] - a \cdot y_0'\} , \text{ pro } |X| < 1$$

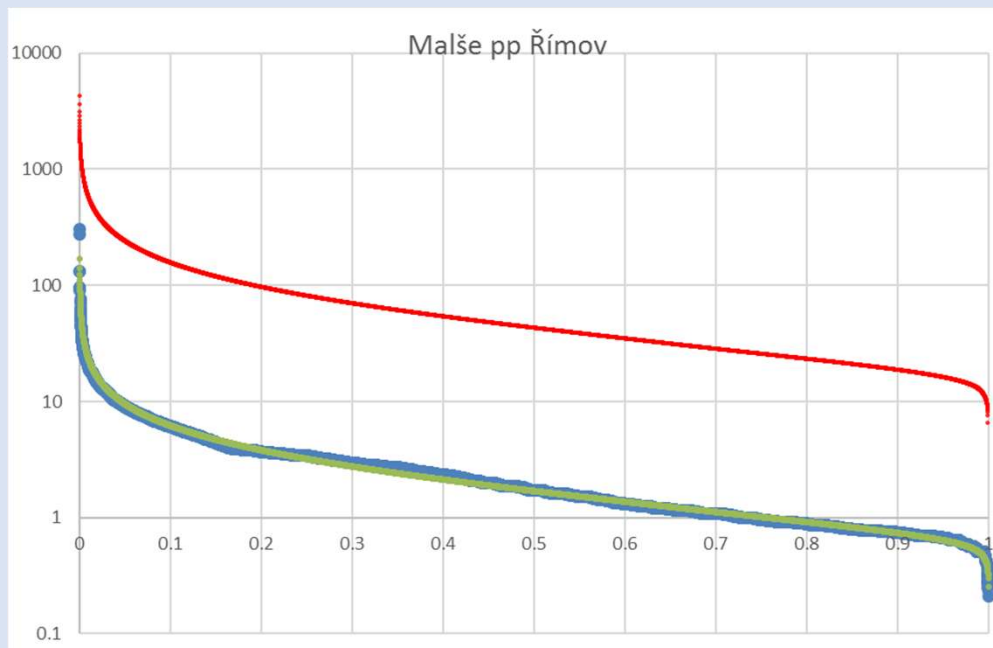
Příklady užití

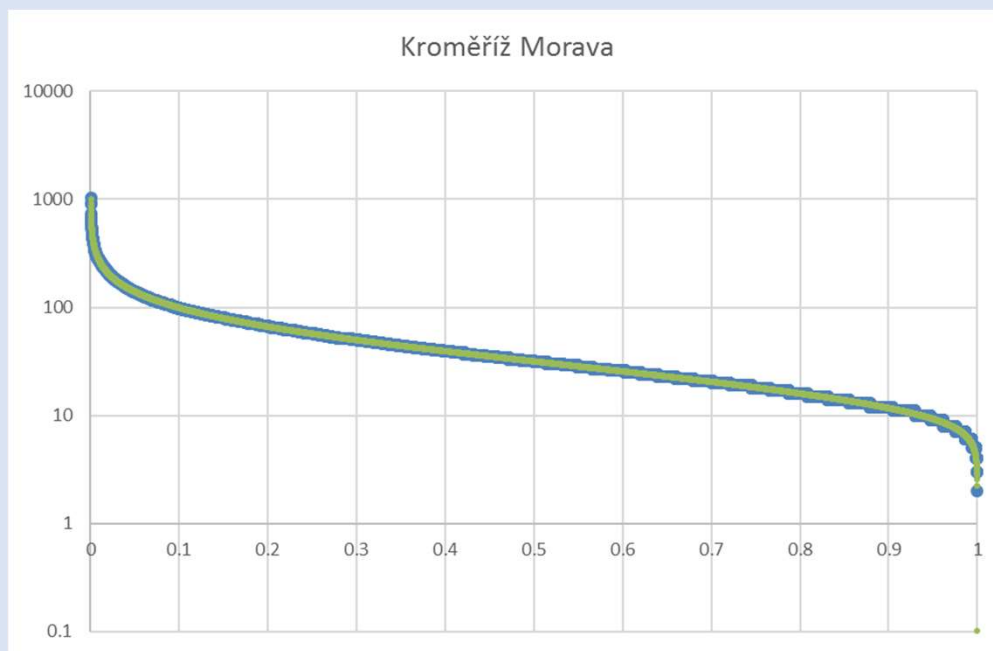


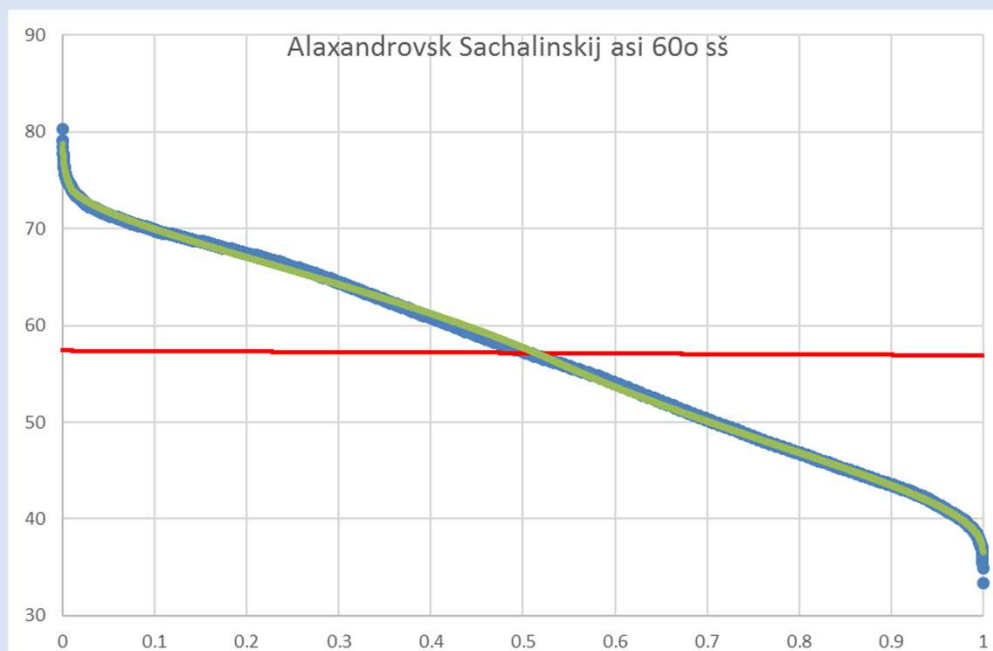
Sloup Punkva průtoky

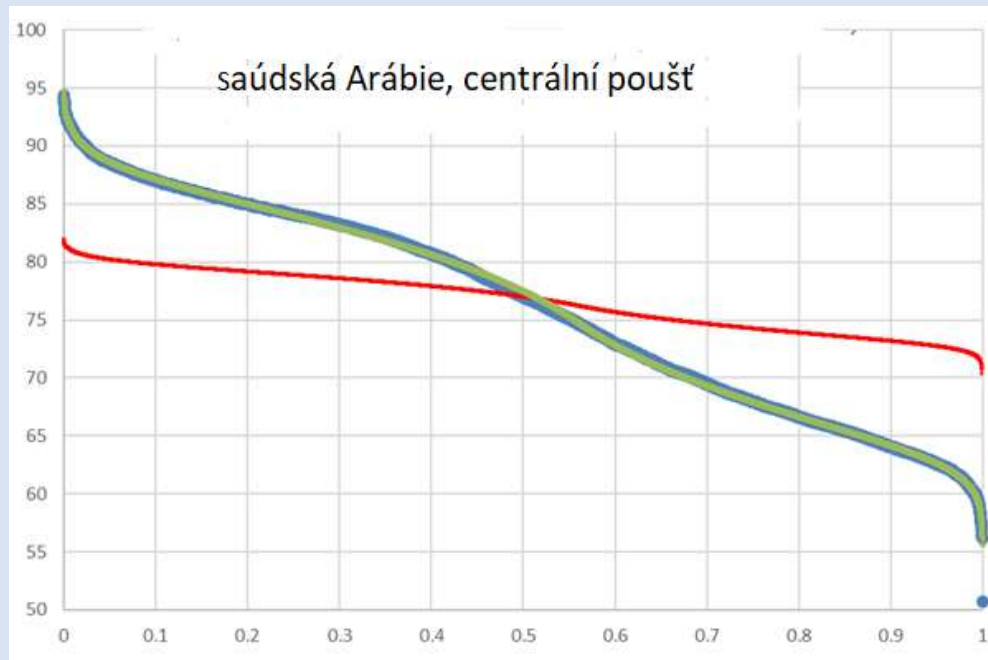


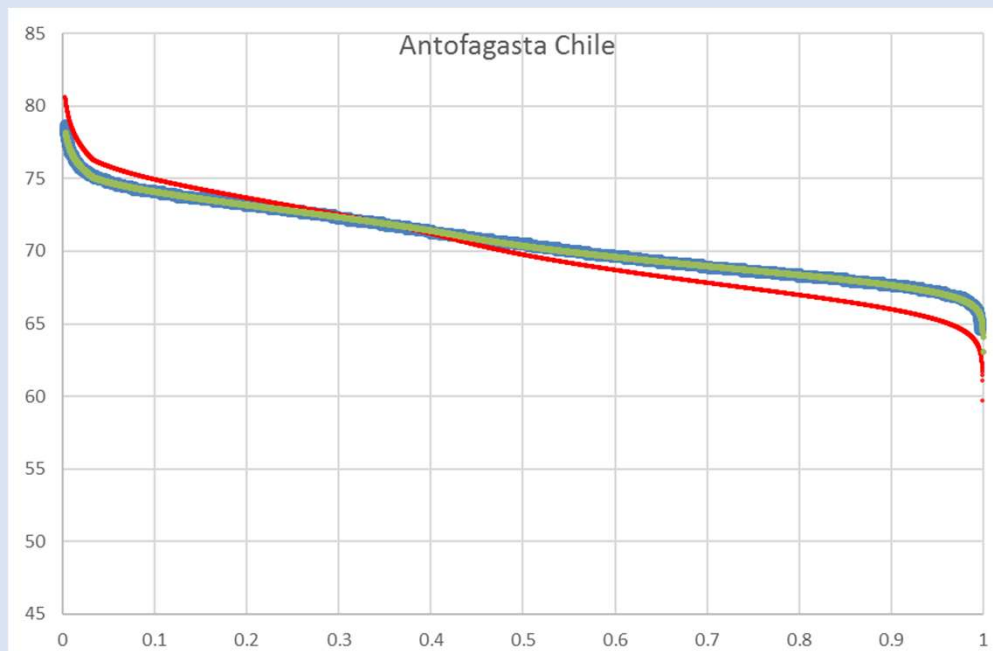


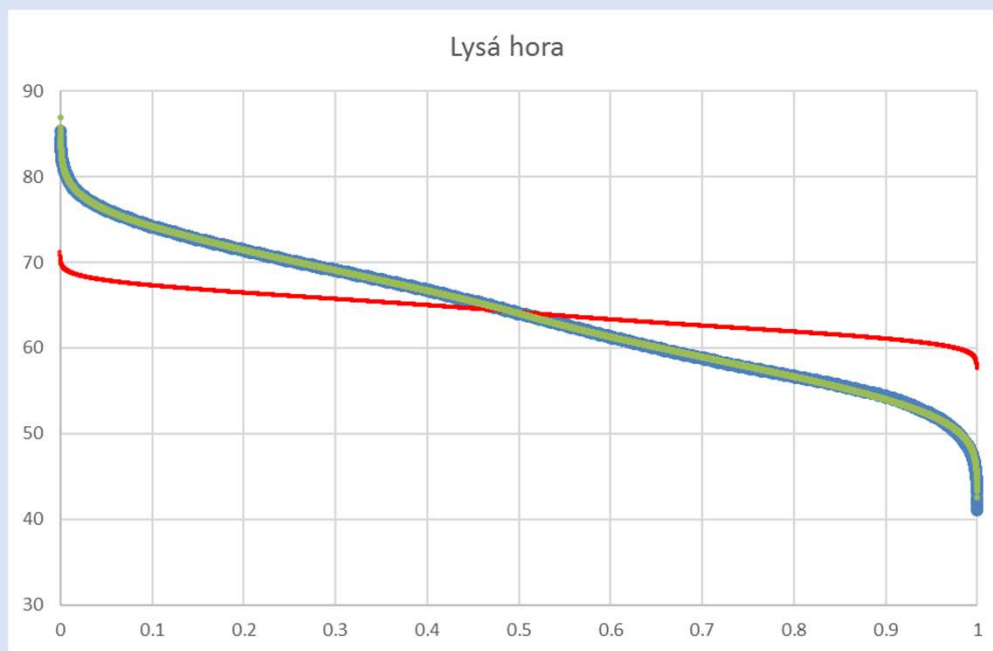




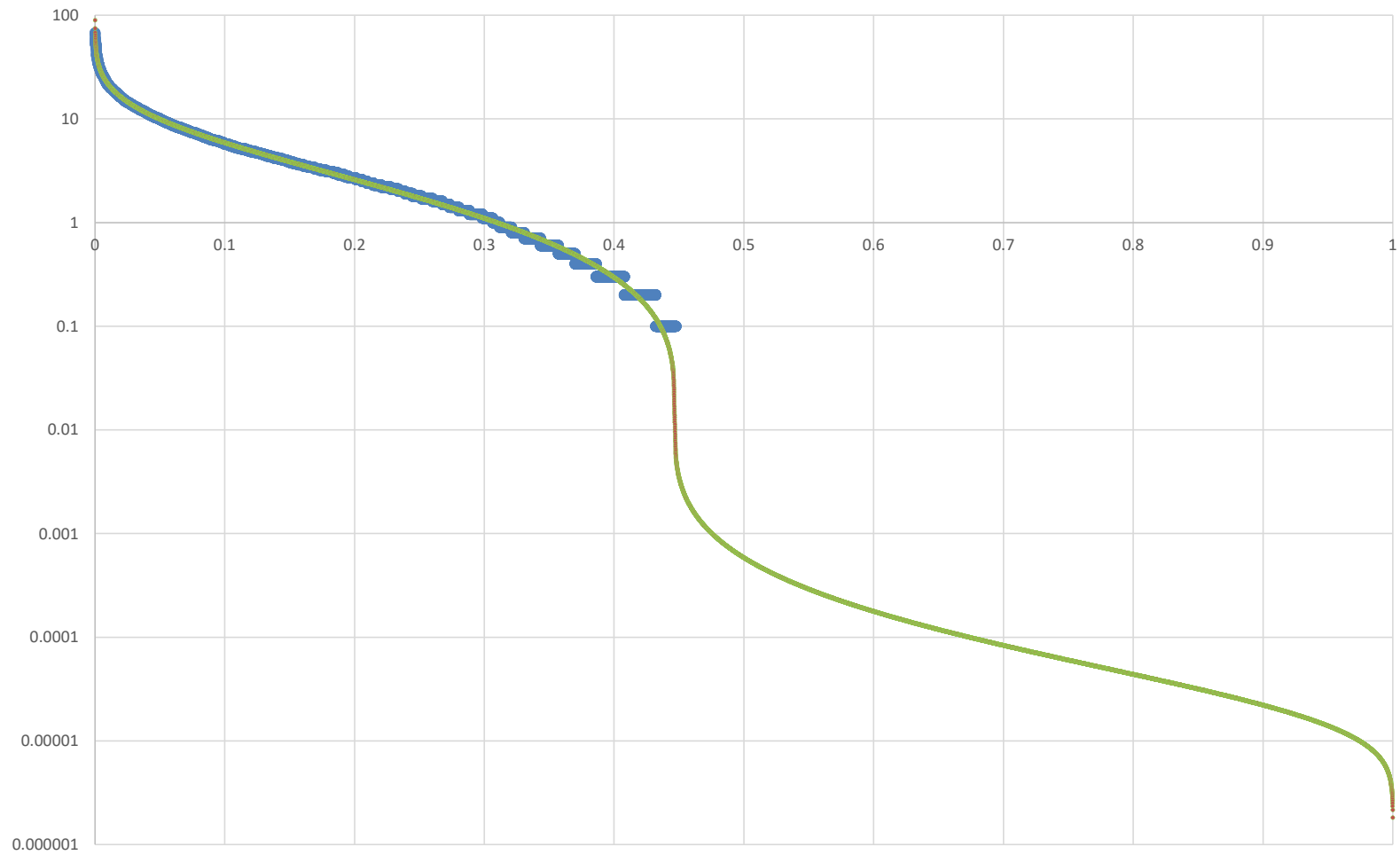




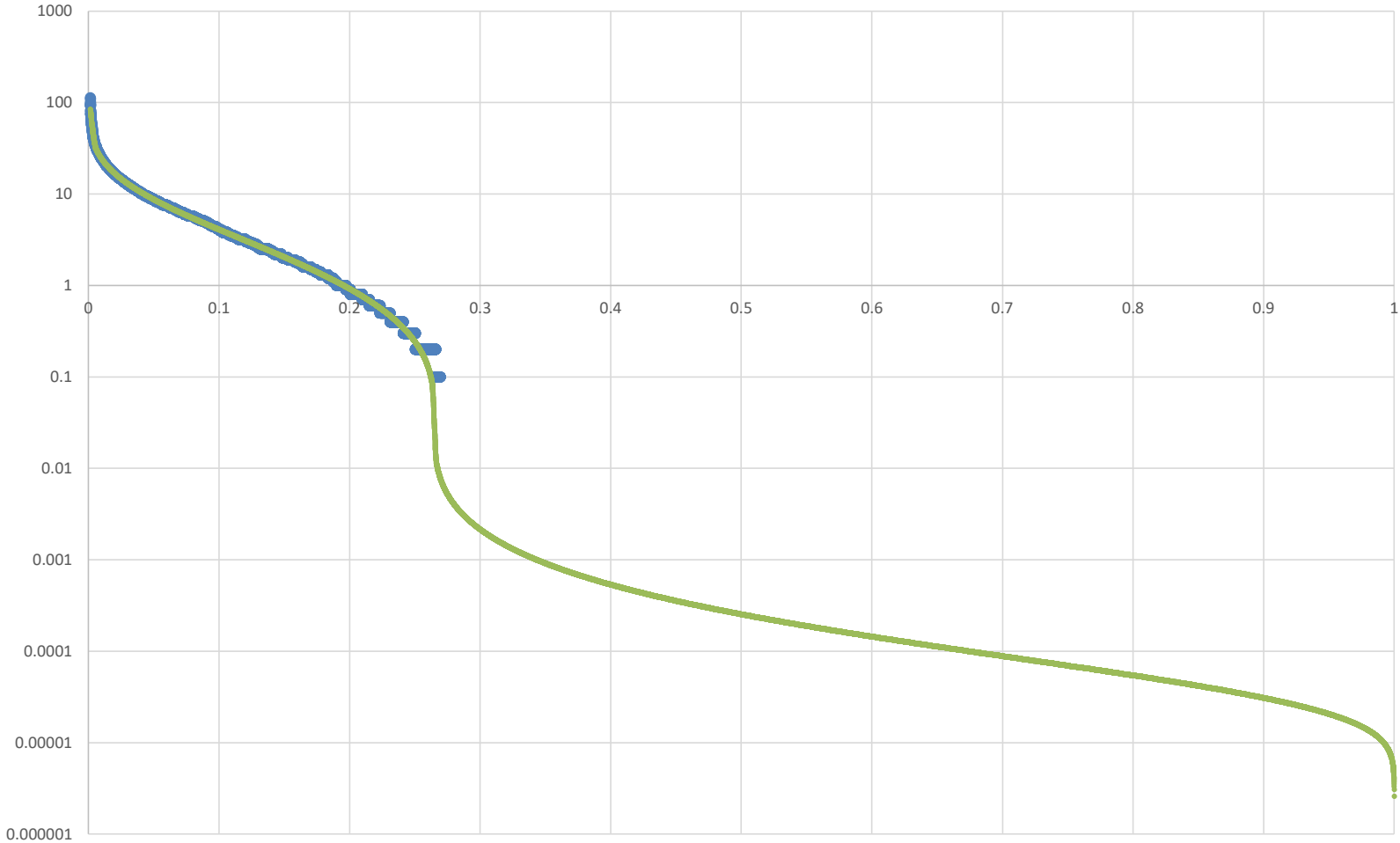




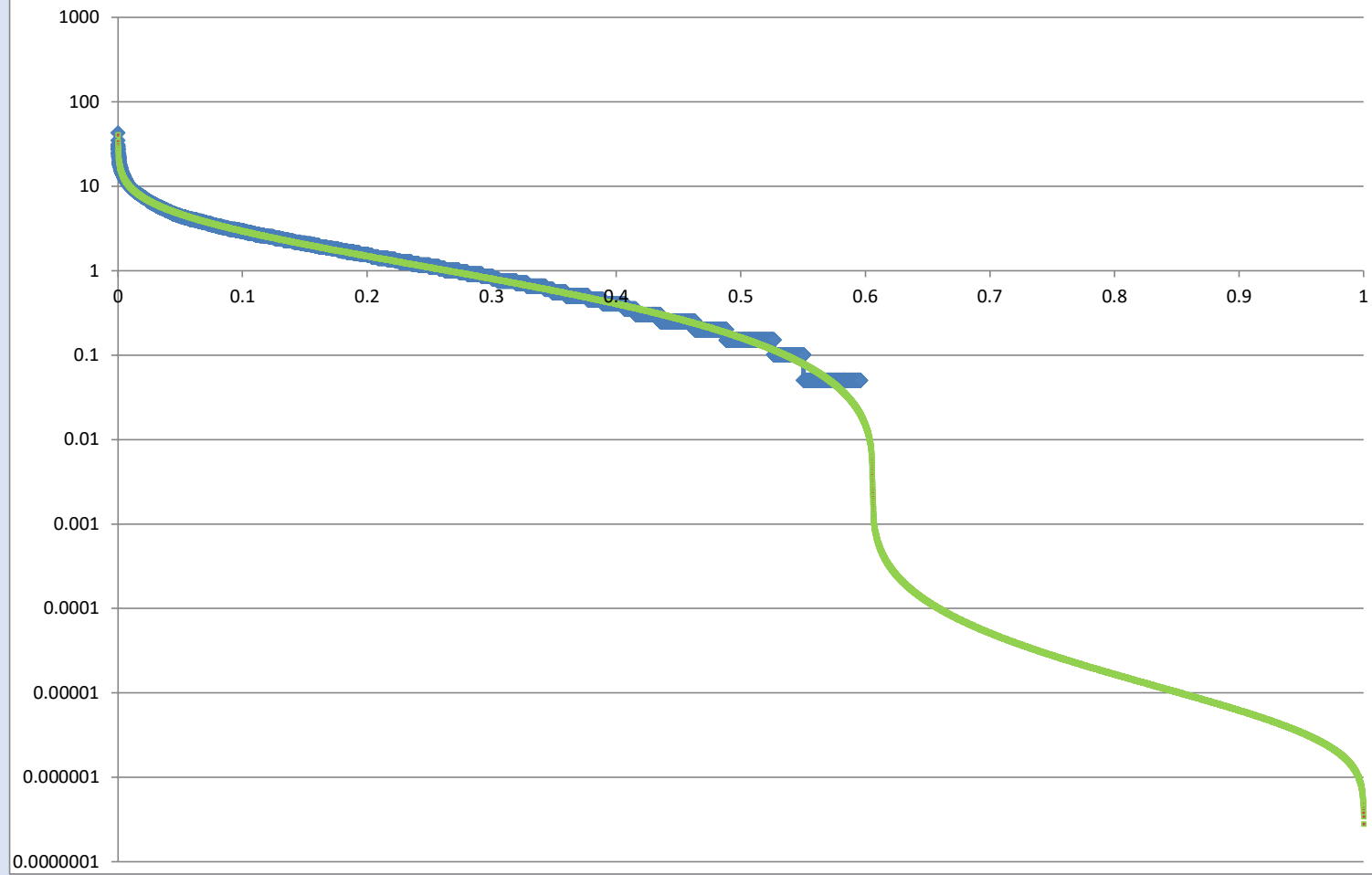
Polička, denní srážky

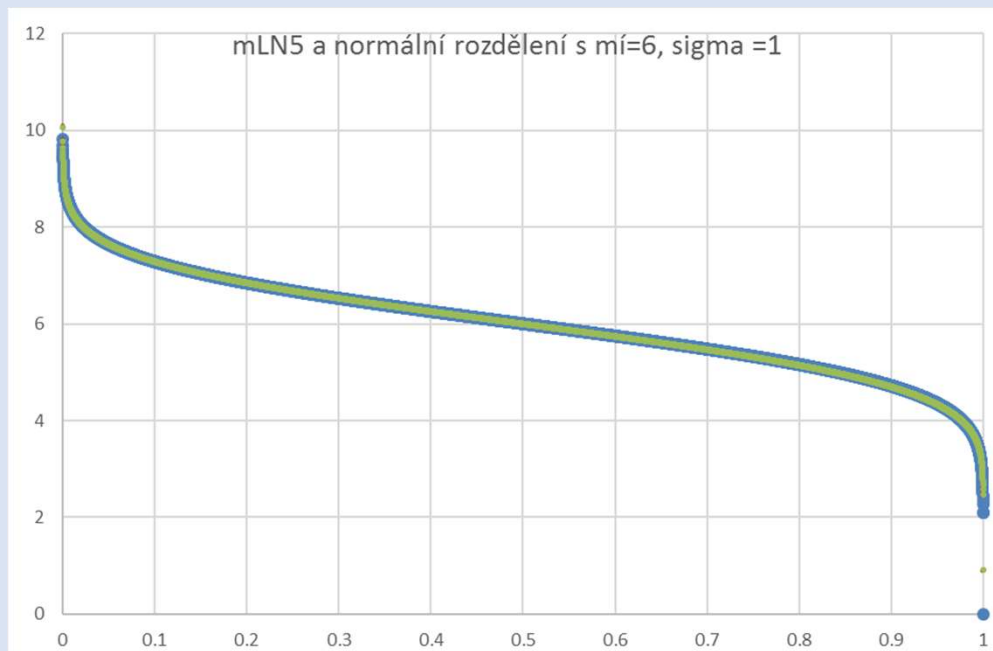


Palermo, denní srážky



Dublin, denní srážky





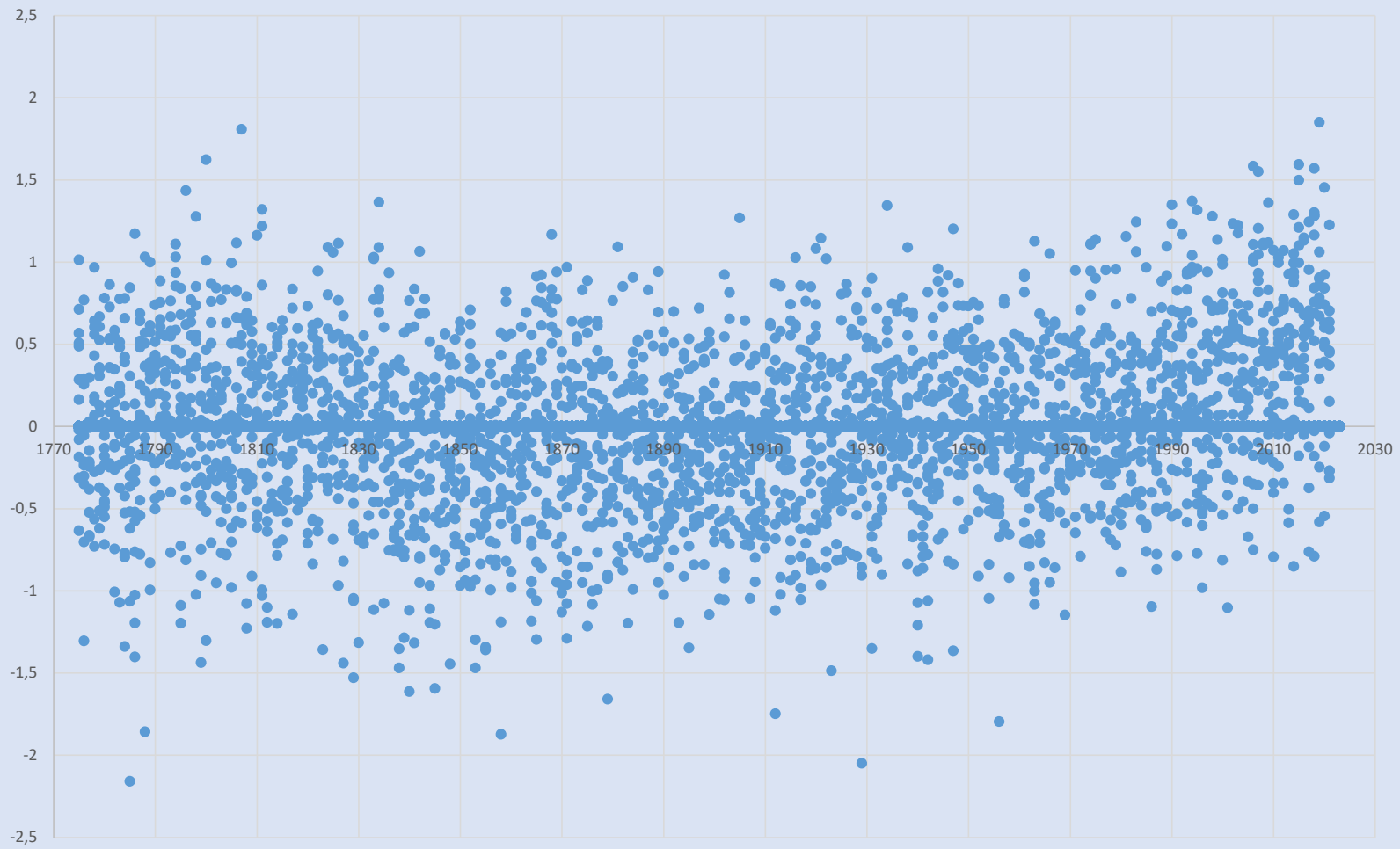
Prodlužování řad

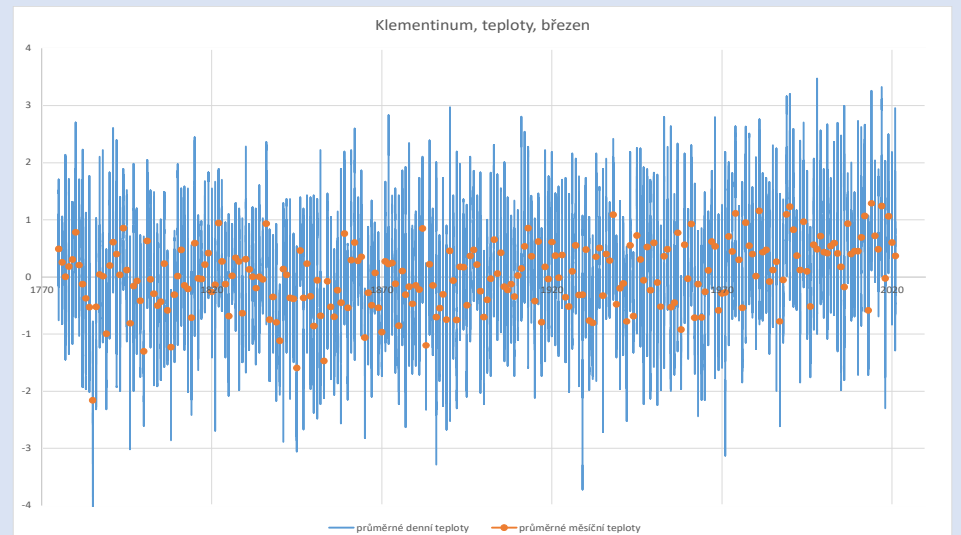
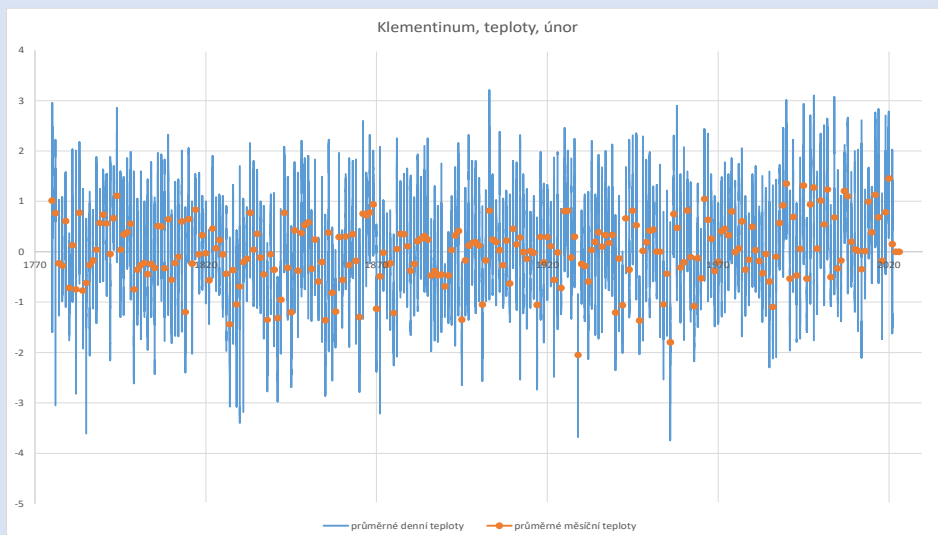
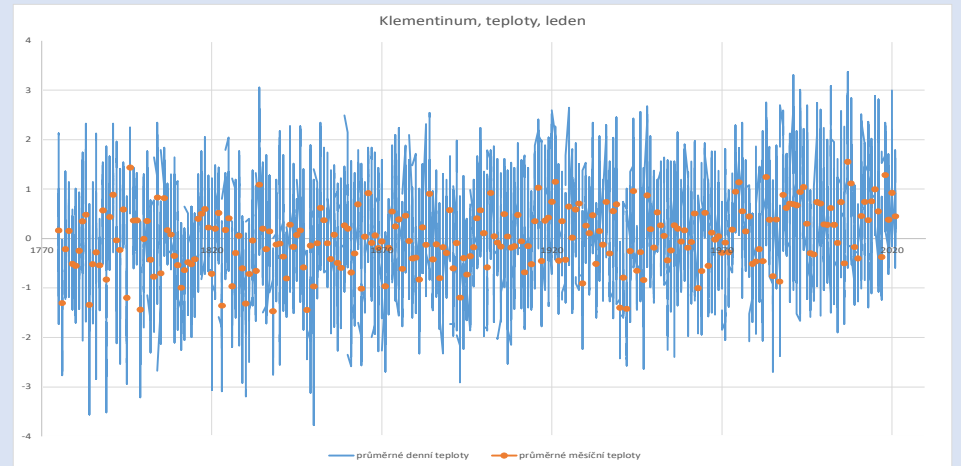
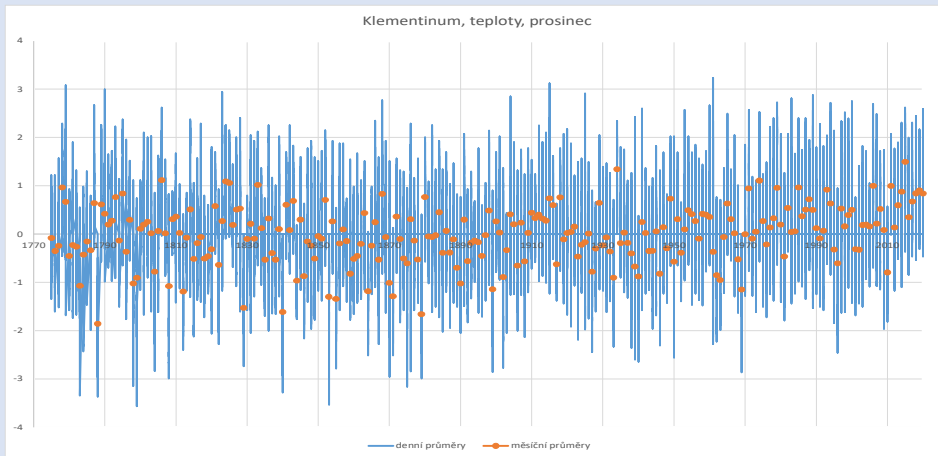
- Tento postup prokládání teoretických křivek umožňuje i prodlužování řad přesněji než dosavadními postupy s pomocí násobení poměrem průměru ze vzorové řady.
- Máme řadu 10 let a vzorovou 30 let. Je třeba těch 20 let doplnit alespoň přibližně.
- Proložíme obě řady teoretickou křivkou překročení ve společném období (10 let).
- Do doplňovaných 20ti let ve vzorové řadě dáme parametry mLN5 z desetiletí. Zpětnou transformací na „normální rozdělení“ získáme odchylky rozptylů mezi obdobími. Musíme v doplňované řadě zachovat poměry průměrů a k takto získanému průměru prodloužení dopočítat správný parametr a .

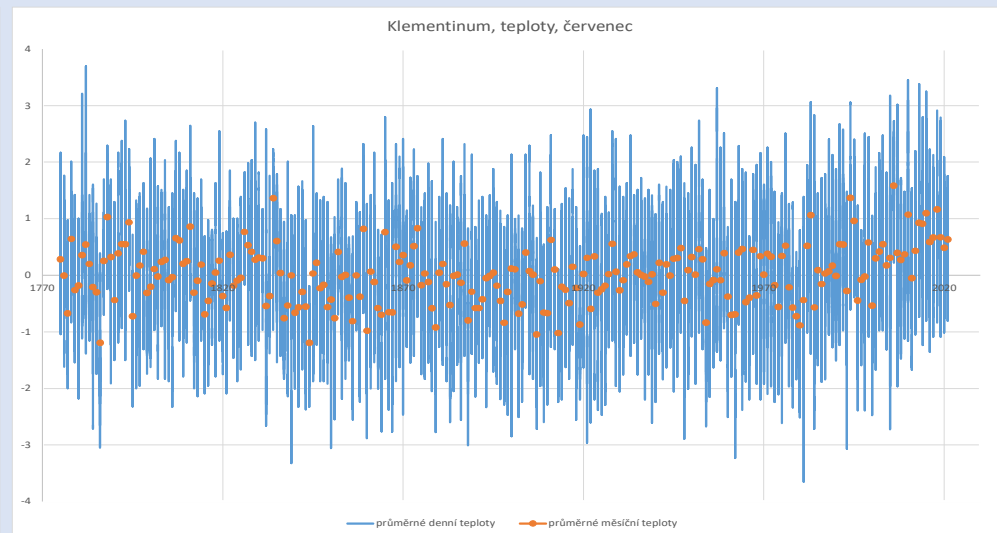
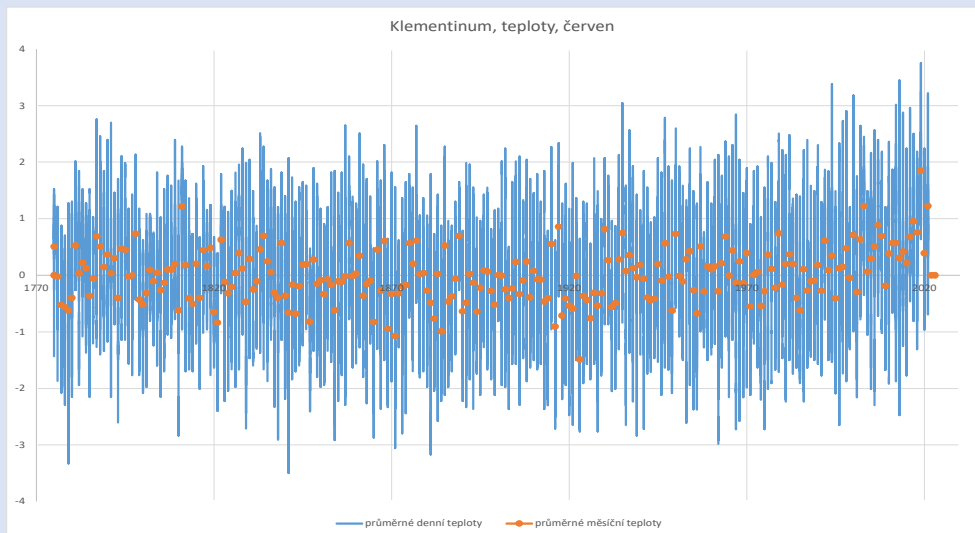
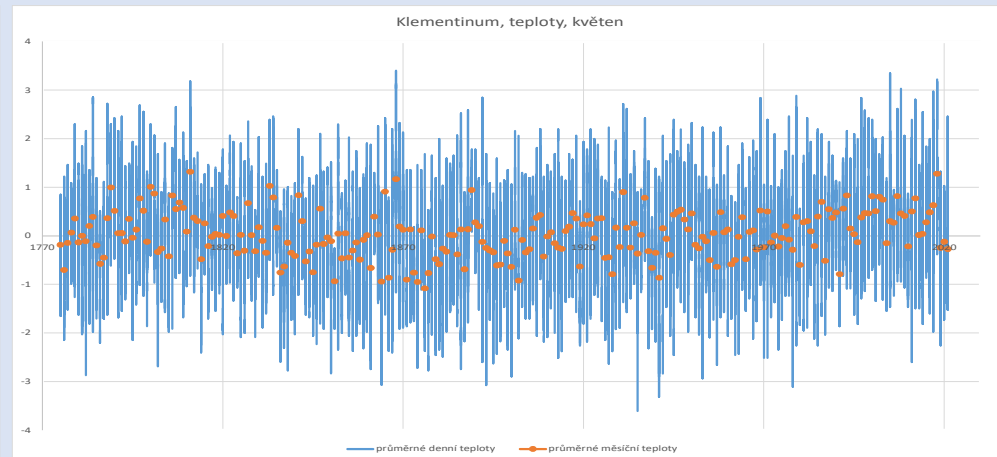
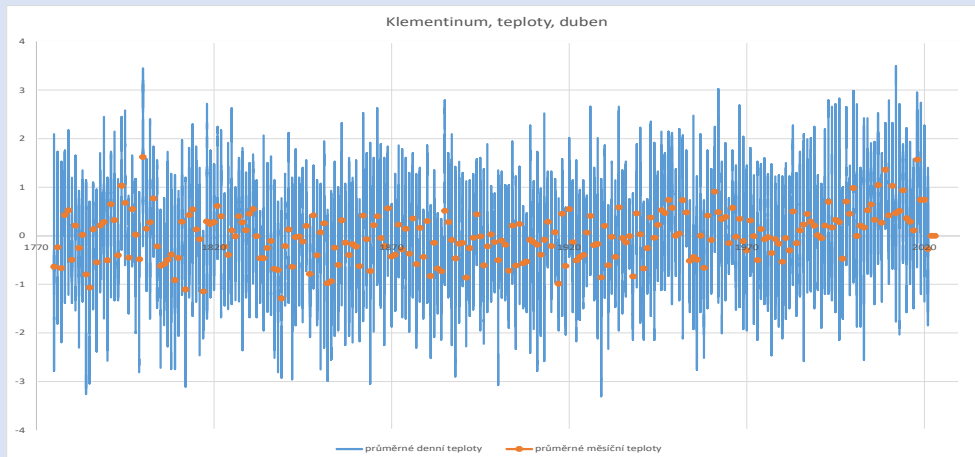
- Toto „normální“ standardizované rozdělení přeneseme jako základ 20ti let doplnění a z něj s pomocí parametrů 10ti let doplňované řady spočteme křivku překročení 20ti let doplňované řady. Spojením obou úseků získáme doplněnou křivku překročení za 30ti letí. Pokud bychom očíslovali pořadí hodnot ve vzorové časové řadě a to dále přenášeli s křivkou překročení v „normálním“ rozdělení ze vzorové časové řady, získali bychom přibližný průběh časové řady doplněné.

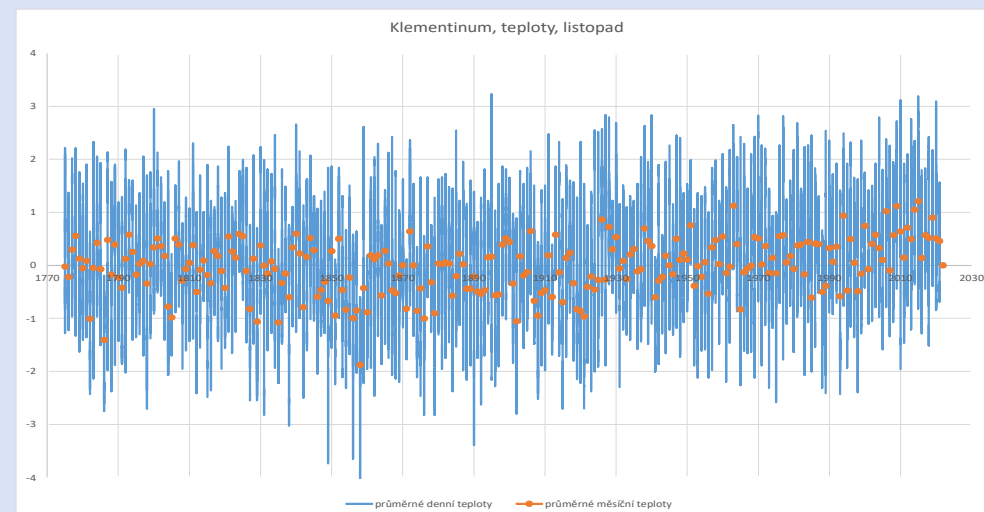
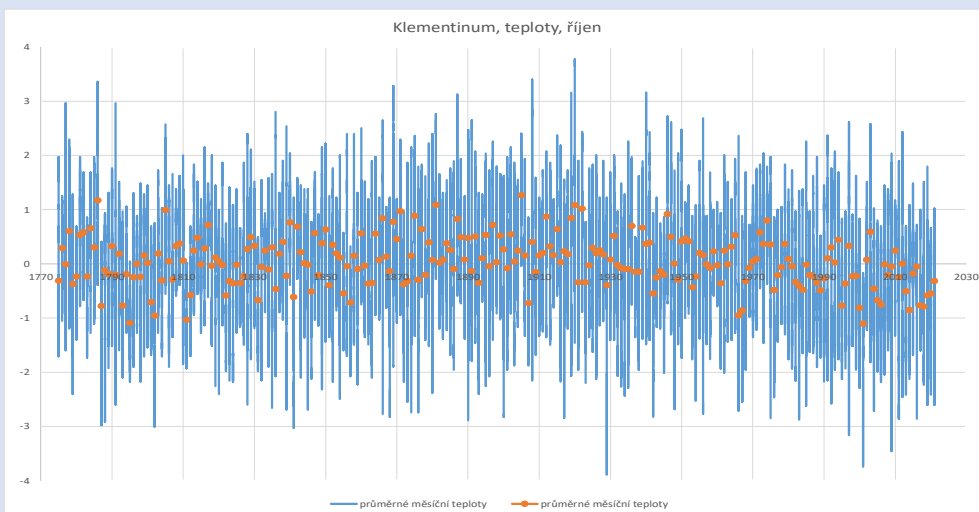
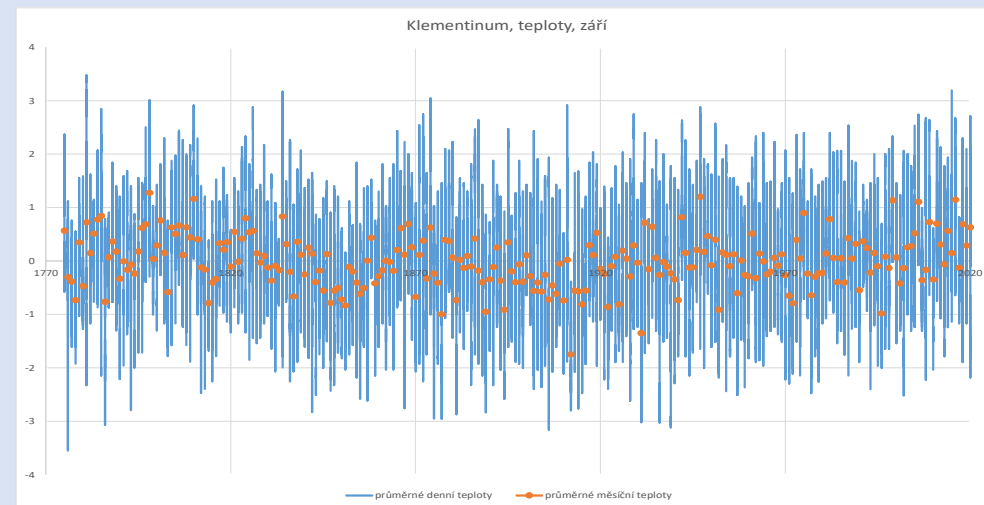
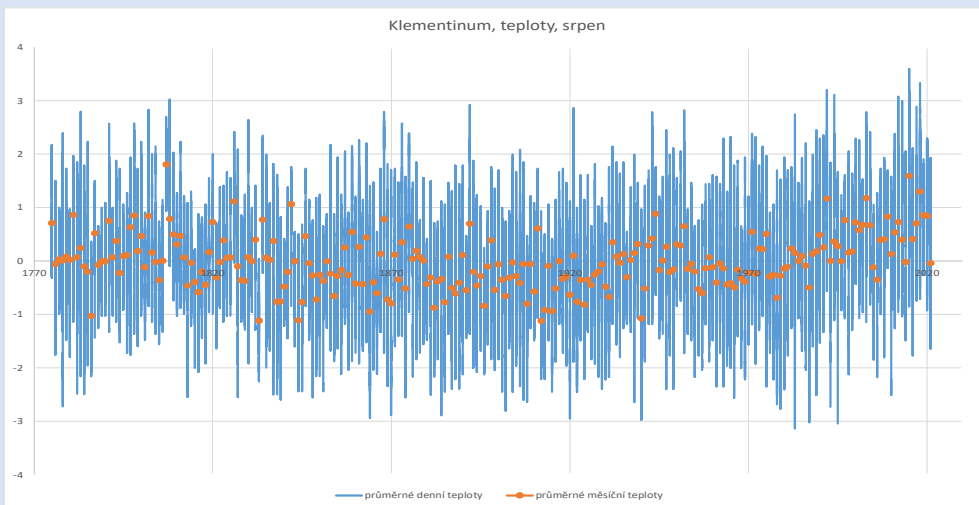
- mLN5 rozdělení je možno použít k poměrně dokonalému převodu dat do standardizovaného normálního rozdělení. Lze tak převést i data, kde to pomocí jiných rozdělení nejde nebo je to nepřesné. Příkladem zde mohou být měsíční data teplotní řady Praha - Klementinum.

Klementinum, teploty měs.průměry standardizované

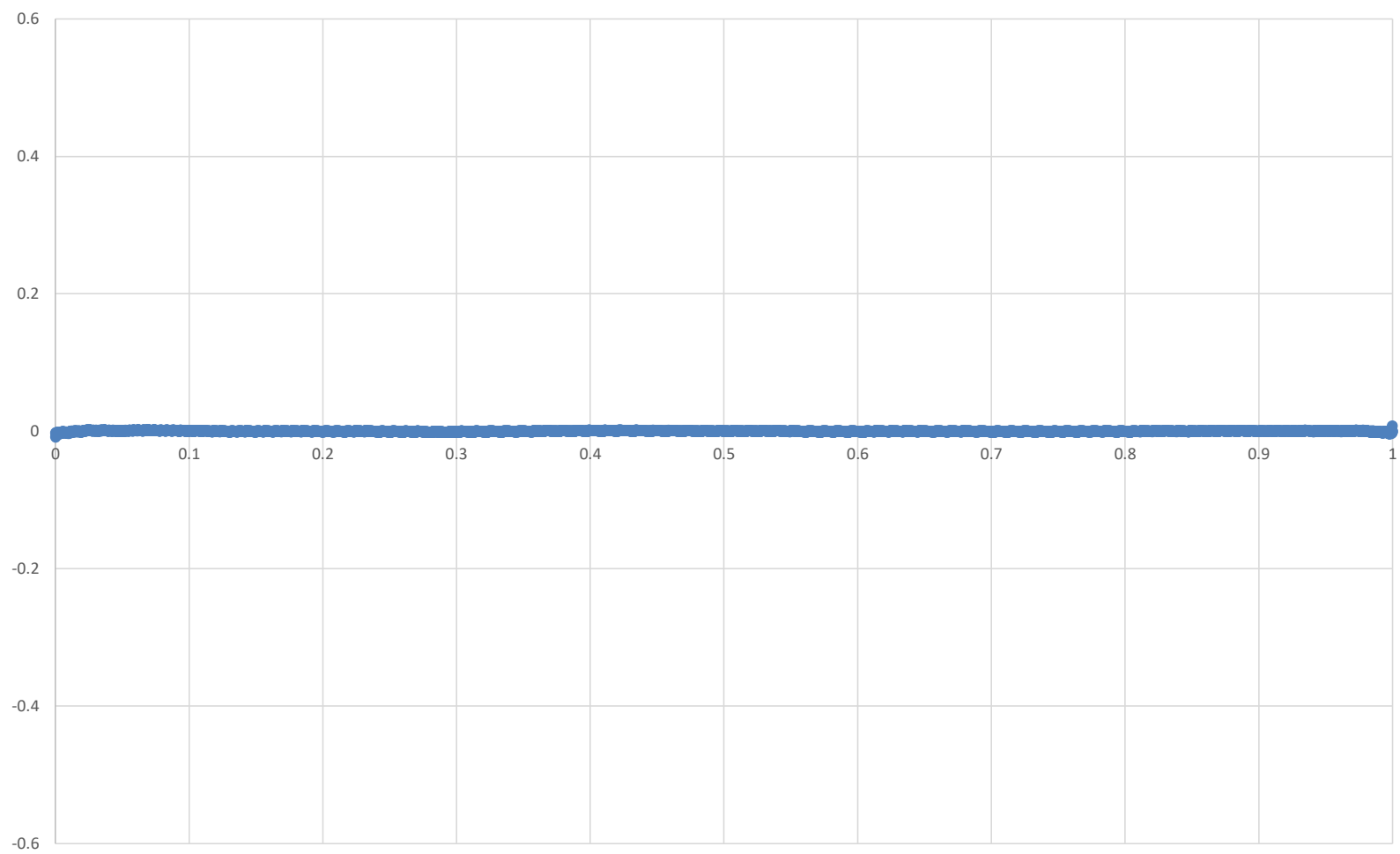








Relativní chyby teor. křivky, červenec



Závěr

Rozdělení mLN5 je velmi plastická transformace normálního rozdělení – umožňuje modelovat rozdělení v rozsahu od normálního rozdělení až po velmi asymetrická rozdělení.

Vzhledem k případné periodičnosti klimatických a hydrologických procesů je vhodné, aby vstupní data měla rozdělení symetrické, nejlépe normální. Tento požadavek splňuje právě zpětná transformace mLN5 rozdělení.

Při analýze reálných i simulovaných dat se ukazuje, že z uvedených metod odhadu parametrů rozdělení LN5 resp. mLN5 nejlepší výsledky poskytuje trojúhelníková metoda, a to včetně extrapolací.

V současnosti se LN5 a mLN5 i nově vyvinutá trojúhelníková metoda používá v ČHMÚ v posudkové činnosti. V Ústavu pro výzkum klimatické změny (CzechGlobe) se připravují analýzy klimatických dat v širokém rozsahu stanic z celé planety.