# ODHAD PARAMETROV VŠEOBECNÉHO PARETOVHO ROZDELENIA SOFTVÉROM EVA V PROSTREDÍ JAZYKA R.

**Abstrakt**

V prípade výskytu extrémnych hodnôt v databáze údajov je možné na ich popísanie zvoliť model prekročenia prahu využitím všeobecného Paretovho rozdelenia (GPD). Príspevok sa zaoberá odhadom parametrov tohto rozdelenia využitím open source systému R prostredníctvom softvéru Extreme value analysis (EVA). Využitím balíka in2extRemes možno po načítaní údajov do softvéru realizovať analýzu vhodnej voľby prahu, odhad spomínaných parametrov. Interpretácia dosiahnutých výsledkov je podporená aj grafickou prezentáciou výsledkov použitých metód. Predikciu rozdelenia extrémnych škôd môže poisťovňa využiť pre realizáciu riadenia rizika v súvislosti so zabezpečením jej solventnosti.

**Kľúčové slová**

Extrémne hodnoty, model prekročenia prahu, všeobecné Pareto rozdelenie, open-source system R, package in2extRemes, Extreme value analysis (EVA), mean residual life plot.

## 1 MODEL PREKROČENIA PRAHU (THRESHOLD MODEL)

V rámci analýzy extrémnych hodnôt sa budeme zaoberať modelom prekročenia prahu, ktorý budeme aplikovať na databázu údajov, ktoré možno považovať pre jednoduchosť za hodnoty náhodnej premennej X. Samozrejme, že je prirodzené považovať za extrémne udalosti tie hodnoty $X$ , ktoré prekračujú určitú hodnotu tzv. ***prah prekročenia*** $u$ (***threshold***).

### Všeobecné Paretovo rozdelenie

Náhodné premenné $X-u/X>u$, resp. $X/X>u$ je možné popísať všeobecným Paretovým rozdelením. Stanovenie vhodnej hodnoty prahu $u$ si vyžaduje samostatnú analýzu, v ktorej sa verifikácia opiera viac menej o grafickú prezentáciu. Pre dostatočne veľké $u$ je distribučná funkcia náhodnej premennej $Y=X-u/X>u$ približne rovná

$$H(y)=1-\left(1+\xi\frac{y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}, \quad \text{pre} \quad y>0 \text{ a } 1+\xi\frac{y}{\tilde{\sigma}}>0 \tag{1}$$

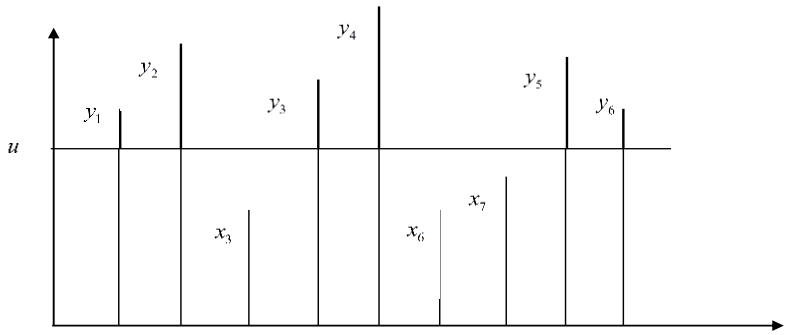
Náhodná premenná $Y=X-u/X>u$ sa riadi ***všeobecným Paretovým rozdelením*** (GPD – generalized Pareto distribution). Ak $\xi<0$, tak podmienené rozdelenie $Y$ má hornú hranicu $u-\frac{\tilde{\sigma}}{\xi}$. Ak $\xi\geq 0$, podmienené rozdelenie $Y$ nemá hornú hranicu. Pre strednú hodnotu náhodnej premennej $Y$ pre $\xi<1$ platí

$$E(Y)=\frac{\tilde{\sigma}}{1-\xi} \tag{2}$$

Keďže poznáme distribučnú funkciu náhodnej premennej $Y=X-u/X>u$ vieme určiť aj distribučnú funkciu náhodnej premennej $X/X>u$. Pretože $X=Y+u$, je distribučná funkcia náhodnej veličiny $X/X>u$ rovná

$$H(x) = 1 - \left(1 + \xi \frac{x-u}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}} \quad x > u \tag{3}$$

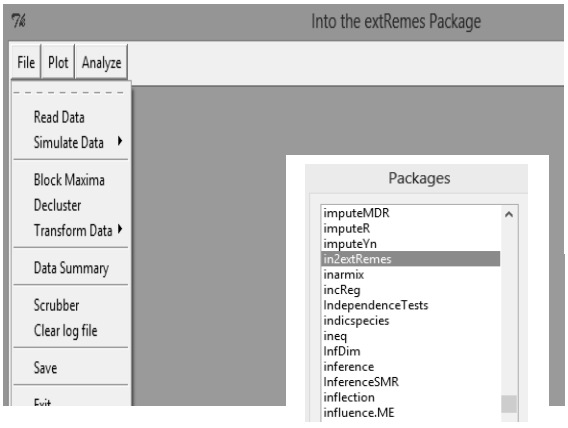
Náhodná premenná $X / X > u$ sa riadi všeobecným Paretovým rozdelením s parametrami $u, \tilde{\sigma}, \xi$, čo môžeme zapísať $GP(u, \tilde{\sigma}, \xi)$. Ak dáta $x_1, x_2, x_3, ..., x_n$, ktoré analyzujeme predstavujú hodnoty náhodnej premennej $X$, potom hodnoty náhodnej premennej $X / X > u$ zapíšeme následovne $\{x_i : x_i > u, \quad i = 1, ..., n\}$, resp. $x_{(1)}, ..., x_{(k)}$. V súvislosti s náhodnou premennou $Y = X - u / X > u$ jej hodnoty označíme $y_1, ..., y_j$, kde $y_j = x_{(j)} - u, \quad j = 1, ..., k$. Opísanú situáciu môžeme graficky zobraziť na obr. 1. [2]

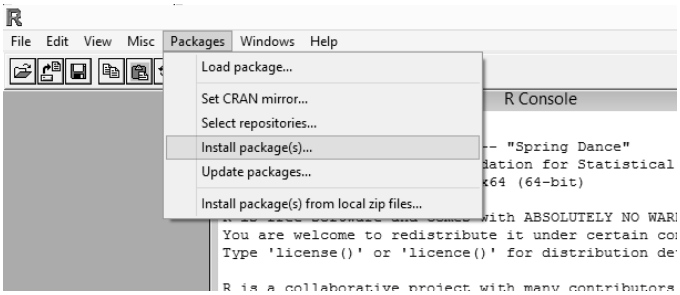



Obr. 1 Model prekročenia prahu.

Na analýzu dát (overenie určenia hodnoty prahu $u$, odhad parametrov všeobecného Paretovho rozdelenia) využijeme niektoré balíky (***packages***) a príkazy systému R a špeciálne softvér ***EVA (Extreme value analysis)***. Softvér je určený pre analýzu extrémnych hodnôt, pričom pre jeho použitie je nutná inštalácia balíka ***in2extRemes***. [6]

```
> in2extRemes()
```




Obr. 2 Inštalácia package in2extRemes a ukážka prostredia softvéru EVA (Extreme value analysis)

**Voľba úrovne prahu $u$ pomocou grafickej analýzy.**

**a)** ***Výskumná metóda - mean residual life plot*** (***graf priebehu priemerných reziduálov***)
(nevyžaduje odhad parametrov)

Uvažujeme o modeli prekročenia prahu $u$ s nameranými hodnotami $x_1, ..., x_n$, ktoré považujeme za hodnoty náhodnej premennej $X$. Zvoľme si úroveň $u_0$ a uvažujeme o

náhodnej premennej ${}_{u_0}Y = X - u_0 / X > u_0$, pre ktorej strednú hodnotu $E\left({}_{u_0}Y\right)$ podľa vzťahu (3) pre $\xi < 1$ [1]

$$E\left({}_{u_0}Y\right) = \frac{\tilde{\sigma}_{u_0}}{1-\xi} \tag{4}$$

Pre úroveň $u > u_0$ potom pre $E\left({}_{u}Y\right)$ platí

$$E\left({}_{u}Y\right) = \frac{\tilde{\sigma}_{u_0}}{1-\xi} + \frac{\xi}{1-\xi}u = \alpha + \beta u\,, \qquad \text{kde } \alpha,\ \beta \text{ sú konštanty.} \tag{5}$$

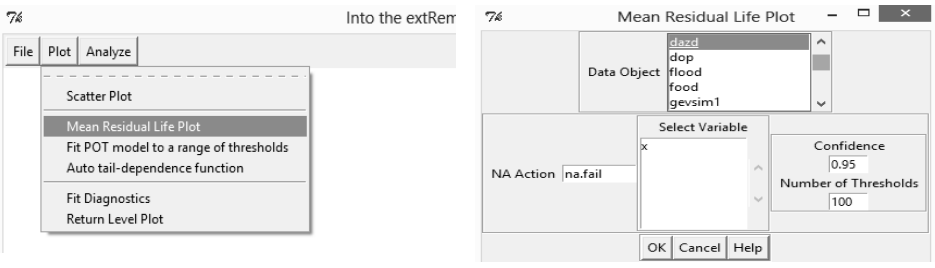
Pre ľubovoľné $u > u_0$ je $E\left({}_{u}Y\right)$ lineárnou funkciou, ktorá závisí od prahu $u$. Preto pre $u \in \langle u_0, x_{\max} \rangle$ ležia body $\left(u,\ E\left({}_{u}Y\right)\right)$ na priamke, za predpokladu, že $u_0$ je „dostatočne veľké“ . Keďže pre odhad $E\left({}_{u}Y\right)$ súvisiaci s rozsahom súboru $n_u$ hodnôt náhodnej premennej ${}_{u}Y$ platí

$${}_{u}\overline{Y} = \frac{1}{n}\sum_{i=1}^{n_u} {}_{u}Y\,, \tag{6}$$

na identifikáciu toho, či už sme určili dostatočne vysokú úroveň $u_0$ použijeme ***mean residual life plot*** (graf priebehu priemerných reziduálov), ktorý pozostáva z bodov

$$\left\{\left(u,\ \frac{1}{n_u}\sum_{i=1}^{n_u} {}_{u}y_i\right) : u < x_{\max}\right\}, \tag{7}$$

kde ${}_{u}y_i$ je hodnota náhodnej premennej ${}_{u}Y$. $E\left({}_{u_0}Y\right) = E\left(X - u_0 / X > u_0\right)$ nazývame aj ***mean excess***. Tie hodnoty $u$, pre ktoré body grafu ležia „ približne “ na priamke sú vhodné pre určenie prahu. Chceme určiť taký prah $u_0$, ktorý je čo najmenší, ale pre príslušné ${}_{u_0}Y$ už môžeme predpokladať, že majú zovšeobecnené Paretovo rozdelenie. Ak totiž zvolíme príliš vysoký prah $u$, potom máme málo hodnôt náhodného výberu ${}_{u}Y_i$ a príslušné odhady pomocou tohto náhodného výberu majú veľký rozptyl. Vyššie uvedená voľba úrovne je ***výskumná*** a realizuje sa pred určením odhadu parametrov. Na zostrojenie mean residual life plot využijeme spomínaný softvér EVA, ktorý graficky zobrazí aj zvolené ***konfidenčné intervaly (confidence imtervals)***, obr. 8.




Obr. 3 Výber databázy pre realizáciu Mean residual life plot v prostredí softvéru EVA.

**b) Voľba prahu $u$ podľa stability parametrov.**

Pomocou mean residual life plot sa dá niekedy ťažko interpretovať vhodná voľba prahu. Iný spôsob grafického určenia vhodnej voľby prahu, ktorý je založený na fitovaní GPD, resp.

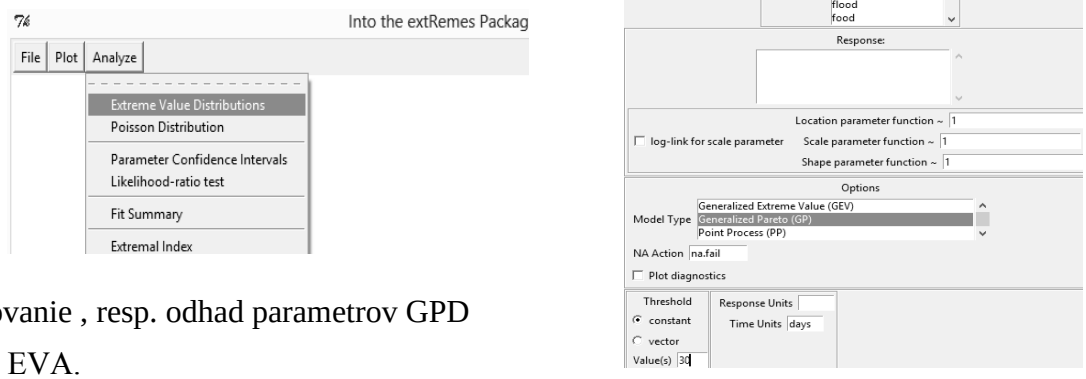
odhade jeho parametrov $\hat{\xi}$ a $\hat{\tilde{\sigma}}_u$ pre rôzne hodnoty prahu $u$ je overovanie ich stability. Ak $u_0$ je „vhodne určená“ hodnota prahu, tak pre $u > u_0$ platí

$$\tilde{\sigma}_u = \tilde{\sigma}_{u_0} + \xi\left(u - u_0\right) = \gamma + \xi u\,, \tag{8}$$

kde $\gamma, \xi$ sú konštanty (pre všetky „vhodné“ $u > u_0$). Ak $u_0$ je vhodný prah, tak pre $u > u_0$ sú parametre $\xi$ aj $\gamma = \tilde{\sigma}_u - \xi \cdot u$ konštantné. [1]

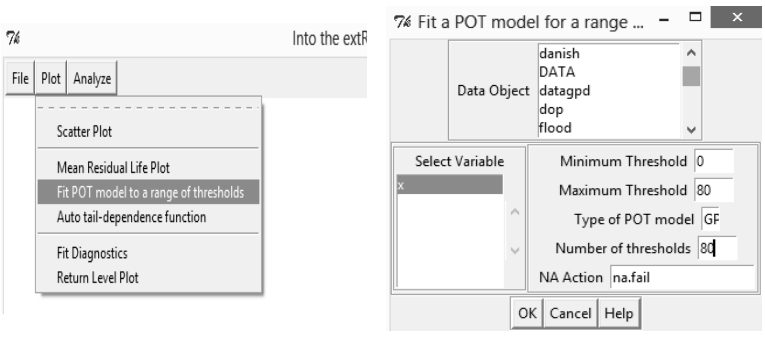
Hodnoty odhadov parametrov $\hat{\xi}$ a $\hat{\tilde{\sigma}}_u$ pre zvolenú úroveň prahu $u$ dostaneme metódou ***maximálnej vierohodnosti*** (***maximum likehood estimates***) využitím softvéru EVA v prostredí systému R na karte ***Analyze*** (Extreme Value distributions), resp. v okne ***Fit EV Model***.

Odhad parametrov je možné realizovať pomocou balíka ***extRemes*** a funkcie ***fevd*** pre fitovanie všeobecného Paretovho rozdelenia (type=“GP“) a hodnotu prahu (treshold=30), ktorú zhodou okolností využíva aj softvér EVA .




Obr. 4 Fitovanie , resp. odhad parametrov GPD v prostredí EVA.

Pre spomínané grafické overenie stability parametrov $\gamma, \xi$ je nutné na karte ***Plot*** zvoliť ***Fit POT models to range of thresholds*** , obr. 4 s výstupom v okne ***R Graphics***, obr.9.[6]

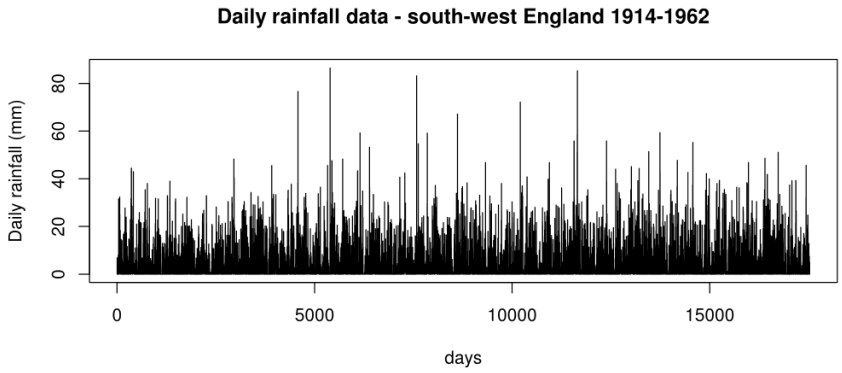



Obr. 5 Zadávanie vstupných údajov pre overenie stability parametrov $\gamma, \xi$ v prostredí EVA.

## 2 PRAKTICKÁ REALIZÁCIA

Metodológiu spracovania extrémnych hodnôt využitím fitovania všeobecným Paretovým rozdelením budeme prezentovať na údajoch týkajúcich sa denných zrážok v juhozápadnom Anglicku v priebehu rokov 1914-1962. Tieto pozorovania sme získali zo systému R v balíku

***"ismev",*** v ktorom sú tieto dáta označené ako "rain" s popisom "Daily Rainfall Accumulations in South-West England".



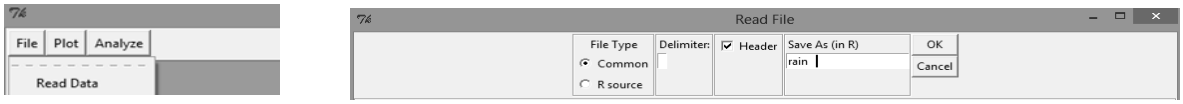

Obr. 6 Denných zrážok v juhozápadnom Anglicku v priebehu rokov 1914-1962.

Časový vývoj spomínaných zrážok počas 17531 dní zobrazíme v systéme R v rámci jeho grafického rozhrania (R Graphics) pomocou príkazu ***Plot***, ktorý zadáme v okne R Console v nasledovnom syntaxe [3],[4]

```
plot(rain,main="Daily rainfall data - south-west England 1914-
1962",xlab="days",ylab="Daily rainfall (mm)",type="l")
```

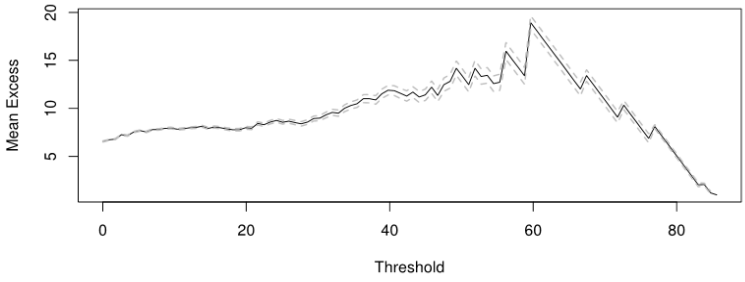
Na analýzu dát "rain" využijeme softvér *EVA* (Extreme value analysis)v grafickom prostredí systému jazyka R, ktorého riešiteľský aparát sme si predstavili v predchádzajúcej kapitole. V tejto časti príspevku si ešte predstavíme načítanie databázy do tohto softvéru a potom zobrazíme výstupy jednotlivých krokov uskutočnenej analýzy.

Na úvod je údaje potrebné načítať a uložiť (Save As (in R)). V našom prípade z lokálneho disku (D), kde sme pre potreby načítania vytvorili z dát "rain" textový súbor rain.txt.[5]




Obr. 7 Načítanie databázy pre sofvér EVA.

Určíme hodnotu prekročenia prahu a odhadneme parametre všeobecného Paretovho rozdelenia v softvéri EVA. Grafickou analýzou pomocou mean residual life plot s 95% konfidenčnými intervalmi a stability odhadnutých parametrov overíme vhodnosť stanovenej hodnoty prahu $u = 30$.




Obr. 8 Mean residual life plot s 95% konfidenčnými intervalmi.

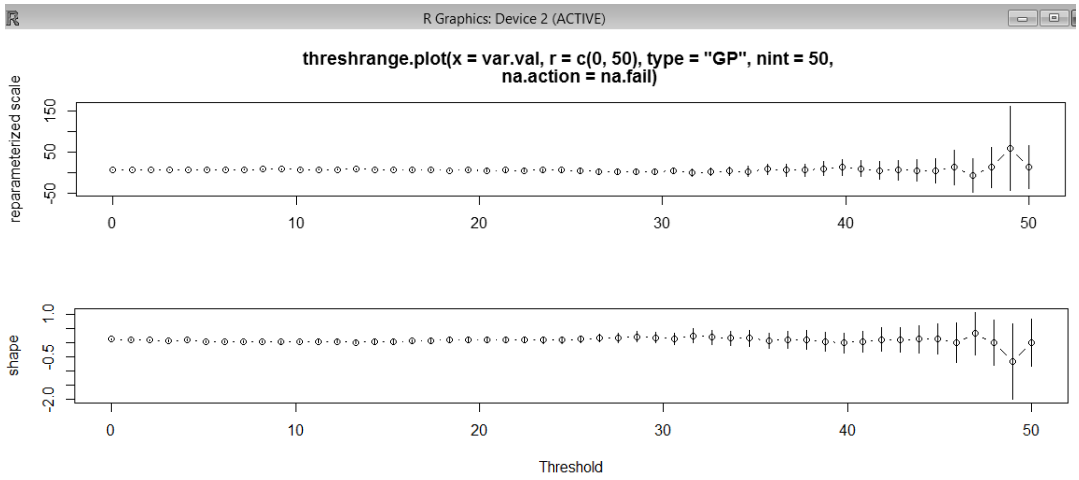
Ako už bolo spomenuté táto metóda nie je pri praktickej realizácii jednoducho interpretovateľná. Vzhľadom na dokázanú lineárnu závislosť a fakt, že nie sú potrebné

odhadnuté hodnoty parametrov musíme brať do úvahy kladné aj záporné hodnoty koeficientu $\beta$ . Vzhľadom na teóriu popísanú v predchádzajúcej kapitole a fakt, že prah musí byť dostatočne veľký si z dvoch alternatív $u = 30$, resp. $u = 60$ vyberáme ako vhodne zvolenú hodnotu prahu (threshold) $u = 30$. V prípade $u = 60$ by bol totiž súbor pozorovaní veľmi malý. Metóda overovania stability parametrov je logicky založená na ich odhade, ktorý sme realizovali meódou MLE (maximum likehood estimates). Výstup softvéru EVA v prostredí Workspace v systéme R

```
[1] "Estimation Method used: MLE"
Estimated parameters:          scale    shape

                           7.440270 0.184499
```

To znamená, že v prípade hodnoty $u = 30$ sú odhadnuté hodnoty GPD $\hat{\sigma}(scale) \approx 7,44$ a $\hat{\xi}(shape) \approx 0,184$. Na základe teórie popísanej v predchádzajúcej kapitole overíme graficky vhodnosť zvoleného prahu tak, že zobrazíme závislosť reparametrizovaného scale t.j. $\gamma$ a parametra shape $\xi$ od hodnoty prahu (threshold) $u$ .




Obr. 9 Výstup softvéru EVA v prostredí R Graphics - grafické overenie stability parametrov .

Hodnotu prahu $u = 30$ možno označiť za vhodne zvolenú vzhľadom na vizuálne potvrdenie stability reparametrizovaného parametra $\gamma$ a parametra $\xi$ .

## Záver

R je jazyk a zároveň prostredie vhodné pre realizáciu štatistických výpočtov a tvorbu grafických výstupov. Relatívne ľahko umožňuje pridávať veľké množstvo balíčkov, ktoré zjednodušujú prácu užívateľa. V prípade potreby vytvorenia užívateľského grafického rozhrania (GUI- graphical user interface) je nutné využiť nadstavbu R Commander. V tejto súvislosti sa vytvára priestor pre využitie tohto prostredia, resp. jazyka R v rámci riešenia rôznych úloh z oblasti teórie rizika v neživotnom poistení.

## Použitá literatúra


1 Coles, S. (2007), An introduction to Statistical Modeling of Extreme Values, London Springer.

2 Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modeling extremal events for insurance and finance*, Berlin, Springer.

3 McCown, Frank. (2006) *Producing Simple Graphs with R*, online (2014) na: <http://www.harding.edu/fmccown/r>

4 W. N. Venables, D. M. Smith, and the R Development Core Team (1990-2013). *An Introduction to R*, online (2014) na < http://cran.r-project.org/manuals.html>

5 R Development Core Team (2000-2013). *R Data Import/Export*, online (2014) na < http://cran.r-project.org/doc/manuals/r-release/R-data.pdf >

6 Gilleland, E. and Katz, R. W. (2011) *New software to analyze how extremes change over time*. Eos. , <http://www.ral.ucar.edu/staff/ericg/extRemes/>

**Kontaktné údaje**

Mgr. Vladimír Mucha, PhD., Katedra matematiky a aktuárstva, Fakulta hospodárskej informatiky, Ekonomická univerzita v Bratislave, Dolnozemská cesta 1, 852 35 Bratislava, tel. +421 2/672 95 810, e-mail: vladimir.mucha@euba.sk