Összefüggőségek vizsgálata extrémérték modellekben

Szakdolgozat

Készítette:

Gáspár Sándor

alkalmazott matematikus szak

Témavezető:

Dr. Zempléni András egyetemi docens

Valószínűségelméleti és Statisztika Tanszék

EÖTVÖS LORÁND TUDOMÁNYEGYETEM

TERMÉSZETTUDOMÁNYI KAR

2002

Tartalomjegyzék

Bevezetés

A dolgozat stacionárius sorozatok extrémérték elméletével foglalkozik. Az extrémérték elmélet bizonyos megfigyelések kiemelkedően nagy vagy éppen kicsi értékeinek modellezésével foglalkozik. A mindennapokban gyakran fordulnak elő ilyen problémák többek között biztosítási, meteorológiai, hidrológiai alkalmazásoknál. Gyakran van szükség olyan extrémumok modellezésére, mint például kiemelkedő biztosítási kárösszegek, nagy csapadékmennyiség, magas vízszint. Egy folyó esetén gazdasági szempontból is lényegi kérdés, hogy mi az a magas vízszint, amit átlagosan T évente egyszer halad meg, így milyen magasra méretezzék a védőgátakat.

Az extrémérték modellezés klasszikus fejezeteiben az a feltevésünk, hogy független, azonos eloszlásból származó megfigyelések állnak rendelkezésünkre. A gyakorlatban viszont azok a folyamatok, amelyekre extrémérték modelleket alkalmazunk gyakran nem tekinthetők független megfigyelés-sorozatnak, például mert erős időbeni összefüggőség figyelhető meg. Ez az extrémumok viselkedése szempontjából azt jelentheti, hogy az extrém értékek csoportokban, egymás után figyelhetőek meg, egy extrém megfigyelés után más kiugró értéket is tapasztalunk. Ez a rövid távú, lokális összefüggőség jelensége. Alkalmazásokban a hosszú távú összefüggőség hatását az extrémumokra gyakran elhanyagolhatónak tekintik.

Példánknál maradva, egy folyó esetében az egymást követő napi vízszintértékek nem egymástól függetlenül alakulnak és bizonyos magas vízszintértékeket ugyancsak magas értékek követnek rövid időn belül. Így a mintánkban nagy lesz a lokális összefüggőség. Ugyanakkor például egy három hónap múlva történő extrém megfigyelés feltehetően már nem hozható összefüggésbe a korábbi kiemelkedő értékkel.

A dolgozat első fejezetében tömören összefoglaljuk a klasszikus extrémérték elmélet fő eredményeit. Látni fogjuk, hogy karakterizációs tételek segítségével egyértelműen meghatározhatjuk független, azonos eloszlású valószínűségi változók maximumának határeloszlását.

A második fejezetben a klasszikus eredményeket általánosítjuk stacionárius sorozatokra. Belátjuk, hogy bizonyos szigorú feltételek fenntartásával a stacionárius sorozatok extrémumai ugyanolyan módon viselkednek, mint a független esetben.

A következő részben a stacionárius sorozatok extrémérték elméletének kulcsfontosságú paraméterét az extremális indexet vezetjük be. Segítségével – valamivel enyhébb feltételek teljesülése esetén – a klasszikus extrémérték modellezés eredményei alkalmazhatók összefüggő sorozatokra.

Lényeges kérdés tehát az extremális index statisztikai becslése. Látjuk majd, hogy a becslések meghatározása újabb problémákat vet fel. Két alapvető módszer létezik, de ezeken kívül ismertetünk más becslési eljárásokat és megvizsgáljuk a klasszikus becslések néhány tulajdonságát is. Az extrémumok csoportképző tulajdonsága ezekben a fejezetekben jut szerephez.

A negyedik fejezetben gyakorlati alkalmazásokra térünk át, az extremális index használatát mutatjuk be extrémérték eloszlás illesztésére. Új módszert ismertetünk extrémérték eloszlás kvantiliseinek meghatározására.

Végül a leírt módszerek alkalmazását konkrét adatsor elemzésével prezentáljuk.

1. fejezet

Klasszikus extrémérték modellezés

Ebben a fejezetben összefoglaljuk a klasszikus extrémérték elmélet azon főbb tételeit, amelyek megfelelőit azután stacionárius sorozatok extrémumaira alkalmazzuk. A bizonyításoknál segítségül hívjuk majd ezen eredményeket is, mindamellett figyelemmel kísérjük a kétféle elmélet hasonlóságait ill. különbségeit is. A klasszikus extrémérték elméletről bővebben [1]-ben olvashatunk.

A fejezetben $X_1, X_2, \ldots, X_n$ független, azonos F eloszlásfüggvénnyel rendelkező valószínűségi változók sorozatát jelöli. A minta maximumaira az $M_n = \max(X_1, \ldots, X_n)$ jelölést alkalmazzuk. Különböző normalizáló konstansok segítségével a

$$P(c_n^{-1}(M_n - d_n) \leq x) \tag{1.1}$$

valószínűségek határértékét $n \to \infty$ esetén és a $c_n > 0$, $d_n \in \mathbb{R}$ sorozatok létezését szeretnénk vizsgálni. ((1.1)-et szokás $P(M_n \leq u_n)$ alakba is írni, ahol $u_n = u_n(x) = c_n x + d_n$.)

A következőkben megvizsgáljuk, hogy az F eloszlásfüggvényre adott mely feltételek biztosítják az $P(M_n \leq u_n)$, $n \to \infty$ határérték létezését, megfelelő u_n normálás esetén?

1. Tétel. (Poisson-approximáció)
Adott $\tau \in [0, \infty]$ és (u_n) valós számsorozat esetén a következők ekvivalensek:

$$n\overline{F}(u_n) \rightarrow \tau \tag{1.2}$$

$$P(M_n \leq u_n) \rightarrow e^{-\tau}. \tag{1.3}$$

□

Megjegyzés. Az elnevezést a következő indokolja. Az $N_n = \sum_{i=1}^{n} I_{\{X_i > u_n\}}$ mennyiség binomiális eloszlású $(n, \overline{F}(u_n))$ paraméterekkel. Poisson határeloszlás tétele alapján: $N_n \xrightarrow{d} \mathrm{Poi}(\tau)$ akkor és csak akkor, ha $EN_n = n\overline{F}(u_n) \rightarrow \tau$, amely éppen (1.2). Továbbá $P(M_n \leq u_n) = P(N_n = 0) \rightarrow \exp\{-\tau\}$ adja (1.3)-t.

A következőkben azt vizsgáljuk, hogy milyen eloszlások léphetnek fel a maximumok határeloszlásaként. Ehhez először is bevezetjük a *max-stabilis eloszlás* fogalmát.

1. Definíció. (Max-stabilis eloszlás, Max-stable)
Azt mondjuk, hogy az X valószínűségi változó (ill. a hozzá tarozó F eloszlás) max-stabilis eloszlású*, ha léteznek olyan $c_n > 0$ és $d_n \in \mathbb{R}$ konstansok, hogy $M_n \stackrel{d}{=} c_n X + d_n$, $n \geq 2$ esetén, ahol $M_n = \max(X_1, \ldots, X_n)$ és $X, X_1, \ldots, X_n$ független azonos eloszlásúak.*

2. Tétel. *A max-stabilis eloszlások családja és a független azonos eloszlású változók normalizált maximumainak lehetséges határeloszlásaként fellépő eloszlások családja egybeesik.*

Bizonyítás. (vázlat)
A bizonyítás azon múlik, hogy léteznek olyan c_n, $c_{nk} > 0$ és d_n, $d_{nk} \in \mathbb{R}$ konstansok, amelyekkel

$$\begin{aligned} \left(\lim_{n\to\infty} F^n(c_n x + d_n)\right)^k &= H^k(x), \ x \in \mathbb{R} \\ \lim_{n\to\infty} F^{nk}(c_{nk} x + d_{nk}) &= H(x), \ x \in \mathbb{R}. \end{aligned}$$

A fenti normalizáló konstansokra vonatkozó tételekből következik, hogy ekkor F max stabilis. □

A következő tétel a klasszikus extrémérték elmélet alaptétele.

3. Tétel. (Fisher-Tippett tétel, maximumok határeloszlása)
Legyen (X_n) független, azonos eloszlású valószínűségi változók sorozata. Ha léteznek $c_n > 0$ és $d_n \in \mathbb{R}$ normalizáló konstansok valamint egy nem-elfajuló H eloszlásfüggvény, amelyre

$$c_n^{-1}(M_n - d_n) \xrightarrow{d} H,$$

akkor H a következő eloszlástípusok egyikébe tartozik:

1. Frèchet eloszlás: $\phi_\alpha(x) = \begin{cases} 0 & , x \leq 0 \\ \exp\{-x^{-\alpha}\} & , x > 0 \end{cases}$

2. Weibull eloszlás: $\psi_\alpha(x) = \begin{cases} \exp\{-(-x)^\alpha\} & , x < 0 \\ 1 & , x \geq 0 \end{cases}$

3. Gumbel eloszlás: $\lambda(x) = \exp\{-e^{-x}\}, \quad -\infty < x < \infty.$

□

A fenti eloszlásokat *standard extrémérték eloszlás*oknak nevezzük. E három eloszlást közösen az úgynevezett *általánosított extrémérték eloszlások* családjával reprezentálhatjuk:

$$H(x) = \begin{cases} \exp\{-(1+\xi x)^{-1/\xi}\} & ,\text{ha } \xi \neq 0 \\ \exp\{-\exp\{-x\}\} & ,\text{ha } \xi = 0 \end{cases} \tag{1.4}$$

ahol $1 + \xi x > 0$.

Bevezetve a μ eltolás és σ skála paramétereket, az általánosított extrémérték eloszlásokat a következő alakban írhatjuk le:

$$H(x) = \exp\left[-\left\{1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right\}_+^{-1/\xi}\right], \tag{1.5}$$

ahol $\{x\}_+ = max(0, x)$, és ξ-t alak paraméternek nevezzük. A továbbiakban ezen eloszláscsaládra GEV(μ, σ, ξ) jelöléssel hivatkozunk.

Ezek után felmerülhet az a kérdés, hogy adott H extrémérték eloszlás mellett mely F eloszlás esetén tart az M_n maximum H-hoz?

2. Definíció. (Max-vonzástartomány, Maximum domain of attraction)
Azt mondjuk, hogy az X valószínűségi változó (ill. a hozzá tartozó eloszlás vagy eloszlásfüggvény) a H extrémérték eloszlás max-vonzástartományába tartozik, ha léteznek olyan $c_n > 0$, $d_n \in \mathbb{R}$ konstansok, hogy $c_n^{-1}(M_n - d_n) \xrightarrow{d} H$.
Jel.: $X \in MDA(H)$ $(F \in MDA(H))$

4. Tétel. (MDA(H) karakterizációja)
Az F eloszlás a H extrémérték eloszlás max-vonzástartományába tartozik a $c_n > 0,\ d_n \in \mathbb{R}$ normalizáló konstansokkal akkor és csak akkor, ha

$$\lim_{n\to\infty} n\overline{F}(c_n x + d_n) = -\ln H(x), \quad x \in \mathbb{R}. \tag{1.6}$$

$H(x) = 0$ esetben a határérték ∞.

□

2. fejezet

Stacionárius sorozatok extrémumai

A gyakorlatban azok a folyamatok, amelyek extrémumai érdekesek számunkra, általában nem tekinthetők független megfigyeléseknek. A független, azonos eloszlású sorozat természetes általánosítása az *(erősen) stacionárius sorozat.* A stacionárius sorozatok összefüggőségi szerkezetét általánosságban leírni nehéz, csak bizonyos feltételek bevezetésével lehet.

3. Definíció. (Erősen stacinárius sorozat)
Azt mondjuk, hogy az $X_1, X_2, \ldots, X_n, \ldots$ valószínűségi változók sorozata (erősen) stacionárius*, ha minden véges dimenziós eloszlása eltolásinvariáns, azaz*

$$(X_{t_1}, \ldots, X_{t_m}) \stackrel{d}{=} (X_{t_1+h}, \ldots, X_{t_m+h}),$$

minden $t_1 < \ldots < t_m$ indexre és h egészre.

Azt mondjuk, hogy $(\tilde{X}_n)$ az (X_n)-hez *tartozó független sorozat*, ha ugyanolyan marginális eloszlással rendelkezik, mint (X_n) és független.
A továbbiakban $X_1, \ldots, X_n$ jelöljön egy stacionárius sorozatból vett megfigyeléssorozatot $F(x) := P(X_1 \le x)$ eloszlásfüggvénnyel. $M_n := \max(X_1, \ldots, X_n)$ legyen a minta maximuma $F_n(x) := P(M_n \le x)$ eloszlásfüggvénnyel.

$(\tilde{X}_n)$ legyen a hozzátartozó független sorozat és $\tilde{M}_n := \max(\tilde{X}_1, \ldots, \tilde{X}_n)$ az első n változó maximuma.

Reálisnak tűnő feltétel az extrémumok szempontjából az, hogy rövid távon egymás után több kiugró értéket is megfigyelhetünk. Ezzel szemben egy későbbi extrém megfigyelés bekövetkezésének valószínűségét már nem befolyásolja a korábbi esemény. Az egymástól távol lévő extrém megfigyelések függetlenségét jelenti a következő feltétel.

Feltétel. ($D(u_n)$ feltétel)
Bármely p, q és n egészek esetén, amelyekre

$$1 \leq i_1 < \ldots < i_p < j_1 < \ldots < j_q \leq n,$$

valamint $j_1 - i_p \geq l$ teljesül, fennáll a következő:

$$\left| P\left(\max_{i \in A_1 \cup A_2} X_i \leq u_n \right) - P\left(\max_{i \in A_1} X_i \leq u_n \right) P\left(\max_{i \in A_2} X_i \leq u_n \right) \right| \leq \alpha_{n,l},$$

ahol $A_1 = \{i_1, \ldots, i_p\}, A_2 = \{j_1, \ldots, j_q\}$ és $\alpha_{n,l} \to 0$, midőn $n \to \infty$ és $l = l_n = o(n)$.

Ezen és hasonló feltételek vizsgálatával bővebben Leadbetter, Lindgren és Rootzén [3] foglalkozik. A $D(u_n)$ feltételből következik például, hogy

$$P(M_n \leq u_n) = P^k(M_{\lfloor n/k \rfloor} \leq u_n) + o(1) \tag{2.1}$$

konstans k-ra, vagy $k_n = o(n)$ esetén. ($\lfloor x \rfloor$ jelöli x egész részét.)

A következő tétel a stacionárius sorozatok maximumainak lehetséges határeloszlásáról szól.

5. Tétel. *Legyen $X_1, \ldots, X_n$ erősen stacionárius sorozat M_n maximummal. Tegyük fel, hogy léteznek olyan $c_n > 0$ és $d_n \in \mathbb{R}$ konstansok, hogy $c_n^{-1}(M_n - d_n) \xrightarrow{d} G$, valamely G eloszlásra. Ha az $(u_n) = (c_n x + d_n)$ sorozat kielégíti a $D(u_n)$ feltételt minden valós x esetén, akkor G extrém érték eloszlás.*

Bizonyítás. Az 1. Definíció valamint az 2. és 3. Tétel alapján tudjuk, hogy G extrémérték eloszlás akkor és csak akkor, ha max-stabilis. (2.1)-ből következik, hogy

$$P(M_{nk} \leq c_n x + d_n) = P^k(M_n \leq c_n x + d_n) + o(1) \to G^k(x)$$

bármely $k \geq 1$ egész esetén, G-nek minden x folytonossági pontjában. Másrészt:

$$P(M_{nk} \leq c_{nk} x + d_{nk}) \to G(x).$$

Eloszlástípusokra vonatkozó tételekből (lásd [1]) következik, hogy léteznek olyan $\tilde{c}_k > 0$ és $\tilde{d}_k \in \mathbb{R}$ konstansok, hogy

$$\lim_{n\to\infty} \frac{c_{nk}}{c_n} = \tilde{c}_k \quad \text{és} \quad \lim_{n\to\infty} \frac{d_{nk} - d_n}{c_n} = \tilde{d}_k,$$

és a független azonos G eloszlású $Y_1, \dots, Y_k$ változókra

$$\max(Y_1, \dots, Y_k) \stackrel{d}{=} \tilde{c}_k Y_1 + \tilde{d}_k$$

$\square$

Megjegyzés. Az előző tételből *nem* következik, hogy ha $c_n^{-1}(M_n - d_n) \xrightarrow{d} G$ és $c_n^{-1}(\tilde{M}_n - d_n) \xrightarrow{d} H$, akkor $G = H$. Amint később látni fogjuk G gyakran H^θ alakú, ahol $\theta \in (0, 1]$.

Látható, hogy ha létezik határeloszlás, akkor az csakis max-stabilitás eloszlás lehet. Ezután a $P(M_n \leq u_n)$ valószínűségek határértékének létezéséhez elégséges feltételt szeretnénk, olyan (u_n) küszöbérték sorozatok esetén, amelyekre fennáll:

$$n\overline{F}(u_n) \to \tau \quad , \tau \in [0, \infty).$$

Ehhez bevezetjük a következő feltételt.

Feltétel. ($D'(u_n)$ feltétel)

$$\lim_{k\to\infty} \limsup_{n\to\infty} n \sum_{j=2}^{\lfloor n/k \rfloor} P(X_1 > u_n, X_j > u_n) = 0$$

Ez a feltétel a stacionárius sorozatok azon tulajdonságát jellemzi, hogy az extrémumok nem egymás után, csoportokba képződve jelentkeznek. Ezek után kimondhatjuk az 1. Tétel analogonját:

6. Tétel. *Tegyük fel, hogy az (X_n) stacionárius sorozat és az (u_n) küszöbérték sorozat kielégíti a $D(u_n)$ és $D'(u_n)$ feltételeket, továbbá legyen $\tau \in [0, \infty)$. Ekkor a következők ekvivalensek:*

$$\begin{aligned} \lim_{n\to\infty} n\overline{F}(u_n) &= \tau && (2.2)\\ \lim_{n\to\infty} P(M_n \le u_n) &= e^{-\tau}. && (2.3) \end{aligned}$$

Bizonyítás. (Itt csak az elégségesség bizonyítását közöljük a $D(u_n)$ és $D'(u_n)$ feltételek használatának szemléltetésére. A szükségesség bizonyítása hasonlóképpen történik.)
Ha $l \ge 1$, akkor fennáll:

$$\begin{aligned} \sum_{i=1}^{l} P(X_i > u_n) &- \sum_{1\le i<j\le l} P(X_i > u_n, X_j > u_n) \\ \le P(M_l > u_n) &\le \sum_{i=1}^{l} P(X_i > u_n). && (2.4) \end{aligned}$$

Az X_n stacionaritásából:

$$\begin{aligned} \sum_{i=1}^{l} P(X_i > u_n) &= l\overline{F}(u_n), \\ \sum_{1\le i<j\le l} P(X_i > u_n, X_j > u_n) &\le l\sum_{j=2}^{l} P(X_1 > u_n, X_j > u_n). \end{aligned}$$

Ezt összevetve (2.4)-el $l = \lfloor n/k \rfloor$-re és rögzített k-ra, alsó és felső becslést kaphatunk $P(M_{\lfloor n/k \rfloor} \le u_n)$-re:

$$\begin{aligned} 1 - \lfloor n/k \rfloor \overline{F}(u_n) &\le P(M_{\lfloor n/k \rfloor} \le u_n) \\ &\le 1 - \lfloor n/k \rfloor \overline{F}(u_n) + \lfloor n/k \rfloor \sum_{j=2}^{\lfloor n/k \rfloor} P(X_1 > u_n, X_j > u_n). \end{aligned}$$

A tétel feltételéből következik, hogy $\lfloor n/k\rfloor \overline{F}(u_n) \to \tau/k \quad (n \to \infty)$, valamint $D'(u_n)$ miatt teljesül:

$$\limsup_{n\to\infty} \lfloor n/k\rfloor \sum_{j=2}^{\lfloor n/k\rfloor} P(X_1 > u_n, X_j > u_n) = o(1/k),\ k \to \infty.$$

Így a következő becsésekhez jutunk:

$$1 - \frac{\tau}{k} \leq \liminf_{n\to\infty} P(M_{\lfloor n/k\rfloor} \leq u_n) \leq \limsup_{n\to\infty} P(M_{\lfloor n/k\rfloor} \leq u_n) \leq 1 - \frac{\tau}{k} + o(1/k).$$

Ebből (2.1) alapján kapjuk:

$$\begin{aligned}\Big(1 - \frac{\tau}{k}\Big)^k &\leq \liminf_{n\to\infty} P(M_n \leq u_n) \\ &\leq \limsup_{n\to\infty} P(M_n \leq u_n) \leq \Big(1 - \frac{\tau}{k} + o(1/k)\Big)^k.\end{aligned}$$

Majd k-val végtelenhez tartva:

$$\lim_{n\to\infty} P(M_n \leq u_n) = e^{-\tau}.$$

□

Most kimondhatjuk az M_n maximumok határeloszlásáról szóló tételt.

7. Tétel. (Stacionárius sorozat maximumának határeloszlása)
Legyen (X_n) egy stacionárius sorozat F eloszlásfüggvénnyel, melyre $F \in MDA(H)$ valamely H extrémérték eloszlásra, azaz léteznek olyan $c_n > 0$ és $d_n \in \mathbb{R}$ konstansok, hogy

$$\lim_{n\to\infty} n\overline{F}(c_n x + d_n) = -\ln H(x), \quad x \in \mathbb{R}. \tag{2.5}$$

Továbbá tegyük fel, hogy az $(u_n) = (c_n x + d_n)$ sorozatok kielégítik a $D(u_n)$ és $D'(u_n)$ feltételeket minden $x \in \mathbb{R}$ esetén. Ekkor (2.5) ekvivalens a következőkkel:

$$c_n^{-1}(M_n - d_n) \xrightarrow{d} H, \tag{2.6}$$

$$c_n^{-1}(\tilde{M}_n - d_n) \xrightarrow{d} H. \tag{2.7}$$

Bizonyítás. (2.5) és (2.7) ekvivalenciája következik a független sorozatok maximumairól szóló 4. Tételből. (2.5) és (2.6) a 6. Tétel alapján ekvivalens. □

A fentiek alapján tehát látható, hogy a $D(u_n)$ és $D'(u_n)$ feltételeknek eleget tevő stacionárius folyamat valamint a hozzá tartozó független sorozat maximumai hasonlóan viselkednek. A tételből az is világos, hogy a c_n és d_n sorozatokat ugyanannak választhatjuk mindkét sorozat esetén. Mindezt persze a meglehetősen szigorú $D(u_n)$ és $D'(u_n)$ feltételek biztosítják, amelyek teljesülésének ellenőrzése nehéz feladat.

1. Példa. Tegyük fel, hogy $Y_1, Y_2, \ldots, Y_n$ független, azonos eloszlású valószínűségi változók sorozata $\sqrt{F}$ eloszlásfüggvénnyel, valamely F eloszlásfüggvényre. Definiáljuk az (X_n) sorozatot a következőképpen:

$$X_n = \max(Y_n, Y_{n+1}), \quad n \in \mathbb{N}.$$

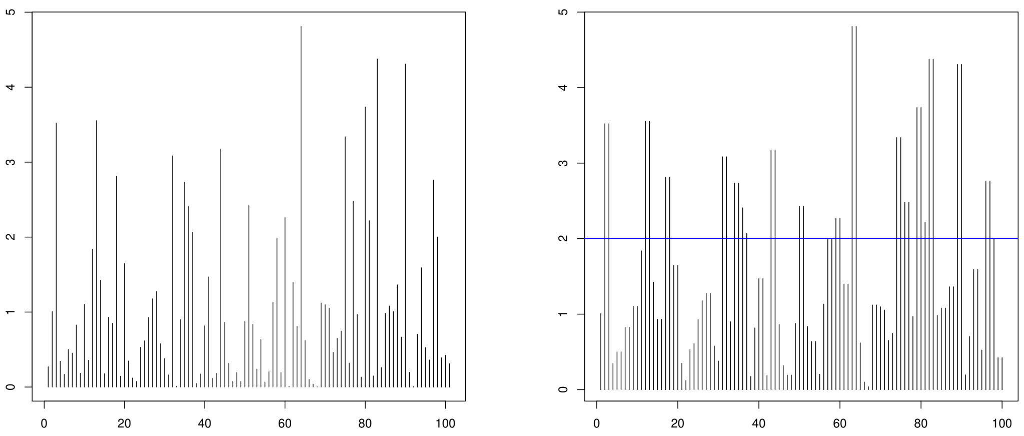
Ekkor (X_n) stacionárius sorozat F marginális eloszlással. A konstrukcióból az is világos, hogy (X_n) maximumai párokban jelentkeznek, egymást követő megfigyelések. Tegyük fel most, hogy $(u_n) \uparrow x_F$ (x_F az F eloszlás felső végpontja), valamint (u_n)-re fennáll (2.2). Ekkor $F(u_n) \to 1$ és

$$nP(Y_1 > u_n) = n\left(1 - \sqrt{F(u_n)}\right) = \frac{n\overline{F}(u_n)}{1 + \sqrt{F(u_n)}} \to \frac{\tau}{2}.$$

Így az 1. Tétel alapján

$$\begin{aligned} P(M_n \le u_n) &= P(\max(Y_1, \ldots, Y_n, Y_{n+1}) \le u_n) \\ &= P(\max(Y_1, \ldots, Y_n) \le u_n) F(u_n) \\ &\to e^{-\tau/2}, \quad n \to \infty. \end{aligned}$$

A $D(u_n)$ feltétel természetesen igaz, hiszen ha $l \ge 2$, akkor Y_i-k függetlensége miatt $\alpha_{n,l} = 0$ választás lehetséges. Minthogy $\mathrm{P}(M_n \le u_n) \to e^{-\tau/2}$, ezért a $D'(u_n)$ feltétel nem teljesülhet. Valóban, mivel X_1 és X_j független



(a) Standard exponenciális független Y_n változók realizációi

(b) Az $(X_n) = \max(Y_n, Y_{n+1})$ folyamat realizációja

2.1. ábra. Az X_n folyamat extrémumai párosával jelentkeznek

$j \geq 2$-re, ezért:

$$
\begin{aligned}
& n \sum_{j=2}^{\lfloor n/k \rfloor} P(X_1 > u_n, X_j > u_n) \\
= \quad & nP(X_1 > u_n, X_2 > u_n) + n(\lfloor n/k \rfloor - 2)P^2(X_1 > u_n) \\
= \quad & nP(\max(Y_1, Y_2) > u_n, \max(Y_2, Y_3) > u_n) + \tau^2/k + o(1) \\
= \quad & nP(Y_2 > u_n \text{ vagy } Y_1 > u_n, Y_3 > u_n) + \tau^2/k + o(1), \quad n \to \infty.
\end{aligned}
$$

Valamint:

$$nP(Y_2 > u_n \text{ vagy } Y_1 > u_n, Y_3 > u_n) \sim nP(Y_1 > u_n) \to \tau/2.$$

Ebből következik, hogy $D'(u_n)$ nem állhat fenn. Az (X_n) sorozatnak klaszter képzési tulajdonsága van, az extrémumok párosával jelentkeznek (2.1. ábra).

3. fejezet

Az extremális index

Az előző fejezetben láttuk, hogy csak meglehetősen szigorú feltételek biztosítják azt, hogy egy stacionárius sorozat maximumai ugyanúgy viselkedjenek, mint a hozzá tartozó független sorozaté. Az extremális index olyan mennyiség, amely egy stacionárius sorozat összefüggőségi struktúrája és extremális viselkedése közti kapcsolatot karakterizálja.

3.1. Az extremális index definíciója, tulajdonságai

4. Definíció. (Extremális index)
Legyen (X_n) egy erősen stacionárius sorozat és θ egy nem-negatív szám. Tegyük fel, hogy bármely $\tau > 0$-hoz létezik olyan (u_n) sorozat, hogy

$$\lim_{n\to\infty} n\overline{F}(u_n) = \tau \tag{3.1}$$

$$\lim_{n\to\infty} P(M_n \le u_n) = e^{-\theta\tau}. \tag{3.2}$$

Ekkor a θ számot az (X_n) sorozat extremális index*ének nevezzük.*

Megjegyzés. 1.) A definíció független az (u_n) sorozat választásától. Pontosabban, ha (X_n) extremális indexe $\theta > 0$, akkor bármely (u_n) valós sorozat és

$\tau \in [0, \infty]$ esetén (3.1), (3.2) és $P(\tilde{M}_n \leq u_n) \to \exp\{-\tau\}$ ekvivalensek, lásd Leadbetter [2]. Következésképpen ha $F \in \mathrm{MDA}(H)$ valamely H extrémérték eloszlásra és $\tilde{M}_n$ a hozzátartozó független sorozat maximuma, akkor

$$c_n^{-1}(\tilde{M}_n - d_n) \xrightarrow{d} H \iff c_n^{-1}(M_n - d_n) \xrightarrow{d} H^\theta \tag{3.3}$$

alkalmas $c_n > 0$ és d_n normalizáló konstansokkal.
2.) Mivel egy extrémérték eloszlás max-stabilis, ezért H^θ ugyanolyan típusú eloszlás, mint H, azaz léteznek olyan $c > 0$ és $d \in \mathbb{R}$ konstansok, hogy $H^\theta(x) = H(cx + d)$.

Látható, hogy ha egy stacionárius sorozatnak létezik extremális indexe, akkor a klasszikus extrémérték elmélet megfelelő eredményei könnyen átültethetők az összefüggő esetre.

2. Példa. Az 1. Példában megmutattuk, hogy az (X_n) sorozatra fennáll $P(M_n \leq u_n) \to e^{-\tau/2}$. Tehát az (X_n) sorozat extremális indexe $\theta = 1/2$.

A definíció kapcsán felvetődhetnek a következő kérdések:
Hogyan értelmezhető az extremális index? vagy
Milyen értékeket vehet fel az extremális index?
Minden erősen stacionárius sorozatnak létezik extremális indexe?
Ez utóbbi kérdésre a válasz *nem*!

3. Példa. Tegyük fel, hogy (X_n) független, F eloszlású valószínűségi változók sorozata, továbbá $F \in \mathrm{MDA}(\phi_\alpha)$ és c_n legyen a maximumot normalizáló konstans. Legyen A olyan pozitív valószínűségi változó, amely független (X_n)-től. Ekkor

$$\begin{aligned} & P(c_n^{-1} \max(AX_1, \dots, AX_n) \leq x) \\ &= \quad P(c_n^{-1} M_n \leq A^{-1} x) \\ &= \quad EP(c_n^{-1} M_n \leq A^{-1} x \mid A) \\ &\to E \exp\left\{-x^{-\alpha} A^\alpha\right\}, \quad x > 0. \end{aligned}$$

Az A változó alkalmas megválasztása mellett az (AX_n) folyamat maximuma nem lehet Frèchet eloszlású, ezért nem létezik határeloszlása.

Megjegyzés. Stacionárius sorozatok széles körére igaz, hogy léteznek olyan $0 \leq \theta' \leq \theta'' \leq 1$ valós számok, amelyekkel

$$e^{-\theta''\tau} \leq \liminf_{n\to\infty} P(M_n \leq u_n) \leq \limsup_{n\to\infty} P(M_n \leq u_n) \leq e^{-\theta'\tau}, \quad \tau > 0,$$

tetszőleges u_n sorozatra, mely kielégíti (3.1)-et. Ennek bizonyítása megtalálható Leadbetter, Lindgren és Rootzen [3]-ban.

1. Állítás. $\theta \in [0, 1]$

Bizonyítás. Minthogy az $e^{-\theta\tau}$ kifejezés valószínűség határértékeként áll elő ezért $e^{-\theta\tau} \leq 1$. Mivel $\tau > 0$ ezért θ szükségképpen nemnegatív. Másrészt:

$$P(M_n \leq u_n) = 1 - P\left(\bigcup_{i=1}^{n}\{X_i > u_n\}\right) \geq 1 - n\overline{F}(u_n).$$

Itt a baloldal ugyancsak $e^{-\theta\tau}$-hoz tar, míg a jobboldal $1 - \tau$-hoz. Így $e^{-\theta\tau} \geq 1 - \tau$ minden $\tau > 0$ esetén, ami csak úgy lehet, ha $\theta \leq 1$. □

Az extremális index definíciójából az 1. és 6. Tételekből egyszerűen következik a következő két állítás:

2. Állítás. *Ha (X_n) független, azonos eloszlású valószínűségi változók sorozata, akkor $\theta = 1$.*

□

3. Állítás. *Ha (X_n) erősen stacionárius sorozat, valamint (u_n) a valós számoknak olyan sorozata, amely kielégíti* (3.1)*-et, valamint a $D(u_n)$ és $D'(u_n)$ feltételeket, akkor $\theta = 1$.*

□

A fenti állításokból következik, hogy ha egy stacionárius sorozat extremális indexe 1, abból nem következik, hogy független sorozat.

Láttuk, hogy ha egy stacionárius sorozatra $\theta > 0$, akkor a független esethez hasonlóan általánosított extrémérték eloszlás lép fel a maximumok határeloszlásaként. Most vizsgáljuk meg, hogy $\theta = 0$ esetén mit mondhatunk a határeloszlásokról.

8. Tétel. *Legyen* (X_n) *erősen stacionárius sorozat, mely extremális indexe* $\theta = 0$. *Jelölje* $\tilde{M}_n$ *a hozzátartozó független sorozat maximumát. Ekkor tetszőleges* (v_n) *számsorozat esetén*

(a) *Ha* $\liminf_{n\to\infty} P(\tilde{M}_n \le v_n) > 0$, *akkor* $P(M_n \le v_n) \to 1$

(b) *Ha* $\limsup_{n\to\infty} P(M_n \le v_n) < 1$, *akkor* $P(\tilde{M}_n \le v_n) \to 0$.

Bizonyítás. Mivel $\theta = 0$, azért az extremális index definíciójából következik, hogy $\forall\ \tau > 0$ esetén $P(M_n \le u_n) \to 1$, midőn $n \to \infty$, ahol u_n olyan, hogy $n\overline{F}(u_n) \to \tau$. Ha $\liminf_{n\to\infty} P(\tilde{M}_n \le v_n) = \rho > 0$ és τ olyan, hogy $e^{-\tau} < \rho$, akkor mivel $P(\tilde{M}_n \le u_n) \to e^{-\tau}$, azért szükségképpen $v_n > u_n$ minden elegendően nagy n esetén. Így

$$\liminf_{n\to\infty} P(M_n \le v_n) \ge \lim_{n\to\infty} P(M_n \le u_n) = 1,$$

ami éppen (a).
Most tegyük fel, hogy $\limsup_{n\to\infty} P(M_n \le v_n) < 1$. Akkor mivel $P(M_n \le u_n) \to 1$, azért $v_n < u_n$, $\forall\ \tau > 0$-ra, elegendően nagy n esetén. Következésképpen:

$$\limsup_{n\to\infty} P(\tilde{M}_n \le v_n) \le \lim_{n\to\infty} P(\tilde{M}_n \le u_n) = e^{-\tau}\ .$$

Ebből τ-val a végtelenhez tartva kapjuk tételünk (b) részét. □

Megjegyzés. A fenti tételből következik, hogy ha egy stacionárius sorozat extremális indexe $\theta = 0$, akkor M_n-nek és $\tilde{M}_n$-nak nem létezhet egyszerre határeloszlása. A $\theta = 0$ elfajult esettel bővebben [3] foglalkozik.

Az extremális index jelentőségét szemlélteti a következő példa:

4. Példa. Tegyük fel, hogy olyan gátat kell létesítenünk egy folyó mellé, amely 95%-os biztonsággal véd az elkövetkezendő 100 évben. Továbbá tételezzük fel, hogy az éves maximális vízszint 99,9% és 99,95%-hoz tartózó percentilise 10m ill. 11m. Ha az éves maximumokat homogén, azonos eloszlásúnak feltételezzük, akkor a gátnak 11m magasnak kell lenni ($0,9995^{100} \approx 0,95$). Míg ha a folyamat stacionárius sorozat például $\theta = 0.5$ extremális indexszel, akkor a gátat elegendő 10m magasra építeni ($0,999^{50} \approx 0,95$).

3.2. Az extremális index értelmezése

Összefüggő folyamatoknak csoportképző tulajdonsága van, az extrém értékek nem egymástól függetlenül, hanem klaszterekbe rendeződve jelentkeznek. Hsing, Hüsler és Leadbetter [4] megmutatta, hogy bizonyos keverő feltételek teljesülése esetén a meghaladások folyamata

$$N_n(\cdot) = \sum_{i=1}^{\infty} \varepsilon_{i/n}(\cdot) I_{\{X_i > u_n\}}$$

összetett Poisson-folyamathoz tart

$$N(\cdot) = \sum_{i=1}^{\infty} \xi_i \varepsilon_{\Gamma_i}(\cdot),$$

midőn $n\overline{F}(u_n) \to \tau > 0$. Itt

$$\varepsilon_x(A) = \begin{cases} 1 & \text{ha } x \in A \\ 0 & \text{ha } x \notin A, \quad A \in \mathcal{A} \ \ \sigma\text{-algebra} \end{cases}$$

Az $N(\cdot)$ hátterében lévő homogén Poisson-folyamat paramétere $\theta\tau$, pontjai a Γ_i-k. A ξ_i-k $(i = 1, 2, \dots)$ – ún. klaszter méretek – egymástól és a homogén Poisson-folyamattól független valószínűségi változók $\pi_j := P(\xi_i = j)$ eloszlással $(j = 1, 2 \dots)$. A meghaladások várható száma:

$$\begin{aligned} EN(0,1] &= E\sum_{i=1}^{\infty} \xi_i \varepsilon_{\Gamma_i}(0,1] \\ &= E\sum_{i=1}^{\infty} \varepsilon_{\Gamma_i}(0,1] E\xi_i \\ &= \theta\tau E\xi_i. \end{aligned} \tag{3.4}$$

Általános feltevések mellett fennáll a következő:

$$\begin{aligned} \pi_j &= \lim_{n\to\infty} \pi_j(n) \\ &= \lim_{n\to\infty} P\left(\sum_{i=1}^{r} I_{\{X_i > u_n\}} = j \mid \sum_{i=1}^{r} I_{\{X_i > u_n\}} > 0 \right), \quad j \in \mathbb{N}. \end{aligned}$$

Itt $r = r(n)$ olyan, hogy $r(n) \to \infty$ és $r(n)/n \to 0$. A folyamat szintmeghaladásait r hosszú blokkokra osztva figyeljük, így különítünk el csoportokat. További feltételek teljesülése esetén (lásd Smith [5]):

$$\lim_{n\to\infty} \sum_{j=1}^{\infty} j\pi_j(n) = \sum_{j=1}^{\infty} j\pi_j = E\xi_1. \tag{3.5}$$

Ezért (3.4) és (3.5) alapján kapjuk a következőt:

$$\begin{aligned}\tau = \lim_{n\to\infty} n\overline{F}(u_n) &= \lim_{n\to\infty} EN_n(0,1] = \lim_{n\to\infty} E\sum_{i=1}^{n} I_{\{X_i>u_n\}} \\ &= EN(0,1] = \theta\tau E\xi_1.\end{aligned}$$

Ezt átrendezve $\theta = (E\xi_1)^{-1}$, vagyis az extremális index reciproka a határértékként fellépő összetett Poisson-folyamat klaszterméret-várható értékének.

A meghaladások folyamatának határértékeként összetett Poisson-folyamatot kaptunk, extrém értékek egy csoportja jön létre véletlen időpontokban. Ha ezt összevetjük az 1. Tétellel és az azt követő megjegyzéssel, akkor láthatjuk, hogy ott éppen a függetlenség biztosította azt, hogy a meghaladások folyamata homogén Poisson-folyamathoz tartson. Az extremális index egy stacionárius sorozat szintmeghaladásainak csoportképző tulajdonságát reprezentálja, független esetben a $\theta = 1$ extremális index megfelel ennek az értelmezésnek. Gyakorlati alkalmazásokban számos folyamat esetén találkozhatunk csoportképző tulajdonsággal. Például egy folyó árvizei nagyobb hullámokban, több napig tartó magas vízszintértékekkel jelentkeznek.

3.3. Az extremális index becslési módszerei

Az extremális index becslésének alapvető megközelítése az előző részben tárgyalt klaszter-képzési tulajdonságon alapszik. Először definiáljunk valamilyen módon csoportokat, majd számoljuk meg a csoporton belüli szintmeghaladásokat. Ezután az átlagos csoport-elemszám reciprokával becsüljük θ-t.

3.3.1. A blokk módszer

Az extremális index definíciójából kiindulva

$$P(M_n \leq u_n) \approx P^{\theta}(\tilde{M}_n \leq u_n) = F^{\theta n}(u_n),$$

azáltal, hogy $n\overline{F}(u_n) \to \tau > 0$, ahol $\tilde{M}_n$ a stacionárius sorozattal megegyező marginális eloszlással rendelkező független sorozat. Így

$$\lim_{n\to\infty} \frac{\ln P(M_n \leq u_n)}{n \ln F(u_n)} = \theta.$$

Minthogy az $F(u_n)$ és $P(M_n \leq u_n)$ mennyiségek ismeretlenek, ezért ezeket kell becsülni. Az $\overline{F}(u_n)$ farokeloszlás tapasztalati becslése:

$$\frac{N}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i > u_n\}},$$

ahol tehát N az u_n szint feletti megfigyelések száma. $P(M_n \leq u_n)$ becslése nem kézenfekvő. A 2. fejezetben említettük, hogy a $D(u_n)$ feltételből következik, hogy

$$P(M_n \leq u_n) \approx P^k(M_{\lfloor n/k \rfloor} \leq u_n), \tag{3.6}$$

alkalmas k konstans, vagy $k = o(n)$ mellett. Ez adja az ötletet a blokk módszerhez. Az egyszerűség kedvéért tegyük fel, hogy $n = rk$, ahol $r = r(n) \to \infty$ és $k = k(n) \to \infty$ egész számok. Másképpen legyen $r = \lfloor n/k \rfloor$. Azaz osszuk fel az $X_1, \ldots, X_n$ mintát k darab r hosszú blokkra:

$$X_1, \ldots, X_r; \ldots ; X_{(k-1)r+1}, \ldots, X_{kr}$$

Minden egyes blokknak számoljuk ki a maximumát:

$$M_r^{(i)} = \max(X_{(i-1)r+1}, \ldots, X_{ir}) \ , \ i = 1, \ldots, k.$$

Így (3.6) a következő közelítést adja:

$$\begin{aligned} P(M_n \le u_n) &= P\left(\max_{1 \le i \le k} M_r^{(i)} \le u_n\right) \approx P^k(M_r \le u_n) \\ &\approx \left(\frac{1}{k}\sum_{i=1}^{k} I_{\{M_r^{(i)} \le u_n\}}\right)^k = \left(1 - \frac{K}{k}\right)^k, \end{aligned}$$

ahol K jelölje azon blokkok számát, amelyekben legalább egy érték meghaladja az u_n szintet. Az előzőekből kapjuk a következő becslést:

$$\hat{\theta}_1 \ = \ \frac{k}{n}\frac{\ln(1 - K/k)}{\ln(1 - N/n)} = \frac{1}{r}\frac{\ln(1 - K/k)}{\ln(1 - N/n)}.$$

A Taylor-sorfejtésből adódó közelítés a

$$\hat{\theta}_2 = \frac{K}{N} = \frac{1}{r}\frac{K/k}{N/n} \approx \hat{\theta}_1$$

Ez a becslés megfelel annak az értelmezésnek, hogy az extremális index a folyamat extrémumainak klaszter-képzési tulajdonságát reprezentálja.

A blokk módszerben klasztereket definiáltunk úgy, hogy a mintát fölosztottuk r hosszú blokkokra. Azt mondjuk, hogy egy *klaszter létrejött az $i-ik$ blokkban*, ha az

$$\left\{M_r^{(i)} > u_n\right\} = \bigcup_{j=1}^{r} \left\{X_{(i-1)r+j} > u_n\right\}$$

esemény bekövetkezik. A klaszter elemszáma a meghaladások száma a blokkban. Ezek az események karakterizálják (X_n) extremális viselkedését, ha $r = r(n)$-et $o(n)$ sebességgel növeljük.

A módszer nehézsége, hogy a becslés függ r és u_n értékétől, így a becslés kulcsfontosságú kérdése:
Adott n mellett hogyan válasszuk meg az r és u_n paramétereket?
Részleges válaszok találhatók a szakirodalomban, lásd például [6].

3.3.2. A futam módszer

O'Brien [7] bebizonyította, hogy bizonyos feltételek mellett fennáll a következő:

$$\begin{aligned} P(M_n \leq u_n) &= (F(u_n))^{nP(M_{2,s} \leq u_n \mid X_1 > u_n)} + o(1) \\ &= \exp\{-nP(X_1 > u_n, M_{2,s} \leq u_n)\} + o(1) \end{aligned}$$

azáltal, hogy $n\overline{F}(u_n) \to \tau$, ahol $M_{2,s} = \max(X_2, \ldots, X_s)$, és s-re különböző feltételek állnak fenn, többek között $s = s(n) \to \infty$ és $s/n \to 0$. Ugyanakkor az extremális index definíciója alapján:

$$P(M_n \leq u_n) = \exp\{-\theta\tau\} + o(1).$$

Így a fenti egyenlőségekből:

$$\lim_{n\to\infty} \theta_n(s(n), u_n) = \lim_{n\to\infty} P(M_{2,s} \leq u_n \mid X_1 > u_n) = \theta \ .$$

Ezzel θ feltételes valószínűségként értelmezhető. $M_{2,s}$ akkor lehet kisebb u_n-nél, ha X_1 az utolsó elem az u_n szintet meghaladó értékek csoportjában. Ha a nagy értékek csoportokban fordulnak elő, akkor lesznek hosszabb szakaszok is a csoportok között, amelyekben kis értékek vannak. A becslés ilyen *futamok* megszámlálásán alapszik:

$$\hat{\theta}_3 = \frac{\sum_{i=1}^{n-r} I_{A_{i,n}}}{\sum_{i=1}^{n} I_{\{X_i > u_n\}}} = \frac{\sum_{i=1}^{n-r} I_{A_{i,n}}}{N},$$

ahol

$$A_{i,n} = \{X_i > u_n, X_{i+1} \leq u_n, \ldots, X_{i+r} \leq u_n\}.$$

Vagyis legalább r egymás utáni megfigyelést keresünk az u_n szint alatt, amelyek két klasztert határolnak. Ezáltal újabb definícióját adtuk a csoportoknak. Két árvizes csoportot különbözőnek tekintünk, ha legalább r darab u_n alatti megfigyelés választja el őket. A klaszter mérete, pedig egy ilyen csoportban az u_n szintet meghaladó értékek száma. A futam-becslés a blokk módszerhez hasonlóan az u_n és r_n paraméterek megválasztásától függ.

3.3.3. Maxima módszerek

A maxima módszerekkel való becslés az extremális index (3.2) alatti karakterizációján alapszik. Azt használja ki, hogy az extremális index összefüggést ad egy stacionárius sorozat maximumainak és az ugyanolyan marginális eloszlású független változók maximumainak extrémérték eloszlása között.

Amint az extremális index bevezetésénél láttuk, ha $X_1, X_2, \ldots, X_n$ valószínűségi változók egy stacionárius sorozata θ extremális indexszel, akkor $n \to \infty$ esetén:

$$F_n(u_n) \to G(x) = \{H(x)\}^\theta, \tag{3.7}$$

ahol u_n megfelelő sorozat és $H(x)$ az (X_n)-hez tartozó független sorozat maximumának határeloszlása: $\mathrm{GEV}(\mu, \sigma, \xi)$. Ebből következik, hogy $G(x)$ is általánosított extrémérték eloszlás, $\mathrm{GEV}(\mu_\theta, \sigma_\theta, \xi_\theta)$ a következő paraméterekkel:

$$\mu_\theta = \mu + \frac{\sigma}{\xi}(\theta^\xi - 1), \quad \sigma_\theta = \sigma\theta^\xi \quad \text{és} \quad \xi_\theta = \xi. \tag{3.8}$$

A $G(x)$ és $H(x)$ eloszlásokat az eredeti stacionárius sorozat és a hozzá tartozó független sorozat maximumaiból becsülhetjük. Ezután a két eloszlás megfelelő összehasonlításból a θ paraméterre kaphatunk becslést.

Általánosított extrémérték eloszlás illesztését stacionárius sorozat maximumaira következőképpen végezhetjük. A klasszikus extrémérték elméletben megszokott módszerekhez hasonlóan a mintát r hosszú blokkokra osztva az egyes blokkok maximumait vizsgáljuk. Jelölje $M_r^{(i)}$ $(i = 1, \ldots, k)$ az i-ik blokk maximumát, ahol $k = \lfloor n/r \rfloor$. Ekkor a $D(u_n)$ feltétel és (3.7) alapján elegendő nagy r mellett az $M_r^{(i)}$ értékek extrémérték eloszlásból vett független megfigyeléseknek tekinthetők. Ezeket $\mathrm{GEV}(\mu_\theta, \sigma_\theta, \xi_\theta)$ eloszlással modellezzük, amihez az eloszlás paramétereit becsüljük. A paraméterbecslésekre számos módszer ismert.

Ancona-Navarette és Tawn [8] alapján az eredeti stacionárius sorozattal megegyező marginális eloszlású független sorozatot úgy kaphatunk, hogy a megfigyelések indexeit véletlenítjük, vagyis véletlenszerűen megkeverjük a mintát. Ez az új sorozat úgy viselkedik, mint a stacionárius sorozathoz tartozó független sorozat approximációja.

Most két különböző módszert ismertetünk a θ paraméter becslésére az így konstruált független sorozat felhasználásával.

Eloszlások összehasonlításán alapuló módszer

Jelölje $\tilde{M}_r^{(i)}$ $(i = 1, \ldots, k)$ a konstruált független sorozat r hosszú blokkjainak maximumait. Ezeket független, általánosított extrémérték eloszlásból vett megfigyeléseknek tekintjük, azaz GEV(μ, σ, ξ) eloszlással modellezzük. Ha az extremális index θ, akkor az eredeti összefüggő sorozat $M_r^{(i)}$ $(i = 1, \ldots, k)$ maximumait GEV$(\mu_\theta, \sigma_\theta, \xi_\theta)$ eloszlással modellezzük, ahol a $\mu_\theta, \sigma_\theta,$ és ξ_θ paramétereket (3.8) adja.
Gomes [9] javaslatára a θ-ra úgy adhatunk becslést, hogy a független minta és az eredeti sorozat maximumaira illesztett általánosított extrémérték eloszlás paramétereinek becsléseit, $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$-t és $(\hat{\mu_\theta}, \hat{\sigma_\theta}, \hat{\xi_\theta})$-t vetjük össze és a (3.8) összefüggés segítségével kifejezzük θ-t. Ehhez először a ξ paramétert becsüljük, mely háromféleképpen történhet. A (3.8) összefüggés alapján az eltolás és skála paraméterekből a ξ paraméterek becslésének megkerülésével

$$\tilde{\xi} = \frac{\hat{\sigma} - \hat{\sigma}_\theta}{\hat{\mu} - \hat{\mu}_\theta} \quad .$$

Minthogy $\xi = \xi_\theta$, ezért külön-külön a két maximum-sorozatból becsülve ξ-t újabb becslést kaphatunk rá $\tilde{\xi} = \hat{\xi}$ ill. $\tilde{\xi} = \hat{\xi}_\theta$. Elméletileg a három érték megegyezik, de a gyakorlatban a becslések eltérnek. Az extremális index becslése ezután

$$\hat{\theta}_4 = \left(\frac{\hat{\sigma}}{\hat{\sigma}_\theta}\right)^{-1/\tilde{\xi}} .$$

Az eljárás lényege tehát, hogy külön-külön becsüljük meg a kétféle GEV paramétervektort, majd a köztük lévő megfelelő összefüggés kihasználásával következtetünk θ-ra.

Maximum likelihood módszer

A maximum likelihood módszer esetén–a fentiekkel szemben– a $(\mu, \sigma, \xi, \theta)$ paramétervektort egyszerre becsüljük. E négy paraméter egyértelműen meghatározza M_r és $\tilde{M}_r$ eloszlásának összes paraméterét, hiszen

$$\mathrm{GEV}(\mu_\theta, \sigma_\theta, \xi_\theta) = \mathrm{GEV}(\mu + \frac{\sigma}{\xi}(\theta^\xi - 1), \sigma\theta^\xi, \xi) = \mathrm{GEV}(\mu, \sigma, \xi, \theta).$$

Ezek szerint az összefüggő sorozat maximumainak extrémérték eloszlását úgy is kezelhetjük, mint egy négyparaméteres $\mathrm{GEV}(\mu, \sigma, \xi, \theta)$ eloszláscsalád. A paraméterek együttes becsléséhez az

$$\underbrace{M_r^{(1)}, \ldots, M_r^{(k)}}_{\mathrm{GEV}(\mu,\sigma,\xi,\theta)}, \underbrace{\tilde{M}_r^{(1)}, \ldots, \tilde{M}_r^{(k)}}_{\mathrm{GEV}(\mu,\sigma,\xi)} \tag{3.9}$$

mintát vesszük alapul és független, de nem azonos eloszlású GEV valószínűségi változóknak tekintjük a fenti marginális eloszlásokkal. Az együttes vektor nem független, de Ancona-Navarette és Tawn [8] alapján ez a hatás elhanyagolható a gyakorlatban. A paraméterek, így θ maximum likelihood becsléséhez (3.9) likelihood-függvényét maximalizálva jutunk el.

A maxima módszerek előnye, hogy a becslés csak egyetlen paramétertől, az r-től függ és nincs szükség szintek definiálására. Ugyanakkor r-t megfelelően kell megválasztani ahhoz, hogy a különböző maximumok függetlennek tekinthetőek legyenek, és emellett elegendő sok megfigyelésünk legyen az eloszlásillesztéshez.

Mivel a fenti eljárások véletlenítést alkalmaznak, ezért a becsléseket úgy érdemes végezni, hogy minden egyes r érték mellett többször keverjük át a sorozatot, majd minden lépésben számoljuk ki θ-t. A becslések tapasztalati eloszlását tekintve, empirikusan becsülhetjük θ-t az átlaggal vagy mediánnal, a szórás figyelembevételével konfidencia intervallum konstruálható. Az r érték megválasztásához a különböző hibák figyelembevételével juthatunk el. Ilyen például a becslések szórása vagy az alakparaméter három különböző becslésének az eltérése. A következő fejezetben azonban látni fogjuk, hogy az extremális index gyakorlati alkalmazásánál például extrémérték eloszlás illesztésére, az r paraméter megválasztása már nem lesz kérdés.

3.4. Az extremális index becsléseinek tulajdonságai

Ebben a részben az extremális index blokk és futam módszerrel történő becslésének határeloszlásáról és konzisztenciájáról lesz szó. Ezek létezéséhez különböző feltételekre lesz szükségünk, ezért vezessük be a következő jelöléseket.

Továbbra is $X_1, \ldots, X_n$ jelöljön egy stacionárius sorozatból vett megfigyelés-sorozatot, M_n pedig az első n mintaelem maximumát.
Ezen kívül a becslési módszereknél bevezetett jelölésekkel:

$$\begin{aligned} M_{2,r} &= \max(X_2, \ldots, X_r), \\ M_r^{(i)} &= \max(X_{(i-1)r+1}, \ldots, X_{ir}) \text{ és} \\ A_{i,n} &= \{X_i > u_n, X_{i+1} \leq u_n, \ldots, X_{i+r} \leq u_n\}. \end{aligned}$$

Az u_n sorozatról feltesszük, hogy $nP(X_1 > u_n) = n\overline{F}(u_n) \to \infty$ valamint a becslésekben szereplő r paraméterről, hogy $r = r_n = o(n)$.
Az itt leírtak alapja az extremális index kétféle megközelítése:

$$\theta^B(r,u) = \frac{P(M_r > u)}{rP(X_1 > u)}$$

ill.

$$\theta^R(r,u) = P(M_{2,r} \leq u \mid X_1 > u)$$

Amint láttuk, ezen értelmezésekből kaphatunk θ-ra becslést. Az extremális index n elemű mintából blokk módszerrel történő becslését $\hat{\theta}_n^B$-vel, a futam módszerrel történő becslését $\hat{\theta}_n^R$-el jelöljük.

Definiáljuk $1 \leq m \leq n$ esetén az $\mathcal{F}_{1,m}(u) = \sigma\left\{I_{\{X_i > u\}} : m \leq i < n\right\}$ σ-algebrát és legyen

$$\varphi(k) := \varphi(k,u) = \sup |P(B \mid A) - P(B)|,$$

ahol a szuprémum az összes olyan $A \in \mathcal{F}_{1,m}(u)$, $B \in \mathcal{F}_{m+k,\infty}(u)$ halmazokon értendő, amelyekre $P(A)P(B) > 0$ és $m > 1$.

9. Tétel. *Tegyük fel, hogy $n \to \infty$ esetén*

$$\theta_n^B := \theta^B(r_n, u_n) = \frac{P(M_r > u_n)}{rP(X_1 > u_n)} \to \theta \tag{3.10}$$

továbbá $r = r_n$ és $r = 1$, valamint $k = \lfloor n/r \rfloor$ esetén

$$\begin{aligned} \delta_n &:= \sum_{i=1}^{k} (1 - i/k)\{P(M_r^{(i+1)} > u_n \mid M_r^{(1)} > u_n) - P(M_r^{(1)} > u_n)\} \\ &= o(n\overline{F}(u_n)). \end{aligned} \tag{3.11}$$

Ekkor $\hat{\theta}_n^B \to \theta$ 1 valószínűséggel, midőn $n \to \infty$.

$\square$

Megjegyzés. Mivel $|\delta_n| \leq \sum_{i=r}^{n} \varphi(1 + (i-1)r)) \leq \sum_{i=1}^{n} \varphi(i)$, ezért (3.11) teljesüléséhez elegendő, ha

$$\sum_{i=1}^{n} \varphi(i, u_n) = o(n\overline{F}(u_n)) \tag{3.12}$$

10. Tétel. *Tegyük fel, hogy $n \to \infty$ esetén*

$$\theta_n^R := \theta^R(r_n, u_n) = P(M_{2,r} \leq u_n \mid X_1 > u_n) \to \theta \tag{3.13}$$

továbbá $r = r_n$ és $r = 1$ esetén

$$\gamma_n := \sum_{i=1}^{n} (1 - i/n)\{P(A_{i+1,n} \mid A_{1,n}) - P(A_{1,n})\} = o(n\overline{F}(u_n)). \tag{3.14}$$

Ekkor $\hat{\theta}_n^R \to \theta$ 1 valószínűséggel, midőn $n \to \infty$.

$\square$

Megjegyzés. Mivel $|\gamma_n| \leq \sum_{i=r}^{n} \varphi(1 + i - r)$, ezért (3.14) fennáll, ha (3.12) teljesül.

Ezek után $\hat{\theta}_n^B$ és $\hat{\theta}_n^R$ határeloszlásának meghatározásához vezetünk be feltételeket. Ehhez szükségünk lesz a következő jelölésekre.

$$Y_i := I_{A_{i,n}} - \theta_n^R I_{\{X_i > u_n\}}, \qquad Z_i := I_{\{M_r^{(i)} > u_n\}} - \theta_n^B \sum_{j=(i-1)r+1}^{ir} I_{\{X_j > u_n\}}$$

Látható, hogy $EY_i = EZ_i = 0$, és $D^2 Y_i = \theta_n^R (1 - \theta_n^R) \overline{F}(u_n)$.

11. Tétel. *Tegyük fel, hogy $\theta < 1$, a 10.Tétel feltételei fennállnak és*

$$\sup_n \varphi(kr_n, u_n) \to 0 \quad (k \to \infty). \tag{3.15}$$

Ha $(D^2 \sum_{i=1}^n Y_i)/(nD^2 Y_1) \to \sigma_R^2$, midőn $n \to \infty$ és $r_n^2 = o(n\overline{F}(u_n))$, akkor

$$\sqrt{n\overline{F}(u_n)} \left(\hat{\theta}_n^R - \theta_n^R \right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_R^2 \theta (1-\theta)\right). \tag{3.16}$$

□

12. Tétel. *Tegyük fel, hogy a 9.Tétel feltételei és* (3.15) *teljesülnek. Továbbá $r_n^4 = o(n\overline{F}(u_n))$ és $(D^2 \sum_{i=1}^n Z_i)/(n\overline{F}(u_n)) \to \sigma_B^2$, midőn $n \to \infty$. Ekkor*

$$\sqrt{n\overline{F}(u_n)} \left(\hat{\theta}_n^B - \theta_n^B \right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_B^2\right). \tag{3.17}$$

□

Megjegyzés. Ha $(\theta_n^R - \theta) = o(n\overline{F}(u_n))$ ill. $(\theta_n^B - \theta) = o(n\overline{F}(u_n))$, akkor (3.16)-ban ill. (3.17)-ben θ_n^R ill. θ_n^B helyett θ írható. A 11. és 12. Tételek alkalmazásával konfidencia intervallum konstruálható θ-ra, amelyhez σ_R^2-re ill. σ_B^2 -re kell becslést adni. Ez meglehetősen nehéz feladat, eredmények a szakirodalomban még nem találhatóak.

A fenti tételek bizonyításai megtalálhatóak Weissman, Novak [10]-ben. Itt utalást találhatunk arra is, hogy az extremális index becsléseinek fenti tulajdonságai alapján a futam módszer tűnik megbízhatóbbnak. A becslések szórását és torzítását figyelembe véve ugyancsak a futam módszert preferálják Smith, Weissman [6]-ban.

4. fejezet

Általánosított extrémérték eloszlás illesztése stacionárius sorozat maximumaira

A 5. Tétel alapján tudjuk, hogy egy stacionárius sorozat maximumainak lehetséges határeloszlásaként általánosított extrémérték eloszlás (GEV):

$$G(x) = \exp\left[-\left\{1 + \xi_\theta \left(\frac{x - \mu_\theta}{\sigma_\theta}\right)\right\}_+^{-1/\xi_\theta}\right], \tag{4.1}$$

léphet fel, ahol $\{x\}_+ = max(0, x)$, $\sigma_\theta > 0$, és μ_θ, σ_θ, ξ_θ az eltolás, skála és alak paraméterek. A stacionárius sorozat extrémumait ezen eloszláscsaládból származó eloszlással szeretnénk modellezni.

4.1. A klasszikus módszer

Amint az extremális index maxima módszerrel való becslésénél írtuk, klasszikus módszerrel könnyen becsülhetjük stacionárius sorozat maximumainak eloszlását. A megfigyelés-sorozatot r hosszú blokkokra osztva az egyes blokkok maximumait független, extrémérték eloszlásból származó megfigyeléseknek tekintjük. Ezután erre az új mintára illesztünk általánosított extrémérték eloszlást. Ebben az esetben a klasszikus extrémérték elméletben megszokott módszerek alkalmazhatók. Az eloszlás illeszkedését a modellből számított kvantilisek ill. valószínűségek és azok empirikus megfelelőinek összehasonlításával vizsgálhatjuk. Az illeszkedést például az ún. PP-plot és QQ-plot segítségével követhetjük nyomon. (Bővebben lásd például [1], [11].)

Minthogy gyakran idősorok maximumait vizsgáljuk, ezért célszerű a blokkok méretét úgy megválasztani, hogy valamilyen időperiódusnak feleljen meg. Hidrológiai alkalmazásokban például gyakran az éves maximumokat vizsgálják. A GEV(μ, σ, ξ) eloszlás paramétereinek többféle becslése ismeretes. Többek között a maximum-likelihoood becslés, amikor a minta együttes sűrűségfüggvényét maximalizáljuk és a három paramétert egyszerre becsüljük.

Az egyik leggyakrabban felvetődő kérdés, hogy mi az a szint, amit átlagosan T évente halad meg a folyamat? Ha éves maximumokat vizsgálunk, akkor ezt a rájuk illesztett GEV eloszlásból már könnyedén meghatározhatjuk. Ha megbecsüljük azon $0 \leq p \leq 1$-hez tartozó q_p kvantilist, amelyre

$$P(M_n \leq q_p) = 1 - p \ ,$$

akkor q_p-t átlagosan $1/p$ évente haladja meg a folyamat. q_p-t (4.1)-ből kifejezve kapjuk:

$$q_p = \begin{cases} \mu_\theta - \frac{\sigma_\theta}{\xi_\theta}\left[1 - \{-\log(1-p)\}^{-\xi_\theta}\right] & ,\text{ha } \xi \neq 0 \\ \mu_\theta - \sigma_\theta \log\{-\log(1-p)\} & ,\text{ha } \xi = 0. \end{cases} \tag{4.2}$$

A GEV eloszlás $\xi > 0$ esetén alulról, $\xi < 0$ esetén felülről korlátos, így a végpontra azonnal kapjuk:

$$x_G = \mu_\theta - \frac{\sigma_\theta}{\xi_\theta}. \tag{4.3}$$

Ha az általánosított extrémérték eloszlás paramétereinek maximum likelihood becslése $(\hat{\mu_\theta}, \hat{\sigma_\theta}, \hat{\xi_\theta})$, akkor a fenti mennyiségek maximum likelihood becslését úgy kapjuk, hogy a (4.2) ill. (4.3) képletekbe behelyettesítjük a paraméterek becsléseit.

Bevezetve az $y_p = -\log(1-p)$ jelölést a kvantilisek és a végpont maximum likelihood becsléseire kapjuk tehát:

$$\begin{aligned}
\hat{q_p} &= \begin{cases} \hat{\mu_\theta} - \frac{\hat{\sigma_\theta}}{\hat{\xi_\theta}} \left[1 - {y_p}^{-\hat{\xi_\theta}}\right] & ,\text{ha } \xi \neq 0 \\ \hat{\mu_\theta} - \hat{\sigma_\theta} \log y_p & ,\text{ha } \xi = 0 \end{cases} \\
\hat{x_G} &= \hat{\mu_\theta} - \frac{\hat{\sigma_\theta}}{\hat{\xi_\theta}} \quad .
\end{aligned}$$

Az $(\log y_p, \hat{q_p})$ pontokat koordináta-rendszerben ábrázolva a különböző visszatérési időkhöz tartozó szinteket szemléltethetjük. Látható, hogy a klasszikus módszer nem használja ki a folyamat összefüggőségi szerkezetét, így például az extremális indexet sem. A klasszikus extrémérték elmélet és gyakorlati alkalmazásai összefoglalva megtalálhatók például [11]-ben.

4.2. Extremális indexet használó módszerek

Az itt leírt módszerek esetében a stacionárius sorozat összefüggőségi struktúrája és az extremális index jut főszerephez. Összefüggő sorozat maximumainak eloszlását alapvetően kétféle megközelítésből határozhatjuk meg.

Az egyik megközelítés, ha ismerjük a stacionárius sorozat marginális eloszlását. Ekkor a karakterizációs tételek segítségével meghatározhatjuk, hogy független esetben milyen H extrémérték eloszláshoz tart a maximum eloszlása. Ha ismerjük az extremális indexet, akkor H-ból meghatározhatjuk az összefüggő folyamat maximumának határeloszlását, hiszen tudjuk, hogy ez H^θ-al lesz egyenlő.

A másik megközelítés esetén már ismertnek tételezzük fel a stacionárius sorozathoz tartozó független sorozat maximumainak határeloszlását. Így az előzőhöz hasonlóan ebből már tudjuk az eredeti folyamat maximumának határeloszlását is.

A gyakorlatban ritkán ismert egy stacionárius sorozat marginális eloszlása, így a következő részben a második megközelítésre mutatunk alkalmazásokat.

4.3. Véletlenítéses módszerek

A véletlenítéses módszerek alapjául a maxima módszereknél leírtak szolgálnak. Egy stacionárius sorozathoz tartozó független sorozatot konstruálunk úgy, hogy az eredeti sorozatot véletlenszerűen megkeverjük. Az összefüggő sorozattal való kapcsolatot pedig az extremális index biztosítja.

Amint említettük, gyakran az éves maximumok vizsgálata az érdekes, ezért a továbbiakban egy stacionárius folyamat éves maximumainak eloszlását szeretnénk meghatározni.

4.3.1. Eloszlás illesztése adott extremális index mellett

Ha a stacionárius sorozat extremális indexét előzetesen megbecsültük, akkor az éves maximumokra való extrémérték eloszlás illesztését a következőképpen végezhetjük.

1. Véletlenítéssel konstruáljunk egy a stacionárius sorozathoz tartozó független sorozatot,

2. Becsüljük meg a kapott független sorozat éves maximumaira illesztett extrémérték eloszlás paramétereit $\longrightarrow$ $\mathrm{GEV}(\mu, \sigma, \xi)$,

3. Határozzuk meg (3.8) alapján az eredeti éves maximumok $\mathrm{GEV}(\mu_\theta, \sigma_\theta, \xi_\theta)$ eloszlásának paramétereit.

– A fenti lépéseket többszöri véletlenítés mellett elvégezve

4. A stacionárius sorozat éves maximumaira illesztett extrémérték eloszlás paramétereit az egyes lépésekben kapott becslések átlagával közelíthetjük.

Láttuk, hogy az extremális index pontos értékének meghatározása nehéz, ezért különböző, reálisnak tűnő θ értékek mellett ajánlatos a becsléseket elvégezni. Majd érdemes megvizsgálni a θ paraméterre való érzékenységet, vagyis az eloszlás illeszkedésének jóságát az egyes θ értékek mellett.

A fenti módszer előnye, hogy az extremális indexen keresztül kihasználja az eredeti folyamat összefüggőségi szerkezetét, mely erősen befolyásolhatja az extrém értékek előfordulását. A klasszikus esetben csupán az éves maximumokat vizsgáljuk és elveszítjük a mögötte lévő eredeti folyamat bizonyos jellemzőit, így a kiemelkedő értékek csoportképző tulajdonságát.

Megjegyzés. A leírt módszer mellékeredménye egy stacionárius sorozat maximumaira illesztett extrémérték eloszlás alak paraméterének (ξ-nek) új becslési módszere. Láttuk, hogy ez megegyezik a hozzátartozó független sorozat maximumaira illesztett extrémérték eloszlás alak paraméterével. Így ha a véletlenítéssel kapott független sorozatból becsüljük az alak paramétert, akkor a többszöri véletlenítés mellett kapott becslésekből empirikusan közelíthetjük ξ-t például az átlaggal, vagy más statisztikai jellemzővel.

4.3.2. Kvantilis becslések

Stacionárius sorozat maximumainak kvantilis becslései véletlenítéses módszerrel az előzőekre épülnek.

1. Véletlenítéssel konstruáljunk egy a stacionárius sorozathoz tartozó független sorozatot,
2. Becsüljük meg a kapott független sorozat éves maximumaira illesztett extrémérték eloszlás paramétereit $\longrightarrow$ $\mathrm{GEV}(\mu, \sigma, \xi)$,
3. Az extremális index segítségével határozzuk meg az eredeti éves maximumok extrémérték eloszlásának paramétereit,
4. Határozzuk meg az így kapott $\mathrm{GEV}(\mu_\theta, \sigma_\theta, \xi_\theta)$ eloszlás kívánt kvantiliseit, ill. végpontját.

- Végezzük el a fenti lépéseket többszöri véletlenítés mellett, végül

5. Az egyes lépésekben kapott becslések tapasztalati eloszlását tekintve a becslések átlagával közelítsük a kvantiliseket, a szórást figyelembe véve pedig konfidencia intervallumot konstruálhatunk.

Megjegyzés. Ha a független sorozat maximumaira illesztett GEV(μ, σ, ξ) extrémérték eloszlás végpontja $x_H = \mu - \sigma/\xi$, akkor a θ extremális indexű folyamat maximumaira illesztett GEV$(\mu_\theta, \sigma_\theta, \xi_\theta)$ végpontja (3.8) alapján

$$x_G = \mu_\theta - \frac{\sigma_\theta}{\xi_\theta} = \mu + \frac{\sigma}{\xi}\left(\theta^\xi - 1\right) - \frac{\sigma}{\xi}\theta^\xi = \mu - \frac{\sigma}{\xi},$$

vagyis nem függ az extremális indextől. Tehát az extremális index ill. az összefüggőség az eloszlás végpontját nem befolyásolja, viszont a gyakorlatban érdekes magas értékek meghaladási valószínűségeit igen. Így a véletlenítéses módszer másik mellékeredményeként új módszert kaptunk extrémérték eloszlás végpontjának becslésére a fent leírt módon.

Az extremális index becslésének nehézségei miatt érdemes a θ különböző értékei mellett kiszámolni a kvantilis becsléseket, és megvizsgálni az extremális indexre való érzékenységét. Erre a következő fejezetben láthatunk példát, ahol a fenti módszerek alkalmazását követhetjük nyomon.

4.3.3. Maximum likelihood módszer

A maximum likelihood módszernél ugyancsak a sorozat véletlenítésével konstruált független minta maximumait használjuk. Az alapötlet a maxima módszereknél leírt maximum likelihood becslés. Vagyis vegyük az

$$(\underbrace{M_r^{(1)}, \dots, M_r^{(k)}}_{\mathrm{GEV}(\mu,\sigma,\xi,\theta)}, \underbrace{\tilde{M}_r^{(1)}, \dots, \tilde{M}_r^{(k)}}_{\mathrm{GEV}(\mu,\sigma,\xi)}) \tag{4.4}$$

vektort, mely elemeit független, de nem azonos eloszlásúnak tekintünk. Együttes eloszlásukat a $(\mu, \sigma, \xi, \theta)$ négyes egyértelműen meghatározza. Ezért felírható (4.4) együttes sűrűségfüggvénye, és a négy paramétert maximum likelihood módszerrel becsülhetjük. Bizonyos regularitási feltételek teljesülése esetén profile likelihood módszerrel (lásd. például [11]) konfidencia intervallum konstruálható az egyes paraméterekre, így θ-ra. Kvantilisek a becsült paraméterekből számolhatók.

A módszer előnye, hogy nincs szükség az extremális index előzetes becslésére, azt egyidejűleg becsüljük a másik három paraméterrel. Ezáltal nincs szükség szintek definiálására és az r érték meghatározására. Az r paraméter valójában az a blokkméret amilyen hosszú blokkok maximumainak szeretnénk az eloszlását meghatározni. Mivel ezekre a blokkmaximumokra illesztünk extrémérték eloszlást, és közben becsüljük θ-t is, ezért ebben a θ becslésben benne van r, azonban nincs szükség egy általánosabb θ érték meghatározására, mely már nem függ tőle. Az extremális indexet tehát a stacionárius sorozatok maximumának határeloszlásaként fellépő extrémérték eloszlás paramétereként kezeljük, így a többi paraméterrel együtt becsülhetjük.

A módszer jellegzetessége a véletlenítéssel konstruált sorozattól való függés. Különböző véletlenítéssel kapott független maximumok különböző becsléseket eredményeznek. Ezért érdemes többszöri átkeverés mellett elvégezni a négy paraméter becslését és empirikusan közelíteni a paramétereket. Továbbá kérdéses, hogy hogyan kapcsoljuk össze az egyes lépésekben kapott esetleges konfidencia határokat és a véletlenítéseket.

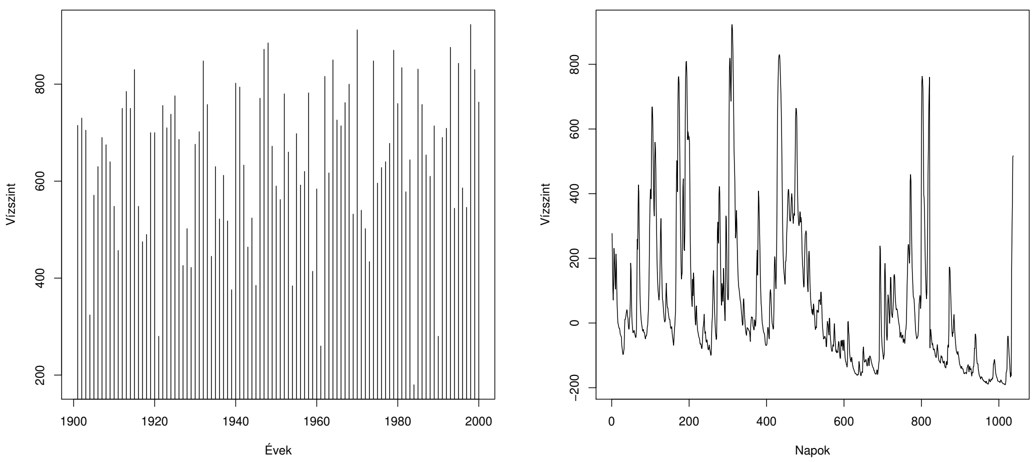
További kérdést vet fel, hogy az alkalmazott módszerrel konstruált független sorozat mennyire jól viselkedik, mint a stacionárius sorozathoz tartozó független sorozat. Egy alternatíva lehet a stacionárius sorozat megritkításával kapott új sorozat, mint független és azonos marginális eloszlású sorozat konstruálása.

Az előzőekben bemutatott véletlenítéses módszerek alkalmazását a következő fejezetben mutatjuk be.

5. fejezet

Alkalmazások

Ebben a fejezetben az eddig leírt becslési módszereket konkrét példára alkalmazva vizsgáljuk. A Tisza folyó Vásárosnaménynél mért vízállás-adatsorát elemezzük, mely 1901-től 2000-ig állt rendelkezésünkre és a napi maximumokat tartalmazza. Az éves maximumok sorozata, ill. az utóbbi három év napi maximumai az 5.1. ábrán láthatóak.



5.1. ábra.
A Tisza folyó Vásárosnaménynél mért éves maximumai 1901-től 2000-ig, és az 1998-2000 évek napi maximumai

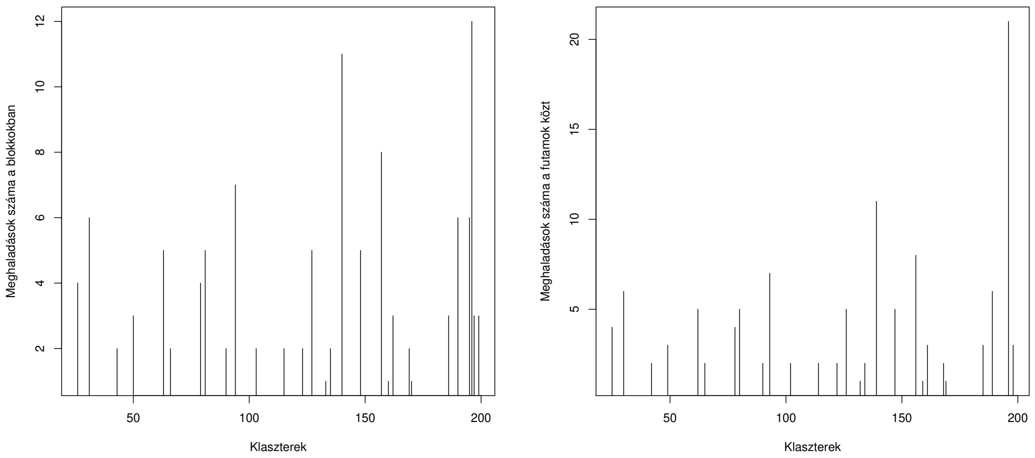
Mivel egy folyó egymást követő napi vízszintértékei összefüggenek, ezért az adatsort nem független megfigyelés-sorozatnak, hanem stacionárius sorozatnak tekintettük. A folyamat fizikai természeténél fogva szezonális ciklus figyelhető meg az adatokban, a trend viszont elhanyagolható. Vizsgálataink azt mutatták, hogy például az extremális index becslését nem befolyásolta jelentősen a ciklus jelenléte, és minthogy számunkra az eredeti adatsor extrémumai a lényegesek, ezért a szezonális komponenst nem távolítottuk el.

Célunk bizonyos visszatérési időkhöz tartozó vízszintek meghatározása, és általánosított extrémérték eloszlás illesztése a vízszint-adatsor éves maximumaira. Ehhez az előző fejezetekben ismertetett módszereket alkalmazzuk és összevetjük a hagyományos eljárásokkal. A különböző számítások elvégzéséhez az $\mathcal{R}$ programcsomagot használtuk.

Elsődleges vizsgálataink a folyamat összefüggőségi struktúrájának feltérképezésére irányultak. Az összefüggőség az extrém értékekben is jelentkezik, tipikusak a több napig tartó magas vízszinttel jelentkező árvizek. Ezért először az extrémumok csoportképző tulajdonságát vizsgáltuk.

Az extremális index becslési módszereinél klasztereket definiáltunk. A blokk módszer esetén a sorozatot r hosszú blokkokra osztva akkor jön létre egy klaszter, ha egy blokkban legalább egy megfigyelés meghaladja az u szintet. A klaszter elemszáma a meghaladások száma a blokkban. A futam módszernél a klasztereket legalább r darab u szint alatti megfigyelés választja el egymástól, a klaszterek elemszáma pedig a meghaladások száma két futam között.

Ezeket a klaszter-definíciókat követve elemeztük az extrémumok csoportképző tulajdonságát. A 100 éves adatsor klasztereinek elemszámai $u = 750$ és $r = 180$ mellett az 5.2 ábrán láthatóak. Ezen r érték mellett legfeljebb 200 klaszter fordulhat elő, ezeket az idő függvényében ábrázoltuk. Látható, hogy a blokk módszer klaszterei előfordulnak a futam módszer klaszterei között is, de az egymáshoz közeli klaszterek ott összeadódnak a futamoknak megfelelően. Összességében azt mondhatjuk, hogy a folyamat extrémumaira nagymértékben jellemző a csoportképző tulajdonság.



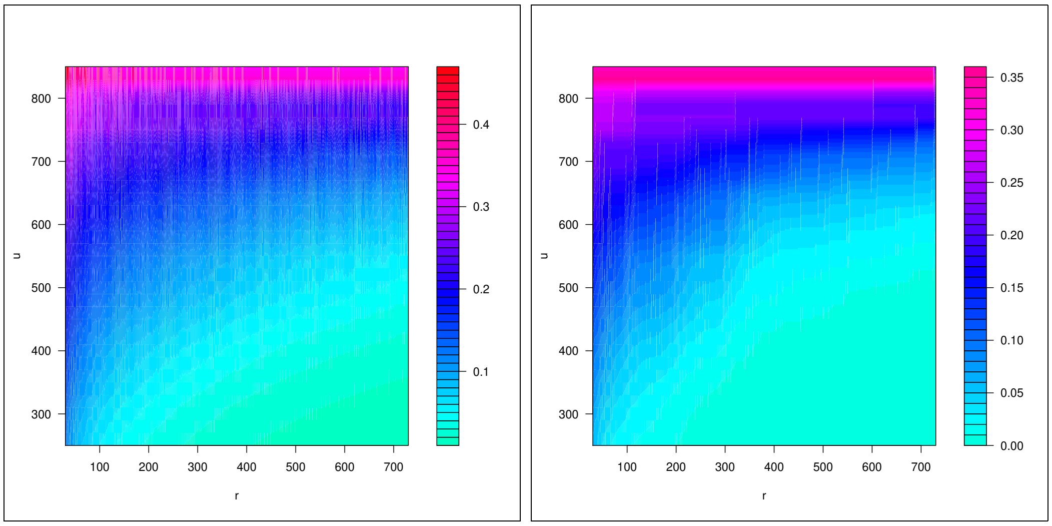
(a) Blokkokkal definiált klaszterek

(b) Futamokkal definiált klaszterek

5.2. ábra.
A Tisza folyó Vásárosnaménynél mért napi maximumainak klaszter-elemszámai $u = 750$ és r=180 mellett

A folyamat összefüggőségi szerkezetének megismerésében további feladatunk az extremális indexének meghatározása volt. Először a blokk és futam módszerekkel való becsléseket vizsgáltuk.

A hiányzó adatok elhagyása után 35885 megfigyeléssel dolgozhattunk. A becslések r paraméterére vonatkozó feltételek miatt r-t 30-tól 730-ig változtattuk. A blokk és futam módszerek u paraméterét úgy kell megválasztani, hogy a folyamat marginális eloszlásának végpontjához tartson. A 100 éves adatsor maximuma 923 cm, és mindössze 15 érték volt 850 cm-nél magasabb, azok 5 blokkban fordultak elő. Ezen meggondolásokból u-t 250 cm-től 850 cm-ig változtattuk. Ilyen paraméterek mellett számoltuk ki a blokk módszerrel kapott kétféle becslést, valamint a futam módszerrel kapott becslést. A blokk módszerből kapott logaritmust tartalmazó becslést $\hat{\theta}_1$, a Taylor-sorfejtésből adódó becslést $\hat{\theta}_2$, míg a futam módszerrel kapott becslést $\hat{\theta}_3$ jelöli. A különböző r és u értékekhez tartozó becsléseket az 5.3. ábra szemlélteti. Az ábrán a vízszintes tengelyen r, a függőleges tengelyen u értékeinek



(a) Blokk módszer (b) Futam módszer

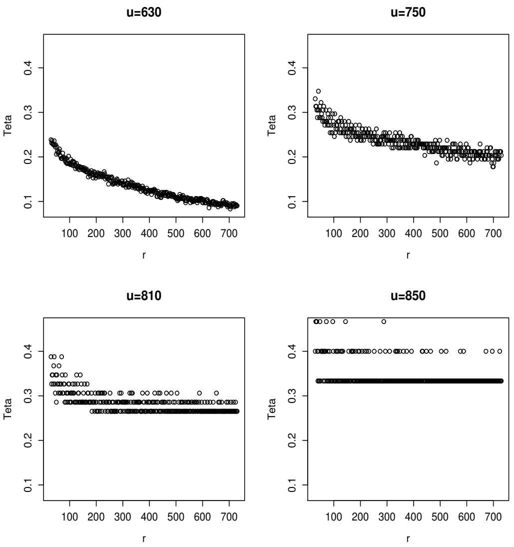
5.3. ábra. Az extremális index becslései különböző r és u értékek mellett

függvényében a különböző becslések láthatók. (Az ábra a blokk módszerrel kapott $\hat{\theta}_2$ és a futam módszerrel kapott $\hat{\theta}_3$ becsléseket tartalmazza.)

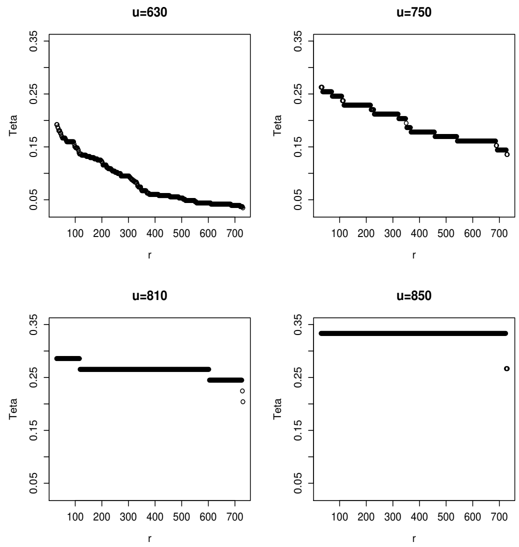
A megfelelő becslést úgy kell kiolvasni, hogy r és u értékét növelve, ezek függvényében nézzük θ értékét. Az extremális index becsléseinél említettük, hogy az r és u paraméterek megválasztása meglehetősen nehéz. A gyakorlatban, azonban bizonyos szint és blokkméret felett a becslések meglehetősen stabilnak tűnnek. Ezt támasztja alá az 5.4. ábra, ahol különböző u értékek mellet r függvényében ábrázoltuk $\hat{\theta}_2$ és $\hat{\theta}_3$ értékeit. (lásd továbbá Függelék)

Alacsony u szint és r érték választása esetén $\hat{\theta}_1$ nem értelmezhető, mert a logaritmus argumentumában 0 áll. Ennek oka az, hogy minden egyes blokkban legalább egy megfigyelés túllépi az u szintet. Azonban mivel a becslést u ill. r értékét növelve kell meghatározni, ez nem jelent problémát.

A grafikonokról is leolvasható, hogy az u és r paraméterek növelésével az extremális index becslései hogyan változnak. Azt tapasztaltuk, hogy a becsléseket nem befolyásolja jelentősen az r paraméter megválasztása, a paraméterekre vonatkozó feltételek szem előtt tartásával könnyen meghatározható az extremális index értéke. Az is látható, hogy mind a blokk, mind a futam



(a) Blokk módszer



(b) Futam módszer

5.4. ábra.

Az extremális index becslései különböző u értékek mellett r függvényében

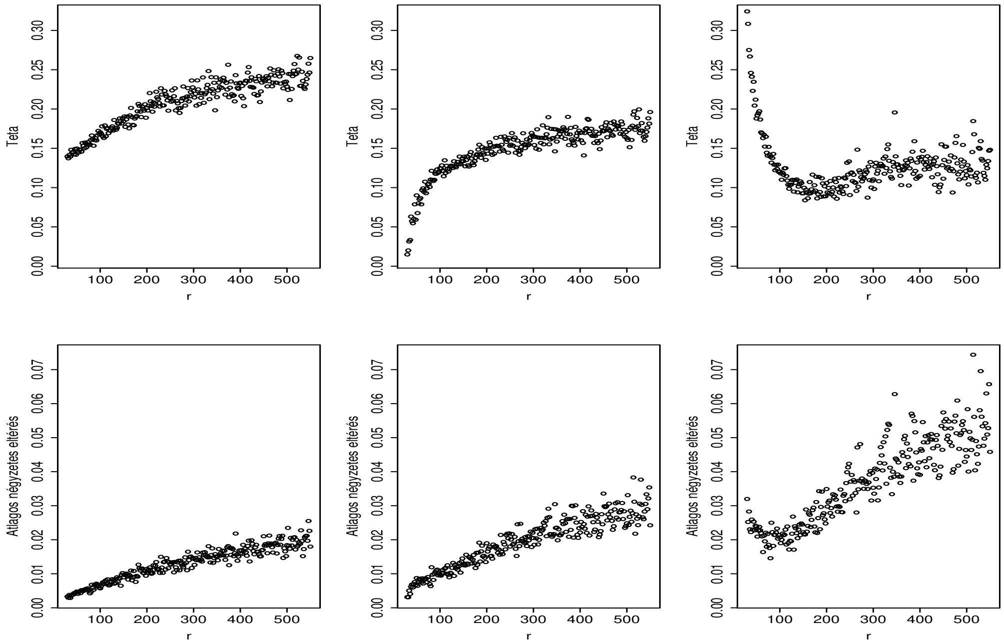
módszer ugyanazt a becslést adja θ-ra.

A blokk módszerrel kapott kétféle becslést és a futam módszerrel kapott becslést, különböző u és r értékek mellett az 5.1. táblázat mutatja.

A fenti vizsgálatok eredményeképpen arra következtethetünk, hogy a Tisza folyó vásárosnaményi szakaszán mért napi maximum vízállások folyamatának extremális indexére $\theta = 0.33$ elfogadható becslés. Az extremális index értelmezése szerint ez azt is jelenti, hogy az igen magas szinteknél jelentkező extrémumok átlagosan hármas csoportokban jelentkeznek.

u=810	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	u=850	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
r=90	0.312	0.306	0.286	r=90	0.336	0.333	0.333
r=250	0.302	0.286	0.265	r=250	0.341	0.333	0.333
r=500	0.296	0.265	0.265	r=500	0.349	0.333	0.333
r=690	0.333	0.286	0.245	r=690	0.350	0.333	0.333

5.1. táblázat. Az extremális index becslései a blokk és futam módszerekkel



5.5. ábra.
Az extremális index becslései és azok átlagos négyzetes eltérései maxima módszerekkel az r paraméter függvényében

Ezek után az extremális index maxima módszerrel való becslését vizsgáltuk. Az r paraméter értékét 30-tól 550-ig változtattuk, és minden esetben 50-szer végeztük el a véletlenítést. Ezután a kapott becslések átlagával közelítettük θ-t. A különböző r hosszú blokkokra illesztett extrémérték eloszlás alakparaméterét a maxima módszereknél leírtak szerint háromféleképpen becsültük, majd azokból számítottuk az extremális indexet. A két eloszlás eltolás– és skálaparaméteréből számolt alakparamétert használó becslés $\hat{\theta}_4$, az eredeti folyamat blokkmaximumainak alakparaméterével számított becslés $\hat{\theta}_5$, a független sorozat maximumainak alakparaméteréből kapott becslés $\hat{\theta}_6$. A becsléseket és azok átlagos négyzetes eltéréseit r függvényében az 5.5. ábra mutatja. A három különböző ξ becslésből adódó extremális index-becslést és zárójelben azok átlagos négyzetes eltérését különböző r értékek mellett az 5.2. táblázat tartalmazza.

	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$
r=100	0.170 (0.0067)	0.126 (0.0110)	0.124 (0.0223)
r=300	0.219 (0.0143)	0.159 (0.0199)	0.100 (0.0354)
r=550	0.262 (0.0245)	0.198 (0.0377)	0.155 (0.0657)

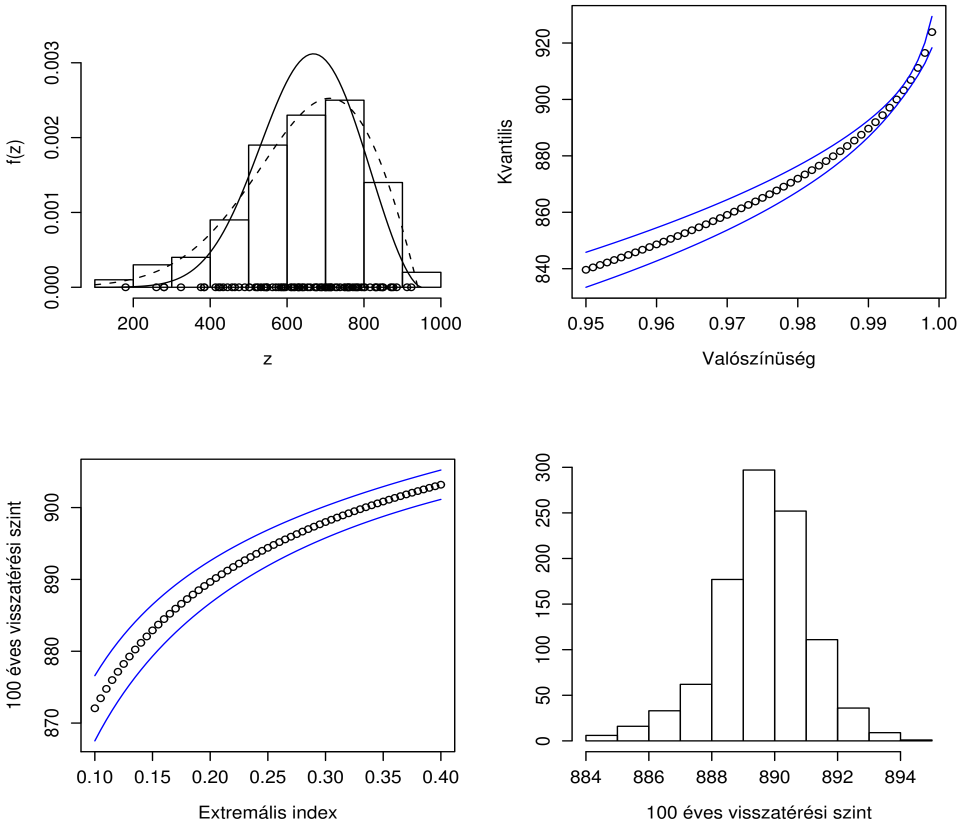
5.2. táblázat. Az extremális index becslései maxima módszerekkel

Tudjuk, hogy a maxima módszerek előnye, hogy a becslés nem függ az u szint megválasztásától. Azonban látható, hogy a becslés nagymértékben függ az r paraméter megválasztásától, valamint látjuk, hogy a háromféle becslés meglehetősen eltér egymástól. A becslések átlagos négyzetes eltérése az r paraméterrel nő. Összességében pedig azt tapasztaltuk, hogy a maxima módszerrel kapott extremális index-becslések kisebbek, mint a hagyományos módszerekkel kapott eredmények. Ancona-Navarette és Tawn [8] szerint ennek a sorozat nem-stacionaritása lehet az egyik oka, ami a mi esetünkben a szezonális komponens jelenléte miatt fordulhat elő.

Az extremális index maximum likelihood módszerrel történő becslésére itt nem térünk ki, ennek majd extrémérték eloszlás illesztésénél lesz jelentősége.

Az extremális index becslése után általánosított extrémérték eloszlást illesztettünk az éves maximumok 100 éves adatsorára. A negyedik fejezetben ismertetett véletlenítéses eljárásokat alkalmaztuk.

Az adott extremális indexet használó módszernél θ-t 0.05 és 0.3 között változtatva vizsgáltuk az eloszlás illeszkedését és a kvantilisek érzékenységét az extremális indexre. A független maximumok GEV paramétereit 1000 véletlenítést elvégezve becsültük. Minden egyes lépésben maximum likelihood módszerrel becsültük a (μ,σ,ξ) paramétervektort, majd az extremális index segítségével határoztuk meg az összefüggő maximumok GEV paramétereit. Az egyes lépésekben kapott paraméterbecslések tapasztalati eloszlása a véletlenítést tekintve normálisnak mondható (lásd Függelék 6.3. ábra), ezért azok átlagával közelítettük a paramétereket. A véletlenítést alkalmazó kvantilis becslési módszert szintén 1000 véletlenítés mellett, a fenti θ értékek mellett alkalmaztuk. Az egyes lépésekben kapott kvantilisek tapasztalati



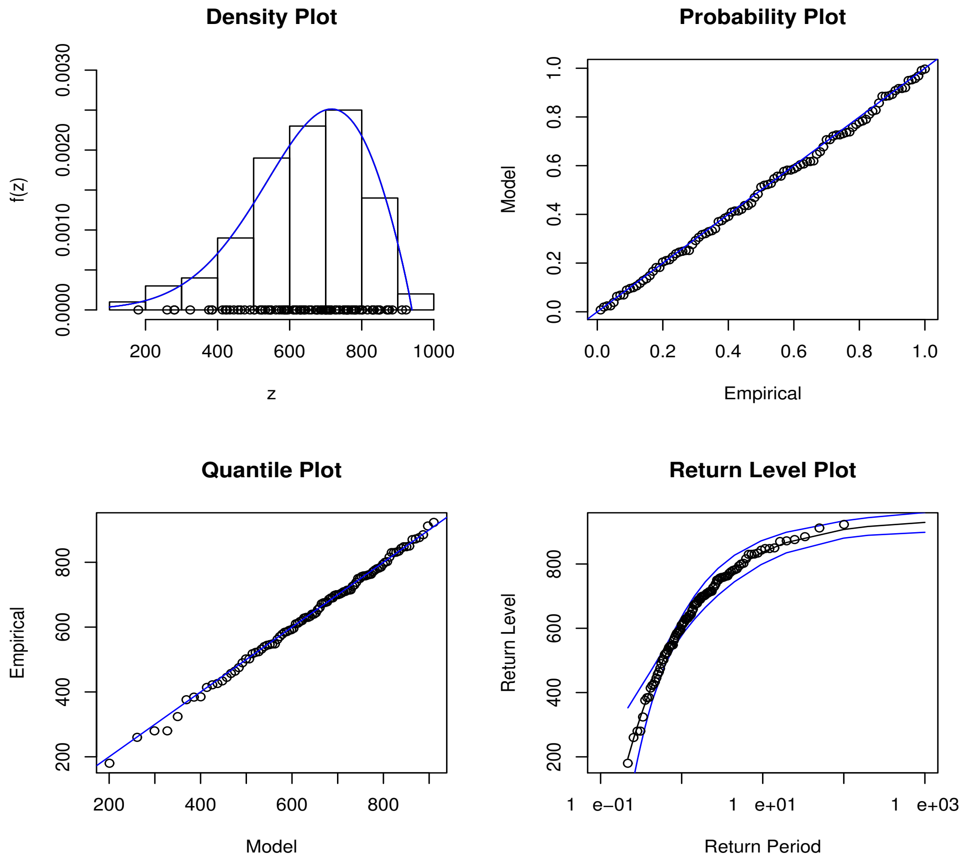
5.6. ábra.

Véletlenítéses módszerrel kapott általánosított extrémérték eloszlás illeszkedése és különböző kvantilisek

eloszlása normálisnak tételezhető fel, konfidencia intervallumot a szórásból számoltunk. Az 5.6. ábra a véletlenítéses módszerrel kapott különböző eredményeket mutatja. A bal felső ábrán a $\theta = 0.2$ extremális index melletti extrémérték eloszlás (folytonos) és a klasszikus módszerrel becsült extrémérték eloszlás (szaggatott) illeszkedése látható. Mellette a különböző valószínűségekhez tartozó kvantilisek és konfidencia határaik vannak feltüntetve ugyancsak $\theta = 0.2$ esetén. A következő grafikon a 99%-os kvantiliseket ábrázolja az extremális index függvényében, végül a véletlenítések során kapott 99%-os kvantilisek hisztogramja látható 0.2 extremális index mellett.

A kapott extrémérték eloszlásokat vizsgálva a fenti θ értékek függvényében, a legjobb illeszkedést 0.2 körüli értékekre kaptuk, amely megfelel a maxima módszereknél kapott extremális index-becslésnek. A kvantilisek becslésének jellegzetessége, hogy a 100 éves visszatérési időhöz tartozó érték körüli konfidencia intervallum a legszűkebb. Ennek magyarázata az, hogy a véletlenítés során kapott független sorozat éves maximumai közt mindig szerepel a teljes adatsor maximuma, ami adott esetben éppen a 100 éves maximum. Így az egyes véletlenítésekkel kapott független GEV eloszlások ezen kvantilisei körül lesz a legkisebb a szórás. Ennek oka tehát a véletlenítéssel konstruált független sorozat jellegzetessége, ezért célszerű lehet más becslés a független maximumok marginális eloszlására. A kvantilisek érzékenységét vizsgálva az extremális indexre, nagy eltérések nem tapasztalhatók a reálisan előfordulható θ értékek esetén, de ezek figyelembe vétele mindenképpen szükséges. A kvantilisek eloszlása normálisnak mondható (5.6. ábra), így az egyes lépésekben kapott értékek átlagát vettük. Összességében azt mondhatjuk, hogy a fenti eljárást alkalmazva olyan modellt illesztettünk az éves maximumokra, amelyből számított kvantilisek kisebbek, mint a klasszikus módszerrel kapott értékek. A különböző visszatérési időkhöz tartozó vízszinteket más módszerekből kapott eredményekkel egybevetve a dolgozat végén prezentáljuk.

Végül a véletlenítéses eljárásoknál ismertetett maximum likelihood módszert alkalmaztuk az éves maximumok eloszlásának meghatározására. Mivel itt egyszerre becsüljük a $(\mu, \sigma, \xi, \theta)$ paramétervektort, ezért az extremális index maximum likelihood becslését is megkapjuk az adott r esetén. Azonban mivel a θ paramétert éppen az adott r nagyságú blokkmaximumok eloszlásának illesztésére szeretnénk felhasználni, ezért nincs szükség annak az r-től független becslésére. A négy paraméter becsléséhez 1000-szer végeztük el a véletlenítést, majd az egyes lépésekben kapott becslések átlagát vettük. A véletlenítések során kapott paraméterek és kvantilisek tapasztalati eloszlása nem mondható szimmetrikusnak (lásd Függelék 6.4. ábra), ezért a mediánnal való közelítést is megvizsgáltuk. Azt tapasztaltuk azonban, hogy az átlagból kapott értékek jobban illeszkedtek az empirikus értékekre.



5.7. ábra.

Extrémérték eloszlás illeszkedése és a visszatérési görbe, a $(\mu, \sigma, \xi, \theta)$ paramétervektor együttes becsléséből

A kvantilisek konfidencia határait az egyes lépésekben kapott értékek szórásából számoltuk. A kapott extrémérték eloszlás illeszkedése a vásárosnaményi éves maximumokra és a visszatérési görbe az 5.7. ábrán látható. A négy paraméter becsült értéke: $\hat{\mu} = 761.05$, $\hat{\sigma} = 93.82$, $\hat{\xi} = -0.526$ és $\hat{\theta} = 0.3096$. A θ paraméter értéke közelítően megegyezik a blokk és futam módszereknél kapott extremális index-becsléssel. A négy paraméterből számított három extrémérték eloszlás-paraméter közel megegyezik a klasszikus módszerrel becsült három paraméterrel, a kvantilisek között kis eltéréseket találtunk. A klasszikus módszerrel becsült extrémérték eloszlás illeszkedése

cm	20 év	50 év	100 év	200 év	500 év	X_G
q_1	840 (6.20)	872 (4.50)	890 (2.90)	903 (2.00)	916 (3.70)	949 (19.6)
q_2	867 (32.1)	894 (27.8)	907 (26.9)	917 (27.4)	925 (29.2)	940 (38.9)
q_3	868 (21.6)	895 (21.0)	909 (22.2)	918 (24.1)	926 (26.7)	940 (34.9)

5.3. táblázat. A különböző visszatérési időkhöz tartozó vízszintek becslései

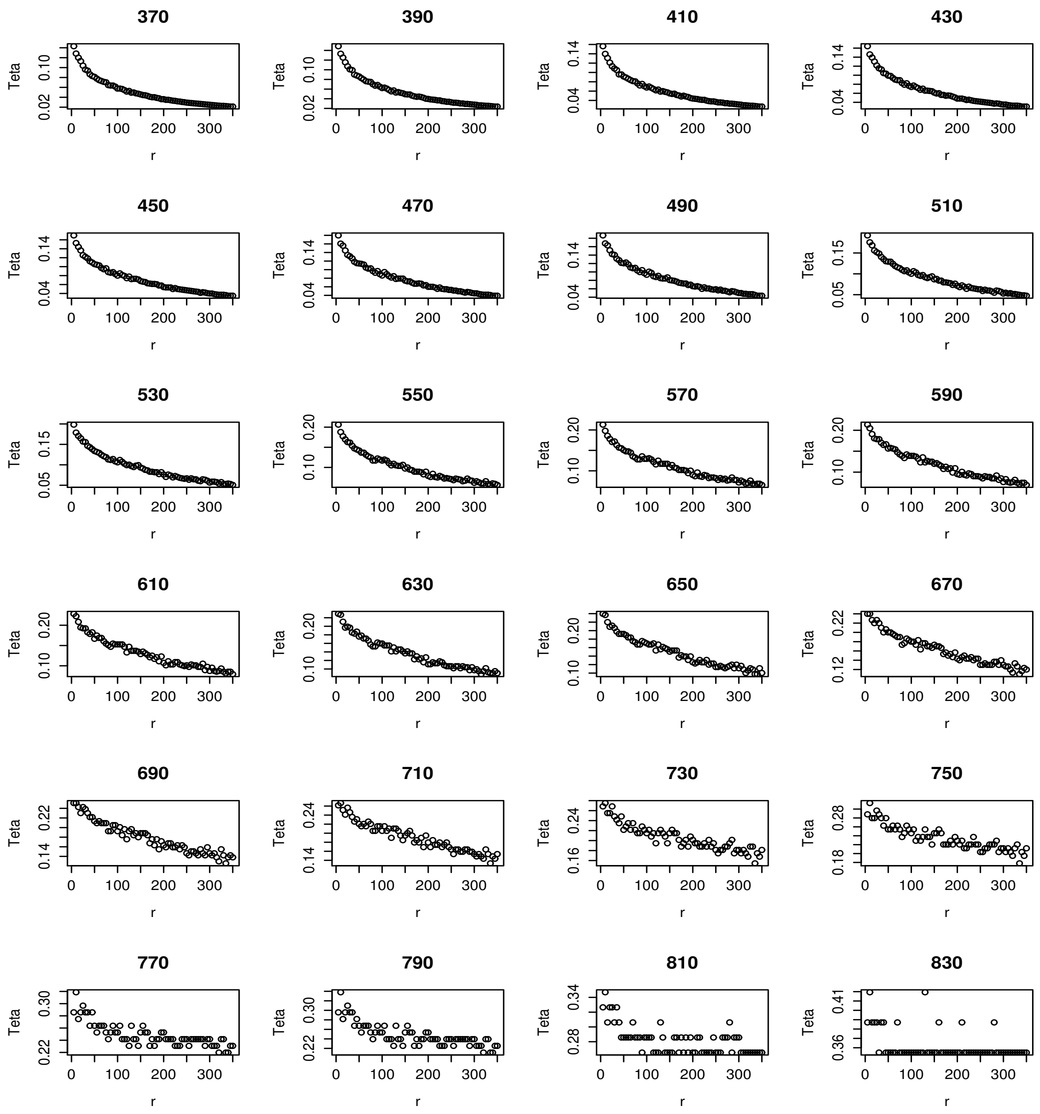
a Függelék 6.6. ábráján látható. Az 6.5. ábrán θ-nak egy adott véletlenítés melletti profile-likelihood becslése látható, valamint a paraméterek átlagából ill. mediánjából kapott eloszlások illeszkedése az éves maximumok hisztogramjára.

A $\theta = 0.2$ extremális indexszel véletlenítéssel kapott kvantilisek (q_1), az előző maximum likelihood módszerrel (q_2), valamint a klasszikus módszerrel becsült kvantilisek (q_3) és azok 95%-os konfidencia határai az 5.3. táblázatban találhatók. A kvantilisek a megfelelő visszatérési idők függvényében vannak feltüntetve, ezen kívül az eloszlások végpontja szerepel a táblázatban. Ha összevetjük a kapott modelleket, láthatjuk, hogy a négyparaméteres maximum likelihood módszerrel kapott modell jól illeszkedik az empirikus adatokra és hasonló eredményt ad, mint a klasszikus módszer. Az adott extremális indexet használó véletlenítéses modell alacsonyabb visszatérési szinteket eredményez, azonban kevésbé jól illeszkedik az adatokra.

Összefoglalva a dolgozatban bemutatott modellek alkalmazását megállapíthatjuk, hogy az összefüggőségek vizsgálata extrémérték modellekben összetett feladat. Rögtön az extremális index becslésénél újabb problémákkal találhatjuk szemben magunkat. Ezen túlmenően azonban talán a gyakorlat szempontjából legfontosabb kérdés az extremális index alkalmazása, használata extrémérték eloszlás illesztésére, kvantilisek meghatározására. Erről a témáról meglehetősen kevés szól a szakirodalomban. A dolgozatunkban bemutatott új modellek fejlesztése, azok további vizsgálatai, valamint újabbak kidolgozása további feladataink közé tartoznak.

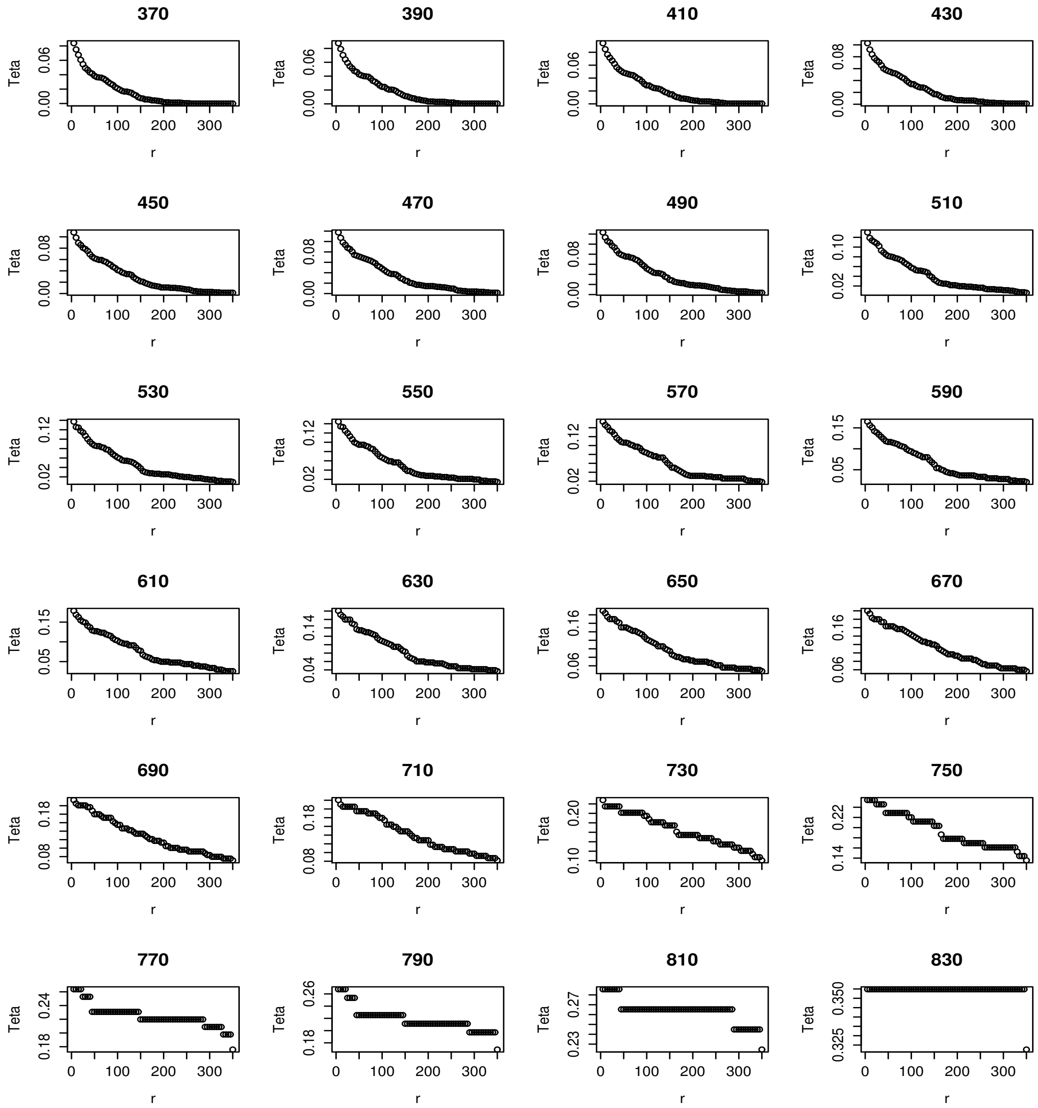
Függelék

A Függelékben azokat az ábrákat sorakoztattuk fel, melyek a dolgozat terjedelménél fogva kimaradtak az előző fejezetből.

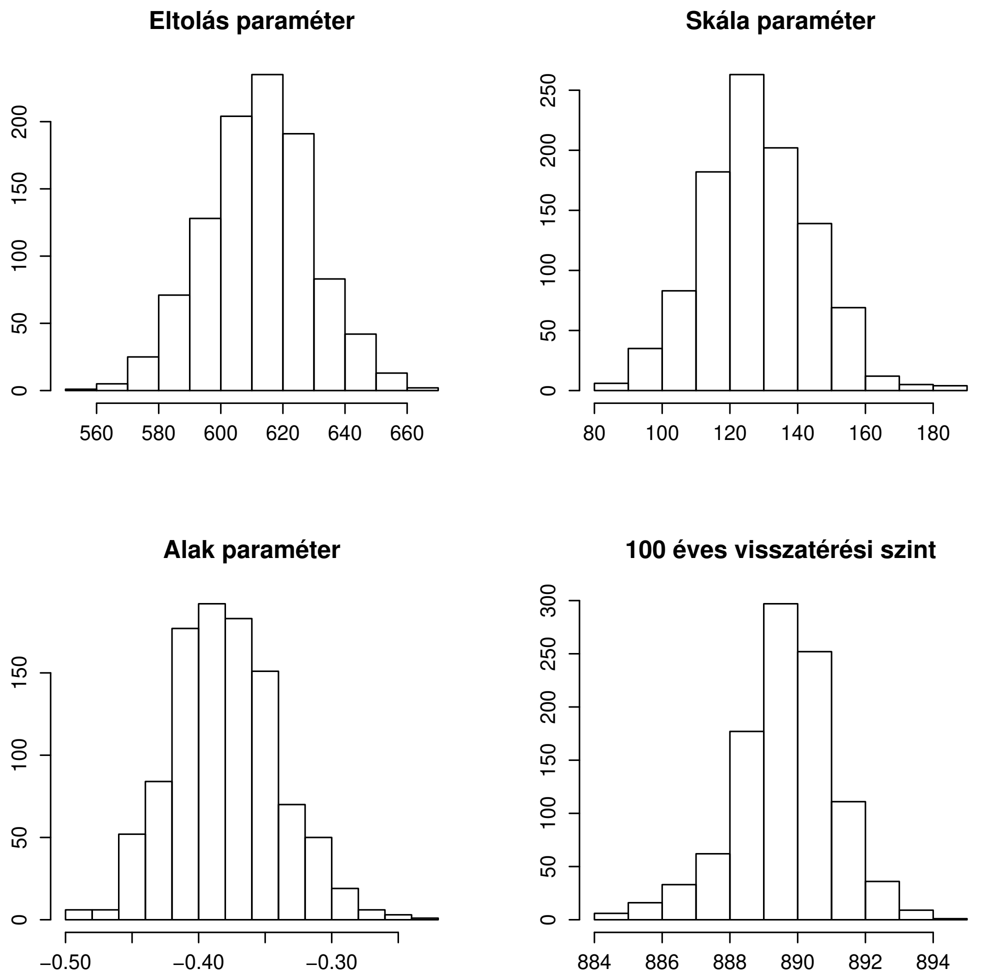


6.1. ábra.

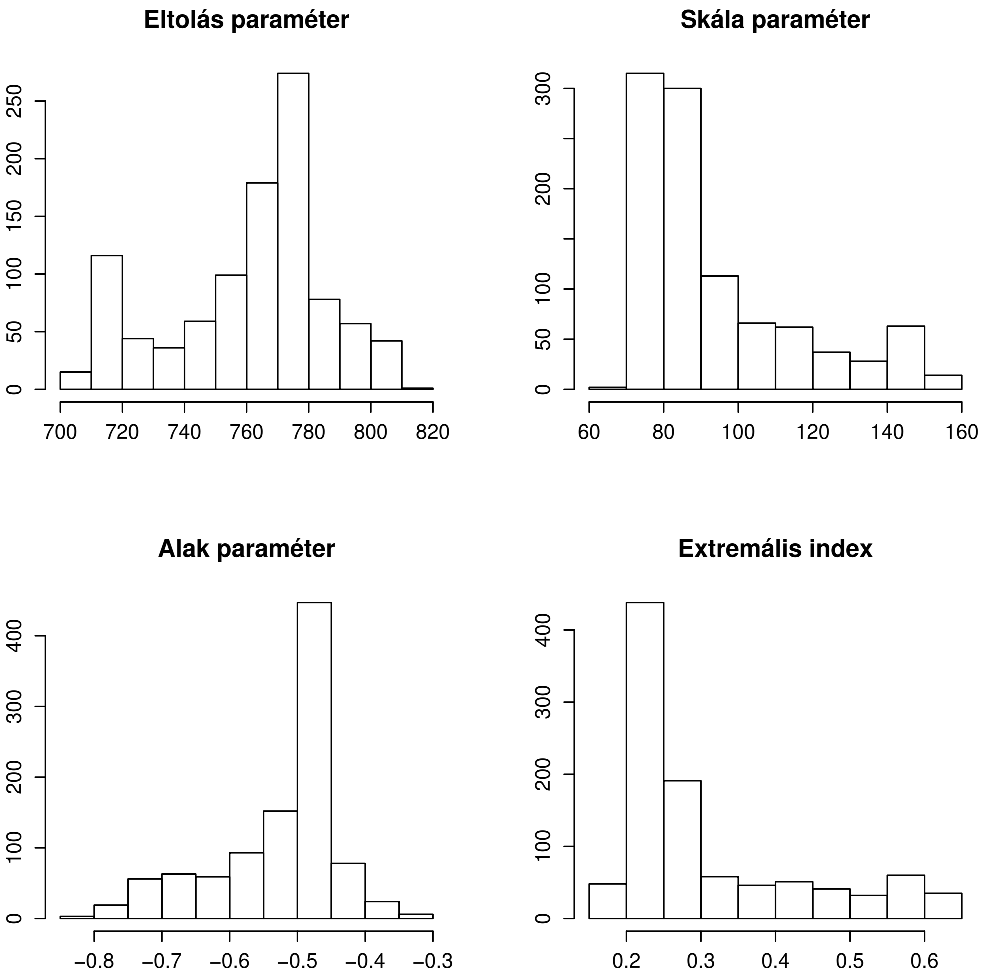
Az extremális index becslése blokk módszerrel különböző u értékek mellett az r paraméter függvényében



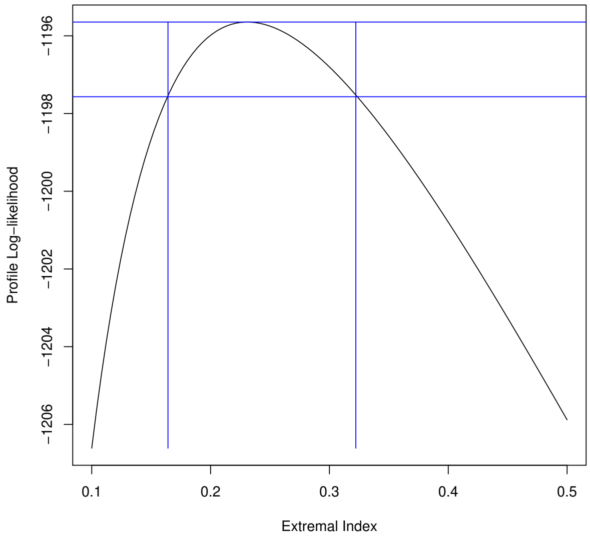
6.2. ábra.
Az extremális index becslése futam módszerrel különböző u értékek mellett az r paraméter függvényében



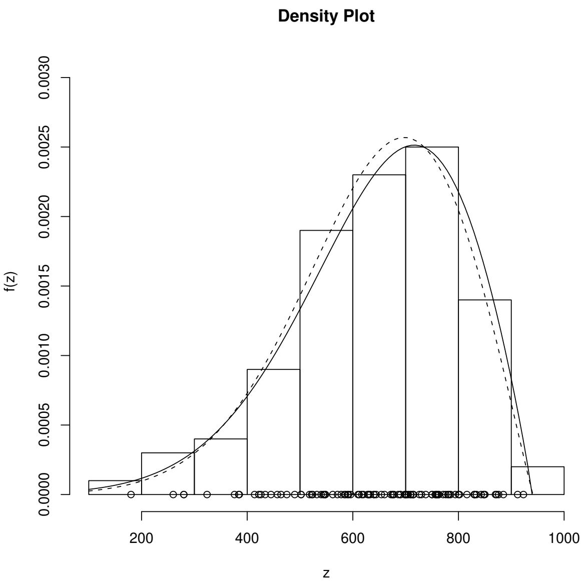
6.3. ábra.
1000 véletlenítés során kapott GEV($\mu_\theta, \sigma_\theta, \xi_\theta$)-paraméterek hisztogramjai 0.2 extremális index mellett, és az ezekből számolt 99%-os kvantilisek tapasztalati eloszlása



6.4. ábra.
1000 véletlenítés során kapott GEV(μ, σ, ξ, θ)-paraméterek hisztogramjai

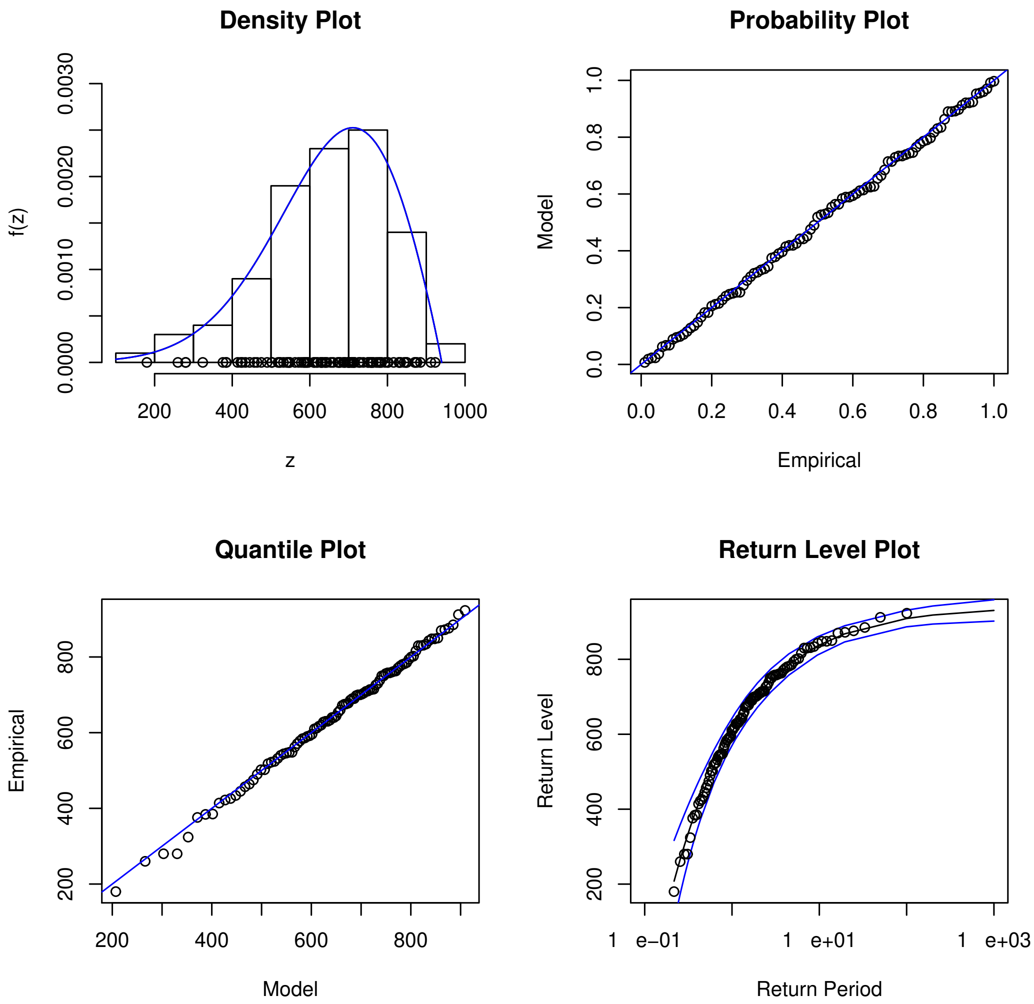


(a) Az extremális index profile likelihood becslése egy adott véletlenítés mellett az éves maximumok esetén



(b) 1000 véletlenítésből becsült eloszlás illeszkedése a paraméterek átlagából számolva (folytonos) ill. a mediánokból kapott paraméterekkel (szaggatott)

6.5. ábra. A négyparaméteres maximum likelihood módszer eredményei



6.6. ábra.

Extrémérték eloszlás illeszkedése és a visszatérési görbe a klasszikus becslési eljárások eredményeként

Irodalomjegyzék

[1] Embrechts, P., Klüppelberg, C., Mikosch, T.: „Modelling Extremal Events for Insurance and Finance”,
Springer (1997).

[2] Leadbetter M.R.: „Extremes and Local Dependence of Stationary Sequences”,
Z. Wahrscheinlichkeitstheorie verw. Gebiete 65, 291-306 (1983).

[3] Leadbetter, M.R., Lindgren, G., Rootzén, H.: „Extremes and Related Properties of Random Sequences and Processes”,
Springer, New York, (1983).

[4] Hsing, T., Hüsler, J., Leadbetter, M.R.: „On the Exceedance Point Process for a Stationary Sequence”,
Prob.Theory and Related Fields 78, 97-112, (1988).

[5] Smith, R.L.: „A counterexample concerning the extremal index”,
Adv.Appl.Probab. 20., 680-683 (1988).

[6] Smith, R.L., Weissman, I.: „Estimating the Extremal Index”,
J.R.Statist.Soc.B 56(3), 515-528 (1994).

[7] O’Brien, G.L.: „Extreme Values for Stationary and Markov Sequences”,
Ann.Probab. 15, 281-291, (1987).

[8] Ancona-Navarette, M.A., Tawn, J.A.: „A Comparison of Methods for Estimating the Extremal Index”, *Extremes 3(1), 5-38, (2000).*

[9] Gomes, M.I.: „On the estimation of parameters of rare events in environmental time series", *Statistics for the Environment 2: Water Related Issues (V. Barnett, K.F. Turkman) John Wiley&Sons, 225-241, (1993).*

[10] Weissman, I., Novak S.Yu.: „On Blocks and Runs Estimators of the Extremal Index",
J.Statist.Planning Inference 66, 281-288, (1998).

[11] Coles, S.: „An Introduction to Statistical Modelling of Extreme Values", *Springer, London (2001).*