

ÁRVIZEK A TISZÁN ÉS NÉHÁNY MELLÉKFOLYÓJÁN EXTRÉMÉRTÉK-MODELLEZÉS A GYAKORLATBAN

BOZSÓ DÁVID – RAKONCZAI PÁL – ZEMPLÉNI ANDRÁS

A tanulmányban bemutatjuk az extrémérték-elemzés módszereit, így különösen a blokkmaximumok vizsgálatánál használatos GEV- (általánosított extrémérték) eloszlásokat és a szint feletti adatok elemzésére alkalmas GP- (általánosított Pareto) eloszlásokat. Ismertetjük a paraméterbecsléshez nélkülözhetetlen maximum likelihood eljárást és a konfidenciaintervallumok konstrukciójára, valamint modellszelekcióra használt profil-likelihood módszert. Az extrémérték modellek illeszkedésének vizsgálatára alkalmas statisztikai próbákat is bemutatunk. Mindezeket felhasználva becsléseket adunk a Tisza vízgyűjtőjéből származó vízállás- és vízhozam adatok alapján az eloszlások kvantiliseire (ezek éppen az adott visszatérési idejű árvizek) és vizsgáljuk az eredményeink időfüggését.

TÁRGYSZÓ: Extrémérték-modellezés. Maximum likelihood becslés. Visszatérési szintek. Illeszkedés-vizsgálat.

Az Eötvös Loránd Tudományegyetem Természettudományi Karának Valószínűség-elméleti és Statisztika Tanszéke aktív résztvevője a „Az árvízi kockázatok meghatározásához szükséges műszaki és tudományos alapok megteremtése, új árvízi gyakorisági- és kockázatbecslési módszerek kidolgozása” c. FKFP (Felsőoktatási Kutatási és Fejlesztési Pályázat) projektnek, melynek koordinátora a Vízgazdálkodási Tudományos Kutató Részvénytársaság (VITUKI Rt). Ennek keretében nagyrészt tiszai, kisebb részben a Tisza mellékfolyóira vonatkozó adatsorokat dolgoztunk fel.

A projektnek külön aktualitást adtak az elmúlt öt évben előfordult árvizek. Így az elemzéseink során arra is megpróbáltunk választ adni, hogy vajon megfigyelhető-e az árvizek magasságának emelkedése az utóbbi évtizedekben. Az elemzések módszere az elméleti tételekkel is alátámasztott extrémérték-modellezés, melynek fő eszköze az extrémérték- (Extreme Value – EV) eloszlások, illetve az általánosított Pareto- (Generalized Pareto – GP) eloszlások illesztése.

A következőkben először az elméleti eredményeket foglaljuk össze a modellektől a becslési módszereken át az illeszkedés vizsgálatáig. Ezután pedig a kapott eredményeket mutatjuk be.

1. MODELLEK

Először a vizsgálat elméleti kereteként szolgáló modelleket, nevezetesen a blokkmaximumok, illetve a szint feletti maximumok módszerét mutatjuk be.

1.1. Blokkmaximumok

Ennek a megközelítésnek nagy előnye az egyszerűség, valamint az, hogy rendelkezésre állnak független azonos eloszlású megfigyelések maximumainak határeloszlására vonatkozó elméleti eredmények. Ezzel indokolható az ún. általánosított extrémérték- (Generalized Extreme Value – GEV) eloszlások alkalmazása az eredeti megfigyeléseinkből képzett blokkok maximumaira. (A tétel első megjelenése *Fisher–Tippett* [1928].)

Tétel (Fisher–Tippett): Legyenek X_1, X_2, \dots, X_n független, azonos eloszlású valószínűségi változók. Ha vannak a_n, b_n normáló konstansok, amelyek esetén igaz az, hogy $[\max(X_1, X_2, \dots, X_n) - a_n]/b_n$ nem-elfajuló határeloszláshoz közelít, akkor ez a határeloszlás szükségképpen max-stabilis vagy ún. extrémérték-eloszlás. A három extrémérték-eloszlás eloszlásfüggvénye:

1. *Frechet-eloszlás*:

$$F_\alpha(x) = \exp(-x^{-\alpha}),$$

($x > 0$), és α pozitív paraméter.

2. *Weibull-eloszlás*:

$$F_\alpha(x) = \exp(-(-x)^{-\alpha}),$$

ahol $x < 0$, és α ugyancsak pozitív paraméter.

3. *Gumbel-eloszlás*:

$$F_\alpha(x) = \exp(-\exp(-x)).$$

Megjegyzendő, hogy nem minden esetben lehet normálni: diszkrét eloszlásokra is oszcillálhat a maximum eloszlása. Folytonos eloszlásokra az eloszlásfüggvény reguláris viselkedése szükséges a felső végpont közelében (teljesül minden fontos eloszlásra). A részletes feltételek megtalálhatók például *Embrechts–Kluppelberg–Mikosch* [1997] monográfiájában, ahol szinte az egész anyagunkkal kapcsolatosan bőségesen található további információk is.

A három extrémérték-eloszlást közös paraméteres formába is írhatjuk:

$$F_\alpha(x) = \exp\left\{-\left[1 + \xi \frac{x - \mu}{\sigma}\right]^{-1/\xi}\right\}, \quad /1/$$

ha

$$1 + \xi \frac{x - \mu}{\sigma} > 0.$$

(Feltéve, hogy $\xi \neq 0$. Az előző paraméterezésnél szereplő α -t most $1/\xi$ alakba írtuk át.) $\xi = 0$ esetén határértékként éppen a Gumbel-eloszlást kapjuk.

Az /1/-ben már az általános esetet írtuk fel, ahol μ az eltolás-, σ skála-, és ξ az alakparaméter. A modellezésnél tehát három paraméter illesztésére van szükség. Ezt az eloszláscsaládot nevezzük általánosított extrémérték-eloszlásnak (*Jenkinson* [1955]).

A becslésekhez feltétlenül javasoljuk konfidenciaintervallumok hozzárendelését is, hiszen ezek alapján tudjuk – legalább közelítőleg – számszerűsíteni a becslésünkben rejlő bizonytalanságot. A konfidenciaintervallumok konstrukciójára a 2. pontban térünk vissza.

A gyakorlatban az évi maximum értékekre illeszthetjük a GEV-eloszlást, feltételezve, hogy alkalmazható rájuk a tétel, azaz tekinthetők független azonos eloszlású megfigyelések maximumainak. Bár ezen tanulmány kereteit meghaladja, megjegyezzük, hogy a tétel állítása a függetlenségnél gyengébb feltételek teljesülése esetén is bizonyítható (például *Leadbetter–Lindgren–Rootzen* [1983]). Ugyanakkor az évi maximumok használatával elveszíthetünk fontos információkat, hiszen ha egy évben például két vagy több jelentős árvíz is volt, akkor azok közül csak a legnagyobb szerepel ebben az elemzésben. Ezért lehet érdekes az alábbi alternatív megközelítés.

1.2. Szint fölötti maximumok

A módszer lényege az, hogy egy magas u küszöb feletti megfigyeléseket tekintve, a kapott meghaladási értékek – a 1.1. pontban leírt feltételek teljesülése esetén – általánosított Pareto- (GP) eloszlással modellezhetők. Ennek eloszlásfüggvénye:

$$P(X - u < y \mid X > u) \approx 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \quad /2/$$

ha $y > 0$, továbbá

$1 + \xi y / \tilde{\sigma} > 0$ és $\xi \neq 0$. $\xi = 0$ esetén határértékként éppen az exponenciális eloszlást kapjuk. A Pareto-eloszlás paraméterei az alábbi összefüggés alapján határozhatók meg az előző pontban ismertetett GEV-eloszlás paramétereiből:

$$\tilde{\sigma} = \sigma + \xi(\mu - u)$$

Ha $\xi < 0$, akkor az eloszlás jobb oldali végpontja véges, egyébként pedig végtelen. Ez a módszer, és a legnagyobb elem mellett további mintaelemek alkalmazásának ötlete is először *Pickands* [1975] cikkében jelent meg.

Mivel az eredeti (például naponkénti) megfigyelések természetesen erősen összefüggők, ezért a modellezésnél az u szintet meghaladó blokkokból csak a legnagyobb elemet vesszük figyelembe.

Ez is mutatja, hogy a GP-eloszlások alkalmazásakor több probléma merül fel, mint a 1.1. esetben. Különösen azért, mert az elméleti eredmény duplán aszimptotikus, azaz nemcsak a minta elemszámának, hanem a küszöbnek is végtelenhez kell tartania. A célszerű kompromisszumot jelentő szint (ahol már elég jó közelítéssel érvényes az aszimptotikus eredmény, de a szórás még nem túl nagy, tehát kellő számú adaton alapul a becslés) megtalálása nem mindig lehetséges egyértelműen. Ezt a küszöbértéket, aminél nagyobb értékekre már megfelelően illeszkedik az eloszlás, hagyományos illeszkedésvizsgálati eljárásokkal tesztelhetjük (5. pont). A gyakorlatban a becsült értékek értelemszerűen függenek a küszöb választásától. Erre a kérdésre az alkalmazási részben még visszatérünk. Már itt megjegyezzük azonban, hogy adott küszöbérték esetén a becslés maximum likelihood módszerrel (2. pont) könnyen elvégezhető. A pontbecslés mellett GEV-eloszlás illesztéséhez hasonlóan itt is célszerű konfidenciaintervallumok hozzárendelése az egyes becslésekhez

További probléma, hogy milyen távoli csúcsokat tekintünk független megfigyeléseknek. Ezzel a kérdéssel az elemzéseket bemutató részben foglalkozunk. Egy lehetséges, automatikus deklaszterezési eljárást ad *Ferro és Segers [2003]* cikke.

Némi könnyebbéget jelent ugyanakkor az, hogy nem kell bármely két egymást követő megfigyelésnek függetlennek lennie (ez várhatóan nem is teljesül a gyakorlatban), az elméleti eredmény teljesüléséhez elég az is, hogy ha időben elég távol vannak egymástól, hiszen akkor az összefüggés közöttük kicsi.

2. MAXIMUM LIKELIHOOD BECSLÉS

A szakirodalomban leginkább elterjedt módszer a maximum likelihood, melynek lényege, hogy azt a paraméterértéket tekintjük becslésnek, melyre a megfigyelésvektor együttes sűrűségfüggvénye maximális. Az elmélete részletesen kidolgozott, és reguláris esetekben aszimptotikusan hatásos és aszimptotikusan normális eloszlású a becslés. A becslés tulajdonságai nemreguláris esetekben természetesen mások; ezekről *Smith [1985]* cikkében olvashatunk részletesen. Ebből kiderül, hogy vannak ugyan a vizsgált eloszláscsaládban irreguláris elemek (ha $\xi < -1$), de ezek igen ritkán fordulnak elő a gyakorlatban.

Az 1. pontban vizsgált esetekben nem adható ugyan meg zárt alakban a maximum likelihood becslés, de numerikus maximalizálás révén könnyen megkaphatjuk a sűrűségfüggvény maximumhelyét. Nagy előnye ennek a becslési eljárásnak, hogy az aszimptotikus tulajdonságai ismertek, és így segítségével a pontbecslés mellett konfidenciaintervallumok is konstruálhatók, erről részletesebben a 3. pontban lesz szó.

3. KONFIDENCIAINTERVALLUMOK KONSTRUKCIÓJA

A legáltalánosabban használt módszer a maximum likelihood becslés aszimptotikus viselkedése alapján határozza meg a konfidenciaintervallumokat. Ezúttal azonban mégsem ezt alkalmaztuk, mert úgy tapasztaltuk, hogy a normális határeloszláson alapuló szimmetrikus konfidenciaintervallumok nem a legalkalmasabbak az extrémérték-

elemzésre. Ezt a hátrányt kiküszöböli a következő, ún. profil-likelihood eljárás, mely ugyancsak aszimptotikus, és a likelihood függvény regularitása esetén érvényes.

Legyenek X_1, X_2, \dots, X_n független valószínűségi változók, $l(\theta)$ pedig a log-likelihood függvény. θ a paramétervektor, melyet $\theta = (\theta_i, \theta_{-i})$, alakba írhatunk, azaz különvesszük a vizsgálandó paramétert (θ_i). A módszer elméleti háttérét kissé általánosabban fogalmazzuk meg, megengedve, hogy ne csak egydimenziós paraméterre keressünk konfidenciatartományt: legyen tehát $\theta = (\theta^{(1)}, \theta^{(2)})$ és itt a $\theta^{(1)}$ k dimenziós paramétervektor a vizsgálatunk tárgya.

Tétel: Ha X_1, X_2, \dots, X_n független megfigyelések, a maximum likelihood becslésre vonatkozó regularitási feltételek mellett, nagy n -re

$$2\{l(\hat{\theta}) - l_p(\theta^{(1)})\} \sim \chi_k^2,$$

azaz az ún. devianciastatisztika közelítőleg k szabadságfokú eloszlású, ahol $\hat{\theta}$ a likelihood becslés, $l_p(\theta^{(1)})$ pedig az ún. profil-likelihood értéke:

$$l_p(\theta^{(1)}) = \max_{\theta^{(2)}} l(\theta^{(1)}, \theta^{(2)}).$$

Következmény: Az egydimenziós esetben θ_i -re a

$$C_\alpha = \{\theta_i : 2\{l(\hat{\theta}) - l_p(\theta_i)\} \leq c_\alpha\}$$

$(1-\alpha)$ megbízhatósági szintű konfidenciaintervallum, ahol c_α a χ_1^2 eloszlás $(1-\alpha)$ -kvantilise.

Ez az eljárás egymásba ágyazott modellek esetén alkalmas a modellválasztásra is, az alábbiak szerint. Legyen M_1 modell θ paramétervektorral, M_0 modell pedig ennek része: M_0 esetén a θ paramétervektor első k komponense 0. Legyenek $l_1(M_1)$, illetve $l_0(M_0)$ a maximalizált log-likelihood értékek és

$$D = 2\{l_1(M_1) - l_0(M_0)\} \quad /3/$$

az eltérésstatisztika (devianciastatisztika). Az M_0 modellt a szignifikanciaszinten elutasítjuk az M_1 modellel szemben, ha $D > c_\alpha$, ahol c_α a χ_k^2 eloszlás $(1-\alpha)$ kvantilise.

4. VISSZATÉRÉSI SZINTEK

A visszatérési szintek a gyakorlati szempontból talán legfontosabb számított mennyiségek. Ha megkaptuk az évi maximum-adatsorra illesztett paraméteres modellt, akkor ennek a magas kvantilisei (azaz azok az értékek, amelyeket adott, kicsi p valószínűséggel halad meg az adott eloszlás) úgy is tekinthetők, mint olyan értékek, amelyeket meghaladó árvíznek az első bekövetkezése $1/p$ év múlva várható. Az éves maximumoknál például a

99/100-os kvantilis (a 99. percentilis) az az érték, amelynél nagyobbat várható értékben száz év múlva kapunk. Megjegyzendő, hogy ebben az esetben is, a geometriai (Pascal-) eloszlás tulajdonságaiból adódóan, $\frac{1}{2}$ -nél nagyobb annak a valószínűsége, hogy az adott szintet meghaladó első árvíz a konkrét időpont (például száz év) előtt bekövetkezzék.

A vizsgált két modell esetében ezek a visszatérési szintek a következőképp számíthatók.

4.1. GEV-eloszlások

Az 1. pontban bemutatott GEV- (általánosított extrémérték) eloszlások $(1-p)$ kvantilisei:

$$z_p = \mu - \frac{\sigma}{\xi} (1 - y_p^{-\xi}),$$

ha $\xi \neq 0$, és

$$\mu - \sigma \log y_p,$$

ha $\xi = 0$, ahol

$$y_p = \log(1-p).$$

4.2. GP-eloszlások

A szint fölötti meghaladásokra alkalmazható GP (általánosított Pareto) -eloszlás $(1-p)$ kvantilise az alábbiak szerint kapható meg:

$$x_p = u + \frac{\sigma}{\xi} \left[\left(\frac{1}{p} \right)^\xi - 1 \right],$$

ha $\xi \neq 0$,

$$x_p = u + \sigma \log \left(\frac{1}{p} \right),$$

ha $\xi = 0$. Itt is, mint korábban, u az a küszöb, ami feletti adatokra a GP-eloszlást illesztettük. Ha az n elemből álló, m év alapján kapott adatsorra általánosított Pareto-eloszlást illesztettünk, akkor ennek az eloszlásnak az $1-1/k$ kvantilise (tehát az az érték, amelyet egy véletlenszerűen kiválasztott mintaelem $1/k$ valószínűséggel halad meg) úgy

is tekinthető, mint az az érték, amelyet először várhatóan $\frac{km}{n}$ év múlva fog meghaladni az árvíz. (Coles, [2001])

5. ILLESZKEDÉSVIZSGÁLAT

Számos klasszikus statisztikai teszt foglalkozik az illeszkedésvizsgálattal. Az egyik legrészletesebb mű a témában *D'Agostino* és *Stephens* [1986] munkája. Az ebben a könyvben bemutatott tesztek természetesen itt nem részletezzük, csak megemlítjük legfontosabb tulajdonságaikat.

A Kolmogorov–Szmirnov-próba, mely az elméleti és a tapasztalati eloszlásfüggvény közötti maximális eltérést vizsgálja, pontosan specifikált nullhipotézis vizsgálatára alkalmas. A mi esetünk becsléses, hiszen csak az eloszlás típusát képzeljük adottnak. Ekkor szimulációval lehet meghatározni a közelítő kritikus értékeket, de a próba közismerten nem túl erős, így mi ezúttal nem alkalmaztuk.

A χ^2 -próba talán a leggyakoribb, illeszkedésvizsgálatra használt eljárás. Előnye, hogy a becsléses esetre is alkalmazható, és noha aszimptotikus, az adatok mennyisége az árvízi elemzéseknél nem lehet akadálya alkalmazásának. Ugyanakkor ez, hasonlóan a Kolmogorov–Szmirnov-próbához, nem érzékeny az extrémumoknál megfigyelhető esetleges eltérésekre. Ráadásul a χ^2 -próbánál az osztályok meghatározása nem egyértelmű, így az eredmények sem azok.

A klasszikus statisztikai tesztek közül a mi céljainknak leginkább az Anderson–Darling-teszt felel meg, amely az

$$A^2 = \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x)$$

statisztikára épül, ahol F az illesztett eloszlás, F_n pedig a tapasztalati eloszlásfüggvény. Ennek alkalmazása esetünkben azért előnyös, mert az extrémumok esetében megjelenő eltérések nagy súllyal szerepelnek a képletben. Kiszámítása a következő módon történhet:

$$A^2 = -n - \sum_{i=1}^n (2i - 1)(\log z_i + \log(1 - z_{n+1-i})) / n,$$

ahol $z_i = F(X_i)$ (X_i a rendezett minta i -edik eleme). A próba aszimptotikus kritikus értékeit ismert F esetére a határeloszlásból könnyen meg lehet határozni, *Choulakian–Stephens* [2001] cikkükben a Pareto-eloszlás becsléses esetére is megadták. Ugyanakkor szimulációval a kisebb mintaelemszámokra is megkaphatjuk a kritikus értékeket, melyek függnek az alakparaméter becsült értékétől. Hasonló, egyoldalón érzékeny próbát ismertet *Zempléni* [2004].

Mindenképpen célszerű grafikusán is vizsgálni, hogy a kapott becslés valóban megfelelő-e. Erre a leggyakrabban használt módszer az adatok hisztogramjának és a kapott modell sűrűségfüggvényének összevetése. Ennek korlátja azonban az, hogy a hisztogram függ az osztályok számától és a végpontjaiktól, valamint nehezen ábrázolható kiugró értékek fellépése esetén. Ezért, ezen hagyományos ábrák mellett, érdemes a Q-Q ábra elkészítése is, melyen az n elemű adathalmaz pontjait a tapasztalati eloszlás $\frac{k}{n+1}$

($k=1, \dots, n$) kvantiliseinek tekintjük, és ezeket vetjük egybe a modell megfelelő kvantiliseivel. Ideális esetben ezek a pontok illeszkednek a diagram átlójára. Nagy előnye ennek az ábrázolásnak a hasonló jellegű P-P ábrával szemben, melynél a $\frac{k}{n+1}$ értékekkel az $F(X_k^{(n)})$ értékeket állítjuk szembe, hogy itt különösen szembeötlők az extrémumoknál adódó esetleges eltérések, amelyek a mi esetünkben kiváltképp fontosak (Coles [2001]).

6. AZ IDŐFÜGGÉS MODELLEZÉSE

Az eddigiekben azt tételeztük fel, hogy a mintaelemek azonos eloszlásúak. Amennyiben az adatok hosszú időt fognak át – ahogy ez az árvízi elemzésnél tipikus – jelentős esély van arra, hogy időközben a háttérben zajló fizikai folyamat, és így a mintaelemek eloszlása is megváltozzék.

Általában a legegyszerűbb modellt célszerű választani, mert a túl sok plusz paraméter megbízható becslése nem várható. Most ezeket mutatjuk be a két vizsgált modell esetére.

6.1. A GEV-eloszlás esete

Itt a legegyszerűbb, a szakirodalomban is megtalálható eset az, hogy a helyparamétert az idő függvényében lineárisan változtatjuk:

$$\mu_t = a + b(t - \bar{t}).$$

Ez a paraméterválasztás azért célszerű, mert így az a értéke éppen az átlagos helyparaméter, b pedig az időegységre (év) jutó paraméterváltozás. A modellek közül a 3. pontban bemutatott profil-likelihood teszttel választhatunk: ha nem szignifikáns a loglikelihood függvény maximumának növekménye, akkor azt mondhatjuk, hogy nincs elegendő bizonyítékunk a paraméter ilyen változására.

Az a, b paramétereket az időtől nem függőnek feltételezett σ, ξ paraméterekkel együtt a szokásos maximum likelihood eljárással becsülhetjük. A modell ellenőrzésére a szokásos Q-Q ábrát csak egy kis módosítással tudjuk használni: az $\tilde{X}_t = \frac{1}{\xi} \log(1 + \xi(X_t - \mu_t)/\sigma)$ standardizált változó már minden t -re Gumbel-eloszlású (Coles [2001]), így ezt lehet a szokásos ábrákkal ellenőrizni.

6.2. A GP-eloszlásra épülő modell

Ezúttal azt tesszük fel, hogy a GP-eloszlás alakparamétere az időben nem változik, míg skálaparamétere az idő lineáris függvénye, azaz $\sigma(t) = \alpha \cdot t + \beta$. Mivel tapasztalataink szerint a változás nem nagy mérvű, ezért az a tény, hogy a lineáris függvény negatív értékeket is felvehet, nem okoz problémát a modellezésnél. Szükség esetén természetesen az időfüggés más függvénykapcsolattal is megadható. A skálaparaméter tulajdonsága, hogy adott alakparaméter esetén minél nagyobbak a megfigyelések, annál nagyobb lesz a

skálaparaméter. Tehát ha az igaz, hogy az árvizek növvő tendenciát mutatnak, akkor növekvő lineáris trendet várunk, azaz az α -t pozitívnak sejtjük.

A kérdéses paramétereket itt is maximum likelihood módszerrel becsülhetjük. Ehhez maximalizáljuk a loglikelihood függvényt, ami a következő:

$$l(\sigma(t), \xi) = -n \log(\alpha \cdot t + \beta) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^n \log\left(1 + \frac{\xi Y_i}{\alpha \cdot t + \beta}\right).$$

Ha vizsgálni szeretnénk, hogy mennyire jó ez az illesztés, akkor az 5. pontban ismertetett grafikus diagnosztikákat, a P-P- és a Q-Q-ábrát használhatjuk. Az előző pontban leírtakkal összhangban azonban ezúttal is transzformálnunk kell az adatokat a következőképpen:

$$\tilde{Y}_{t_i} = \frac{1}{\hat{\xi}} \log\left\{1 + \hat{\xi} \left(\frac{Y_{t_i} - u}{\hat{\alpha} \cdot t + \hat{\beta}}\right)\right\}.$$

Az így transzformált értékekről megmutatható, hogy ha teljesül a modellünk, akkor ezek standard exponenciális eloszlásúak, függetlenül az időtől, így ezekre a standard exponenciális eloszlás segítségével már el tudjuk készíteni a P-P-, illetve a Q-Q-diagramot (Coles [2001]).

A trend szignifikanciavizsgálatát, amely két egymásba ágyazott modell összehasonlítását jelenti, a korábban [3] már definiált devianciastatisztikával értékelhetjük.

Tudjuk, hogy ha a két modell ugyanolyan jól írja le az adatokat, a próbastatisztika aszimptotikusan χ_1^2 eloszlású, tehát akkor van szükség a lineáris trendre, ha a D értéke a k szabadságfokú χ^2 eloszlás $1-\alpha$ kvantilisénél nagyobb. A modellhez tartozó p -érték β , ha $D=Q(\beta)$, ahol $Q(\beta)$ a k szabadságfokú χ^2 eloszlás $1-\beta$ kvantilise.

7. FÜGGETLENSÉGVIZSGÁLAT ÉVI MAXIMUM VÍZÁLLÁS ÉS VÍZHOZAM ADATSOROKBAN

Az adatok függetlenségének vizsgálata azért fontos, mert az összes statisztikai elemzésünk során feltételezzük a megfigyelések függetlenségét. Így van ez az extrém értékek becslésénél és a visszatérési szintek meghatározásánál is.

A függetlenséget is hipotézisvizsgálattal ellenőrizhetjük. Legyen az a nullhipotézisünk, hogy az adatok függetlenek, az ellenhipotézis pedig az, hogy van valami összefüggés közöttük. Ezen hipotézis ellenőrzésére szolgál a χ^2 -próba függetlenségvizsgálatra vonatkozó változata. Ahhoz, hogy ezt a próbát el tudjuk végezni, gyakorisági táblázatra van szükségünk.

Ezt a legegyszerűbb módon úgy állíthatjuk elő, hogy párokat képezünk az adatokból. Minden adatot, az első sorszámú kivételével, párba állítjuk az eggyel kisebb sorszámúval. Így, ha n db adatunk van, $n-1$ párt kapunk. Minden egyes pár meghatároz a síkon egy pontot. A síkot az adatok mediánjának mindkét tengellyel párhuzamosan történő behúzásával négy síknegyedre bontjuk. Megszámoljuk, hogy hány pont esik az egyes síknegyedekbe, és

ezekkel a számokkal készítjük el a gyakorisági táblázatot, amire el lehet végezni a próbát. Mivel 2×2 -es mátrixot kapunk, ezért a próbatisztika egyszerű lesz:

$$n(v_{11}v_{22} - v_{12}v_{21})^2 / (v_{1.}v_{2.}v_{.1}v_{.2}),$$

ahol a v_{ij} az i . sor j -edik eleme, és a nullhipotézis teljesülése esetén aszimptotikusan (nagy n -re) χ^2 -négyzet eloszlást követ, melynek szabadságfoka 1. Ahhoz, hogy el tudjuk dönteni, hogy az eltérés szignifikáns-e, kiszámoljuk a p -értéket. A p -érték annak a valószínűsége, hogy a nullhipotézis teljesülése esetén a statisztika értéke legalább annyi legyen, mint amennyit kiszámoltunk. Ha ez a p -érték kicsi ($p < 0,05$), akkor elutasíthatjuk a nullhipotézist. Ha $p > 0,05$, akkor nem tudjuk elutasítani azt, hogy függetlenek az egymást követő megfigyeléseink. Ez természetesen nem bizonyítja a függetlenséget, de ezekben az esetekben feltételezhetjük annak legalább közelítő fennállását, így a többi módszer alkalmazásakor várhatóan nem követünk el durva hibát.

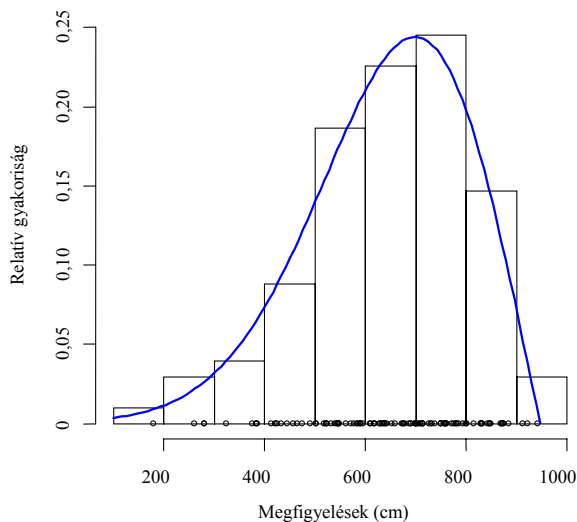
8. AZ EREDMÉNYEK BEMUTATÁSA

A módszerek bemutatása után rátérünk az eredmények ismertetésére. Terjedelmi korlátok miatt csak ízelítőt tudunk adni a kapott eredményekből. Először egyetlen mérőállomásra (Vásárosnaményre) mutatjuk be a legfontosabb alkalmazott módszereket és a kapott grafikonokat, majd a 8.3. pontban az összes eredményt ismertetjük táblázatos formában.

8.1. Vásárosnamény, évi maximális vízállás adatok elemzése

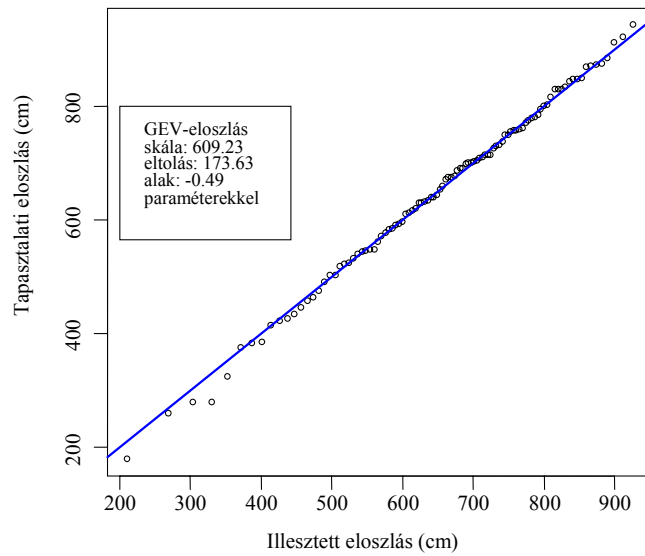
Először tekintsük a GEV-modell eredményeit. Az egyes megfigyelések függetlenségét a 7. pontban bemutatott módszerrel tesztelve nem kaptunk szignifikáns eltérést. Az 1. ábra az éves maximumokat és az azokra illesztett általánosított GEV-eloszlást mutatja.

1. ábra. Évi maximális vízállás adatok és az illesztett GEV-eloszlás Vásárosnaménynél

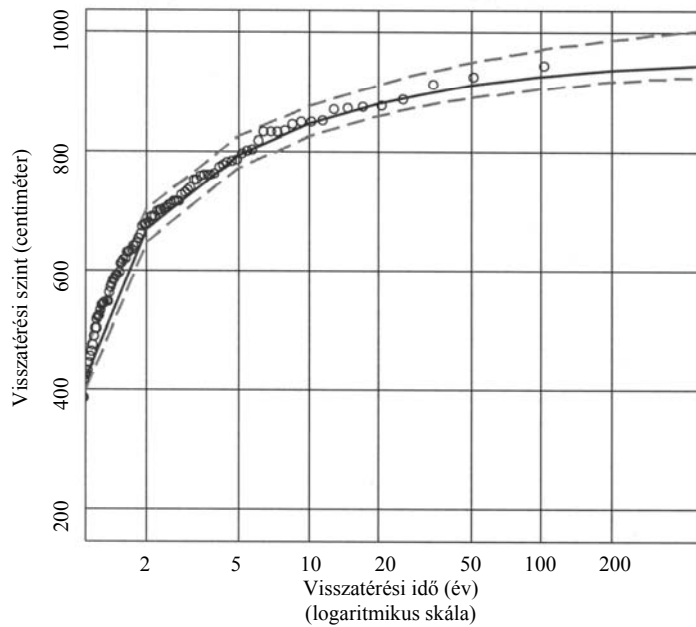


Az illeszkedés szinte tökéletes, amit a 2. ábra Q-Q diagramja is mutat:

2. ábra. Q-Q diagram: évi maximális vízállás adatok és az illesztett GEV-eloszlás (Vásárosnamény)



3. ábra. Az évi maximális vízállás értékeire vonatkozó visszatérési szintek grafikonja (Vásárosnamény)



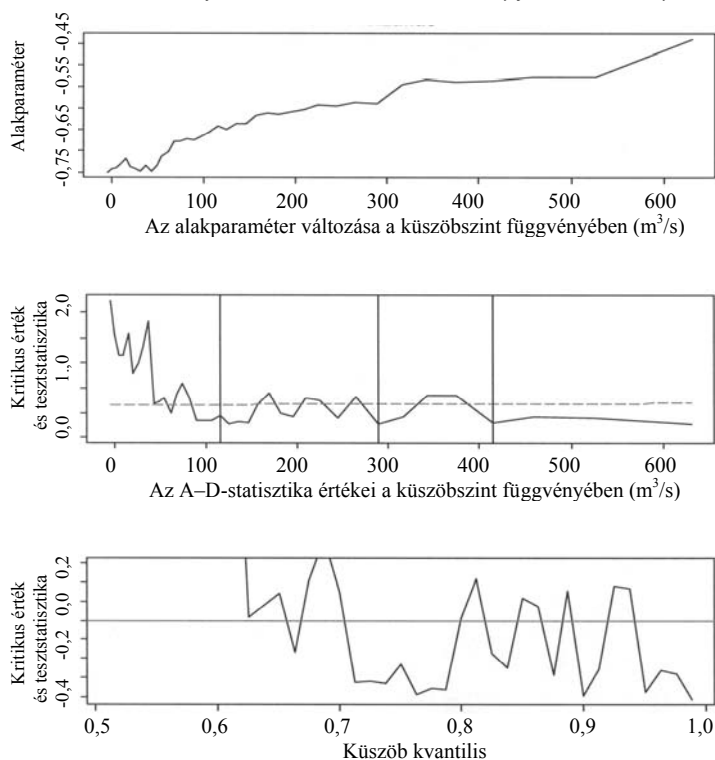
A 3. ábrán a Vásárosnaménynél mért évi maximális vízállás értékeire vonatkozó visszatérési szintgörbét mutatjuk be. A visszatérési szinteket célszerűen logaritmikus skálán ábrázolhatjuk. Tehát az x tengelyen a visszatérési idő logaritmusai változnak, az y tengelyen pedig a becsült visszatérési szintek. Az ábrában a folytonos vonal jelöli magukat az adott visszatérési időhöz tartozó becsült visszatérési szinteket. Az alsó és a felső szaggatott vonal az ehhez a becsült visszatérési szinthez tartozó 95 százalékos megbízhatóságú konfidenciaintervallum alsó és felső határát mutatja. Itt a konfidenciaintervallumot a profil-likelihood módszer segítségével konstruáltuk (3. pont). A rajzon található pontok maguk a megfigyelések, aszerint elhelyezve, hogy például a száz év alatt megfigyelt legnagyobb vízszint megfelel a százéves visszatérési szintnek.

Az ábra talán leglényegesebb tanulsága az, hogy könnyen megeshet: az eddig egyetlen alkalommal megfigyelt legnagyobb árvíz 50 évnél is rövidebb visszatérési idejű (hiszen az 50 évhez tartozó felső konfidenciahatár nagyobb nála).

8.2. Vásárosnamény, a szint feletti adatok elemzése

Az éves maximumok elemzése mellett, mint azt a módszereknél részletesen is megmutattuk, célszerű lehet minden, bizonyos (magas) szint feletti árvíz adatainak elemzése is. Az alábbiakban ezen vizsgálatokból mutatunk be néhányat.

4. ábra. A szint feletti adatok elemzésének néhány jellemző eredménye

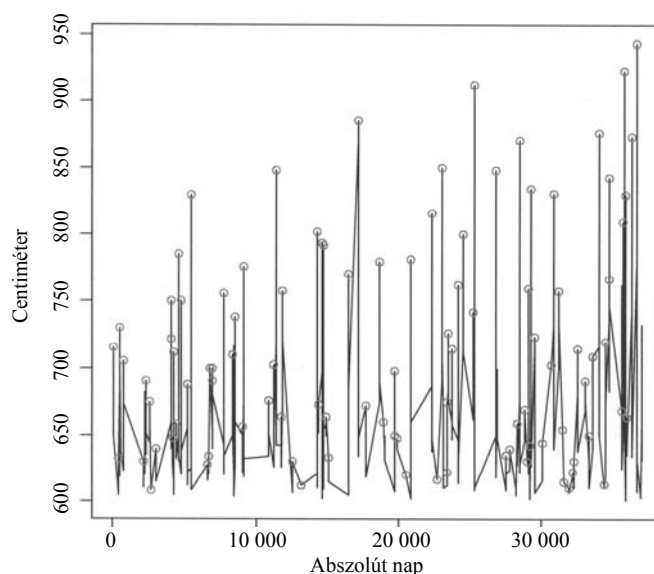


A 4. ábra felső és középső panelje rendre a becült alakparaméter és az Anderson–Darling (A–D)-statisztika értékének változásait mutatja a küszöbszint függvényében. Az alsó panel az A–D-statisztika és a kritikus érték eltérését mutatja a kvantilise függvényében.

A 4. ábráról leolvashatók azok az értékek (az alsó panelről a tapasztalati eloszlás kvantilise függvényében), ahol az Anderson–Darling-féle teszt alkalmazása esetén már elfogadható a Pareto-eloszlásra vonatkozó nullhipotézis. Az elemzéseinket néhány ilyen érték mellett az elsőfokú árvízvédelmi készültségi szinthez tartozó magasságra is elvégeztük.

Az 5. ábrán a Vásárosnaménynél mért, a 600 centiméteres szintet meghaladó napi vízszintadatokat gyűjtöttük ki. Az x tengelyen a megfigyelések ideje látható, abszolút napban, ami annyit jelent, hogy 1901. 01. 01. az első nap, és onnan kezdve számoljuk az abszolút napot. Az ábrán kör jelöli azt a megfigyelést, amit a szint feletti maximumok modelljében használtunk (árvízcsúcsok).

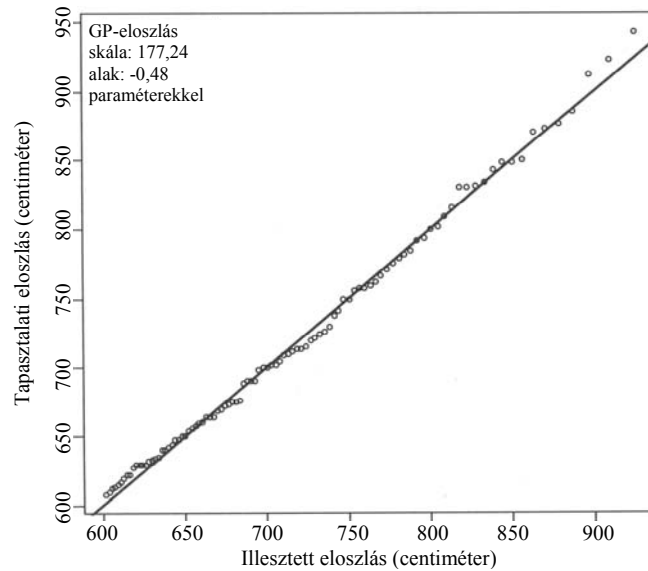
5. ábra. A 600 centiméteres szintet meghaladó napi vízszint adatok Vásárosnaménynél



Az 5. ábrán látható árvízcsúcsokat úgy válogattuk ki, hogy figyelembe vettük az összes, adott szintet meghaladó megfigyelést. Ezek az időtengelyen blokkokat határoznak meg, és két blokkot azonosnak tekintettünk akkor, ha a két blokk között nem telt el 30 olyan nap, amikor a megfigyeléseink végig a szint alatt voltak. Ezeknek a csoportoknak a maximumait tekintettük árvízcsúcsnak (így keletkeztek a kis körök az ábrán). A legmagasabb pontok adatsorát tekintve szembeötlik a felfelé mutató trend, amely mindenképpen figyelmet érdemel.

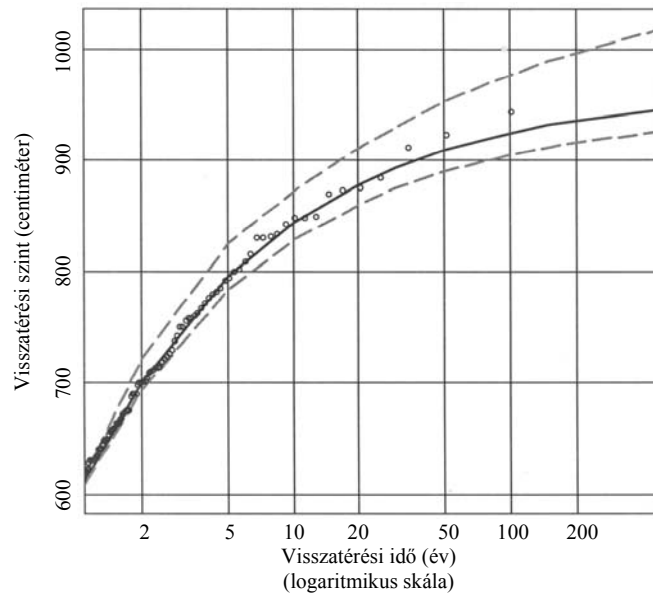
Az 5. ábra kijelölt adataira általánosított Pareto-eloszlást (1.2.) illesztettünk, a paramétereket maximum likelihood módszerrel becslve (2.). Vizsgáljuk meg az illesztést grafikusán!

6. ábra. Q–Q diagram a 600 centiméter feletti árvízcsúcsok általánosított Pareto-eloszlással való illesztéséről



A Q-Q ábra itt is jó illeszkedést mutat. A hisztogramot helykímélés miatt be sem mutatjuk, hanem rátérünk a visszatérési szintekre.

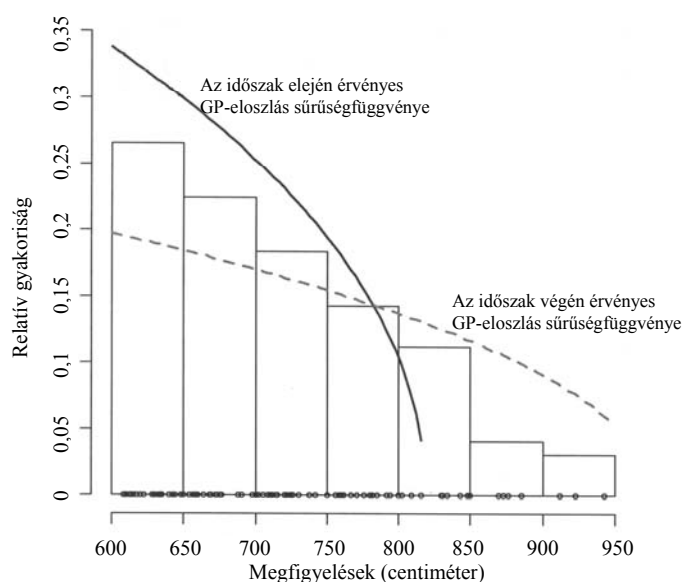
7. ábra. Szint feletti árvízcsúcsokra illesztett, általánosított Pareto-eloszlásból adódó visszatérési görbe (küszöbszint 600 centiméter)



A 7. ábra a szint fölötti árvíz csúcsaira illesztett általánosított Pareto-eloszlásból adódó visszatérési szint görbéjét mutatja. Felépítése teljesen hasonló a 3. ábrához: az időt itt is logaritmusos skálán ábrázoltuk, és feltüntettük az ábrán a 95 százalékos, profil likelihood módszeren alapuló konfidenciaintervallum határait is. A folytonos görbe a becsült visszatérési szint, a szaggatott vonalak az alsó, illetve felső 95 százalékos konfidenciahatárt mutatják.

Már korábban jeleztük, hogy az alkalmazott módszer egyik nehézsége lehet az, hogy a vizsgált folyamat időben nem stabil, így a becsült paraméterek sem rendelkeznek időinvariáns tulajdonsággal. Ezért az időbeli függést külön vizsgálat tárgyává tettük. A 8. ábra ennek a vizsgálatnak néhány eredményét mutatja.

8. ábra. A szint feletti árvízcsúcsok hisztogramja, valamint az időszak elejére és végére becsült eloszlások



Nemcsak a 8. ábrából, de a devianciastatisztika alapján is egyértelműnek tűnik, hogy az idő előrehaladtával egyre gyakoribbak a magas árvizek. Ugyanakkor, ha a vízhozam adatokra is elvégezzük az elemzést, akkor már nem tapasztalunk hasonló tendenciát. Ez arra utalhat, hogy itt (és még néhány más állomás esetén, ahol ugyancsak szignifikáns felfelé mutató tendenciát tapasztaltunk), a meder tulajdonságainak megváltozása eredményezhette a jelenséget.

8.3. Az eredmények összefoglalása

A következő táblákban összefoglaljuk az összes mérőállomásra kapott eredményeket. A közölt eredmények a GP-eloszlásra épülő modelltől származnak, ahol feltételeztük, hogy a skálaparaméter az idővel lineárisan változik. A p -érték megnevezésű oszlop a skálaparaméter trendjének meredekségére vonatkozó teszt p -értékét jelöli. A két utolsó osz-

lop az 1901–1951, illetve az 1952–2002 időszak adatai alapján becült visszatérési szinteket mutatja.

1. tábla

A GP-modellezés összefoglalása vízállásadatokra

Mérőállomás helye	Folyó	Küszöb (centiméter)	Skála- paraméter meredeksége	p-érték	Alak- paraméter	Skála- paraméter	30 éves visszatérési szint	
							az időszak elejéről	az időszak végéről
							(centiméter)	
Tiszabecs	Tisza	100	1,11	0,01	169,80	-0,39	511,52	689,90
Tiszabecs	Tisza	200	1,05	0,01	118,51	-0,35	510,60	685,45
Tiszabecs	Tisza	300	0,94	0,04	72,99	-0,27	516,14	675,40
Garbolc	Túr	150	3,63	0,00	15,47	-0,56	352,19	771,25
Garbolc	Túr	250	2,06	0,02	32,17	-0,45	426,56	688,19
Garbolc	Túr	300	1,82	0,05	15,38	-0,41	439,95	675,46
Ágerdömajor	Kraszna	300	-0,36	0,46	358,15	-0,91	662,28	641,56
Ágerdömajor	Kraszna	500	-0,91	0,10	166,86	-0,58	679,74	610,99
Ágerdömajor	Kraszna	450	-0,83	0,23	201,81	-0,65	674,34	615,08
Csenger	Szamos	200	0,23	0,45	187,85	-0,21	718,40	782,72
Csenger	Szamos	300	0,05	0,85	162,57	-0,18	745,62	761,63
Csenger	Szamos	500	0,69	0,29	122,60	-0,36	702,64	818,17
Vásárosnamény	Tisza	300	1,05	0,00	362,23	-0,69	792,95	938,43
Vásárosnamény	Tisza	500	1,06	0,00	223,26	-0,70	795,84	936,52
Vásárosnamény	Tisza	600	1,04	0,00	147,98	-0,67	796,12	935,18
Záhony	Tisza	150	0,09	0,75	362,47	-0,59	712,45	727,39
Záhony	Tisza	350	0,19	0,50	196,04	-0,47	696,37	730,61
Záhony	Tisza	500	0,17	0,58	106,12	-0,37	693,63	725,17
Polgár	Tisza	400	0,64	0,00	155,76	-0,48	675,58	790,17
Polgár	Tisza	550	0,74	0,00	113,83	-0,70	691,14	783,98
Polgár	Tisza	470	0,67	0,00	133,01	-0,53	680,09	786,26
Szeged	Tisza	500	0,20	0,56	193,17	-0,39	870,09	910,32
Szeged	Tisza	650	0,24	0,53	134,31	-0,42	871,20	911,46

2. tábla

A GP-modellezés összefoglalása vízhozam adatokra

Mérőállomás helye	Folyó	Küszöb (köbméter)	Skála- paraméter meredeksége	p-érték	Alak- paraméter	Skála- paraméter	30 éves visszatérési szint	
							az időszak elejéről	az időszak végéről
							(köbméter)	
Tiszabecs	Tisza	500	1,00	0,71	927,32	-0,28	2853,06	3009,28
Tiszabecs	Tisza	700	1,11	0,69	918,08	-0,32	2840,06	2999,68
Tiszabecs	Tisza	1000	2,94	0,26	781,82	-0,39	2732,57	3093,92
Garbolc	Túr	20	-0,11	0,74	75,09	-0,15	242,05	223,45
Garbolc	Túr	40	-0,11	0,67	55,81	-0,02	245,32	220,32
Garbolc	Túr	60	0,27	0,42	23,81	-0,01	204,84	259,04
Ágerdömajor	Kraszna	35	0,66	0,04	-19,37	0,18	107,79	293,92
Ágerdömajor	Kraszna	50	0,69	0,16	-23,58	0,25	107,15	299,63
Csenger	Szamos	300	4,01	0,00	108,75	0,08	1296,06	2989,18
Csenger	Szamos	400	4,10	0,00	142,18	0,03	1368,19	2784,94
Csenger	Szamos	650	2,49	0,19	204,44	0,09	1710,54	2524,08
Vásárosnamény	Tisza	450	3,61	0,04	999,62	-0,30	2880,27	3767,97
Vásárosnamény	Tisza	900	4,52	0,00	725,68	-0,27	2721,77	3867,96
Vásárosnamény	Tisza	1500	4,50	0,03	609,49	-0,31	2808,19	3775,95
Felsőberecki	Bodrog	450	0,99	0,50	223,03	-0,30	977,60	1098,29
Felsőberecki	Bodrog	500	0,09	0,95	249,20	-0,25	1030,36	1041,95
Polgár	Tisza	1000	2,20	0,66	575,15	-0,02	3468,39	3884,90
Polgár	Tisza	1250	-1,11	0,80	538,53	0,18	4037,70	3737,30
Szolnok	Tisza	1150	0,52	0,72	301,80	0,21	3093,50	3353,47
Szolnok	Tisza	1300	-0,92	0,68	449,78	0,15	3322,37	2979,32

Az adatokból megállapítható, hogy a legtöbb esetben nem szignifikáns az időbeni változás. A néhány 5 százaléknál kisebb p -érték (ezeket emeltük ki sötétebb háttérrel) többnyire a vízállásadatokból kerül ki és növekvő tendenciát mutat. A vízhozam adatoknál még kevesebb helyen szignifikáns a trend, de két esetben: Vásárosnaménynál és Csengernél egyaránt felfelé mutat. Az is megállapítható, hogy a nem szignifikáns változások döntő többsége is felfelé mutat. Nem tűnik alaptalannak tehát az árvizek magasságának emelkedésére vonatkozó tapasztalati megfigyelés, bár az is nyilvánvaló, hogy a vízhozamokra ez a trend sokkal kevésbé egyértelmű, azaz jelentős részben nem a nagyobb és szélsőségesebb csapadék, hanem inkább a meder vízáteresztő képességének változása okozhatja az árvizek magasságának növekedését.

A kapott visszatérési szintek általában jól egybeesnek a különböző szintek esetén, kivételt talán a szolnoki vízhozam adatsor képez.

*

A leírt módszerek felölelik az extrémérték-elemzés témakörének legfontosabb, legegyszerűbb módszereit, melyet a modern szakirodalom ma már rutinszerűnek tekint. Megállapítható, hogy a módszerek kiállták a hazai gyakorlati alkalmazás próbáját, lényegében bármely alkalmas küszöbszint az éves maximumok modelljéhez hasonló eredményt adott, azaz nemcsak a szintválasztás nem, hanem a modellválasztás sem olyan kardinalis kérdés, mint azt a korábbi hidrológiai elemzések sejtették.

IRODALOM

- CHOUŁAKIAN, V. – STEPHENS, M. A. [2001]: Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics*. 43. évf. 478–484. old.
- COLES, S. [2001]: *An introduction to statistical modeling of extreme values*. Springer. London.
- COX, D. R. – HINKLEY, D. V. [1974]: *Theoretical statistics*. Chapman and Hall. London.
- D'AGOSTINO, R. B. – STEPHENS, M. A. [1986]: *Goodness-of-fit techniques*. Marcell Dekker. New York.
- EMBRECHTS, P. – KLÜPPPELBERG, C. – MIKOSCH, T. [1997]: *Modelling extremal events*. Springer. New York.
- FERRO, C. A. T – SEGERS, J. [2003]: Inference for clusters of extreme values. *Journal of the Royal Statistical Society*. Ser. B, 65. évf. 545–556. old.
- FISHER, R. A. – TIPPETT, L. H. C. [1928]: Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of Cambridge Philosophical Society*. 24. évf. 180–190. old.
- JENKINSON, A. F. [1955]: The frequency distribution of the annual maximum (or minimum) values of meteorological events. *Quarterly Journal of the Royal Meteorological Society*. 81. évf. 158–172. old.
- KOTZ, S. – NADARAJAH, S. [2000]: *Extreme value distributions: theory and applications*. Imperial College Press. London.
- LEADBETTER, M. R. – LINDGREN, G. – ROOTZEN, H. [1983]: *Extremes and related properties of random sequences and processes*. Springer. New York.
- PICKANDS, J. III. [1975]: Statistical inference using extreme order statistics. *Annals of Statistics*. 3. évf. 119–131. old.
- SMITH, R. L. [1982]: Maximum likelihood estimation in a class of non-regular cases. *Biometrika*. 72. évf. 67–90. old.
- SMITH, R. L. [1990]: Extreme value theory. In: *Handbook of applicable mathematics, Supplement*. Wiley. Chichester. 437–442. old.
- ZEMPLÉNI A. [1996]: Inference for generalized extreme value distributions. *Journal of Applied Statistical Science*. 4. évf. 2–3. sz. 107–122. old.
- ZEMPLÉNI, A. [2004]: Goodness-of-fit test in extreme value applications. *Discussion paper No. 383, SFB 386, Statistische Analyse Diskreter Strukturen*. TU München.

SUMMARY

In the paper we present the methods of extreme-value analysis, especially the GEV (generalised extreme value) distributions, suitable for analysing block maxima and the GP (generalised Pareto) distributions, applied

to peaks over threshold models. We introduce the maximum likelihood method, needed for estimating the parameters and the profile likelihood method, used for constructing confidence intervals and model selection. We also show statistical tests, suitable for checking the goodness of fit for extreme value models. Based on these methods, we present estimators for the quantiles of the distributions of flood data (both water level and discharge series from the Tisza catchment area are analysed). These quantiles are just the floods with given return periods. We also investigate the time dependence of our results.