Töréspontok keresése meteorológiai idősorokban, és azok hatásainak vizsgálata

Pödör Zoltán
NymE SKK, Informatikai és Gazdasági Intézet
podor@inf.nyme.hu

ÖSSZEFOGLALÓ Az idősorban töréspontként definiálható az a pont, amely mentén az adatsort kettébontva a két részintervallum adatai között statisztikailag is igazolható eltérés mutatható ki. A töréspontok felfedésére több statisztikai eljárás is ismert Az ilyen idősorok közötti kapcsolatok, pontosan a függő vagy független változóban talált töréspont(ok) miatt nem tekinthetőek alapvetően a teljes vizsgálati időszak vonatkozásában stabilnak.

ABSTRACT The point at which the time series is divided in two and differences between the data in the two subintervals can also statistically be confirmed can be defined as a breakpoint. Several statistical procedures are known to detect breakpoints in time series. Such relationships between the time series, exactly because of the breakpoints detected in the dependent or independent variable, may not therefore be considered stable over the whole studied period.

1. Bevezetés

Megfelelő hosszúságú idősorok esetében fennáll annak a lehetősége, hogy a teljes adatsor nem feltétlenül homogén, lehetnek komoly, ugrásszerű változások benne. Ezért az ilyen idősorokat felhasználó elemzések, összefüggés vizsgálatok nem tekinthetőek minden esetben a teljes vizsgálati időszak vonatkozásában stabilnak. Különösen igaz ez a természeti környezetünkkel kapcsolatos vizsgálatokban felhasznált adatokra, hiszen mi magunk is tapasztalhatjuk, hogy időjárásunkban, klímánkban változások zajlanak le. A klimatikus komponensek közül leggyakrabban a hőmérsékletváltozás különböző eseteit szokták vizsgálni, mert ez egy olyan környezeti jellemző, aminek változásait a hétköznapi emberek is közvetlen módon érzékelhetik.

A megfelelő hosszúságú meteorológiai adatsorokban a töréspontok vizsgálatának módszerét egy meteorológiai adatsor, a töréspontok vizsgálatának fontosságát fák éves növekedési adatai és a klimatikus komponensek közötti összefüggéseken keresztül mutatjuk be. Az összefüggés vizsgálatok módszertani hátterének pontos bemutatása, definiálása és az eredmények bemutatása nem része jelen munkának, így erre nem térünk ki. Arra kívánunk rávilágítani, hogy egy-egy ilyen töréspontnak milyen fontos szerepe lehet ezekben a vizsgálatokban, mert sok esetben azok erősségét, de akár irányát is befolyásolhatja.

2. Anyag és módszer

2.1. A felhasznált adatok

Egy mintaterület havi meteorológiai adatait (csapadékösszeg mm-ben és átlag hőmérséklet °C-ban mérve), felhasználva mutatjuk, be a töréspont keresés módszertanát és annak néhány eredményét. A meteorológiai adatok mellé ugyanezen terület éves átlagos fanövekedési adatait hozzávéve és ezek összefüggéseit – a kapott töréspont figyelembe vételével – vizsgálva támasztjuk alá a töréspontok fontosságát az összefüggés vizsgálatokban. A meteorológiai és növekedési adatok 1983-2007 vonatkozásában állnak rendelkezésre, így 25 éves idősorokat vizsgáltunk. Megjegyezzük, hogy a meteorológiai adatok ennél hosszabb intervallumban is rendelkezésre álltak, azonban a növekedési adatok miatt egységesítettük a felhasznált idősorokat.

2.2. Töréspont keresés néhány módszere

Töréspontok kimutatására többféle módszer is ismert a statisztikában [3], mint például

- a szórások minimalizálásának módszere,
- az anomáliák kumulatív összegzése,
- a Pettitt-féle nemparaméteres próba,
- a jel-zaj arány vizsgálata,
- a részátlagok összevetése Student-féle t-próbával.

A módszerek mindegyikének a lényege, hogy a rendelkezésre álló teljes adatsort egy-egy adott év mentén kettébontjuk és az alkalmazott módszertől függő statisztikai próbával, módszerrel vizsgáljuk, hogy az így kapott két részidőszak megfelelő jellemzői között kimutatható-e statisztikailag is igazolható, szignifikáns eltérés. Az idősorok felbontása során figyelemmel kell lenni arra, hogy a kapott részidőszakok hossza egy előre definiált minimum értéket elérjen a statisztikailag korrekt összevethetőség miatt. Ezen minimumérték meghatározása részben függ a rendelkezésre álló adatsor hosszától, részben pedig statisztikai értelemben is megfelelő kell, hogy legyen. A továbbiakban röviden ismertetjük az egyes módszerek lényegét.

2.2.1. Szórások minimalizálásának módszere

Az eljárás lényege, hogy a potenciális töréspontok mindegyikére vonatkozóan meghatározzuk az általuk definiált két-két (1-es és 2-es indexszel jelölve) részidőszak közös szórásnégyzetét:

$$S^2 = \frac{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2}{n_1+n_2-2}. \quad (1)$$

Ezt követően kiszámítjuk a teljes, rendelkezésre álló adatsor szórásnégyzet értékét is, $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ és ezt vetjük össze χ^2 -próba segítségével az aktuálisan vizsgált két részidőszak közös szórásával. Az alkalmazott statisztika

$$\chi^2 = (n-1)\frac{\sigma^2}{S^2} \quad (2)$$

alakú, ahol σ^2 a teljes, rendelkezésre álló időszak, S^2 pedig a két részidőszak közös szórásnégyzete, n pedig a minta elemszáma. Az így definiált próbastatisztika megközelítőleg $n-1$ szabadsági fokú χ^2 eloszlást követ [1].

Azt a pontot tekinthetjük a jelöltek közül tényleges töréspontnak, ahol a χ^2 -érték maximális (lokálisan), mert ezekben a pontokban tér el a két részidőszak közös szórása lefele a teljes időszak szórásához képest.

2.2.2. Anomáliák kumulatív összegzése

Általánosan megfogalmazva azt mondhatjuk, hogy anomáliának tekinthetjük azt, ami az átlagostól eltér. Természetesen nem mindegy, hogy egy adat mennyire tér el az átlagtól, például outlier, kiugró adatok definiálásának egy jól bevált módja az átlagtól legalább az interkvartilis terjedelem (IQR) másfélszeresével való eltérés mérték.

Meteorológiai adatsorok esetében – általános anomáliavizsgálat kapcsán – is az anomáliának az átlagtól való eltérését tekintjük, ezt azonban jelentősen befolyásolhatja a klimatikus idősorokban tapasztalható havi, évszakos trendek és ciklusok jelenléte. Utóbbiakat megfelelő technikákkal ki kell küszöbölni az adatsorokból az anomáliák definiálása előtt, illetve az éves bontás mellett a vizsgálatokat havi szinten is célszerű elvégezni.

Általános értelemben az x_i $(1 \leq i \leq n)$ adathoz tartozó anomália az alábbi egyszerű módon definiálható: $a_i = x_i - \bar{x}$, $1 \leq i \leq n$, ahol *n* az adatsor hossza, $\bar{x}$ pedig a teljes adatsor átlaga. Ezek alapján minden potenciális töréspontra meghatározzuk az anomáliák kumulált értékét:

$$A_{szum} = \sum_{j=1}^{k} a_j,\ k < n. \tag{3}$$

Azt a pontot tekinthetjük töréspontnak, ahol ez az összeg extrém nagy vagy éppen extrém kicsi. A módszernek két nagy hátránya is van, az egyik hogy az idősorok részletesebb és bonyolultabb előkészítést igényelnek, mint a többi módszer esetében, illetve nem teljesen definiált az sem, hogy a kumulált anomáliák esetében mit tekintünk extrém nagynak, illetve extrém kicsi értéknek.

2.2.3. Pettitt-féle nemparaméteres próba

A próba lényege, hogy minden potenciális töréspont esetében kiszámítjuk az

$$X_k = 2R_k - k(n-1) \tag{4}$$

próbastatisztika értékét, ahol $R_k = \sum_{j=1}^{k} r_j$ és r_j az x_j elem – rangkorellációk esetéhez hasonlóan definiált – rangszáma. A teljes minta elemszáma *n*, és $k = 1,2,\dots,n$. A töréspont az a *k* pont lehet, ahol az X_k értéke maximális vagy minimális, attól függően, hogy a tapasztalt ugrásszerű változás fel- vagy lefele irányul-e.

Annak eldöntésére, hogy a módszer által adott X_k érték valóban tényleges, statisztikai értelemben is kimutatható töréspontot határoz-e meg ki kell számítani az annak valószínűségét adó, az X_k értékhez tartozó szignifikanciaszintet. Ezt az alábbi összefüggés szolgáltatja [2]:

$$\propto = 2e^{\frac{-6(X_{extrem})^2}{n^3+n^2}}, \tag{5}$$

ahol x_{extrem} az X_k extrém értékéhez tartozó elemszámnak felel meg. Ha az α értéke kisebb volt, mint 0,05, az x_{extrem} értékhez tartozó pont által elválasztott részek eltérésének valószínűsége meghaladta a 95%-ot, azaz szignifikáns a felfedett töréspont. A módszer a 2.2.5. nemparaméteres változatának tekinthető.

2.2.4. Jel-zaj arány vizsgálata

A determinisztikus modellezés esetében feltételezünk az idősor vonatkozásában egyfajta állandósságot, ez a tartósnak tekinthető tendencia, a szezonalitás és a ciklikusság, míg a véletlent egy zavaró tényezőnek, zajnak tekintjük. Ez a fajta megközelítés – szemben a sztochasztikus modellezéssel – ezért inkább leíró jellegű, a hosszabb távú hatások feltérképezésben erős, a véletlennel kevéssé foglalkozik.

A statisztikai szakirodalomban a hosszú távú, potenciálisan előre jelezhető változásokat jelnek (ami mérhető, tervezhető), míg a természetes változékonyságot zajnak tekintjük.

Töréspont kereshető ezen két komponens idősorban történő előfordulási arányainak vizsgálatával is. A jel–zaj arány becslésére többféle módszer létezik, köztük például az alábbi, Yamamoto által ajánlott és alkalmazott [2]

$$\propto = 2e^{\frac{-6(X_{extrem})^2}{n^3+n^2}}, \tag{6}$$

ahol $\bar{x}_1$, $\bar{x}_2$ és σ_1 és σ_2 a két vizsgált részidőszak átlagai és szórásai. Töréspontnak a jel–zaj arány maximumát produkáló választópontot tekinthetjük, vagyis ahol $A_{jel/zaj}$ a maximális értéket veszi fel. A töréspontot a gyakorlatban akkor tartjuk reálisnak, ha $A_{jel/zaj} \geq 0{,}5$.

2.2.5. Részátlagok összevetése Student-féle t-próbával

A módszer alkalmazhatóságának feltétele, hogy a vizsgált adatsorok normális eloszlásúak legyenek, amit például a Kolmogorov-Szmirnov, vagy a Shapiro-Wilk teszttel vizsgálhatunk.

Az átlagok összevetésére a próba során generált *t*-érték alkalmazható:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \tag{7}$$

ahol $\bar{x}_1$ és $\bar{x}_2$ a két minta átlagát, n_1 és n_2 azok elemszámát, S pedig a két minta közös szórását jelenti. Utóbbi a két részminta szórásának (σ_1 és σ_2) ismeretében az alábbi módon könnyen számítható:

$$S = \sqrt{\frac{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2}{n_1+n_2-2}}. \tag{8}$$

Minden egyes potenciális töréspont által meghatározott két részszakaszhoz kiszámítjuk a Student-féle *t*-próbához tartozó *t*-értéket, a választópontot pedig az első lehetséges értéktől egy-egy évvel mozgatjuk előre az utolsó lehetséges értékig. Az így kapott $|t|$ értékek sorozatából a maximális mutatja azt a pontot, évet, amely mentén az idősort kettéválasztva, a két részsorozat átlaga között maximális az eltérés mértéke. Amennyiben ez az eltérés statisztikailag kimutathatóan szignifikáns is az adott megbízhatósági szinten, akkor mondhatjuk, hogy ez a pont egy töréspont a vizsgált adatsor tekintetében.

Amennyiben a vizsgált idősorban statisztikailag alátámasztva trendszerű és töréspontszerű változás is kimutatható, akkor feltétlenül kérdés, hogy hogyan tudjuk meghatározni, hogy a kimutatott változás inkább egy folyamat következménye, vagy valóban egy hirtelen, ugrásszerű változás történt-e a kijelölt évben.

A kérdés eldöntéséhez össze kell hasonlítani a vizsgált folyamatot leíró lineáris trend és a kapott töréspont jellemzőit, pontosabban az eltérés négyzetösszegeket (SS). Illesszünk az idősorunkra regressziós egyenest, legyen ez $x_{trend} = mt + b$ alakú trendvonal. Határozzuk meg a trendegyenes adott pontokhoz tartozó helyettesítési értékeit, jelöljük ezeket $x_{trend,i} = mt_i + b$-vel, $1 \leq i \leq n$ az n adatot tartalmazó adatsor esetében. Képezzük ezek eltérését az adott pont tényleges x_i értékeitől, majd számítsuk az eltérés négyzetösszeget,

$$SS_{trend} = \sum_{i=1}^{n}\left(x_i - x_{trend,i}\right)^2. \tag{9}$$

Adott ekkor továbbá az idősorban kimutatott töréspont által meghatározott két részidőszak és azok átlagai ($\bar{x}_1$ és $\bar{x}_2$). Az első részidőszak elemei $x_1, x_2, \dots, x_k$, míg a másodiké $x_{k+1}, x_{k+2}, \dots, x_n$. Állítsuk elő a két részidőszakra külön-külön az egyes tényleges mérési adatok és a részidőszak átlagának különbségeit és képezzük ebből az eltérés négyzetösszegeket:

$$SS_{tp} = \sum_{i=1}^{k}(x_i - \bar{x}_1)^2 + \sum_{j=k+1}^{n}(x_i - \bar{x}_1)^2. \tag{10}$$

Az illeszkedések jóságának, pontosságának összehasonlítására képezzük a két eltérés négyzetösszeg hányadosát: $h = \frac{SS_{trend}}{SS_{tp}}$-t. A kapott h hányados értéke alapján meghatározható, hogy melyik eljárás ad pontosabb eredményt, hiszen ha $h > 1$, akkor a töréspontok által meghatározott átlagok, ha $h < 1$, akkor a teljes adatsorra illesztett trendvonal illeszkedik jobban a vizsgált adatsorra.

3. Eredmények, következtetések

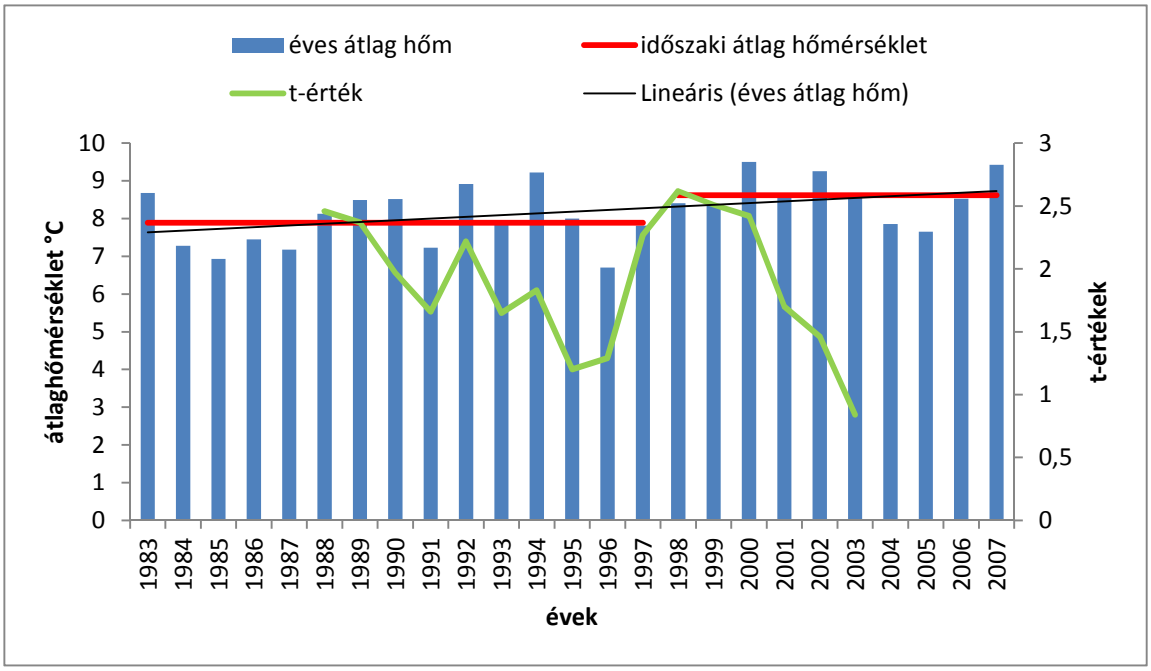
3.1. Töréspont keresés

A 2. fejezetben bemutatott módszerek közül a részátlagok *t*-próbával történő összevetését alkalmaztuk (az alapadatok normalitásának ellenőrzését követően). A töréspont keresést mind a csapadék, mind a hőmérséklet adatsorokra elvégeztük, illetve éves, havi és évszakos bontásban is vizsgáltuk az idősorokat. A minimális intervallum hosszát 5 évben definiáltuk, így a potenciális töréspontok száma 22. A munkában csak az éves adatokra vonatkozó eredmények szerepelnek, mert az összefüggés vizsgálatok is éves szinten kerültek végrehajtásra.

Az éves átlaghőmérsékletek vonatkozásában a töréspont 1998-ra adódott (1. táblázat), mely statisztikailag valóban egy ugrásszerű, hirtelen változás.

év	átlag előtte (°C)	átlag mögötte (°C)	eltérés (°C)
1998	8,01	9,05	1,04

1. táblázat. Töréspont az éves átlaghőmérséklet adatsorban



1. ábra. Töréspont az éves átlaghőmérséklet vonatkozásában

A csapadék adatsor esetében a maximális *t*-érték (1,46) 2000-ben adódott, azonban $p = 0{,}16$ érték mellett, így ezt nem tekinthetjük szignifikáns eltérésnek az általunk kijelölt 95%-os szinten (de még 90%-os szinten sem). Az éves csapadékösszeg tekintetében nem jellemző a komoly változása, azonban annak éven belüli eloszlásában már komoly átrendeződés érzékelhető. Utóbbit támasztják alá a hónapok és évszaki csapadékösszegek vizsgálatára vonatkozó eredmények.

3.2. Összefüggés vizsgálatok

Az éves átlagos növekedési adatsor és a hőmérséklet közötti összefüggésekkel kapcsolatos eredményeket jelenítünk meg. Az elemzések teljesebbé tételéhez alkalmaztuk a CReMIT módszert [4] speciális hőmérséklet adatsorok képzéséhez. Ennek lényege, hogy a szimpla havi adatok mellett különböző időszaki hőmérséklet átlagokat is képeztünk és ezeket is összevetettük a növekedési adatokkal, mint függő paraméterrel. Ehhez a továbbiakban az alábbi jelöléseket alkalmazzuk: *a* az aktuális, *p* az előző, *pp* a kettővel korábbi éve összefüggéseire, szám pedig a hónapra utal. Például *p*10-*a*2: előző év októberétől adott év februárig tartó időszak átlaghőmérséklete. Az alkalmazott összefüggés vizsgálat szignifikancia vizsgálattal egybekötött lineáris korreláció-analízis.

mettől	meddig	1983-2007	1985-1998	1999-2007
H.p8	***H.p11***	-0,53		-0,69
H.p10	***H.p11***	-0,65		-0,73
H.p5	***H.p6***	-0,43		
H.a4	***H.a5***		0,65	
H.a6	***H.a8***	-0,41		-0,6
H.a7	***H.a8***			-0,69

2. táblázat. Összefüggés a részidőszakokban, hőmérséklet

Az összefüggés vizsgálatokat a teljes időszak (1985-2007) és a töréspont keresés eredményeként kapott év (1998) mentén kettébontott időszakokra (1985-1998 és 1999-2007) is elvégeztük. A táblázatban csak a 90%-os szinten statisztikailag szignifikáns eredményeket jelenítettük meg. Azt vizsgáltuk, hogy mennyire hasonlóak, vagy eltérőek a kapott eredmények. A 2. táblázatban szereplő korrelációs értékek alapján jól látható, hogy az első részidőszakban alig adódott szignifikáns kapcsolat. A teljes időszak tekintetébe már több, bár inkább közepes erősségűnek nevezhető összefüggés látszik. Ugyanakkor a második részidőszakban már erősnek mondható összefüggéseket láthatunk jellemzően ugyanazon időszakokra, mint a teljes időszak vonatkozásában. Fontos megjegyezni, hogy a 2. táblázat csak néhány, önkényesen kiemelt időszakot mutat, de a többi vizsgált intervallum tekintetében is hasonló eredmények fogalmazhatóak meg.

4. Összefoglaló

A dolgozat az idősorokban töréspontok keresésére vonatkozó jellemző statisztikai módszereket mutatja be röviden. Egy konkrét példán alkalmaztuk a részátlagok összevetésén alapuló eljárást és vizsgáltuk ennek jelentőségét fák növekedésére vonatkozó összefüggés vizsgálatokban. A bemutatott eredmények arra hívják fel a figyelmet, hogy az alkalmazott elemzési módszer tekintetében (lineáris korrelációanalízis) a független paraméterekben bekövetkező változások komoly hatással bírnak az összefüggés vizsgálatok kimenetére. Azt is

látható, hogy a hosszabb vizsgálati időszakra vonatkozó összefüggések, kapcsolatok nem feltétlenül állnak fent annak egy-egy rövidebb részintervallumára, sőt akár az összefüggések irányát tekintve is bekövetkezhetnek változások.

Irodalomjegyzék

[1] **Dévényi, D., Gulyás, O.,** Matematikai statisztikai módszerek a meteorológiában. Tankönyvkiadó, Budapest (1988) p. 443.

[2] **Mares, I., Mares, C.,** Statistical methods for estimating the signal to noise ratio. *In: Contemporary Climatology (Proceedings of the meeting of the Commission on Climatology of the International Geographical Union)*. Brno. (1994) 373–379.

[3] **Molnár, J.,** A légnyomási mező szerkezete és módosulása a Kárpát-medence térségében. Doktori értekezés, Debrecen (2003) p. 170.

[4] **Pödör, Z., Edelényi, M., Jereb, L.,** Systematic Analysis of Time Series – *CReMIT. Infocommunication Journal*, VI(1) (2014) 16-22.