MTA DOKTORI ÉRTEKEZÉS

MATYASOVSZKY ISTVÁN

NÉHÁNY STATISZTIKUS MÓDSZER AZ ELMÉLETI ÉS ALKALMAZOTT KLIMATOLÓGIAI VIZSGÁLATOKBAN

BUDAPEST, 2013. JANUÁR

Tartalomjegyzék

BEVEZETÉS

Mivel az éghajlat alapvetően statisztikus természetű, ezért vizsgálata a valószínűség-számítás és matematikai statisztika eszközeit igényli. Ezekkel a kérdésekkel foglalkozik a statisztikus klimatológia.

Az értekezésben modern matematikai statisztikai eszközöket mutatunk be és használunk fel elméleti vagy alkalmazott klimatológiai vizsgálatokban. Ennek során természetesen nem törekedünk teljességre, részben egyéni érdeklődésünk végessége, részben terjedelmi korlátok folytán. Meglehet, a tartalomjegyzék alapján talán csak a negyedik rész tűnik újszerűnek, ám a többi fő fejezetben néhány régóta ismert probléma modern eszközeit tárgyaljuk. Ez oly annyira igaz, hogy több eljárást jelen tanulmány szerzője alkalmazott először a nemzetközi meteorológiai irodalomban. Munkánk célja néhány széles körben felhasználható, ám kevéssé elterjedt módszer megismertetése, majd egy-egy alkalmazásának bemutatása. Az önmagukban is értékes eredményekkel egyben érzékeltetjük a bennük rejlő további gyakorlati lehetőségeket.

Részben az általános megismerés, részben a jövő éghajlatának becslési lehetőségeinek értékelése szempontjából fontos a már lezajlott, illetve zajló éghajlatváltozás detektálása és becslése. Az éghajlatváltozást legegyszerűbben a várható érték időbeli változásával szokás leírni, vagyis a trendelemzéssel. Ezzel foglalkozik az első rész.

Az elméleti, de talán még inkább az alkalmazott klimatológiai problémák során fontos igény a különböző változók közötti statisztikai kapcsolat feltárása. Ilyen módszerek tárgyalása és gyakorlati alkalmazása történik az újabb részben.

Az egyik legrégebbi statisztikus klimatológiai eszköz a spektrálanalízis, amikor egy szóban forgó idősor mögött meghúzódó sztochasztikus folyamatot véges sok (diszkrét spektrum) és megszámlálhatatlanul sok (folytonos spektrum) periodikus összetevő

szuperpozíciójaként tekintünk. A szerteágazó problémakör néhány speciális vetületét vizsgáljuk a harmadik részben.

Végezetül meg kell említeni, hogy mivel az éghajlat nagyszámú nemlineáris kölcsönhatás eredményeként jön létre, célszerű az éghajlati idősorokat a jól ismert lineáris idősor modellek helyett nemlineáris idősor modellek segítségével elemezni. Egyebek mellett ilyen lehetőséget tárgyal a negyedik rész.

Mivel nem matematikai műről van szó, a matematikai eszközök tárgyalása csak olyan mértékben történik, ami feltétlenül szükséges a problémák és az alkalmazások megértéséhez. Az elmélet és alkalmazás remélt egyensúlyának megtalálásában talán leginkább a nagyszámban felhasznált statisztikai próbák bemutatásának mélysége jelentette a legnagyobb gondot. Végül úgy döntöttünk, hogy az olyan alapvető próbákat, mint például a khí-négyzet-próba vagy a Kolmogorov-Szmirnov-próba ismertnek feltételezzük és még irodalmi hivatkozást sem adunk. A még mindig alapvető, de talán már nem annyira elterjedt próbák leírása helyett csupán irodalmi hivatkozást nyújtunk. A kevéssé ismert próbák közül a röviden bemutathatókat ismertetjük, míg a hosszú tárgyalást igénylőket ismét csak irodalmi hivatkozással látjuk el. Ez utóbbi megfontolás oka az, hogy nem kívánjuk a fő gondolat követhetőségét kockáztatni egy próba hosszadalmas taglalásával.

A fő fejezetek fent említett csoportosításával a viszonylag egyszerűbb módszerektől a bonyolultabbak felé haladunk. Ennek következtében az autoregresszív folyamatok bizonyos általánosításainak tárgyalására a negyedik fejezetben kerül sor, ám ezt megelőzően több alkalommal kell utalnunk magukra az autoregresszív folyamatokra. Ezért e modell értelmezését most kell megtennünk. Legyen Y_t egy diszkrét paraméterű sztochasztikus folyamat. Ekkor az a_0 konstanssal és az $a_1,...,a_p$ autoregresszív együtthatókkal képezett

$$Y_t = a_0 + a_1 Y_{t-1} + ... + a_p Y_{t-p} + e_t$$

formát *p*-edrendű autoregresszív (AR(p)) folyamatnak nevezzük, ahol az e_t folyamatra teljesül, hogy fehérzaj, vagyis zérus várható értékű, konstans szórású korrelálatlan valószínűségi változók sorozata. Ezen kívül e_t és Y_{t-j}, $j = 1,..., p$ korrelálatlanok. Egy ilyen folyamat akkor és csak akkor stacionárius, ha az

$$1 - a_1 u - ... - a_p u^p = 0$$

ún. karakterisztikus egyenlet gyökei abszolút értékben egynél nagyobbak. Olykor előfordul, hogy túlságosan nagy *p* szükséges adott idősor AR folyamattal történő kielégítő modellezéséhez. Ilyenkor célszerű lehet a gazdaságosabb, tehát kevesebb paraméter becslését igénylő

$$Y_t = a_0 + a_1 Y_{t-1} + ... + a_p Y_{t-p} + e_t - b_1 e_{t-1} - ... - b_q e_{t-q}$$

p-edrendű autoregresszív – *q*-adrendű mozgó átlag (ARMA(p,q)) folyamat alkalmazása, ahol $b_1,...,b_q$ a mozgó átlag együtthatók.

1. TRENDELEMZÉS

A trendet (a várható érték időbeli menetét leíró függvényt) még ma is leggyakrabban az idő lineáris függvényének tekintik. A függvényben szereplő paraméterek becslését a legkisebb négyzetek (ordinary least squares: OLS) módszerével végzik, tehát az aktuális megfigyelések és becslésük különbségének négyzetes összegének minimalizálásával. A trend létét vagy hiányát a lineáris közelítésből fakadó egyenes becsült meredekségével vizsgálják, vagyis ha ez a becslés nem különbözik statisztikailag szignifikánsan zérustól, akkor nincsen trend. Az ekkor alkalmazott *t*-próba ismertetése egyebek mellett megtalálható például Matyasovszky (2002) kötetében. A valóságos trend azonban rendszerint eltér a lineáristól, és így az egész eljárás hibás eredményre vezethet. A következőkben ezért a trendbecslés jóval általánosabb megközelítésével foglalkozunk.

1.1 MÓDSZER

A trend linearitását általában nem, de bizonyos simaságát feltételezhetjük. Ez utóbbi azt jelenti, hogy a trendfüggvény adott időpont nem túl szűk környezetében jól közelíthető egy alacsony fokú polinommal. Ezekben az esetekben előnyösen alkalmazhatók a nemparaméteres becslési technikák. Ezek matematikai irodalma rendkívül gazdag, ám meteorológiai alkalmazásuk még ma is alig elterjedt. Ez meglepő, mert a jól ismert heurisztikus mozgó átlagolási módszerek (binomiális simítás, gaussi simítás, stb.) elméletileg jól megalapozott általánosításának is tekinthetők. Ismereteink szerint meteorológiai célra Matyasovszky (1992) javasolta először az eljárást.

A $t_1,...,t_n$ időpontokban rendelkezésre álló $y_1,...,y_n$ adatsort

$$y_i = f(t_i) + e_i\,,\ i = 1,...n \tag{1.1}$$

alakban tekintjük, ahol $f(t)$ a trendfüggvény, és az $\{e_i\}$ zajra teljesül, hogy fehérzaj. A módszer a trendfüggvény minden időpont körüli lokális polinomiális közelítésén alapszik. Több változata ismeretes, melyek közül a súlyozott lokális regresszió (weighted local regression: WLR) tekinthető a leguniverzálisabb eljárásnak (Fan, 1992; Fan, 1993; Fan and Gijbels, 1992). Az $\hat{f}(t) = \hat{a}_0 = \hat{a}_0(t)$ becslés a

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} a_j (t_i - t)^j \right)^2 K\left(\frac{t_i - t}{b}\right) \tag{1.2}$$

mennyiség adott *t* melletti, $a_0,...,a_p$ szerinti minimalizálásával nyerhető, ahol $K(u)$ itt most nem részletezett tulajdonságokkal rendelkező ún. magfüggvény. Ilyenkor a trend becslésének torzítását alapvetően $f^{(p+1)}(t)$, tehát a trendfüggvény (*p*+1)-edik deriváltja határozza meg. A magfüggvény adja meg, hogy a t pont környezetéhez tartozó négyzetes hibákat milyen ütem szerint vesszük figyelembe egyre kisebb súllyal, ahogy a t időponttól távolodunk. A *b*-vel jelölt sávszélesség a figyelembe veendő környezet szélességét definiálja.

Kimutatható, hogy az alkalmazások során a sávszélesség megválasztása lényegesen nagyobb fontossággal bír, mint a polinom foka és a magfüggvény. Ezért általában elegendő a lokálisan lineáris közelítéssel élni (p =1), amihez a $K(u) = 3/4(1-u^2), |u| < 1, K(u) = 0, |u| \geq 1$ ún. Epanechnikov-féle magfüggvény tartozik. Ez a magfüggvény a sávszélesség optimális választása mellett biztosítja a becslés minimális átlagos négyzetes hibáját p =1 esetén. Optimális sávszélességnek azt tekintjük, ami a becslés átlagos négyzetes hibáját, tehát a becslés varianciájának és torzítás négyzetének összegét minimalizálja. Kis sávszélesség ugyanis kis torzítást, de nagy szórást eredményez (a kapott görbe nem kellően sima), míg nagy sávszélesség nagy torzítást, de kis szórást szolgáltat (a kapott görbe túlságosan sima). A nemparaméteres módszerek tehát a sávszélesség választásán keresztül megteremtik a trendbecslés torzításának és szórásának optimális viszonyát úgy, hogy közben a trendfüggvény teljes formájára semmilyen feltevéssel nem élnek. Ez az eljárás rendkívül értékes tulajdonsága a paraméteres módszerhez képest. Hátránynak vélhető viszont, hogy a trend hiányára vonatkozó null-hipotézis nem ellenőrizhető. E hátrány valójában mégsem jelentkezik, mert lehetőség van a sávszélesség becslésére is, és a trend hiányában p =1 esetén ilyenkor rendkívül nagy, gyakorlatilag végtelen sávszélesség adódik optimálisnak. A végtelen sávszélesség abból fakad, hogy a trend lokális lineáris közelítése ilyenkor globálisan lineárisba megy át. Ekkor viszont a lineáris trend jelenléte ellenőrizhető a fejezet elején említett *t*-próbával.

A b sávszélesség becslése nagy gyakorlati jelentőséggel bír. Legegyszerűbb esetben $\hat{b}$ a

$$CV(b) = \sum_{i=1}^{n} (y_i - \hat{f}_i(t_i))^2 \qquad (1.3)$$

mennyiség b szerinti minimalizálásával nyerhető, ahol $\hat{f}_i(t_i)$ $f(t_i)$-nek olyan becslése, hogy a t_i időponthoz tartozó y_i megfigyelést nem vesszük figyelembe. Ha a trend meglehetősen

komplex formájú (éles csúcsok és lassú változások egyaránt jellemzik) és/vagy a $\{t_i\}$ időpontok nem ekvidisztánsak, akkor célszerű a sávszélességet is időfüggőnek tekinteni (lokális sávszélesség). Ez a helyzet viszonylag ritkán áll elő a klimatológiai alkalmazásokban, ezért általában megelégedhetünk a globális sávszélességgel. A kérdés további tárgyalása helyett csupán utalunk Matyasovszky (1998) és Matyasovszky (2002) munkájára. Gyakori ellenben, hogy $\{e_i\}$ szórása nem állandó, ami szerencsére nem jelent nehézséget, ha a szórás időbeli változása hasonlóan sima, mint a várható érték időbeli változása. Komoly probléma ellenben, ha $\{e_i\}$ nem korrelálatlan, mert ez azt vonja maga után, hogy (1.3) minimalizálásával megbízhatatlan becsléshez jutunk a sávszélességre nézve. Például erős pozitív korrelációk esetében $\hat{b}$-ra rendkívül kis értéket, gyakorlatilag zérust kapunk. A probléma kezelése természetesen a korrelációk figyelembevételével történhet, amire számos eljárás ismeretes (Fernandez and Fernandez, 2004). Munkánk során a Fernandez and Fernandez (2004) által javasolt alábbi eljárást alkalmazzuk. A

$$TSCV(b) = 1/(n-P)\sum_{t=P+1}^{n}(y_t - \hat{y}_t)^2 \qquad (1.4)$$

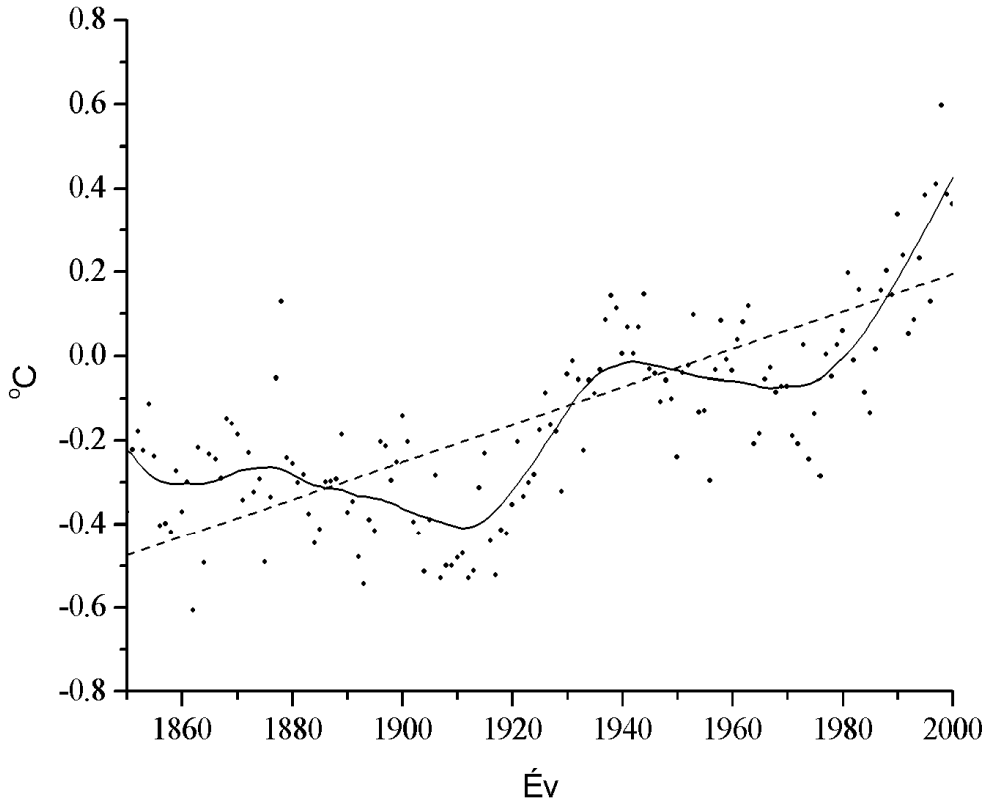
mennyiséget minimalizáljuk *b* szerint, ahol *P*>*p* és $\hat{y}_t$ egy *p*-edrendű autoregresszív becslése y_t-nek az $\hat{y}_t = \tilde{f}(t) + \sum_{j=1}^{p}\hat{a}_j(y_{t-j} - \tilde{f}(t-j))$ alakban. Az $a_1,...,a_p$ együtthatókat az OLS módszerrel becsüljük úgy, hogy $\tilde{f}(t-j)$ az $f(t-j), j = 0,..., p$ olyan, a *b* sávszélesség melletti WLR becslése, amely csak az $y_1, y_2,..., y_{t-p}$ adatokat használja fel. Mivel TSCV csak a *t* időpontot megelőző adatokat veszi figyelembe, ám a trend végső becslése a *t* időpont utáni adatokat is használja majd, ezért a TSCV minimalizálásával nyert becsült sávszélességnek egy korrekcióját kell végrehajtani. Ez végső soron egy konstanssal való szorzást jelent, s a konstans választásáról például Müller (1991) tanulmánya tájékoztat.

A nemparaméteres trendbecslés mélyebb matematikai részleteiről - a már korábban közölt forrásokon kívül - számos helyen olvashatunk, melyek közül Simonoff (1996) átfogó munkáját emeljük ki.

1.2. ALKALMAZÁSOK

1.2.1. Északi Hemiszféra átlaghőmérséklete

Tekintsük az Északi Hemiszféra évi középhőmérsékleteinek az 1961-1990 időszakhoz képesti anomáliáit az 1850-2009 évek időszakára (Jones et al., 2010). Lineáris trend illesztésével arra az ismert következtetésre juthatunk, hogy az Északi Hemiszféra átlaghőmérséklete az 1850-2009 közötti periódusban összességében 0,73 ^{o}C emelkedést mutat, ami 0,045 ^{o}C/10év növekedési rátának felel meg. A WLR eljárás azonban az 1914-1942 időszakra összességében



1. ábra
Az Északi Hemiszféra hőmérsékleti anomáliái (pontok) és trendje WLR (folytonos vonal) és lineáris (szaggatott vonal) közelítéssel

0,133 ^{o}C/10év, sőt az 1975-2009 periódusra 0,183 ^{o}C/10év emelkedő rátát nyújt úgy, hogy közben az 1942-1975 évek során 0,057 ^{o}C/10év ütemű csökkenést jelez (Matyasovszky,

2011). A lineáris és a WLR módszerrel nyert trendek közötti különbség világosan látszik az 1. ábrán.

Az említett évszámok, illetve időszakok nem véletlenül lettek kiválasztva. Az utóbbi időben ugyanis növekvő figyelem fordul az ún. hirtelen éghajlatváltozások (abrupt climate changes) felé. A hirtelen változást mutató időpontok azonosítása tradicionálisan azon alapszik, hogy a trendfüggvényt szakaszonként konstansnak tekintik, mely szakaszokat a trend ugrásai határolják. Szentimrey et al. (1992) egy korai munkáját számos további hasonló tanulmány követte (Fraedrich et al., 1997; Jiang et al., 2002; Smadi, 2006; Zhao et al., 2007), melyeket Feng et al. (2010) próbált áttekinteni. Ez a megközelítés azonban fizikailag tarthatatlan, hiszen nem képzelhető el az a helyzet, hogy az éghajlat valameddig változatlan, majd azonnal egy másik és egy ideig ismét állandó éghajlatba megy át. A következő fejezetben ezért kísérletet teszünk a hirtelen éghajlatváltozások megalapozottabb statisztikai detektálására.

1.2.2. Hirtelen éghajlatváltozások

Hirtelen éghajlatváltozás történik, amikor az éghajlati rendszer valamilyen ok hatására átlép egy küszöböt, ami egy új állapotba történő átmenetbe kényszeríti az éghajlati rendszer által determinált és a kiváltó oknál nagyobb sebességgel (National Research Council, 2002). Az Egyesült Államok Éghajlatváltozás Tudományos Programja szerint hirtelen éghajlatváltozásról akkor beszélünk, ha a változás néhány évtized (vagy rövidebb idő) alatt következik be és legalább néhány évtizedig fenn áll, ami jelentős hatással van az emberi és természeti környezetre. Sajnos e két definíció voltaképpen nem definiál, mert olyan bizonytalan szavakkal operál, mint „valamilyen ok”, „nagyobb sebességgel”, vagy „néhány évtized”, „jelentős hatás”. Mivel bármilyen tulajdonságnak egy adatsorban való detektálása

statisztikai eszközöket igényel, a hirtelen éghajlatváltozás értelmezése is statisztikai alapú kell, legyen.

Ezért Matyasovszky (2011) alapján a trendfüggvény deriváltjának ugrásait tekintjük hirtelen változásnak. A feladat a WLR módszerrel oldható meg, mert (1.2) minimalizálásával $f^{(k)}(t)$ becslése $\hat{f}^{(k)}(t) = \hat{a}(t)k!$ lesz, ahol $k \le p$. Az iménti definíció persze egy kompromisszum. Egyrészt azért, mert csak a várható érték időbeli változására épít. Másrészt azért, mert a trend ugrásainak megengedése - mint láttuk - elfogadhatatlan, míg a második vagy magasabb deriváltjai ugrásának értelmezése már túl sima trendet adna ahhoz, hogy hirtelen változásról beszélhessünk. A detektálás módszere épít Wishart (2009) eljárására, de jelentősen különbözik tőle.

Ha a trend nem sima, tehát deriváltjának vannak ugrásai, akkor

$$f(t) = f_S(t) + \sum_{k=1}^{K} c_k \varphi_k(t) \ , \ \ \varphi_k(t) = \begin{cases} 0, & t < \tau_k \\ t, & t \ge \tau_k \end{cases} , \tag{1.5}$$

ahol $c_k, k = 1,...,K$ az $f'(t)$ ugrásainak nagysága a $\tau_1 < \tau_2 < ... < \tau_K$ időpontokban. Ismerve *K* értékét (az ugrások számát) és a $\tau_1,...,\tau_K$ pontokat (az ugrások időpontját), akkor a $c_k, k = 1,...,K$ ugrások nagysága az OLS módszerrel becsülhető a

$$\sum_{t=1}^{n} \left(y_t - \hat{f}(t) - \sum_{k=1}^{K} c_k \varphi_k(t) \right)^2 \tag{1.6}$$

mennyiség minimalizálásával, ahol $\hat{f}(t)$ az $f(t)$ WLR becslése (Cline et al., 1995), amivel voltaképpen $f(t)$ sima részét, azaz $f_S(t)$–t kívánjuk becsülni. A feladat *K* és $\tau_1,...,\tau_K$ megadása, és annak eldöntése, vajon $\hat{c}_k$ statisztikailag szignifikánsan különbözik-e nullától. Mindez a következőképp történik. Ha $f'(t)$-nek ugrása van a *t*=*s* pontban, akkor $\left|\hat{f}''(s)\right|$ nagy, mert $\left|f''(s)\right|$ végtelen. Ezért $\left|\hat{f}''(t)\right|, 1+b < t < n-b$ segít a $\tau_1,...,\tau_K$ időpontok

behatárolásában. Ha $\left|\hat{f}''(t)\right|$ maximális a t=s időpontban, akkor képezhető $f(t)$-nek egy $\tilde{f}(t)$ becslése K=1 és $\hat{\tau}_1 = s$ mellett az (1.6) minimalizálásával. Ha $\hat{c}_1$ szignifikánsan különbözik zérustól, akkor az eljárás folytatódik $\left|\hat{f}''(t)\right|$ második, harmadik, stb legnagyobb értékével, amíg az ugrások különböznek nullától. A $\hat{c}_1,...,\hat{c}_K$ ugrások szignifikánsan különböznek zérustól, ha $\tilde{f}(t)$ $\hat{c}_1,...,\hat{c}_K$ mellett jobban közelíti az $y_t, t = 1,...,n$ adatsort, mint $\tilde{f}(t)$ $\hat{c}_1,...,\hat{c}_{K-1}$ mellett (vagy $\hat{f}(t)$ K=1 esetén). A közelítés jóságát a

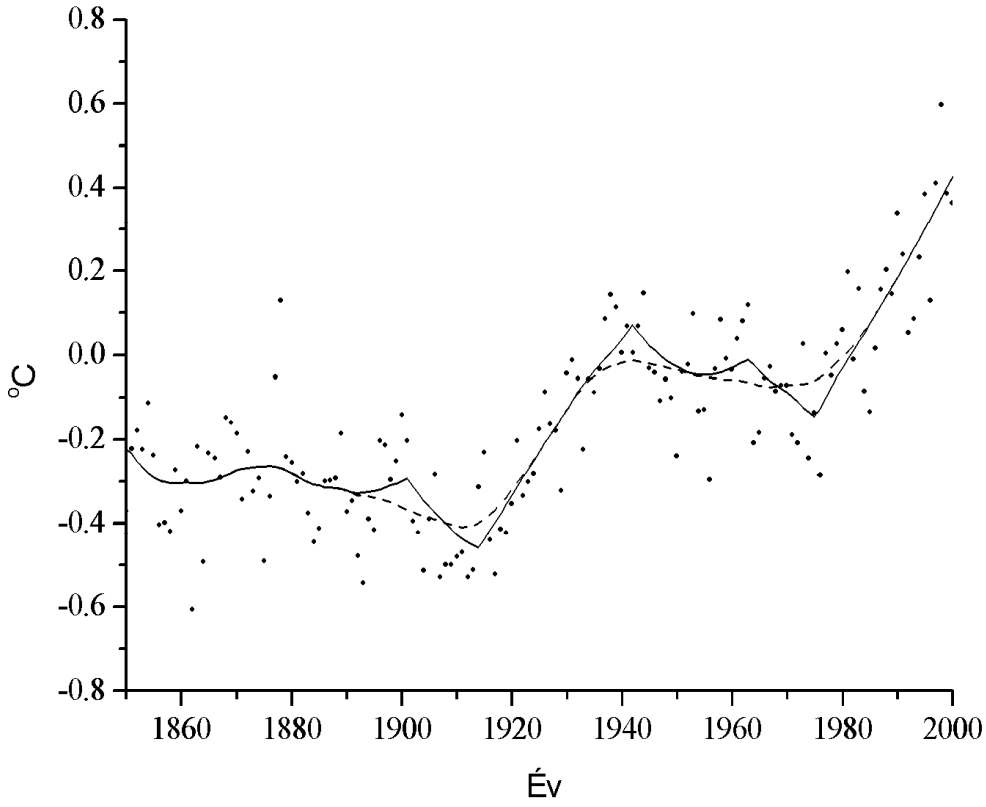
$$GCV(K) = \frac{n}{(n - h_{K,b})^2} \sum_{t=1}^{n} (y_t - \breve{f}(t))^2 \qquad (1.7)$$

mennyiség méri (Craven and Wahba, 1979), ahol $\breve{f}(t) = \tilde{f}(t)$ K>1 mellett, és $\breve{f}(t) = \hat{f}(t)$ K=1 esetén. Itt $h_{K,b}$ a $\underline{\underline{H}}$ mátrix nyoma (főátlóbeli elemeinek összege). Az $\breve{f}(t)$ ugyanis végső soron az adatsor elemeinek lineáris kombinációja, tehát $\underline{\breve{f}} = \underline{\underline{H}}\,\underline{y}$, ahol a $\underline{\underline{H}}$ mátrix (1.6) minimalizálása során nyert lineáris egyenletrendszerből származtatható, továbbá $\underline{\breve{f}} = (\breve{f}(1),..., \breve{f}(n))^T$, $\underline{y} = (y_1,..., y_n)^T$, és a T felső index a transzponálást jelöli. Látható, hogy (1.7) K szerinti minimalizálása a becsült trend illeszkedése és a modell komplexitása között teremt egyensúlyt. Nagy K esetén ugyanis kicsi a szumma értéke, de nagy a szumma előtti szorzó, míg kis K esetén nagy a szumma értéke, de kicsi az előtte lévő szorzó. Csupán megjegyezzük, hogy ha $\underline{y}$ elemei korreláltak, akkor (1.7)-ben $\breve{f}(t)$ helyett olyan $\hat{y}_t$ szerepel, ami y_t-nek a becsült trendet is magában foglaló autoregresszív becslését tartalmazza. A $\underline{\underline{H}}$ mátrix ekkor az $\underline{\hat{y}} = \underline{\underline{H}}\,\underline{y}$ egyenletnek felel meg, ahol $\underline{\hat{y}} = (\hat{y}_1,..., \hat{y}_n)^T$.

Még megválaszolandó kérdés a sávszélesség megadása. Alapvető szempont, hogy az ugrások felderítése során a becsült trend kis torzítása az elsődleges (Wishart, 2009), tehát viszonylag kis sávszélesség választása a szerencsés. A sávszélesség becslési tulajdonságainak ismeretében viszont nyilvánvaló, hogy a deriváltjában ugrásokkal rendelkező trendre

vonatkozó sávszélesség nem lesz nagyobb, mint a sima trendre kapott becsült sávszélesség. Ezért ismét alkalmazható Fernandez and Fernandez (2004) módszere akár van ugrás, akár nincs. A viszonylag kis sávszélesség más szempontból is szerencsés. Az így választott *b* esetében ugyanis az (1.7)-ben lévő szumma meghaladhatja minimális értékét, a szumma előtt lévő szorzó - $h_{K,b}$-nak *b*-től való függése folytán - pedig szintén meghaladhatja az optimális sávszélességhez tartozó értékét. Mindezzel elkerülhetjük a reálisnál több ugrás beazonosítását.

Az Északi Hemiszféra hőmérsékletének hirtelen változásaira az 1901, 1914, 1942, 1963 és 1975 éveket adta a módszer (2. ábra). Az átfogó tendencia természetesen a melegedő,

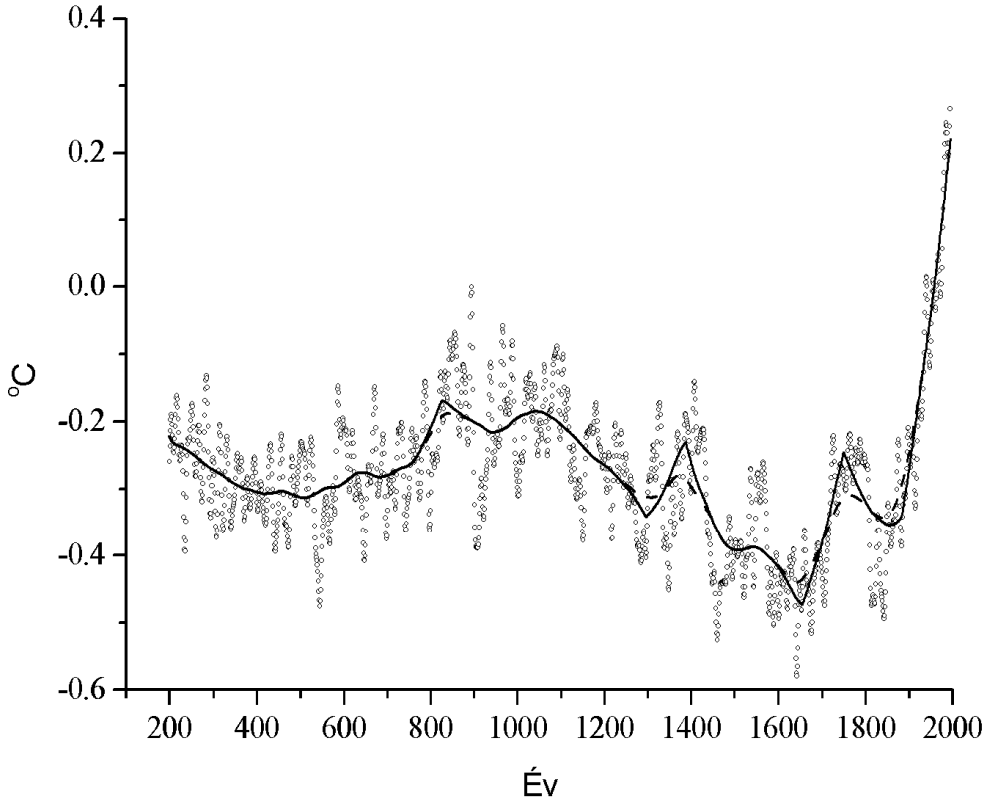


2. ábra
Az Északi Hemiszféra hőmérsékleti anomáliáinak sima trendje (szaggatott vonal) és trendje hirtelen változással (folytonos vonal)

ám három hirtelen hűlési időpont is látható (1901, 1942 és 1963). Korábbi munkák, melyek a kritizált módszerrel vagy egyszerű vizuális megfigyelés konklúziójaként születtek csak az

1940-es, 1970-es évek hirtelen változásairól beszélnek (Thompson et al., 2010). Kisebb területek vagy rövidebb időszakok vizsgálata némileg különböző időpontokat azonosítottak, illetve Ivanov and Evtimov (2000) munkája még az 1963-as évet jelölte meg hirtelen változásként. Nem találkoztunk azonban olyan tanulmánnyal, mely a hirtelen változások ilyen finom szerkezetét tárta volna fel.

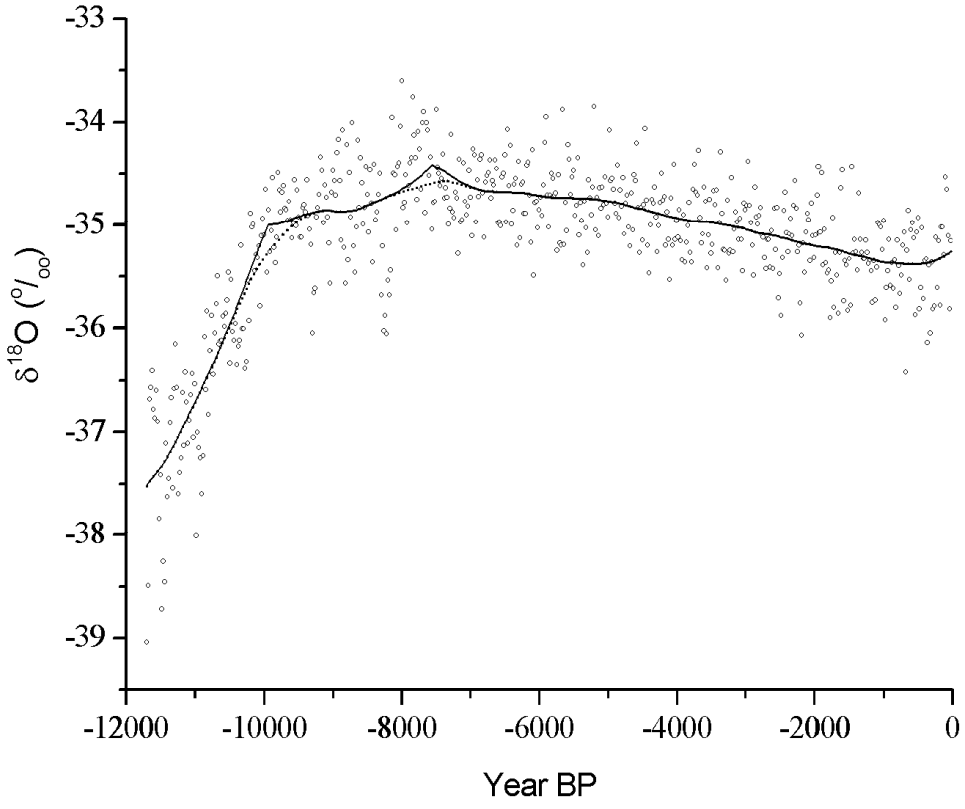
Az eljárást alkalmaztuk az Északi Hemiszféra rekonstruált évi hőmérsékleti sorára is az i.sz. 200-1995 évekre (3. ábra). A hőmérsékleti értékek most az 1856-1995 időszakhoz képesti anomáliákat jelentik. Az adatsor műszeres megfigyelt, nagy felbontású proxy adatok és éghajlati modell eredmények szintéziseként jött létre (Jones and Mann, 2004). A 950-1200



3. ábra
Az Északi Hemiszféra hőmérsékleti anomáliáinak sima trendje (szaggatott vonal) és trendje hirtelen változással (folytonos vonal)

közötti (Lamb, 1977) Középkori Meleg Periódus (Medieval Warm Period: MWP) és az 1450-1850 közötti (Grove, 1988) Kis Jégkorszak (Little Ice Age: LIA) ezen időszak legjellemzőbb éghajlati epizódja. Ezek a kifejezések széles körben elterjedtek, jóllehet nincs is egyértelműen elfogadott definíciójuk (Bradley et al., 2003). Egy gyenge ugrás 825 körül jelzi az utóbbi 1800 év legmelegebb időpontját az 1920-as évek utáni időszakot nem számítva. Egy közeli másik lokális hőmérsékleti csúcs (nem ugrásszerű) jelentkezik 1040 táján, míg a két időpont között 945 körül relatíve alacsony, de még mindig magas hőmérsékletek fordulnak elő. Ezért ha megvizsgáljuk, hogy a trendnek a 945-ös évhez tartozó értéke korábban és később hol fordul ismét elő, akkor a 795-1120 időszak bontakozik ki. Az ily módon definiált MWP némiképp korábbi és hosszabb, mint Lamb (1977) időszaka. A LIA-val kapcsolatos lehűlés két fázisban jelentkezik: egy erősebb és hosszabb periódus 1387-1656 között, és egy gyengébb és rövidebb időszak 1749-1883 között. A LIA ezért (kerekítve) az 1390-1880 évekre tehető, ami hosszabb, mint Grove (1988) időszaka. A legnyilvánvalóbb változás azonban a 19. század végén (1883) hirtelen meginduló nagyon intenzív melegedés.

Végül az elmúlt 11700 évre vonatkozó oxigén izotóp adatokra, pontosabban az NGRIP (North Greenland Ice Core Project, 2004) jégfurat O^{18}/O^{16} izotóparányával kapcsolatos Holocén $\delta^{18}O$ adatokra alkalmaztuk az eljárást (4. ábra). Ezek az adatok 20 éves átlagokként álltak rendelkezésre. Az időszak legjellemzőbb éghajlati epizódja a Holocén Éghajlati Optimum (Holocene Climate Optimum: HCO). Mivel ez a meleg periódus fokozatos lehűléssel ért véget, behatárolása eléggé bizonytalan. Például Johnsen et al. (2001) NGRIP oxigén izotóp adatok alapján 8600-4300 évvel ezelőttre teszi az időszakot, míg Kaufman et al. (2004) más paleo adatok vizsgálatával, 9000-6000 évvel ezelőttre datálja. Módszerünk 9940 és 7560 évvel ezelőttre jelez hirtelen változást. Az első a korábbi igen intenzív melegedés ugrásszerű lassulását, de nem megszűnését jelzi. Sőt hamarosan ismét gyorsul a melegedés, aminek befejeződését és a fokozatos hűlésbe való átmenetét 7560 évvel



4. ábra
Az elmúlt 11700 év NGRIP $\delta^{18}O$ értékeinek sima trendje (pontozott vonal) és trendje hirtelen változással (folytonos vonal)

ezelőttre azonosíthatjuk. Ez egybe esik az időszak során tapasztalható legmelegebb évvel. Az első hirtelen változás a relatíve stabil éghajlat kezdetének, vagyis HCO kezdetének vehető. A trend ugyanezen értéke időben előre haladva 3320 évvel ezelőtt fordul elő ismét, ami ezért HCO végének tekinthető. A teljes HCO tehát értelmezésünk szerint (kerekítve) 9900-3300 évvel ezelőttre datálható, ami jóval hosszabb, mint a korábban definiált időszakok. Meg kell még jegyezni, hogy a HCO időszaka egyértelműen melegebb, mint napjaink éghajlata (4. ábra). Egy nagyjából 8200 évvel ezelőtt bekövetkezett jelentős, de rövid lehűlést számos forrás említ (Cheng et al., 2009). Eljárásunk ezt az epizódot nem tudta detektálni, aminek oka az, hogy az esemény túlzottan rövid az adatsor 20 éves felbontásához képest, így ez a rövid hőmérsékleti visszaesés nem tud a trendben feltűnni. Ellenben, ha ez az esemény valóban

létezik, akkor az ezzel kapcsolatos adatok hirtelen erősen eltérnek a trendtől, ami a zaj varianciájának megnövekedéseként jelentkezik. Ezért előállítottunk egy új adatsort az eredeti adatok trendjétől való eltérésének négyzeteként. Az eljárás ezen új adatsorra való alkalmazásával azt nyertük, hogy a variancia 8140 évvel ezelőtt hirtelen változást mutat. Mivel a változás erőteljes csökkenés, ezért az említett 8140 évvel ezelőtti időpont a hideg periódus végét jelzi, ami jó összhangban van Thomas et al. (2007) vagy Kobashi et al. (2007) által talált hideg időszakkal.

1.2.3. Allergén pollenek

Az elmúlt néhány évized során számos növényi faj pollenje által kiváltott allergiás tünetek és allergiás légúti betegségek számának erőteljes növekedése figyelhető meg világszerte (Damialis et al., 2007). Az ezzel párhuzamosan zajló globális éghajlatváltozás miatt logikus felvetés, hogy a pollenszezon fenológiai jellemzői (a pollenszezon kezdete, vége, tartama) és mennyiségi jellemzői (évi összes pollenszám, napi pollenszámok éves maximuma: éves csúcspollen) is változást mutatnak. Fontos tényező, hogy a pollen és az egyéb légszennyező anyagok közötti kölcsönhatások úgy módosíthatják a pollenek tulajdonságait, hogy az érzékeny egyének még könnyebben válhatnak érzékenyekké (D'Amato, 2011). Ezért a pollenszezon feltételezett változásai nem feltétlenül magyarázzák a lakosság növekvő allergiás megbetegedéseit. Mégis fontos megvizsgálni, hogy az allergén pollenek karakterisztikái (fenológiai és mennyiségi jellemzői) mutatnak-e trendet a megfigyelt napi pollenszámok tükrében. Természetesen számos ilyen vizsgálat történt már, de tudomásunk szerint ez idáig mindössze három tanulmány adott átfogó képet a regionális pollenflóráról, nevezetesen Clot (2003), Damialis et al. (2007) és Cristofori et al. (2010) munkái rendre 25, 16 és 23 taxon figyelembevételével.

A hazai viszonyok jellemzésére 19 taxon napi pollenszámait vizsgáltuk a rendelkezésünkre álló 1997-2007 közötti 11 éves időszakban Szegedre (Makra et al., 2011a). Ezek a taxonok (*Alnus* (éger), *Ambrosia* (parlagfű), *Artemisia* (üröm), *Betula* (nyír), *Cannabis* (kender), Chenopodiaceae (libatopfélék), *Juglans* (dió), *Morus* (eperfa), *Pinus* (fenyő), *Plantago* (útifű), *Platanus* (platán), Poaceae (fűfélék), *Populus* (nyár), *Quercus* (tölgy), *Rumex* (lórom), *Taxus* (tiszafa), *Tilia* (hárs), *Ulmus* (szil) és *Urtica* (csalán)) a vizsgált időszak összes pollenmennyiségének 93,2%-át adják. A három legnagyobb pollenszámokat mutató taxon az *Ambrosia* (32,3%), Poaceae (10,5%) és *Populus* (9,6%). Az adatokat a Szegedi Tudományegyetem Bölcsészettudományi Kari épületének a tetején 20 m nagasságban üzemelő Hirst-típusú pollencsapda (Hirst, 1952) szolgáltatta.

Két lényeges körülményre kell felhívni a figyelmet. Az egyik, hogy a vizsgált pollen karakterisztikák valószínűleg nem tekinthetők normális eloszlásúaknak (például az éves csúcspollen), ezért a *t*-próba helyett (mely feltételezi a normalitást) egy nemparaméteres próbát, a Mann-Kenndall-tesztet (Önöz and Bayazit, 2003) alkalmaztunk. A másik körülmény, hogy a trend létének igazolása mindössze 11 adat felhasználásával igen kevéssé ígérkezhet sikeresnek. Valóban, a 19 taxon 5 karakterisztikájára elvégzett összesen 95 vizsgálat mindössze 16, 10 és 3 esetben jelzett trendet a 10, 5 és 1 %-os szignifikancia-szinten, illetve pusztán az évi összes pollenszám esetében a 19 taxonra csupán 4, 1 és 0 esetben mutatkozott szignifikáns trend az említett szinteken. Ezért a napi pollenszámokra a 11 éves időszak 11 adata alapján az év összes napjára (a pollenszezon idejére) külön-külön elvégeztük a Mann-kendall-tesztet (MK-tesztet). A nagyszámú próbastatisztika egyedi kiértékelése értelmetlen lenne, ezért kihasználtuk, hogy az MK-próbastatisztika a trend hiányára vonatkozó null-hipotézis teljesülése esetén (aszimptotikusan) standard normális eloszlású. A trend létezéséről szóló döntés így azonos azzal a problémával, hogy a napi MK-teszt értékek évi átlaga szignifikánsan különbözik-e nullától. Az ezzel kapcsolatos klasszikus

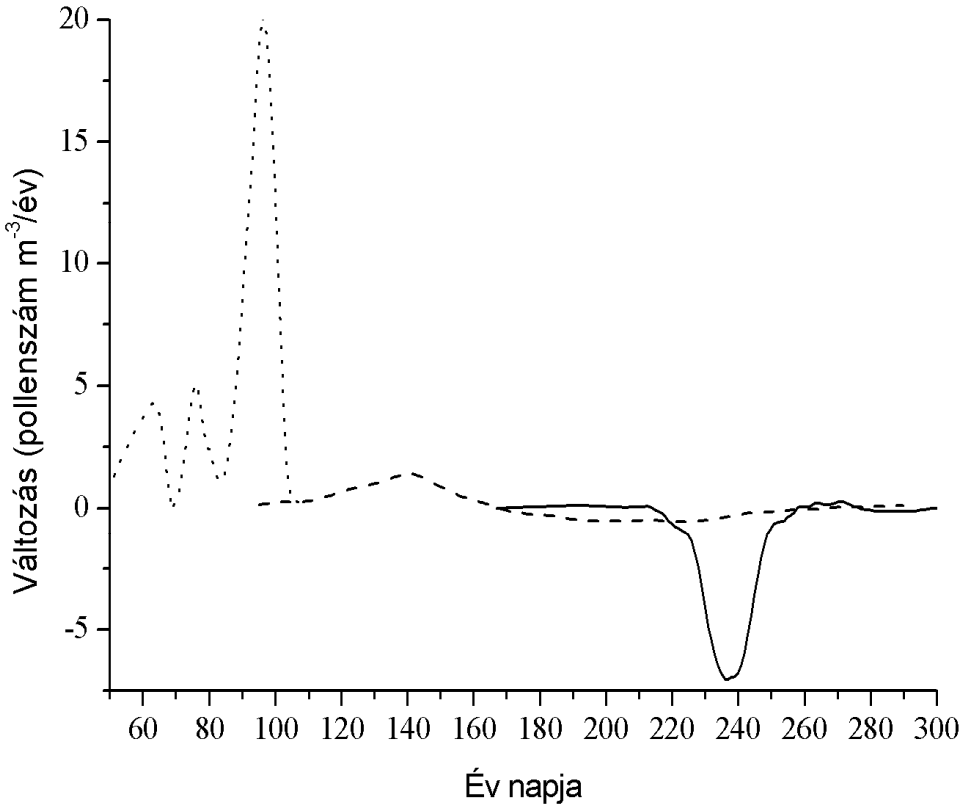
t-próba most az *u*-próbába megy át (Dévényi és Gulyás, 1988), mivel a szórás ismert (eggyel egyenlő az MK próbastatisztika standard normális eloszlása miatt), ugyanakkor módosítottuk azt, mert az egymást követő MK-teszt értékek közötti korrelációk miatt az évi átlag szórása nagyobb, mint korrelálatlan adatok esetén. Elsőrendű autoregresszív (AR(1)) modellt illesztettünk a napi MK-teszt értékekhez, és az illesztett AR(1) modell felhasználásával (Matyasovszky, 2002, 71. oldal) adtuk meg az említett szórást. Ekkor az 5%-os valószínűségi szinten már 11 taxon évi összes pollenszáma mutat szignifikáns trendet, s e 11-ből 7 jelez növekedést (1. Táblázat). Megtörténhet azonban, hogy a pollenszezon pozitív és negatív trendeket mutató időszakokból áll össze és az MK-teszt értékek átlaga emiatt nem ad átfogó (teljes évre számított) trendet 5, 8 és 11 taxonra a 10, 5 és 1 %-os szinten.

*1. Táblázat. A napi lineáris trendekből számított évi összes pollenszám változás (ÉÖPV) 10 évre vonatkozóan (pollenszám m^{-3}/ 10 év). A *, **, *** szimbólumok rendre a 10, 5, 1%-os szignifikancia-szintre utalnak.*

Taxon	ÉÖPV
Alnus	-214
Ambrosia	-1170
Artemisia	-60
Betula	-60
Cannabis	47*
Chenopodiaceae	-175**
Juglans	253***
Morus	400***
Pinus	-194***
Plantago	91**
Platanus	271**
Poaceae	176
Populus	2981***
Quercus	236*
Rumex	-505***
Taxus	678***
Tilia	-65*
Ulmus	-160***
Urtica	1183***

Ezt a lehetőséget azzal lehet vizsgálni, hogy a napi MK-teszt értékeket a WLR módszerrel simítjuk, megkapva ezzel a próbastatisztika éves menetét. Ha egyetlen napon sincs

trend, akkor a becsült sávszélesség rendkívül nagy (gyakorlatilag végtelen), ami egy közel nulla meredekségű egyenest eredményez, mivel a napi trendek évi ciklusához történő lokálisan lineáris közelítés globálisan lineáris lesz. A simítás során azonban minden egyes taxonra jól definiált véges sávszélességet nyertünk, ami az összes taxonra trendet jelez. A WLR módszer a napi bontású MK-statisztikákon keresztül tehát összehasonlíthatatlanul finomabb képet nyújt az allergén pollenek trendjéről, mint a mások által követett szokványos eljárás. Az 5. ábra a korábban említett három legnagyobb pollenszámokat mutató taxon változásának (a lineáris trend meredekségének) évi menetét mutatja.



5. ábra
Az Ambrosia (folytonos vonal), Poaceae (szaggatott vonal) és Populus (pontozott vonal) változásának (a lineáris trend meredekségének) évi menete

Különböző meteorológiai változók (minimum hőmérséklet, maximum hőmérséklet, középhőmérséklet, globálsugárzás, relatív nedvesség, szélsebesség és a csapadékösszeg) napi

értékeire hasonló vizsgálatot végeztünk, s ezek 11 éves trendjének évi menetét kapcsolatba hoztuk az egyes taxonok 11 éves trendjének évi menetével. A meteorológiai adatokat egy Szeged belvárosában található monitoring állomás szolgáltatta. Azt találtuk, hogy az egyes meteorológiai változók trendjei, figyelembe véve a taxonok klimatikus igényeit, igen jól magyarázzák a pollenkoncentrációk trendjeit. Ez jól kivehető a meteorológiai változók trendjeinek évi menete és a taxonok trendjének évi menete közötti korrelációból. A taxonok trendjének évi menete és a meteorológiai változók trendjeinek évi menete közötti többszörös korreláció egészen meglepően nagynak adódott. A legnagyobb többszörös korreláció az *Artemesia* esetén 0,998, de az *Ambrosia* és *Urtica* esetén fellépő legalacsonyabb 0,827-es érték is igen magas. A részleteket lásd Makra et al. (2011a) tanulmányában.

2. REGRESSZIÓ, KVANTILIS REGRESSZIÓ

A WLR azokban az esetekben is alkalmazható, amikor a független változó (prediktor) értékei nem időpontok vagy egyéb determinisztikus mennyiségek, hanem maga is valószínűségi változó, sőt változók (prediktorok) realizációi (Fan, 1992). Ilyenkor a becsülendő változó (prediktandusz) feltételes várható értékének becslése a cél a *p*-számú prediktor adott $\underline{x}$ értéke mellett. A következőkben ezt a regressziós eljárást, illetve az ennek általánosításaként is értelmezhető kvantilis regressziót tekintjük át, majd mutatunk be egy alkalmazást.

2.1. MÓDSZER

Az $\hat{y}(\underline{x}) = \hat{a} = \hat{a}(\underline{x})$ becslés a

$$\sum_{i=1}^{n}\left(y_i - a - \sum_{j=1}^{p} c_j (x_j - x_{ij})\right)^2 \left|\underline{\underline{B}}\right|^{-1} K\left(\underline{\underline{B}}^{-1}(\underline{x}_i - \underline{x})\right) \qquad (2.1)$$

mennyiség adott $\underline{x}$ melletti, $a, c_1, ..., c_p$ szerinti minimalizálásával nyerhető. Itt $\underline{x} = (x_1, ..., x_p)^T$, míg $\underline{x}_i = (x_{i1}, ..., x_{ip})^T$ az y_i-hez tartozó prediktorok vektora, továbbá $\underline{\underline{B}}$ a sávszélesség mátrixa, $\left|\underline{\underline{B}}\right|$ a $\underline{\underline{B}}$ determinánsa, végül $K(\underline{u})$ itt most nem részletezett tulajdonságokkal rendelkező *p*-változós magfüggvény. Mivel *p*>1 esetén $\underline{\underline{B}}$ becslése meglehetősen bonyolult volna, ezért a magfüggvényre és a sávszélesség mátrixára vonatkozó különböző egyszerűsítések mellett oldják meg (2.1) minimalizálását (Hardle and Müller, 2000).

A (2.1) formula a $\rho(u) = u^2$ jelölés bevezetésével természetszerűleg átírható a

$$\sum_{i=1}^{n}\rho\left(y_i - a - \sum_{j=1}^{p} c_j (x_j - x_{ij})\right)\left|\underline{\underline{B}}\right|^{-1} K\left(\underline{\underline{B}}^{-1}(\underline{x}_i - \underline{x})\right) \qquad (2.2)$$

alakba. Ha e helyett (2.2)-ben a $\rho(u) = |u|$ függvényt alkalmazzuk, akkor a súlyozott abszolút hibák összegét minimalizáljuk, aminek megoldása a prediktandusz feltételes mediánjának becslését nyújtja a *p*-számú prediktor adott $\underline{x}$ értéke mellett. Az így értelmezett medián regresszió akkor igazán hasznos, amikor a prediktandusz valószínűségi eloszlása erősen aszimmetrikus és így a medián és a várható érték jelentősen különbözik. A gyakorlati feladatok során ugyanis nem annyira a minél kisebb négyzetes hiba, hanem a minél alacsony abszolút hiba biztosítása a cél.

Mivel a medián a τ =0,5 valószínűségi értékhez tartozó kvantilis, a medián regresszió általánosításával bármely zérus és egy közötti τ-ra értelmezhető az ún. kvantilis regresszió (Koenker and Bassett, 1978). Ez a prediktandusznak a prediktorok adott értéke melletti

feltételes kvantilis becslését végzi. Megjegyezzük, hogy a τ-kvantilis az a szám, amelynél kisebb értéket a szóban forgó valószínűségi változó τ valószínűséggel vesz fel. Ilyenkor (2.2)-ben a $\rho(u) = (1-\tau)|u|$, $u<0$ és $\rho(u) = \tau|u|$, $u \geq 0$ választással kell élni (Koenker, 2005). Számos τ esetében végrehajtva a kvantilis regressziót, képet kaphatunk a prediktandusznak a prediktorok melletti feltételes valószínűségi eloszlásáról is.

2.2. ALKALMAZÁS: Napi parlagfű pollenkoncentráció

Példaként bemutatjuk a hazánkban nagyon elterjedt parlagfű erősen allergén pollenjének napi koncentráció becslését. Szeged, Legnano és Lyon napi parlagfű pollenkoncentrációit hoztuk kapcsolatba (Makra et al., 2011b) a megelőző napi koncentrációval és a megelőző napi átlaghőmérséklettel, csapadékösszeggel és átlagos szélsebességgel az 1997-2006 időszakban. A Szegeden kívüli további két város bevonására azért került sor, mert a Kárpát-medencén kívül még a Pó-alföld (Legnano) és a Rajna völgye (Lyon) Európa erősen parlagfüves területei (Makra et al., 2011b).

Mivel a pollenkoncentrációk (és természetesen a meteorológiai változók is) jelentős évi menettel rendelkeznek, ezért az imént bemutatott eljárást az időtől függővé kell tenni, hiszen a rendelkezésre álló adatok a különböző időpontokhoz tartozó különböző valószínűségi változókból származnak. Az időfüggést még az is indokolja, hogy a meteorológiai változók adott értékéhez a pollenszezon különböző szakaszaiban szisztematikusan eltérő pollenkoncentrációk tartozhatnak. Például egy október elején fellépő viszonylag magas hőmérsékletre más pollen produkcióval reagál a növény, mint ugyanezen hőmérsékletre augusztus-szeptember fordulóján, ami a maximális koncentrációk időszaka. Ezért a koncentrációk becslését az

$$\hat{y}_i = a_0(t_i) + \sum_{j=1}^{p} a_j(t_i) x_{ij} \qquad (2.3)$$

időfüggő lineáris regresszió formájában keressük. Az általános (2.2) becslés időfüggő általánosítását azért célszerű elvetni, mert még az időfüggetlen esetben is a prediktorok növekvő száma mellett exponenciális ütemben növekvő számú adatra van szükség a regressziós felület adott sűrűségű pontokkal történő reprezentációjához. Más szóval, ha *p* nem kifejezetten kicsi, akkor a (2.2) becslés által nyújtott regressziós felület megfelelő pontosságú reprezentálása irreálisan sok adat esetén volna biztosítható. A jelenség Bellman (1961)

nyomán „dimenzió átok” néven ismeretes. Az időfüggő regressziós együtthatók a t_i időpontra Cai (2007) nyomán a

$$\sum_{k=1}^{n} \rho\left(y_k - a_0 - c_0(t_k - t_i) - \sum_{j=1}^{p} \left(a_j + c_j(t_k - t_i)\right) x_{kj} \right) K\left(\frac{t_k - t_i}{b} \right) \tag{2.4}$$

mennyiség a_j, c_j, j=0,…,p szerinti minimalizálásával becsülhetők $1 \le i \le n$ mellett. Ez az ún. time-varying coefficient model a WLR módszer természetes általánosítása, amelynek korábbi meteorológiai alkalmazásáról nincsen tudomásunk. Itt a $\rho(u)$ függvényt a szerint választjuk, hogy regressziót vagy kvantilis regressziót hajtunk-e végre. Az időpontok értelmezésénél ügyelni kell az évi menetre, ezért az év egy adott napja ugyanazt az időpontot viseli minden évben. Megjegyezzük, hogy a csapadék ún. intermittent jelenség, tehát nincsen mindig, ezért a (2.4) formula bizonyos módosítása szükséges (Li and Racine, 2004), amire terjedelmi korlátok miatt most nem térünk ki. A részleteket lásd Makra et al. (2011b) tanulmányában. Megemlítjük azonban, hogy a napi adatok jelentős autokorrelációval rendelkeznek, ezért egy, az (1.3)-mal analóg kritérium használata nem alkalmas a sávszélesség becslésére, de a feladat különbözősége folytán (1.4) sem jöhet szóba. Ezért úgy jártunk el, hogy minden évre az adott év pollenkoncentrációinak becslésekor az adott év összes adatát kirekesztettük (2.4)-ből, és az így kapott $\tilde{y}_i$ becslésekkel értelmezett

$$\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \tag{2.5}$$

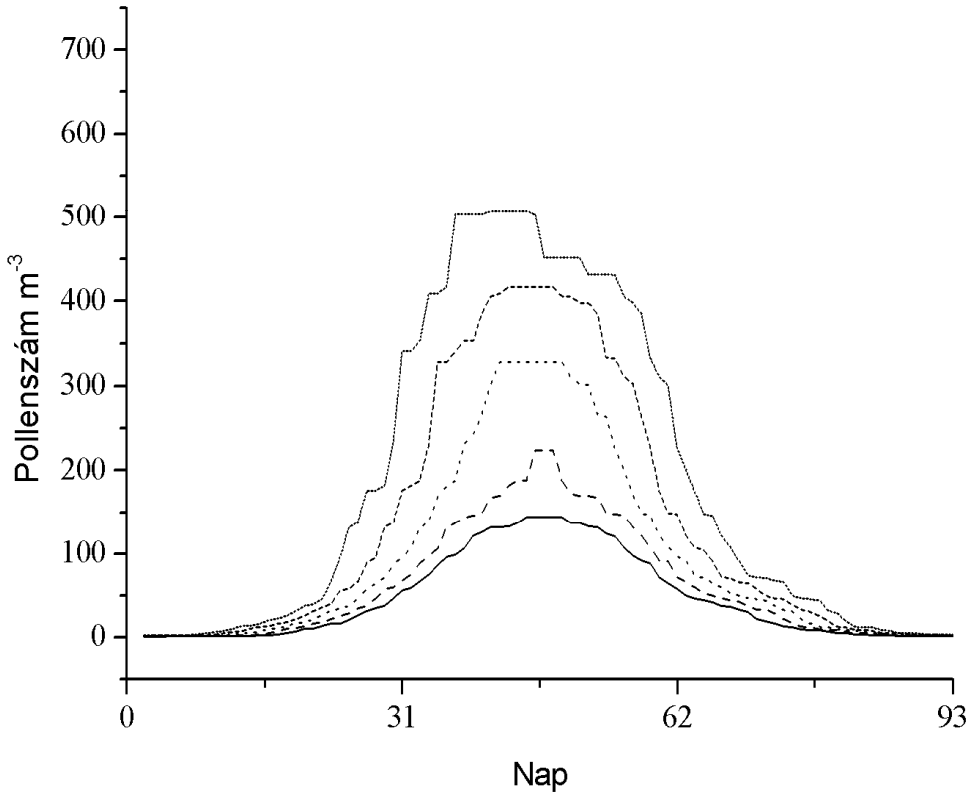
mennyiséget minimalizáltuk b szerint. Ekkor minden időpontra a tízévnyi adat helyett csak kilencévnyit használunk fel, ezért - figyelembe véve az optimális sávszélességnek az adatsor hosszától való függését (Cai, 2007) - a sávszélesség végső becslése $\hat{b} = (9/10)^{1/5} \tilde{b}$ lett, ahol $\tilde{b}$ minimalizálja (2.5)-öt.

2. Táblázat
A napi parlagfű koncentráció 1 napos előrejelzésének hibája (RMSE: átlagos négyzetes hiba gyöke, MAE: abszolút hibaátlag) időfüggő lineáris regresszióval

Város	Lyon		Legnano		Szeged	
Hiba (poll.szám m^{-3})	RMSE	MAE	RMSE	MAE	RMSE	MAE
Előrejelzés	36.3	13.3	34.1	13.3	73.0	26.6
Éves trend	43.2	16.8	38.6	15.5	105.6	42.5

Első pillantásra a 2. táblázat azt mutatja, hogy az előrejelzés a legnagyobb koncentrációkkal rendelkező Szeged esetében a legkevésbé sikeres. Valójában azonban éppen ellenkező a helyzet, ha az előrejelzési hibákat pusztán az évi menettel történő becslési hibákhoz hasonlítjuk. Az éves trendet úgy kaphatjuk meg, hogy a prediktorokat figyelmen kívül hagyjuk, vagyis a (2.3) és (2.4) egyenletekben p=0. Ekkor a becslés által megmagyarázott relatív variancia ($1-RMSE^2_{Előlőrejels}/RMSE^2_{Évestrend}$) Szegedre a legnagyobb (52,2%) és Legnanora a legkisebb (22%), tehát a legpontosabban Szeged napi parlagfű pollenkoncentrációja becsülhető a három hely közül. A legfontosabb meteorológiai változónak a napi középhőmérséklet (Szeged és Legnano) és a napi csapadék (Lyon) bizonyult. Az optimális prediktorok fontossági sorrendjének és számának kiválasztása a jól ismert stepwise regresszióhoz (Draper and Smith, 1981) hasonló módon történt. Az alapgondolat a következő. Tegyük fel, hogy valahány prediktor szerepel már a becslési formában. Mivel újabb prediktor bevonása magasabb dimenziós becslési felületet jelent a prediktorok terében és ez a magasabb dimenziós felület nagyobb mennyiségű adattal reprezentálható (lásd „dimenzió átok”), ezért az optimális sávszélesség nagyobbnak várható, mint az alacsonyabb dimenziós esetben. A nagyobb sávszélesség azonban a becslés nagyobb torzítását eredményezi. Ezért a magasabb számú prediktorhoz tartozó megmagyarázott variancia és az alacsonyabb számú prediktorhoz tartozó megmagyarázott variancia viszonya attól függ, hogy az újonnan bevont prediktor a megnövekedett sávszélesség mellett is tartalmaz-e annyi információt a prediktanduszra nézve, ami ellensúlyozza a torzítás négyzetének növekedését.

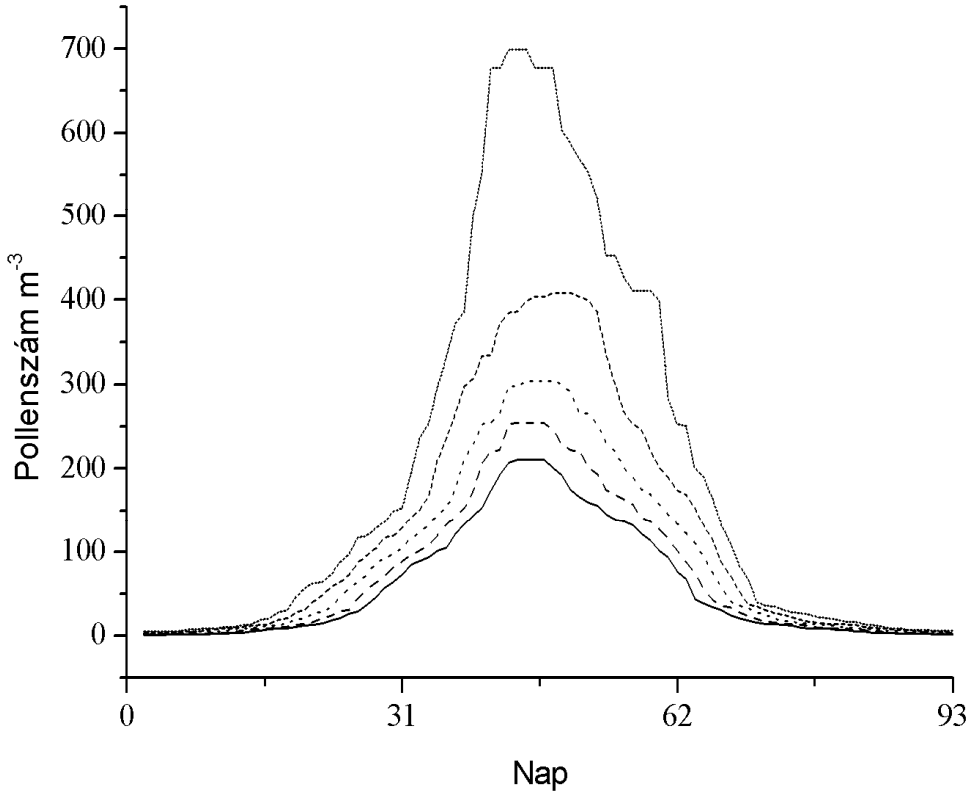
A kvantilis regresszióval kapcsolatos eredményeink Szegedre (Makra and Matyasovszky, 2010) a következőképp foglalhatók össze. A medián regresszió természetesen kisebb abszolút hibaátlagot hozott, nevezetesen 21,2 pollenszám m^{-3} értéket, ami 20,9%-kal kisebb, mint a 2. táblázat megfelelő MAE értéke. A kvantilis regressziót ezúttal az esős és száraz napokra szétválasztva külön-külön értelmeztük, mert a már említett intermittencia nehézkessé teszi a csapadék kezelését. Ugyanakkor részben a pollenszórásra gyakorolt hatása, de még inkább a pollen részecskék kimosódása folytán hasznos a csapadék figyelembevétele. Az előző napi koncentráción kívül az esős napokon a napi globálsugárzás, a száraz napokon a



6. ábra
Napi parlagfű kvantilisek éves trendje a csapadékos napokon a 0,5 (folytonos), 0,6 (szaggatott), 0,7 (pontozott), 0,8 (sűrű szaggatott), 0,9 (sűrű pontozott) kvantilisekre. A horizontális tengelyen lévő 93-as szám a pollenszezon (július 15 – október 15) hosszára utal.

napi középhőmérséklet bizonyult fontos prediktornak. Ha elhagyjuk az összes prediktort (p=0 (2.3)-ban), akkor a kvantiliseknek pusztán az időtől való függéséhez jutunk. A számítások

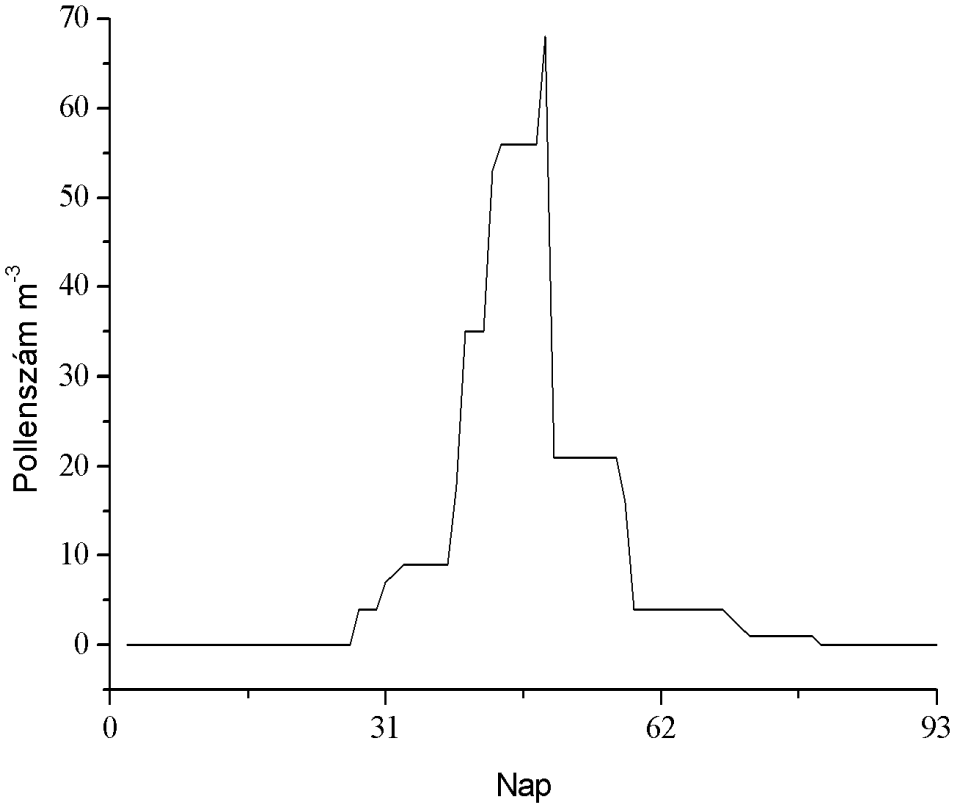
szerint a napi parlagfű pollenkoncentráció kvantilisei általában kisebbek az esős, mint a száraz napokon, továbbá a napi koncentráció valószínűségi eloszlása sokkal elnyújtottabb a magas koncentrációk felé a száraz napokon. Az esős napokhoz tartozó kvantilisek azt jelzik, hogy a pollenkoncentrációk jóval kisebb változékonyságúak a csapadékos napokon (6. és 7. ábra). Mindez világosan jelzi a csapadék koncentrációcsökkentő hatását.



7. ábra
Napi parlagfű kvantilisek éves trendje a csapadékmentes napokon a 0,5 (folytonos), 0,6 (szaggatott), 0,7 (pontozott), 0,8 (sűrű szaggatott), 0,9 (sűrű pontozott) kvantilisekre. A horizontális tengelyen lévő 93-as szám a pollenszezon (július 15 – október 15) hosszára utal.

A hazai parlagfű pollenterhelés súlyosságára jól rávilágít a kvantilis regressziónak a $\tau = 0$ valószínűség melletti alkalmazása is. Ezzel tulajdonképpen a koncentrációk alsó határát lehet meghatározni, mert a $\tau = 0$ melletti kvantilis az a legnagyobb érték, amelynél egy

valószínűséggel nagyobb koncentráció fordul elő. Az érzékeny egyének körülbelül 20 parlagfű pollenszám m^{-3} koncentrációnál már számottevő allergiás tüneteket mutatnak, ezért az egészségi kockázatot jelentő kritikus parlagfű koncentrációnak a 20 pollenszám m^{-3} tekinthető (Jäger, 1998). Megjegyezzük, hogy ezt a küszöböt a parlagfűvel erősen érintett országokban hozták, másutt 5-10 pollenszám m^{-3}-nek veszik, mert az érzékeny egyéneknél már ekkor kezdenek jelentkezni a tünetek. A 8. ábra világosan jelzi, hogy a lehetséges legkisebb koncentráció is csaknem 20 napon át bizonyosan meghaladja az említett küszöbértéket még úgy is, hogy most a prediktorok értékét, tehát például egy előző napi esetlegesen magas koncentrációt figyelembe sem vesszük (p=0 a (2.3)-ban). Ezúttal az összes



8. ábra
Napi parlagfű koncentráció alsó határának éves trendje. A horizontális tengelyen lévő 93-as szám a pollenszezon (július 15 – október 15) hosszára utal.

nap együtt szerepel, mivel a csapadékos és csapadékmentes napok közötti különbség elenyészőnek mutatkozott. Meg kell említeni, hogy a kvantilis regresszió $\tau = 0$ esetén (extrém kvantilis regresszió) a korábbiaktól eltérően történik (Chernozhukov, 2005). Nevezetesen, a (2.3)-ban szereplő regressziós együtthatók $1 \le i \le n$ esetén (2.4)-nek a $\rho(u) = u$ választás melletti a_j, c_j, j=0,…,p szerinti minimalizálásával becsülhetők azon feltétel mellett, hogy

$$a_0 + c_0(t_k - t_i) + \sum_{j=1}^{p} \left(a_j + c_j(t_k - t_j)\right) x_{kj} < y_k . \qquad (2.6)$$

3. SPEKTRÁLANALÍZIS

Az éghajlati idősorok spektrálanalízisének irodalma hihetetlenül gazdag. Ennek áttekintésére terjedelmi korlátok folytán kísérletet sem teszünk, hanem néhány speciális kérdést tárgyalunk. Az alapfogalmak tisztázása érdekében tekintsünk egy diszkrét paraméterű Y_t stacionárius sztochasztikus folyamatot. Ez egy X_t és egy Z_t, egymástól független stacionárius folyamatok összegeként áll elő, ahol X_t diszkrét spektrumú, tehát a kovarianciafüggvénye J számú periodikus tag összege, azaz

$$B_X(k) = \frac{1}{2}\sum_{j=1}^{J} A_j{}^2 \cos(\omega_j k), \tag{3.1}$$

míg Z_t folytonos spektrumú, tehát megszámlálhatatlan periodikus tag összege és a kovarianciafüggvénye és az ún. spektrális sűrűségfüggvénye között a

$$B_Z(k) = \int_0^{\pi} g(\omega)\cos(\omega k)d\omega, \quad g(\omega) = \frac{B_Z(0)}{\pi} + \frac{2}{\pi}\sum_{k=1}^{\infty} B_Z(k)\cos(\omega k) \tag{3.2}$$

relációk állnak fenn. A spektrálanalízis feladata az $\{\omega_j\}$ frekvenciák (esetleg a hozzájuk tartozó $\{A_j\}$ amplitúdókkal) és a $g(\omega)$ spektrális sűrűségfüggvény becslése. A feladat teljes általánosságban szinte megoldhatatlan, mert két tag összegére vonatkozó $y_1,...,y_n$ megfigyelt idősor birtokában kell a két, egyenként nem megfigyelhető összetevőre következtetni.

3.1. MÓDSZEREK

A módszerek rendszerint az

$$I(\omega) = \frac{1}{\pi n}\left[\left(\sum_{j=1}^{n} y_j \cos(\omega j)\right)^2 + \left(\sum_{j=1}^{n} y_j \sin(\omega j)\right)^2\right] \tag{3.3}$$

periodogramból indulnak ki, melyet az $\omega = \lambda_i = 2\pi i / n, i = 1,..., L$ pontokban szokás kiszámolni, ahol *L* az *n*/2 egész része (az *n*/2-nél nem nagyobb legnagyobb egész). A (3.3) várható értéke nagy *n* esetén jó közelítéssel $g(\lambda_i)$, ha λ_i nincs közel egyik ω_j-hez sem, illetve $n/(4\pi)A_j^{\ 2} + g(\omega_j)$, ha $\lambda_i = \omega_j$. Ez utóbbi kifejezésben az első tag dominál, ha az amplitúdó nem túl kicsi és az idősor nem túl rövid. Tegyük fel egy időre, hogy diszkrét spektrum nincs jelen a folyamatban. Ekkor tehát a periodogram a spektrális sűrűségfüggvény aszimptotikusan torzítatlan becslése. Sajnos a becslés konzisztenciájáról nem beszélhetünk, mert a periodogram varianciája *n* növekedésével nem tart zérushoz, hanem nagy *n*-re jó közelítéssel $Var[I(\omega)] = g^2(\omega), \omega \neq 0, \pi$ írható. Az $I(\lambda_i), I(\lambda_j), i \neq j$ periodogram elemek korrelálatlanok (meglehetősen általános feltételek mellett függetlenek is), és $I(\lambda_i)$ aszimptotikusan exponenciális eloszlású $1/g(\lambda_i)$ paraméterrel (Kokoszka and Mikosch, 2000). A spektrális sűrűségfüggvény becslése analóg a trendfüggvény becslésével, csak most az időt a körfrekvencia helyettesíti. Nevezetesen a $\hat{g}(\omega) = \hat{a} = \hat{a}(\omega)$ becslés a

$$\sum_{k=1}^{L} \rho\big(I(\lambda_k) - a - c(\lambda_k - \omega)\big)K\left(\frac{\lambda_k - \omega}{b}\right) \tag{3.4}$$

mennyiség adott ω melletti, *a* és *c* szerinti minimalizálásával nyerhető, ahol $\rho(u) = u^2$. Ennek megoldása kielégíti a

$$\sum_{k=1}^{L}\psi\big(I(\lambda_k)-a-c(\lambda_k-\omega)\big)K\left(\frac{\lambda_k-\omega}{b}\right)=0\,,$$

$$\sum_{k=1}^{L}(\lambda_k-\omega)\psi\big(I(\lambda_k)-a-c(\lambda_k-\omega)\big)K\left(\frac{\lambda_k-\omega}{b}\right)=0 \qquad (3.5)$$

egyenletrendszert, ahol $\psi(u)=\rho'(u)$.

Az éghajlati adatsorok azonban általában nem mentesek a diszkrét periódusoktól (gondoljunk például az évi menetre), sőt a spektrálanalízis talán legfontosabb feladata éppen az ilyen diszkrét periódusok detektálása. Ezért ha λ_i közel van valamelyik ω_j-hez, akkor nem teljesül az aszimptotikus $E\big[I(\lambda_i)\big]=g(\lambda_i)$ reláció, vagyis az ilyen periodogram elemek kiugróak a többiekhez képest. A cél tehát olyan, ún. robusztus eljárást értelmezni a spektrális sűrűségfüggvény becslésére, amely gyakorlatilag nem vesz tudomást az ilyen kiugró értékekről. Ezt követően - a spektrális sűrűségfüggvény ismeretében - van mód a diszkrét periódusok jelenlétének tesztelésére. A probléma lényege az, hogy a diszkrét frekvenciákban (és környezetükben) az OLS eljárást generáló $\rho(u)=u^2$ veszteségfüggvény révén a becsült spektrális sűrűség erősen idomul a kiugró periodogram értékekhez, és így a becslés megbízhatatlanná válik, mert jelentős torzítás lép fel. A megoldás az, hogy (3.4)-ben olyan $\rho(u)$ veszteségfüggényt alkalmazunk, ami a becsült spektrális sűrűség görbéje és a periodogram elemek közötti nagy eltéréseket nem bünteti túlságosan erősen, és ennek következtében a kapott görbe elkerüli a kiugró értékeket. E feladat a robusztus becslések elméletét igényli, amivel kapcsolatban a legkézenfekvőbb hivatkozás Huber (1981) munkájának említése.

3.1.1. Robusztus becslés

Az említett, Janas and von Sachs (1995) nevéhez fűződő robusztus becslést Matyasovszky (2010a) alkalmazta éghajlati adatsorokra. Az eljárás lényege a következő. Legyen $a = a(\omega)$ olyan, ami kielégíti a

$$\sum_{k=1}^{L} \psi\left(\frac{I(\lambda_k) - a}{a}\right) K\left(\frac{\lambda_k - \omega}{b}\right) = 0 \qquad (3.6)$$

egyenletet, és ekkor $\hat{g}(\omega) = a = a(\omega)$. A robusztusság érdekében $\psi(u)$ pozitív és negatív értékeket egyaránt felvevő, monoton növő, korlátos függvény, továbbá $\hat{g}(\omega)$ aszimptotikus torzítatlanságához a $\psi(J-1)$ valószínűségi változó zérus várható értékű. Itt J egy-paraméterű (standard) exponenciális eloszlású valószínűségi változó. Ehhez $\psi(u)$ a $\psi_H(u) = \max\{-c, \min\{c, u\}\}$ Huber-függvénnyel (Huber, 1981) a $\psi(u) = (1-e^{-1})\psi_H(u), u < 0$, $\psi(u) = \psi_H(u), u \geq 0$ formában áll elő c=1 mellett. A (3.5) egyenletekhez képesti különbség egyrészt abból fakad, hogy a lokális lineáris közelítés helyett lokálisan konstans közelítés történik, másrészt a ψ argumentumában történő a-val való osztás (a lokálisan konstans közelítés helyessége esetén) azonosan egy szórásúvá teszi az $(I(\lambda_k) - a)/a$ mennyiségeket ha $a = a(\omega) = g(\omega)$. Ez, továbbá a ψ függvény fenti választása a periodogram elemek exponenciális eloszlásából fakad. Végül a Huber-függvény választása kézenfekvő a robusztus becslések körében betöltött szerepének fontossága révén (Huber, 1981).

Janas and von Sachs (1995) bizonyította a vázolt becslés konzisztenciáját, de nem foglalkozott olyan gyakorlati kérdéssel, mint a becslés aszimptotikus szórása, vagy a sávszélesség megadása. Az előző kérdés a becslés konfidencia-intervallumának származtatásához szükséges, ami pedig a diszkrét periódusok detektálása szempontjából fontos. Matyasovszky (2010a) ezért előállította $\hat{g}(\omega)$ aszimptotikus szórását, de a diszkrét periódusok tesztelésére hasznosabbnak bizonyult az alábbi szimulációs eljárás. Mivel a

$J(\lambda_i) = I(\lambda_i) / g(\lambda_i)$ normalizált periodogram standard exponenciális eloszlást követ, ezért 1. Szimulálunk *L* számú standard exponenciális eloszlású véletlen számot, mely egyben egy periodogramnak tekinthető. 2. A sávszélesség ismeretében a fent bemutatott eljárással (lásd (3.6) egyenlet) előállítjuk ezen periodogramot generáló spektrális sűrűségfüggvény becslését, majd a hozzá tartozó normalizált periodogramot. 3. Az 1.-2. lépés *N*-szeri ismétlésével (*N* nagy, például *N*=10000) minden λ_i pontra külön-külön előállítjuk az *N* számú normalizált periodogram elemek empirikus eloszlásfüggvényét. Adott $\varepsilon\,100\%$ szignifikancia-szint esetére az empirikus eloszlásfüggvények $1-\varepsilon$ kvantilisei szolgáltatják azon null-hipotézis elfogadási tartományának határait, hogy az adott frekvenciákban nincsen diszkrét periódus. A null-hipotézis tesztelésekor tehát a $\hat{J}(\lambda_i) = I(\lambda_i) / \hat{g}(\lambda_i)$ értékeket a 3. lépésben nyert elfogadási tartománnyal vetjük össze.

A sávszélesség becslésére (1.3) kívánkozna úgy, hogy y_i szerepét $I(\lambda_i)$, $\hat{f}_i(t_i)$ szerepét $\hat{g}_i(\lambda_i)$ veszi át. Diszkrét frekvencia környékén azonban rendkívül nagy négyzetes hibák léphetnek fel, ezért túlzottan kis sávszélesség adódna optimálisnak. Matyasovszky (2010a) ezért eredetileg a nagyon nagy és a nagyon kicsi négyzetes hibák elhagyásával módosította (1.3) minimalizálását. Erre azért volt szükség, mert Fan and Jiang (1999) vagy Heng and Leung (2005) által a robusztus nemparaméteres regresszióra nyert sávszélesség becslési technikájának konzisztenciája csak szimmetrikus eloszlású valószínűségi eloszlásokra bizonyított, és az exponenciális eloszlás nem ilyen. Későbbi tapasztalatunk szerint azonban eljárásuk jelen esetben is működik, vagyis a sávszélesség a

$$\sum_{i=1}^{L} \rho\left(\frac{I(\lambda_i) - \hat{g}_i(\lambda_i)}{\hat{g}_i(\lambda_i)}\right) \tag{3.7}$$

mennyiség *b* szerinti minimalizálásával nyerhető.

A diszkrét periódusok tesztelése a következőképp történhet. Igen általános feltételek mellett diszkrét periódusok hiányában a $\max_{1\le i\le L}\{J(\lambda_i)\}-\ln L$ próbastatisztika határeloszlása a standard Gumbel-eloszlás (Kokoszka and Mikosch, 2000). A gyakorlatban természetesen a $\hat{J}(\lambda_i)$-t írjuk $J(\lambda_i)$ helyébe, és a fenti próbastatisztika konkrét értékét a standard Gumbel-eloszlásból nyert, adott szignifikancia-szint melletti kritikus értékkel hasonlítjuk össze azon null-hipotézis mellett, hogy nincsen diszkrét frekvencia. Ha a null-hipotézis elvetésre kerül, akkor vizsgáljuk a második legnagyobb normalizált periodogram értéket az ő határeloszlásával, és az eljárás addig folytatódik, amíg találunk diszkrét periódust. Természetesen egy valóságos diszkrét frekvencia általában nem esik egybe egyik λ_i–vel sem, ezért a szignifikánsnak ítélt λ_j frekvencia pontosítása szükséges. Nevezetesen, $\hat{\lambda}_j$ az a frekvencia lesz, ami λ_j bizonyos környezetében maximalizálja a periodogramot (Chen et al., 2000).

3.1.2. Nem ekvidisztáns időpontokban rendelkezésre álló adatsor

A spektrálanalízis során széles körben követett gyakorlat, hogy elsőrendű autoregresszív (AR(1)) modellt illesztenek az adatsorhoz, és ha a periodogram adott frekvenciánál vagy a frekvenciák egy tartományánál meghalad egy küszöböt, akkor a spektrum ebben a pontban vagy tartományon különbözik az AR(1) spektrumtól. A küszöb nyilvánvalóan függ az AR(1) spektrumtól és a választott szignifikancia-szinttől. Az eredmények interpretációjakor természetesen azok a frekvenciák a fontosak, amelyeknél a modelltől való különbözőség megjelenik. Az eljárás alapja az, hogy az éghajlati adatsorokat jelentős részben vörös zaj jellemzi, amit az adatsorhoz illesztett AR(1) modellel közelítenek. A vörös zaj azt jelenti, hogy az egyre kisebb frekvenciák egyre fontosabb szerepet játszanak a folyamat kialakításában, tehát a spektrális sűrűség a magas frekvenciák irányába monoton csökkenő.

Legyen $y(t_1),...,y(t_n)$ egy stacionárius idősor a $t_1,...,t_n$ időpontokban megfigyelve. Általában az adatsorok ekvidisztánsan állnak rendelkezésre, tehát $\delta_i = t_i - t_{i-1} \equiv \delta, i = 2,...,n$. A δ értékét egységnek tekintve $t_i = i$ írható, és az adatsor az egyszerűbb $y_1,...,y_n$ jelöléssel látható el. Olykor azonban az adatsor nem ekvidisztáns időpontokban áll rendelkezésre, melyre jó példát szolgáltatnak a paleoklíma adatok. Ilyenkor mind az AR modell illesztése, mind a periodogram definíciója módosításra szorul.

Legyen Y_t zérus várható értékű (az egyszerűség kedvéért), stacionárius sztochasztikus folyamat. Egy ilyen AR(1) folyamatot az $Y_t = aY_{t-1} + e_t$ egyenlet definiál, ahol e_t fehérzaj folyamat σ_e^2 varianciával. A spektrális sűrűségfüggvénye

$$g(\lambda) = (\sigma_e^2 / \pi)/(1 + a^2 - 2a\cos(\lambda)), \quad \sigma_e^2 = (1 - a^2)\sigma^2, \tag{3.8}$$

ahol σ^2 az Y_t varianciája. A nem ekvidisztáns időpontok esetére Mudelsee (2002) az OLS eljárást javasolta az *a* autoregresszív paraméter becslésére az

$$y(t_i) = a^{\delta_i/\Delta} y(t_{i-1}) + e(t_i), i = 2,...,n, \tag{3.9}$$

AR(1) modellben, ahol $\Delta = (\delta_2 + ... + \delta_n)/(n-1)$ az átlagos időlépcső. Az $\hat{a}$ a

$$\sum_{i=2}^{n} (y(t_i) - a^{\delta_i/\Delta} y(t_{i-1}))^2 \tag{3.10}$$

mennyiség *a* szerinti minimalizálásával nyerhető. A nemlineáris legkisebb négyzetek módszerének elméletére alapítva (Nielsen, 2011) belátható, hogy $\hat{a}$ aszimptotikusan normális eloszlású

$$\frac{\sum_{i=2}^{n} \delta_i^2 / \Delta^2 \hat{a}^{2(\delta_i/\Delta - 1)} (1 - \hat{a}^{2(\delta_i/\Delta)})}{\left(\sum_{i=2}^{n} \delta_i^2 / \Delta^2 \hat{a}^{2(\delta_i/\Delta - 1)} \right)^2} \tag{3.11}$$

varianciával. Megjegyezzük, hogy Mudelsee (2002) a fenti analitikus forma helyett egy Monte-Carlo-szimulációs technikát javasolt $\hat{a}$ varianciájának meghatározására.

A vázolt eljárást számos paleoklimatológiai vizsgálat során felhasználták már, noha az OLS módszer erősen kritizálható, mert az $e(t_i)$ hiba $(1-a^{2\delta_i/\Delta})\sigma^2$ varianciája nagyobb a nagyobb δ_i időlépcsőknél. Az OLS módszerrel nyert $\hat{a}$ ezért elsősorban azon időszakokra van szabva, ahol az adatok időben ritkán helyezkednek el. Mivel a ritka mintavételezésű adatok nem tartalmazhatják a nagy frekvenciás ingadozásokat, az ilyen adatok túl erős perzisztenciát mutatnak, és ezért az *a* paraméter szisztematikus felülbecslése várható.

Ennek kiküszöbölésére a súlyozott legkisebb négyzetek (WLS) módszerét javasoljuk (Matyasovszky 2012b). Ekkor $\hat{a}$ a

$$\sum_{i=2}^{n}(y(t_i)-a^{\delta_i/\Delta}y(t_{i-1}))^2/(1-a^{2\delta_i/\Delta}) \tag{3.12}$$

mennyiség minimalizálásával nyerhető. Most $\hat{a}$ aszimptotikus varianciája Nielsen (2011) nyomán

$$\left(\sum_{i=2}^{n}\frac{\delta_i^2/\Delta^2\hat{a}^{2(\delta_i/\Delta-1)}}{1-\hat{a}^{2(\delta_i/\Delta)}}\right)^{-1} \tag{3.13}$$

lesz. Ekvidisztáns időlépcsők esetében természetesen (3.11) és (3.13) is a jól ismert (Box and Jenkins, 1970) $(1-\hat{a}^2)/(n-1)$ varianciába megy át.

Legyen $y_1,...,y_n$ egy stacionárius folyamatból származó ekvidisztáns időlépcsőkben megfigyelt idősor. A korábban már megismert periodogram a $\lambda_i = 2\pi i/n$, $i=1,...,L$ pontokban az

$$I(\lambda_i) = n/(4\pi)(a_i^2+b_i^2) \tag{3.14}$$

formában is megadható, ahol az a_i, b_i Fourier-együtthatók az OLS módszerrel nyerhetők a

$$(\underline{\underline{Z}}^T\underline{\underline{Z}})\underline{c} = \underline{\underline{Z}}^T\underline{y} \tag{3.15}$$

lineáris egyenletrendszer megoldásával. Itt a $\underline{\underline{Z}}$ mátrix és a $\underline{c}$ vektor elemei

$$z_{ij} = \begin{Bmatrix} \cos(\lambda_j i),\ j = 1,...,L \\ \sin(\lambda_j i),\ j = L+1,...,2L \end{Bmatrix},\ c_i = \begin{Bmatrix} a_i, i = 1,...,L \\ b_{i-L}, i = L+1,...,2L \end{Bmatrix}, \tag{3.16}$$

továbbá $\underline{y} = (y_1,...,y_n)^T$. A λ_i frekvenciák fenti választásával az egyenletrendszer különálló egyenletekre bomlik, aminek megoldása

$$a_i = 2/n \cdot \sum_{j=1}^{n} y_j \cos(\lambda_j j),\quad b_i = 2/n \cdot \sum_{j=1}^{n} y_j \sin(\lambda_i j),\quad i = 1,...,L. \tag{3.17}$$

A $\underline{c}$ komponensei korrelálatlanok és (3.14) aszimptotikusan $g(\lambda_i)J(\lambda_i)$, ahol $J(\lambda_i), i = 1,...,L$ független standard exponenciális eloszlású valószínűségi változók sorozata, ha nincsen λ_i közelében diszkrét frekvencia. Vagy, ahogy korábban már említettük, $I(\lambda_i)$ aszimptotikusan exponenciális eloszlású $1/g(\lambda_i)$ paraméterrel. Az aszimptotikus tulajdonságok teljesülésének *n*-nel való kapcsolatát jellemezhetjük a periodogram várható értékének *n*-től való függésével:

$$E[I(\lambda_i)] = \int_{-\pi}^{\pi} f(\omega) K_n(\omega - \lambda_i) d\omega, \tag{3.18}$$

ahol $K_n(\omega)$ a Fejér-féle magfüggvény, és $f(\lambda) = g(\lambda), \lambda \geq 0$, $f(\lambda) = g(-\lambda), \lambda < 0$ (Priestley, 1981). Mivel az egyenlet jobb oldala $g(\lambda_i)$-hez tart midőn $n \to \infty$, látható, hogy a periodogram a spektrális sűrűségfüggvény aszimptotikusan torzítatlan becslése, ha λ_i nincs közel egyetlen diszkrét frekvenciához sem. (3.18) szemléletes jelentése az, hogy véges hosszúságú idősor esetén az egyes frekvenciák nem választhatók szét teljesen, hanem adott frekvencia megjelenése keveredik az összes további frekvenciával. E keveredés mértéke persze egyre csökken, ahogy az idősor hossza növekszik.

Nem ekvidisztáns időlépcsők esetében a Lomb-Scargle (L-S) periodogram (Lomb 1976; Scargle 1982) használatos. Ez (3.15)-tel definiálható, de *L*=1 és $\lambda_1 = \lambda$, $a_1 = a$, $b_1 = b$, továbbá

$$z_{ij} = \begin{Bmatrix} \cos(\lambda t_i), j = 1 \\ \sin(\lambda t_i), j = 2 \end{Bmatrix}, \; c_i = \begin{Bmatrix} a, i = 1 \\ b, i = 2 \end{Bmatrix}, \; \underline{y} = (y(t_1), ..., y(t_n))^T \tag{3.19}$$

mellett. Ha $\underline{y}$ fehérzaj folyamatból származik, akkor $\underline{c}$ kovarianciamátrixa $\sigma^2 \underline{\underline{D}}^{-1}$ lesz, ahol $\underline{\underline{D}} = \underline{\underline{Z}}^T \underline{\underline{Z}}$. A $\underline{c}^T (\underline{\underline{D}}^{-1})^{-1} \underline{c} = \underline{c}^T \underline{\underline{D}} \underline{c}$ kvadratikus formával az

$$I(\lambda) = 1/(2\pi) \underline{c}^T \underline{\underline{D}} \underline{c} = 1/(2\pi)(d_{2,2} a^2 + d_{1,1} b^2 - 2 d_{1,2} ab) \big/ (d_{1,1} d_{2,2} - d_{1,2}^2) \tag{3.20}$$

alakban értelmezhető az L-S periodogram, ahol $d_{i,j}$ a $\underline{\underline{D}}^{-1}$ (*i*,*j*)-edik elme. Megjegyezzük, hogy az L-S periodogram fenti értelmezése formailag eltér az eredeti definíciótól, ám ekvivalens vele. A (3.20) formát a később bevezetendő teljes legkisebb négyzetek (TLS) módszere indokolja. Nem nehéz belátni (Kokoszka and Mikosch, 2000), hogy $I(\lambda)$ aszimptotikusan exponenciális eloszlású, ekvidisztáns időlépcső esetében π / σ^2 paraméterrel, ami épp a fehérzaj folyamat spektrális sűrűségfüggvényének reciproka. Nem fehérzaj esetében a $\underline{c}$ kovarianciamátrixa az OLS becslés mellett $\underline{\underline{D}}^{-1} \underline{\underline{Z}}^T \underline{\underline{B}} \underline{\underline{Z}} \underline{\underline{D}}^{-1}$ (Grenander and Rosenblatt, 1957). Kimutatható (Matyasovszky, 2012b), hogy $(\underline{\underline{D}}^{-1} \underline{\underline{Z}}^T \underline{\underline{B}} \underline{\underline{Z}}) \underline{\underline{D}}^{-1}$ tart egy $\pi k(\lambda) \underline{\underline{D}}^{-1}$ alakú kifejezéshez, midőn *n* tart végtelenhez, ahol $k(\lambda)$ ekvidisztáns esetben $g(\lambda)$. Az $I(\lambda)$ tehát most is aszimptotikusan exponenciális eloszlású, ekvidisztáns időlépcső esetében $1/g(\lambda)$ paraméterrel. Jóllehet λ bármilyen frekvencia lehet egy $[\lambda_{\min}, \lambda_{\max}]$ intervallumon, tanácsos őket a $2\pi i/(n\Delta), i = 1, ..., L$ pontokban venni, ahol $\lambda_{\max} = \pi / \Delta$ az ún. átlagos Nyquist-frekvencia (Stoica et al 2009). A $[\lambda_{\min}, \lambda_{\max}]$ intervallumnak a $[2\pi / n, \pi]$ intervallumra való átskálázásával a nem ekvidisztáns időpontok esetén értelmezett L-S periodogram úgy mutatkozik, mint a Δ időlépcsőnként ekvidisztánsan észlelt idősor periodogramja. Fontos különbség azonban, hogy az L-S periodogram elemek egyrészt korreláltak, másrészt az L-S periodogram nagyobb torzítottsággal rendelkezik, mint a periodogram (Vio et al. 2010; Matyasovszky, 2012b), mert az L-S periodogramban az idősort

jellemző valódi periodikus összetevők és a mintavételezés időbeli eloszlásával kapcsolatos periodikus összetevők együttesen jelennek meg (Deeming, 1975).

A probléma kezelésére az ún. teljes legkisebb négyzetek (TLS) módszerét javasoljuk (Matyasovszky, 2012b), vagyis a szóba jövő frekvenciák együttes kezelését az L-S periodogramnál látott egyedi kezelésük helyett. A TLS periodogram a (3.15) egyenletrendszer

$$z_{ij} = \left.\begin{cases}\cos(\lambda_j t_i),\ j = 1,...,L \\ \sin(\lambda_j t_i),\ j = L+1,...,2L\end{cases}\right\},\ \underline{y} = (y(t_1),...,y(t_n))^T \tag{3.21}$$

melletti megoldása után az

$$I(\lambda_j) = 1/(2\pi)(d_{j+L,j+L}a_j^2 + d_{j,j}b_j^2 - 2d_{j,j+L}a_j b_j)\big/(d_{j,j}d_{j+L,j+L} - d_{j,j+L}^2),\ j = 1,...L \tag{3.22}$$

formában értelmezhető. (3.22) a (3.20) nyilvánvaló általánosítása, amiből kifolyólag ugyanazokkal az aszimptotikus tulajdonságokkal rendelkezik, ám egy fontos különbséggel. Jóllehet minden λ_j frekvencia megjelenése keveredik az összes további, a becslési eljárásban nem szereplő frekvenciával, a TLS periodogram torzítása kisebb lesz, mint az L-S periodogramé. Ennek az az oka, hogy a TLS periodogram a $\lambda_j, j = 1,...,L$ frekvenciák együttesére és nem egyedi λ frekvenciákra van értelmezve. A részletek Matyasovszky (2012b) tanulmányában találhatók.

Az OLS-AR(1) spektrális sűrűségfüggvény konfidencia-intervallumának megadásához az alábbi szimulációs eljárást javasoljuk. 1. (3.8)-ból az $\hat{a}$ (3.11) varianciájának felhasználásával szimulálunk egy $g(\lambda_i), i = 1,...,L$ spektrális sűrűséget. 2. Szimulálunk egy periodogramot *n* számú független véletlen számmal, amelyek $1/g(\lambda_i), i = 1,...,L$ paraméterű exponenciális eloszlásból származnak. 3. Az 1 és 2 lépést ismételjük, mondjuk 10000-szer. 4. Minden λ_i-re meghatározzuk az előző lépésből nyert periodogram elemek $(1-\varepsilon)$-kvantilisét. E λ-tól függő kvantilis lesz azon null-hipotézis elfogadási tartománya az ε 100% szignifikancia-szint mellett, hogy az idősorból nyert periodogram az OLS-AR(1) spektrális

sűrűségből származik. A WLS-AR(1) spektrális sűrűség esetén ugyanígy járunk el, csak (3.11) helyett (3.13)-at alkalmazzuk.

3.1.3. Vörös zaj becslése

A vörös zaj spektrumnak AR(1) spektrummal való közelítésének van egy olyan problémája, hogy voltaképpen nem tudjuk, mit is közelítünk és milyen pontossággal. Ennek érzékeltetésére tekintsük az

$$Y_t = \sum_{j=0}^{\infty} b_j e_{t-j}, \quad \sum_{j=0}^{\infty} b_j^2 < \infty \tag{3.23}$$

lineáris folyamatot, ahol e_t fehérzaj σ_e^2 varianciával. A (3.23) folyamat a stacionárius folyamatok rendkívül tág körét öleli fel (Priestley, 1981). A (3.23) szokásos AR(1) közelítése $Y_t^{AR(1)} = a^{AR(1)} Y_{t-1}^{AR(1)} + e_t^{AR(1)}$ az $a^{AR(1)} = R(1)$ paraméterrel, ahol $R(1)$ az egylépéses autokorreláció. Tekintsünk egy tetszőleges $Y_t^{(a)} = aY_{t-1}^{(a)} + e_t^{(a)}$ AR(1) folyamatot. Ennek (3.23)-hoz való hasonlósága mérhető a $\Delta = E\left[(Y_t^{(a)} - Y_t)^2\right]$ mennyiséggel, ami Galbraith and Zinde-Walsh (2002) nyomán a

$$\Delta = \sum_{j=0}^{\infty} (b_j - (1-a^2)^{1/2} \rho^{1/2} a^j)^2 \sigma_e^2 \tag{3.24}$$

alakot ölti, ahol $\rho = \sum_{j=0}^{\infty} b_j^2$ (Matyasovszky, 2013b). A (3.23) folyamatot ezért az az AR(1) modell közelíti a legjobban, amihez tartozó a paraméter minimalizálja (3.24)-et. Mivel azonban ez különbözik $R(1)$-től, nem az $a^{AR(1)} = R(1)$ paraméterű AR(1) folyamat közelíti optimálisan - legalábbis átlagos négyzetes hibában - a szóban forgó folyamatot. Más megfogalmazással: az a folyamat, amit az AR(1) modellek körében az $Y_t^{AR(1)}$ folyamat átlagos négyzetes hibában optimálisan közelít, az nem az Y_t folyamat lesz.

Most felejtkezzünk el erről, és csak azt vizsgáljuk meg, hogy Y_t spektrális sűrűsége miképp közelíthető $Y_t^{AR(1)}$ spektrális sűrűségével. Ismeretes, hogy egy AR(1) folyamat spektrumának konzisztens becslése nyerhető, ha egy idősor birtokában a és σ_e^2 konzisztens becslését írjuk (3.8)-ba (Mann and Wald, 1943). Az azonban szinte kizárható, hogy egy adott feladat során fellépő valóságos folyamat éppen AR(1) folyamat lenne. Ugyanakkor egy lineáris folyamat spektruma konzisztensen becsülhető egy AR(p) folyamattal, ha $p \to \infty, p^3 / n \to 0$ midőn $n \to \infty$ (Berk, 1974). Ha ellenben rögzítjük a $p = 1$ választást, akkor semmit nem tudunk a becslés pontosságáról a valóságos spektrális sűrűség ismerete nélkül. Ezzel szemben az alább bemutatandó, izoton regresszión alapuló eljárás semmilyen analitikus formát nem tételez fel a spektrális sűrűségről. Az eljárás az OLS technikán alapul, ami esetleg nem túl hatékony a normális eloszlástól erősen eltérő eloszlások esetében. Az OLS izoton regresszió azonban bizonyos esetekben megegyezik a maximum-likelihood (ML) izoton regresszióval. Ez a helyzet exponenciális eloszlás esetén is, ezért az alábbi módszer a vörös zaj spektrumnak a periodogramon alapuló ML becslésének tekinthető, hiszen a periodogram elemek aszimptotikusan exponenciális eloszlásúak. A következőkben rátérünk az éghajlati adatsorokra korábban még nem alkalmazott izoton regresszió ismertetésére (Matyasovszky, 2013b).

Legyen az $x_1,...,x_n$ idősor a $t_1,...,t_n$ időpontokban adva az $x_i = f(t_i) + e_i$ formában, ahol e_i nulla várható értékű, σ^2 varianciájú minden *i*-re, továbbá az $\{e_i\}$ sorozat gyengén függő (Zhao and Woodroofe, 2012). Becsüljük az $f(t)$ trendfüggvényt az $\hat{f}(t) \le \hat{f}(s), t < s$ feltétel mellett. Az egyszerűség kedvéért tegyük fel, hogy az időpontok ekvidisztánsan helyezkednek el a $[0,1]$ intervallumon. A

$$\min\left\{\sum_{k=1}^{n}\left(x_k - \hat{f}(t_k)\right)^2\right\}, \quad \hat{f}(t_k) \le \hat{f}(t_l), k < l \tag{3.25}$$

OLS probléma megoldása

$$\hat{f}(t_k) = \max_{i \le k} \min_{k \le j} \frac{x_i + ... + x_j}{j - i + 1},\ k = 1,...,n \tag{3.26}$$

lesz, és $\hat{f}(t)$ balról folytonos a többi pontban. Az $\hat{f}(t)$, $t \in (0,1)$ aszimptotikus viselkedése

$$\hat{f}(t) = f(t) + \frac{2}{n^{1/3}} \left(\frac{1}{2} \sigma^2 f'(t) \right)^{1/3} \cdot \eta \tag{3.27}$$

formában adható meg, ahol $f'(t)$ az $f(t)$ deriváltja, és az η valószínűségi változó Chernoff-eloszlású (Groeneboom and Wellner, 2001).

Legyen $y_1,...,y_N$ egy idősor, melynek spektrális sűrűségét vörös zajként kívánjuk közelíteni. Az izoton regresszió (IR) a következőképp alkalmazható. Legyen $t = (\pi - \lambda)/\pi$, $x_i = I(\lambda_{n-i+1})$ és $f((\pi - \lambda)/\pi) = g(\lambda)$, ahol $g(\lambda)$ a spektrális sűrűség. Végül

$$I(\lambda_i) = \frac{1}{\pi N} \left[\left(\sum_{j=1}^{N} y_j \cos(\lambda_i j) \right)^2 + \left(\sum_{j=1}^{N} y_j \sin(\lambda_i j) \right)^2 \right] \tag{3.28}$$

a periodogram a $\lambda_i = (2\pi i)/N, i = 1,...,n$ frekvenciákban, ahol n az $N/2$ egész része. Azon frekvenciáknál azonban, ahol a spektrum eltér a vörös zajtól, a periodogram elemek viselkedése - mint már láttuk - különbözik a többitől, tehát kiugró értékként kezelendő, ami az IR robusztus változatával tehető meg. Nyilvánvaló, hogy (3.25) a

$$\min \left\{ \sum_{k=1}^{n} \rho\left((x_k - \hat{f}(t_k))/\sigma \right) \right\}, \quad \hat{f}(t_k) \le \hat{f}(t_l),\ k < l \tag{3.29}$$

feladat speciális esete a $\rho(u) = (\sigma u)^2$ függvény mellett, és $\rho(u)$ egyéb választásával biztosítható a becslés robusztussága. Álvarez and Yohai (2011) alapján:

$$\hat{f}(t_k) = \max_{u \le k} \min_{k \le v} \{\hat{m}(u,v)\}, \tag{3.30}$$

ahol $\hat{m}(u,v)$ a

$$\sum_{j\in C(u,v)}\psi\left((x_i-m)/\sigma\right)=0 \tag{3.31}$$

egyenlet megoldása $C(u,v)=\left\{j;1\le j\le n,u\le t_j\le v\right\}$ és $\psi(u)=\rho'(u)$ mellett. A (3.27) egyenletben ekkor σ^2 helyett $r\sigma^2$ szerepel, ahol

$$r=\frac{E\left[\psi^2(e/\sigma)\right]}{\left(E\left[\psi'(e/\sigma)\right]\right)^2}$$

és *e* az $\{e_i\}$ sorozatot generáló valószínűségi változó. Az eljárásnak a spektrális sűrűségre történő alkalmazásakor figyelembe kell venni, hogy σ nem konstans, hanem $\sigma_i=g(\lambda_i)$. Ezért (3.31) a

$$\sum_{j\in C(u,v)}\psi\left(x_i/m-1\right)=0 \tag{3.32}$$

formát ölti, ahol $\psi(u)$ a 3.3.1 fejezetben már megismert függvény. Ekkor aszimptotikusan:

$$g(\lambda)=\hat{g}(\lambda)-\frac{2}{n^{1/3}}\left(-\frac{\pi}{2}g^2(\lambda)g'(\lambda)r\right)^{1/3}\cdot\eta\,. \tag{3.33}$$

Az eljárás alkalmazásához meg kell határozni a periodogram konfidencia-intervallumát azon null-hipotézis mellett, hogy a periodogram vörös zaj spektrumból származik. Ehhez a (3.33)-ban szereplő még ismeretlen mennyiségek megadása szükséges. A Chernoff-eloszlás jól közelíthető egy zérus várható értékű, 0,52 szórású normális eloszlással, de a pontos eloszlás is megadható Groeneboom and Wellner (2001) szerint. A következők Matyasovszky (2013b) tanulmányában találhatók. Nevezetesen, a $\psi(u)$-hoz tartozó *r* értéke 0,75, míg $g^2(\lambda)$ helyettesíthető a $\hat{g}^2(\lambda)$ becslésével. Végül a $g'(\lambda)$ nemparaméteres regressziós technikával becsülhető. Képezzük a $\Delta x_i=N/(2\pi)\cdot(\hat{g}(\lambda_i)-\hat{g}(\lambda_{i+1})),i=1,...,n-1$ adatsort, mely $g'(\lambda_i),i=1,...,n-1$ véges különbséges közelítése. A WLR feladat *p*=0 melletti választásával $g'(\lambda)$-nak a

$$\hat{g}'(\lambda) = \sum_{j=1}^{n-1} \Delta x_j K\left(\frac{\lambda_j - \lambda}{b}\right) \Big/ \sum_{j=1}^{n-1} K\left(\frac{\lambda_j - \lambda}{b}\right) \tag{3.34}$$

ún. Nadaraya-Watson-becsléséhez jutunk. A Nadaraya-Watson-becslésről részletesen például Simonoff (1996) kötetében olvashatunk. A *b* sávszélesség a

$$CV(b) = \sum_{i=1}^{n-1} (\Delta x_i - \hat{g}_i'(\lambda_i))^2 \tag{3.35}$$

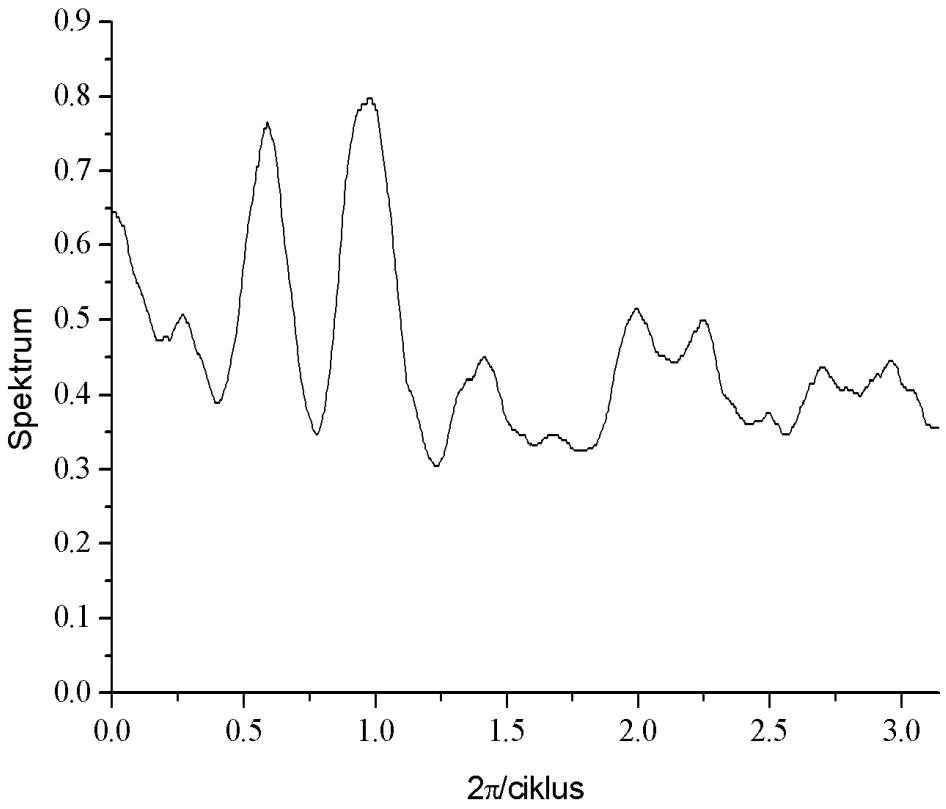
mennyiség *b* szerinti minimalizálásával nyerhető, ahol $\hat{g}_i'(\lambda_i)$ (3.34)-ből jön $\lambda = \lambda_i$ mellett, azzal a különbséggel, hogy most a $\Delta x_j, |j - i| \le 1$ adatok nem szerepelnek (3.34)-ben.

A konfidencia-intervallum megadásához az alábbi szimulációs eljárást javasoljuk. 1. (3.33) alapján szimulálunk egy $g(\lambda_i), i = 1,...,n$ spektrális sűrűséget. 2. Szimulálunk egy periodogramot *n* számú független véletlen számmal, amelyek $1/g(\lambda_i), i = 1,...,n$ paraméterű exponenciális eloszlásból származnak. 3. Az 1 és 2 lépést ismételjük, mondjuk 10000-szer. 4. Minden λ_i-re meghatározzuk az előző lépésből nyert periodogram elemek $(1-\varepsilon)$-kvantilisét. E λ-tól függő kvantilis lesz az $(1-\varepsilon)$ 100%-os konfidencia-intervallum határa.

3.2. ALKALMAZÁSOK

3.2.1. NAO index

Egy tanulmányunkban (Matyasovszky, 2010a) egyebek mellett a NAO index (Ponta Delgada és Stykkisholmur/Reykjavik havi tengerszinti átlagos légnyomáskülönbsége) havi adatsorát vizsgáltuk az 1865-2002 időszakra a spektrális sűrűség robusztus becslésére alapozva. Több szerző számos periódus fontosságára hívta fel korábban a figyelmet, így például Goodman (1998) és Benner (1999) 2-2,3, 3-3,5, 6-10, 20-23 és 50-70 éves periódusidőknél talált lokális csúcsokat a spektrumban.



9. ábra
A havi NAO index spektrális sűrűségének robusztus becslése az 1865-2002 évek alapján

Diszkrét periódust nem mutatott ki a fent leírt robusztus becslési eljárásunk, ám az általunk nyert spektrális sűrűségfüggvény számos lokális csúccsal rendelkezik (9. ábra). A legnagyobb csúcsok a 10,7 és 6,4 hónapnál jelentkeznek, ami logikusan az évi menettel és az évi menet aszimmetriáját tükröző féléves hullámmal kapcsolatos. Egy további jól értelmezhető csúcs 2 év körül mutatkozik a Kvázi-kétéves Oszcillációnak megfelelően. Több egyéb, nehezen magyarázható csúcs is megfigyelhető a magas frekvenciákon. Külön megvizsgáltuk a téli NAO index (havi értékek átlaga decembertől márciusig) adatsorát, mert a NAO és az érintett területek éghajlat-ingadozásainak kapcsolata ezen időszakban jelentkezik a legvilágosabban (Hurrell and van Loon, 1997). Diszkrét periódus detektálhatósága nélkül a spektrális sűrűségfüggvény a 2,4 és 4,7 éves periódusoknál mutat lokális maximumot, ami jó összhangban van a 2-5 éves ciklusok szerepével.

Nicolay et al. (2008) wavelet alapú spektrálanalízissel 11 hónapos és 2,5 éves periodikus komponenst talált a NAO index idősorában. Egy 6,6 hónapos periodikusság jóval gyengébben, míg a 4,7 év körüli ciklus alig-alig jelentkezett náluk. Ezek alapján elmondhatjuk, hogy módszerünk kiállja az összehasonlítást, például a rendkívül hatékonynak tartott wavelet alapú spektrálanalízissel is.

Luterbacher et al. (1999, 2002) által rekonstruált NAO index adatsor 1659 óta áll rendelkezésre (http://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/RNAO/). Itt az éves és féléves hullám a spektrum diszkrét összetevőjeként jelentkezik, igaz csak gyengén szignifikánsan (9%-os szint). A 3.1 fejezetben láttuk, hogy egy diszkrét periódusnak a periodogramhoz való hozzájárulása az adatsor hosszának növekedésével egyenes arányban nő. Mivel a rövidebb megfigyelési adatsorban az említett ciklusok nem tűntek a diszkrét spektrumhoz tartozónak, míg a jelenlegi jóval hosszabb adatsorban már jelentkeznek (igaz csak gyengén szignifikánsan), arra következtetünk, hogy az éves és féléves diszkrét összetevő nem erős, de létező. A harmadik legnagyobb periodogram elem a 65,4 évnél jelentkezik. Ez

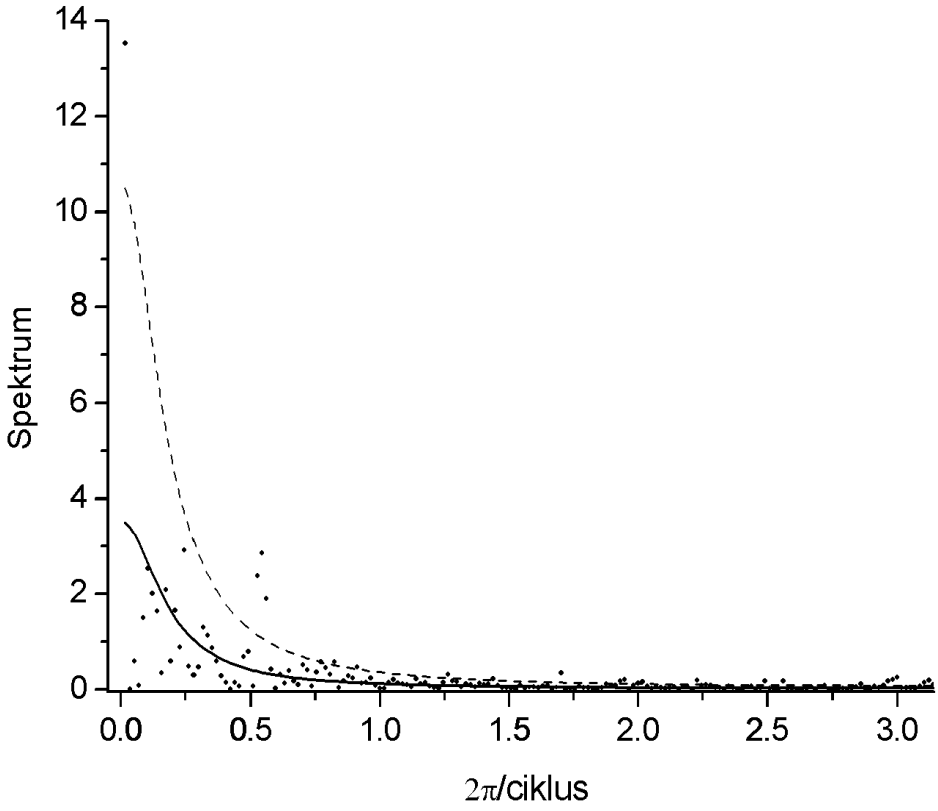
nem tekinthető diszkrét összetevőnek, de a vörös zaj spektrumtól való eltérése 2%-os szinten szignifikáns. Ez jó összhangban van azokkal a korábbi vizsgálatokkal, melyek az 50-70 éves periódus tartomány fontosságát mutatták ki több adatsorban (Loehle and Scafetta, 2011), így például a NAO indexben is (Mazzarella and Scafetta 2012). A spektrális sűrűségfüggvény 5,4 évnél jelentkező 1%-os szinten szignifikáns csúcsa megerősíti Box (2002) korábbi tapasztalatát.

3.2.2. GISP2 Oxigén izotóp adatok a 15000 - 60000 évvel ezelőtti időszakra

A grönlandi GISP2 jégfurat O^{18}/O^{16} izotóparány adatsorát (Groots and Stuvier, 1997) vizsgáltuk (Matyasovszky, 2012b), hogy a 3.1.2 fejezetben leírt módszerrel kapcsolatos tapasztalatainkat összehasonlíthassuk Schulz and Mudelsee (2002) eredményeivel. A mélyebb rétegekből származó jégfurat egyre erőteljesebb összenyomódása folytán az adatsor időbeli felbontása nem ekvidisztáns; az időlépcső 68 évtől 257 évig terjed a Δ=125,8 átlagos értékkel. Az adatokat (N=358) standardizáltuk, tehát a transzformált adatsor nulla átlaggal és egy szórással rendelkezik. Mivel a különböző időpontokból származó O^{18}/O^{16} izotóparány jó indikátora az aktuális hőmérsékletnek, ezért az adatsor elemzése a jelzett időszak hőmérsékleti ingadozásáról nyújt információt.

A becsült autoregresszív együttható $\hat{a}$=0,501±0,101, illetve $\hat{a}$=0,835±0,060 a WLS és az OLS módszer esetén. A ± szimbólum után szereplő értékek a 95%-os konfidencia-intervallumot reprezentálják. Előzetes sejtésünknek megfelelően az együttható kétféle becslése igen erősen eltér. A 10. ábra az L-S periodogramot mutatja az OLS-AR(1) spektrumhoz tartozó 95%-os konfidencia-intervallummal. Három periodogram elem haladja meg a konfidencia-intervallumot, melyek a jól ismert Dansgaard-Oeschger-eseményekkel kapcsolatosak (lásd 4.2.3 fejezet). A jó közelítéssel 1470 éves periódusidejűnek detektált

ciklus teljes összhangban van Schulz and Mudelsee (2002) tanulmányának eredményével. Három további gyenge, de statisztikailag szignifikáns periódus látszik a lényegesen magasabb

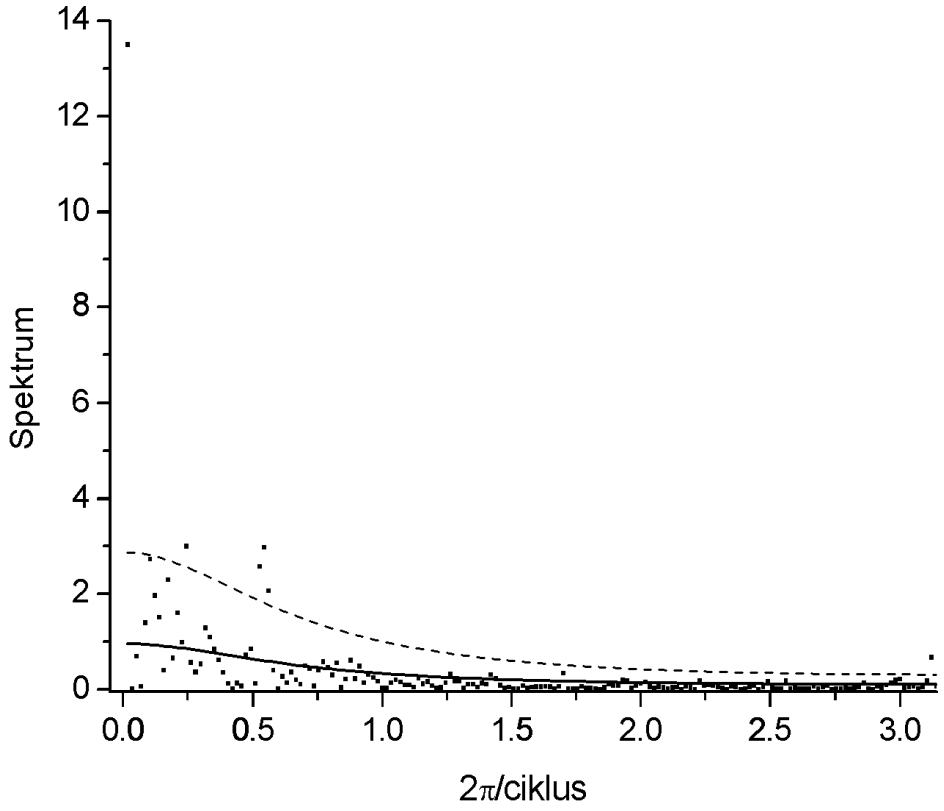


10. ábra
L-S periodogram (pontok), OLS AR(1) spektrális sűrűség (folytonos vonal) és az 5%-os szignifikancia-szinthez tartozó kritikus érték azon null-hipotézis mellett, hogy a periodogram az illesztett AR(1) folyamatból származik (szaggatott vonal) a GISP2 oxigén izotóp adatsorra a 15000-60000 évvel ezelőtti időszakra. A ciklus egysége Δ *=125,8 év.*

frekvenciákon, melyek minden bizonnyal az L-S periodogram viszonylag nagy torzításának melléktermékeként tekinthető.

A 11. ábra a TLS periodogramot és a hozzá tartozó WLS-AR(1) spektrális sűrűség szerinti 95%-os konfidencia-intervallumot tartalmazza. Jóllehet az L-S és a TLS periodogram igen hasonló, a WLS és OLS spektrális sűrűség között igen nagy különbség tapasztalható. Ezért most egy körülbelül 3200 éves ciklus is jelentkezik, ami jó összhangban van Matyasovszky (2010b) által teljesen más módon detektált 3400 éves ciklussal. A

legszignifikánsabb periodogram csúcs azonban az ekliptika körülbelül 41000 éves periódusának felel meg, ami meglepő módon teljesen hiányzik Schulz and Mudelsee (2002)



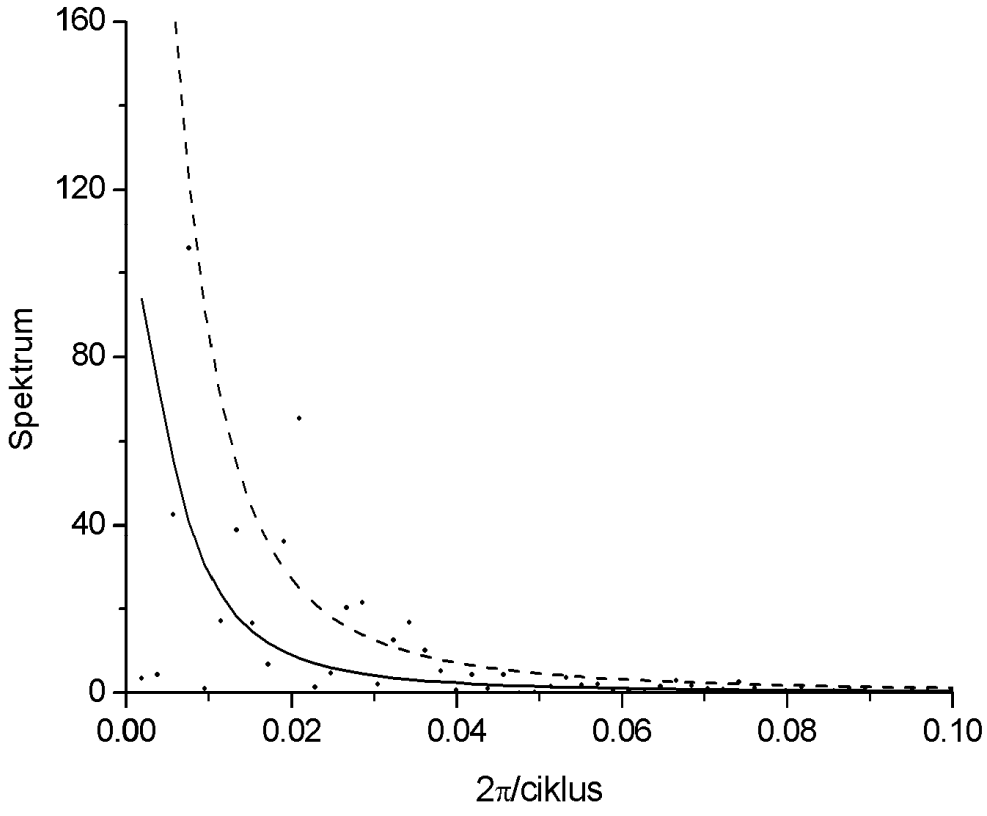
11. ábra
TLS periodogram (pontok), WLS AR(1) spektrális sűrűség (folytonos vonal) és az 5%-os szignifikancia-szinthez tartozó kritikus érték azon null-hipotézis mellett, hogy a periodogram az illesztett AR(1) folyamatból származik (szaggatott vonal) a GISP2 oxigén izotóp adatsorra a 15000-60000 évvel ezelőtti időszakra. A ciklus egysége Δ *=125,8 év.*

vizsgálatában. Az említett tanulmány az L-S peridogramnak az OLS-AR(1) spektrális sűrűség alapján történő korrekcióját is erősen javasolja, mivel úgy találták, hogy az L-S periodogram jelentősen felülbecsli a nagy frekvenciák szerepét. Ez azonban nem így van. Láttuk ugyanis, hogy az OLS AR(1) spektrum az *a* autoregresszív együttható felülbecslése révén túlbecsli az alacsony frekvenciák szerepét és ennél fogva alulbecsli a nagy frekvenciák szerepét. Végső soron tehát a fő probléma nem az L-S periodogramnak, hanem az OLS-AR(1) spektrális sűrűségnek a pontatlanságában keresendő. Ezért a hivatkozott korrekció nem csak nem

szükséges, de kifejezetten káros. Ezt azért fontos megemlíteni, mert a Schulz and Mudelsee (2002) tanulmányában bemutatott REDFIT néven ismeretes helytelen módszert számosan alkalmazták már paleoklíma adatsorokra. A 11. ábrához visszatérve, látható még egy enyhe, de szignifikáns csúcs 252 évnél, ami a naptevékenység hasonló periódusával hozható kapcsolatba (Damon and Sonnett, 1991).

3.2.3. Vostok deuterium tartalom adatsora az elmúlt 422766 évben

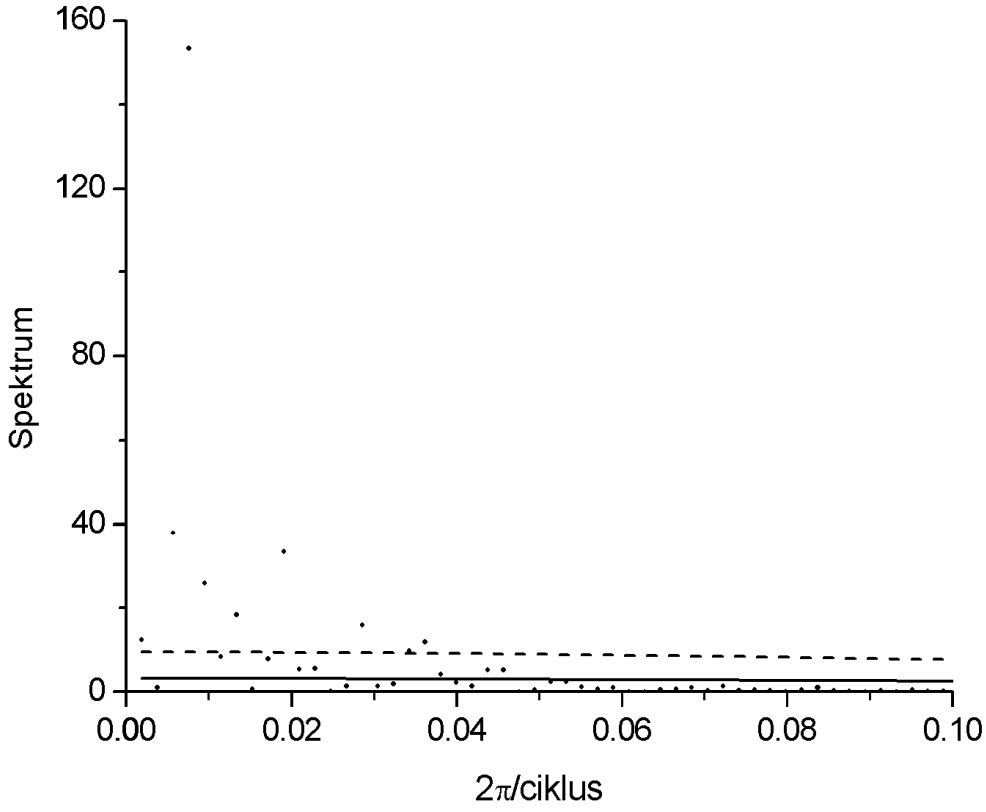
A nem ekvidisztánsan megfigyelt idősorokra vonatkozó kétféle eljárás összehasonlítását az antarktiszi Vostok állomás deuterium tartalom adatsorán is elvégeztük. Mivel a különböző



12. ábra
L-S periodogram (pontok), OLS AR(1) spektrális sűrűség (folytonos vonal) és az 5%-os szignifikancia-szinthez tartozó kritikus érték azon null-hipotézis mellett, hogy a periodogram az illesztett AR(1) folyamatból származik (szaggatott vonal) a Vostok deuterium adatsorra az elmúlt 422766 évre. A ciklus egysége Δ *=127,8 év.*

időpontokból származó deutérium tartalom jó indikátora az aktuális hőmérsékletnek, ezért az adatsor elemzése a jelzett időszak hőmérsékleti ingadozásáról nyújt információt. Az adatsor az elmúlt 422766 évet öleli fel 20 évtől 664 évig terjedő időbeli felbontásban (Δ=127,8 év). A becsült autoregresszív együttható $\hat{a}$=0,818±0,021, illetve $\hat{a}$=0,994±0,005 a WLS és az OLS módszer esetén. Az ezekhez társuló spektrális sűrűségek között óriási különbség lép fel az alacsony frekvenciáknál. Az L-S periodogram és OLS-AR(1) spektrális sűrűség, továbbá a TLS periodogram és WLS spektrális sűrűség között a legfontosabb eltérések a következők. A 12. ábra alapján hét L-S periodogram elem haladja meg a 95%-os konfidencia-intervallumot az OLS-AR(1) spektrális sűrűség mellett (csak az alacsony frekvencia tartományt mutatjuk be, a nagy frekvenciákon nincs említésre méltó tanulság). Ezek 40000-41000 év, 28000-30000 év és 22200-24800 év körül jelennek meg az ekliptika és a precesszió változásinak megfelelő periódusidők mellett. Nagy meglepetésre az excentricitással kapcsolatos 100000 év körüli periódus elmarad, ami gyökeresen ellentmond a korábbi elemzéseknek. Régóta ismert ugyanis, hogy a 100000 éves időskálán a paleoklíma adatokban olyan közelítőleg periodikus ingadozások mutatkoznak, melyek periódus ideje jó közelítéssel megegyezik a Föld pályaelemei (excentricitás, ekliptika szöge, precesszió) változásainak periódus idejével. Az első igazán meggyőző ezzel kapcsolatos eredmény Hays et al. (1976) tanulmányában található. A 12. ábra és 13. ábra összevetése jól jelzi, hogy az L-S és TLS periodogramok között jelentős, míg az OLS-AR(1) és WLS-AR(1) sűrűségek között óriási különbség mutatkozik. Nyolc TLS periodogram elem haladja meg a WLS-AR(1) spektrális sűrűség konfidencia-intervallumát. A legszignifikánsabb periódus az excentricitáshoz változásaihoz kapcsolható 105,000 éves. További periódusok jelennek meg 141000 és 84000 évnél, amik a 105000 éves periódus mesterséges melléktermékei lehetnek a viszonylag kis spektrális felbontásból fakadóan. Érdemes megemlíteni azonban, hogy ezek a ciklusok összhangban vannak egy előző vizsgálatunkkal (Matyasovszky, 2010b), ahol a szóban forgó periódusidő

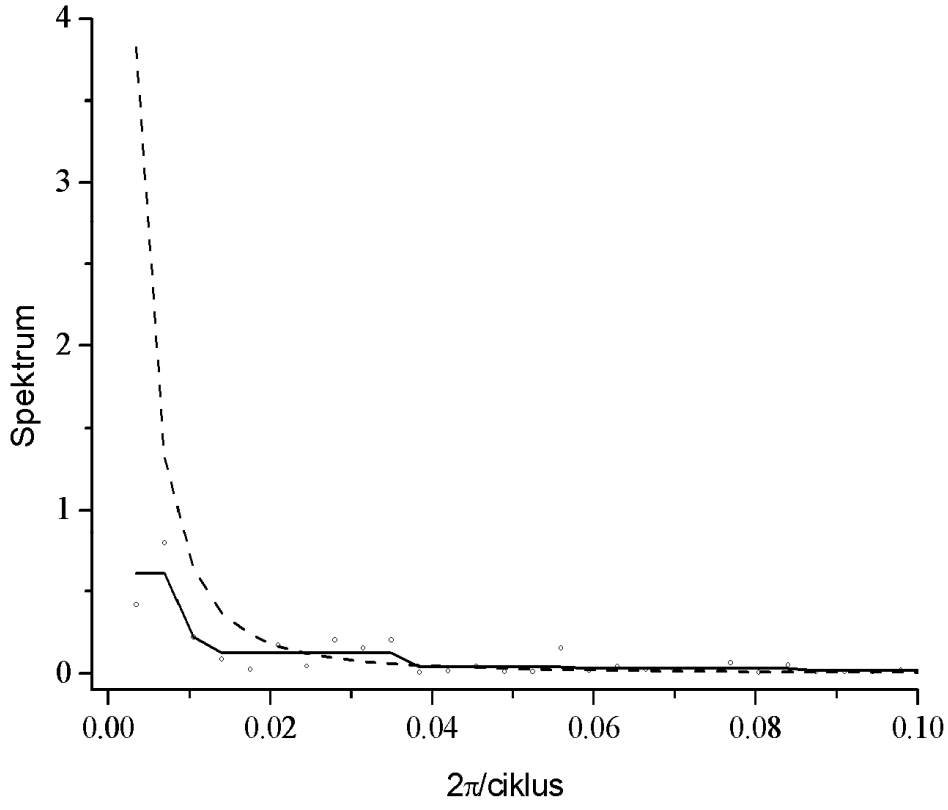
időbeli változását elemeztük. A 41000 éves periódus ezúttal is szignifikáns, de lényegesen gyengébb, mint ahogy azt az L-S periodogram esetében láttuk. Ez megerősíti például Matyasovszky (2010c) korábbi eredményeit. Gyenge, de statisztikailag szignifikáns egy bő 400000 éves ciklus. Ez igen érdekes tanulság, mert a szintén az excentricitás időbeli változásai alapján várható körülbelül 400000 éves periódus csak igen kevés paleoklíma adatsorban mutatható ki (Nie et al., 2008).



13. ábra
TLS periodogram (pontok), OLS AR(1) spektrális sűrűség (folytonos vonal) és az 5%-os szignifikancia-szinthez tartozó kritikus érték azon null-hipotézis mellett, hogy a periodogram az illesztett AR(1) folyamatból származik (szaggatott vonal) a Vostok deuterium adatsorra az elmúlt 422766 évre. A ciklus egysége Δ *=127,8 év.*

3.2.4. Északi Hemiszféra hőmérséklete a 200-1995 évekre

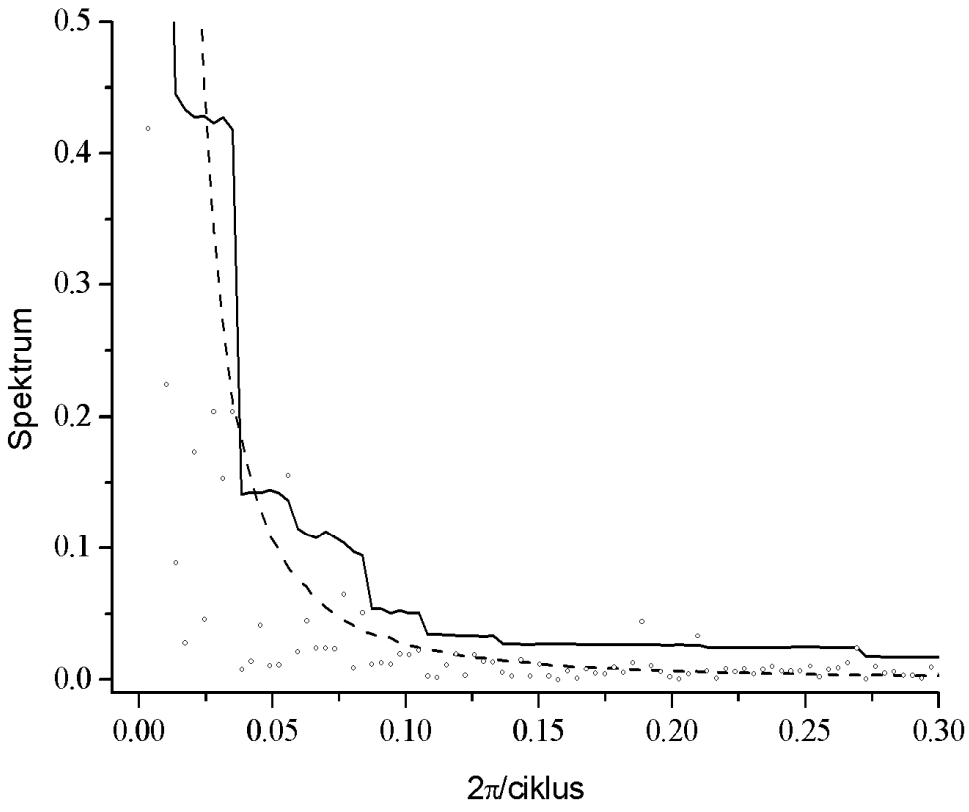
Az Északi Hemiszféra 200-1995 évek közötti rekonstruált hőmérsékleti sorát (lásd 1.2.2 fejezet) az izoton regresszión alapuló spektrálanalízissel is elemeztük (Matyasovszky, 2013b). Mivel a spektrális sűrűség hosszan elnyúló, igen enyhe változást mutat a nagy frekvenciákon és éles csúcsot az alacsony frekvenciáknál, ezért a 14. ábra csak a 60 évesnél hosszabb ciklusoknak megfelelő frekvenciákat tünteti fel. Szembetűnő, hogy az AR(1) spektrális sűrű-



14. ábra
Periodogram (pontok) és a spektrális sűrűség az Északi Hemiszféra rekonstruált hőmérsékleti adatsorára a 200-1995 évekre: robusztus izoton regresszió (folytonos vonal), AR(1) modell (szaggatott vonal)

ség igen jelentősen meghaladja az IR spektrális sűrűséget az alacsony frekvenciáknál. Az AR(1) spektrális sűrűség alapján a spektrum a 114, 81, 23 és 11 éves ciklusoknál különbözik a vörös zaj spektrumtól legalább 5%-os szignifikancia-szinten. Ezek a ciklusok a naptevékenység periodicitásaival hozhatók kapcsolatba (Dammon and Sonett, 1991). A 81 és

11 éves ciklus léte azonban nem igazolható az IR spektrális sűrűség alapján (15. ábra), ellenben egy 33 és 38 év körüli periodicitás detektálható mind az AR(1) mind az IR sűrűségek esetében. E spektrális csúcsok az AMO (Atlantic Multidecadal Oscillation) jelenséggel magyarázhatók. Ez a jelenség ugyan általánosan egy 50-70 éves periodikussággal rendelkező oszcillációként ismert, de más proxy adatok és modell szimulációk jóval komplexebb képet festenek 30 évtől akár 100 évig terjedő periodicitásokkal (Knight et al., 2005). Megjegyezzük, hogy az említett ciklusok egyike sem tekinthető diszkrét spektrumból származónak, csupán a spektrális sűrűség vörös zajtól való eltéréséről van szó. Hadd emeljük ki ismét, hogy az AR(1) és az IR spektrális sűrűségek között óriási különbség mutatkozik az alacsony frekvenciáknál (14. ábra), ami a konfidencia-intervallumoknál csupán a 15. ábra vertikális tengelyének skála-



15. ábra
Periodogram (pontok) 95%-os konfidencia-intervalluma azon null-hipotézis mellet, hogy az Északi Hemiszféra rekonstruált hőmérsékleti adatsorának (200-1995 évek) spektrális sűrűsége vörös zajból származik: robusztus izoton regresszió (folytonos vonal), AR(1) modell (szaggatott vonal)

beosztása miatt nem látható. Jó példa ez arra, hogy a vörös zaj spektrumnak AR(1) modellel történő közelítése esetenként rendkívül megbízhatatlan lehet.

A spektrális sűrűségnek az IR és ennek általánosításaként értelmezhető NIR (nearly isotonic regression: NIR) módszerrel (Tibshirani et al., 2011) történő becslésére további példákat láthatunk Matyasovszky (2013b) tanulmányában. A NIR eljárás, szemben az IR-rel, az OLS becslés során megengedi a spektrális sűrűségnek a monotonitástól való eltérését, ám bünteti ezen eltéréseket. A büntetés mértékét egy κ-val jelölt paraméter szabályozza. A $\kappa = 0$ esetén (nincs büntető tag) a kapott spektrális sűrűség görbéje átmegy az összes periodogram elemen, míg a $\kappa \to \infty$ esetében az IR-hez jutunk, vagyis a görbe monoton. Közöttes κ olyan nemmonoton görbét eredményez, ami egyensúlyt teremt a görbének a periodogram elemkhez való illeszkedési jósága és a monotonitástól való eltérés mértéke között. Nyilvánvaló tehát, hogy κ választása alapvetően befolyásolja a becsült spektrális sűrűség alakját. Tibshirani et al. (2011) algoritmusa azonban gondoskodik e paraméter optimális választásáról is. Ezek után azt gondolhatnánk, hogy a tényleges spektrumnak a vörös zajtól való eltérésének vizsgálata értelmét veszti, hiszen a NIR automatikusan kiadja, hogy ezek az eltérések (ha vannak, mert a spektrális sűrűség nem monoton) mely frekvenciákon fordulnak elő. Sajnos azonban a NIR két hiányossággal rendelkezik. Az egyik, hogy nem ismeretes a robusztus változata. Ez némiképp kezelhető probléma, mert ha valamilyen előzetes elemzés alapján gyanítjuk, hogy bizonyos frekvenciák a diszkrét spektrumhoz tartoznak, akkor elhagyhatjuk ezeket a frekvenciákat és a hozzájuk tartozó periodogram elemket, ugyanis a NIR képes az ily módon nem ekvidisztánsan elhelyezkedő periodogram elemeket is kezelni. A nagyobb gond az, hogy nem áll rendelkezésre a NIR-rel végrehajtott becslés pontosságára, tulajdonságaira utaló, az IR-nél látott (3.33) egyenlettel analóg formula.

4. AUTOREGRESSZÍV IDŐSOR MODELLEZÉS ÁLTALÁNOSÍTÁSAI

Tekintsünk egy stacionárius

$$Y_t = a_0 + a_1 Y_{t-1} + ... + a_p Y_{t-p} + e_t \qquad (4.1)$$

p-edrendű autoregresszív (AR(p)) folyamatot. Ez a stacionárius folyamatok igen tág körét tetszőleges pontossággal közelíti megfelelő *p* mellett abban az értelemben, hogy a (4.1) által generált kovarianciafüggvény tetszőlegesen közel van a modellezni kívánt valóságos folyamat kovarianciafüggvényéhez (Priestley, 1981). Ha egy $y_1,...,y_n$ idősorhoz (4.1) modellt kívánunk illeszteni, akkor az $a_0,...,a_p$ autoregresszív együtthatók becslése az OLS vagy ML módszerrel történik. Ez utóbbi esetben az e_t zajt Gauss-folyamatnak tekintjük, ami $p << n$ esetén lényegében ekvivalens az OLS becsléssel. Ha e_t Gauss-folyamat, akkor Y_t is az, és (4.1) minden statisztikai jellemző (és nem csak a kovarianciafüggvény) szempontjából jó közelítése az $y_1,...,y_n$ idősort generáló folyamatnak. Ha azonban a modellezendő folyamat nem gaussi, akkor ez az utóbbi megállapítás általában nem érvényes. A következőkben ezért a szokásos autoregresszív modellezésnél általánosabb lehetőségeket mutatunk be.

4.1. MÓDSZEREK

Nem-gaussi folyamat estében két megközelítést tárgyalunk. Az első esetben Y_t stacionárius eloszlásának rögzítése mellett keressük e_t eloszlását, vagy Y_t-nek az $Y_{t-1},...,Y_{t-p}$ melletti feltételes eloszlását (4.1.1 fejezet). A második esetben a (4.1) lineáris forma helyett az

$$Y_t = f(Y_{t-1},...,Y_{t-p},e_t,e_{t-1},...,e_{t-q}) \qquad (4.2)$$

nemlineáris folyamat két speciális, de nagy jelentőségű alakját vizsgáljuk (4.1.2 fejezet).

4.1.1. Nem-gaussi AR modell

Az egyszerűség kedvéért a továbbiakban vegyük az AR(1) modellt. Ha a folyamat nem gaussi, meglehetősen nehéz akár az e_t sűrűségfüggvényét, akár az Y_t-nek az $Y_{t-1} = x$ melletti $f(y|x)$ feltételes sűrűségfüggvényét megadni. Ráadásul az Y_t eloszlására kirótt minden konkrét esetben, külön-külön kell a feladatot elvégezni. Például Lawrance (1982) az e_t eloszlása alapján értelmezett gamma-eloszlású AR(1) folyamatot. Azt találta, hogy pozitív valószínűséggel fordul elő az $e_t = 0$ eset, ami azt vonja maga után, hogy bizonyos időpontokban az Y_t és Y_{t-1} közötti kapcsolat determinisztikus. Ez a tulajdonság a gyakorlatban nem igazán realisztikus, ezért ennek kiküszöbölésére Gourieroux and Jasiak (2006) az $f(y|x)$ feltételes sűrűség alkalmas választásával értelmezte a gamma-eloszlású AR(1) folyamatot, illetve Grunwald et al. (2002) a feltételes sűrűség egy széles osztályával általánosabb keretbe helyezte a problémát.

A gamma-eloszlású AR(1) modell azért keltette fel érdeklődésünket, mert a napi parlagfű pollenkoncentráció AR modellezését tűztük ki célul. Comtois (2000) szerint ugyanis a pollenkoncentráció, mint valószínűségi változó, gamma-eloszlást követ. A 4.2.1 fejezetben azonban látni fogjuk, hogy a gamma helyett célszerűbb a lognormális közelítés, ami

szerencsés módon az AR modellezést is megkönnyíti. Ezért a továbbiakban a lognormális AR(1) modellre koncentrálunk.

Tegyük fel, hogy Y_t és Y_{t-1} együttes eloszlása két-dimenziós lognormális. A megfelelő sűrűségfüggvény abból a tényből származtatható, hogy $Z_t = \ln(Y_t)$ és $Z_{t-1} = \ln(Y_{t-1})$ együttes eloszlása definíció szerint két-dimenziós normális. Könnyen belátható ekkor, hogy az $y_1,...,y_n$ idősorban az y_t-nek az y_{t-1}-re vonatkozó feltételes sűrűségfüggvénye

$$f(y|y_{t-1}) = \frac{1}{\sqrt{2\pi}\sigma y}\exp\left(-\frac{1}{2\sigma^2}(\ln(y)-\mu(t))^2\right) \tag{4.3}$$

lesz, ahol $\mu(t) = a_0 + a_1\ln(y_{t-1})$. A feladat az a_0, a_1 és σ becslése. Ez könnyen megoldható, mert Z_t Gauss-folyamat, és ekkor az OLS módszer alkalmazásával a

$$\Delta(a_0,a_1) = \sum_{i=2}^{n}(z_i - a_0 - a_1 z_{i-1})^2 \tag{4.4}$$

mennyiség a_0, a_1 szerinti minimalizálására van szükség, továbbá

$$\hat{\sigma}^2 = \Delta(\hat{a}_0,\hat{a}_1)/(n-3). \tag{4.5}$$

A lognormális eloszlás alapvető tulajdonságainak kihasználásával y_t-nek y_{t-1}-re vonatkozó feltételes várható értékének, illetve feltételes mediánjának becslése

$$\exp(\hat{\mu}(t))\exp(\hat{\sigma}^2/2),$$
$$\exp(\hat{\mu}(t)) \tag{4.6}$$

lesz, ahol

$$\hat{\mu}(t) = \hat{a}_0 + \hat{a}_1\ln(y_{t-1}). \tag{4.7}$$

E két mennyiségre a 4.2.1 fejezetben tárgyalt alkalmazásnál lesz majd szükség.

4.1.2. Nemlineáris AR modell

A (4.2) nemlineáris AR folyamat szükségessége talán a következő megfontolással érzékeltethető a legegyszerűbben. Tegyük fel, hogy egy Y_t folyamatra rendelkezésünkre áll egy $y_1,...,y_n$ idősor, aminek birtokában szeretnénk a *t* időpontban a folyamat olyan *l*-időlépéses $\hat{Y}_{t+l}$ előrejelzését megadni, ami a $\Delta_l = E\left[(Y_{t+l} - \hat{Y}_{t+l})^2\right]$ átlagos négyzetes hibát minimalizálja. Ez az $\hat{Y}_{t+l} = E\left[Y_{t+l}\middle| y_t,...,y_1\right]$ feltételes várható érték lesz, ami Gauss-folyamat esetében az idősor elemeinek lineáris kombinációja. Nem-gaussi esetben azonban $\hat{Y}_{t+l}$ nem feltétlenül lesz lineáris. A (4.1) modell esetében $\hat{Y}_{t+l} = E\left[Y_{t+l}\middle| y_t,...,y_1\right] = E\left[Y_{t+l}\middle| y_t,...,y_{t-p+1}\right]$, ami a $g_l(y_t,...,y_{t-p+1}) = \hat{Y}_{t+l}$ jelölés mellett a

$$g_m(y_t,...,y_{t-p+1}) = a_0 + \sum_{j=1}^{p} a_j g_{m-j}(y_t,...,y_{t-p+1}), \quad m = 1,...,l \tag{4.9}$$

rekurzív formulából számolható, ahol

$$g_0(y_t,...,y_{t-p+1}) = y_t \ ,..., \quad g_{1-p}(y_t,...,y_{t-p+1}) = y_{t-p+1}. \tag{4.10}$$

Az előrejelzés tehát az idősor elemeinek lineáris kombinációja, ami nem-gaussi esetben nem feltétlenül szerencsés. Ilyenkor hasznos a (4.2) nemlineáris modell, sőt az a helyzet is elképzelhető, hogy bár Y_t gaussi, de az e_t zaj nem az, ami csak nemlineáris folyamat esetében lehetséges.

A gyakorlati alkalmazások során természetesen az *f* függvényre vonatkozóan valamilyen feltevéssel kell élni. A függvény választásától függően számos nemlineáris idősor-modell ismeretes. Ezek tárgyalásától eltekintünk, saját munkáinkra való tekintettel csupán két igen jelentős modellt mutatunk be. További részletek például Tong (1990) vagy Fan and Yao (2003) könyvében találhatók.

4.1.2.1. TAR modell

Nemlineáris modellt legegyszerűbben úgy képezhetünk, hogy (4.1) autoregresszív együtthatóit függővé tesszük a folyamat múltbeli értékeitől. Az első pillantásra igen speciálisnak tűnő, de a gyakorlati alkalmazások során gyakran hasznos ún. küszöbmodell vagy TAR (threshold autoregressive: TAR) modell az

$$Y_t = a_0^{(k)} + a_1^{(k)} Y_{t-1} + ... + a_p^{(k)} Y_{t-p} + b^{(k)} e_t \ , \quad r_{k-1} \le Y_{t-d} < r_k \ , \quad k = 1,.., K \tag{4.11}$$

alakban definiálható. Az autoregresszív paraméterek tehát *K* számú értéket vehetnek fel, vagyis a folyamat *K* számú rezsimből tevődik össze. A *t* időpontban fellépő aktuális *k* rezsimet az Y_{t-d} küszöbváltozó határozza meg annak megfelelően, hogy melyik, az $r_k, k = 0,.., K$ küszöbparaméterek által definiált intervallumba esik az értéke.

A $p, K, d, \{r_k\}$ ismeretében az autoregresszív együtthatók az OLS módszerrel, tehát

$$Q = \sum_{k=1}^{K} \sum_{i=p+1}^{n} \delta_{ki} (y_i - a_0^{(k)} - a_1^{(k)} y_{i-1} - ... - a_p^{(k)} y_{i-p})^2 \tag{4.12}$$

minimalizálásával becsülhetők, ahol δ_{ki} egy vagy nulla értéket vesz fel attól függően, hogy az idősor az *i*-edik időpontban a *k*-adik rezsimben van vagy sem. Az autoregresszió rendje (*p*) és a rezsimek száma (*K*) Akaike (1974) nyomán becsülhető. Nevezetesen az a *p* és *K* az optimális, amelyre

$$AIC(p, K) = (n - P) \ln \hat{\sigma}_e^2 + 2K(p + 1) \tag{4.13}$$

minimális, ahol

$$\hat{\sigma}_e^2 = Q/(n - p) \tag{4.14}$$

és *P* az illesztett *p*-edrendű modellek közül a maximális rendet jelöli. A *d* késleltetési paraméter és az $\{r_k\}$ küszöbparaméterek Tsay (1989) szerint adhatók meg. Ezek ugyanis szinte mellékesen nyerhetők annak az igen fontos kérdésnek a vizsgálata során, hogy a szóban forgó idősorunk származhat-e TAR modellből vagy inkább lineáris folyamat realizációjaként

tekintendő. Ez a feladat természetesen statisztikai próba végrehajtását igényli. Ha vannak olyan késleltetési paraméterek, melyeknél a linearitásra vonatkozó null-hipotézis elvetése szükséges, akkor a legmegfelelőbb *d* az, amelyre a próbastatisztika a legnagyobb. Az említett hipotézisvizsgálat azon alapul, hogy rekurzívan, az adatsor újabb és újabb adatát bevonva végezzük el az autoregresszív modell illesztését a küszöbváltozó értékeinek növekvő sorrendje szerint. Ily módon a küszöbváltozó *i*-edik legkisebb értékéhez tartozó $\hat{a}_0(i), \hat{a}_1(i), ..., \hat{a}_p(i)$ becsült autoregresszív paraméterek rendkívül informatívak. Lineáris folyamat esetén ugyanis a rekurzívan becsült autoregresszív paraméterek a tényleges autoregresszív paraméterek körül szóródó ponthalmazt alkotnak. TAR folyamat esetén azonban a küszöbparaméterek környékén törések láthatók a rekurzívan becsült autoregresszív paraméterek menetében, s e törések tájékoztatnak a küszöbparaméterek körülbelüli elhelyezkedéséről. Mivel különböző *i*–kre különböző hosszúságú adatsorral számolunk, az adott aktuális autoregresszív paraméter értékeit célszerű a becslés aktuális szórásával normálni. Az eljárás (Tsay, 1989) részletesebb áttekintése Matyasovszky (2002) munkájában található, míg a küszöbparaméterek pontosabb becslésére Matyasovszky (2001) ismertet eljárást.

Végezetül természetesen szükséges annak vizsgálata, hogy az optimálisnak tűnő TAR modell statisztikailag valóban szignifikáns javulást hoz-e, és így indokolt-e használata a lineáris modell ellenében. Ez a likelihood-hányados-próba segítségével végezhető el. Nevezetesen, a linearitásra vonatkozó null-hipotézis teljesülése esetén az

$$(n-P)\ln\left(\hat{\sigma}^2_{e,L} / \hat{\sigma}^2_{e,NL}\right) \qquad (4.15)$$

próbastatisztika aszimptotikusan khí-négyzet eloszlású $(K-1)(p+1)$ szabadsági fokkal, ahol a (4.15)-ben szereplő *L* és *NL* index a lineáris, illetve a nemlineáris modellre vonatkozik.

Megjegyezzük, hogy a TAR modell meteorológiai alkalmazásaival kapcsolatban saját korábbi tanulmányaink (Matyasovszky, 2001; Matyasovszky, 2003) kivételével csupán Zwiers and von Storch (1990) egy igen egyszerű módszere említhető.

4.1.2.2. ARCH modell

Egy ARCH modell szórása, szemben az AR modellel, nem állandó, és a nemlinearitás a szórás változásának speciális formájából fakad. A *q*-adrendű ún. autoregressive conditional heteroscedastic (ARCH(q)) folyamat Engle (1982) nyomán a következőképp definiálható. Legyen e_t az $e_t = g(t)z_t$ alakban értelmezett valószínűségi változók sorozata, ahol

$$g^2(t) = b_0 + \sum_{j=1}^{q} b_j e_{t-j}^2 \tag{4.16}$$

és z_t, az ún. innováció, zérus várható értékű, egy szórású, független és azonos eloszlású valószínűségi változók sorozata. Nyilvánvaló, hogy e_t zérus várható értékű, továbbá korrelálatlan, de nem független, hiszen a $g^2(t)$ feltételes variancia függ a folyamat múltjától. A variancia pozitív voltához szükséges, hogy $b_0 > 0$, $b_j \geq 0, j = 1,...,q$ legyen. Könnyű látni, hogy e_t^2 AR(q) folyamat, hiszen $e_t^2 = b_0 + b_1 e_{t-1}^2 + ... + b_q e_{t-q}^2 + r_t$, melyben $r_t = g^2(t)(z_t^2 - 1)$. Feltételes várható értéke $g^2(t)$, míg (feltétel nélküli) várható értéke, vagyis e_t (feltétel nélküli) varianciája $\sigma_e^2 = b_0/(1 - b_1 - ... - b_q)$. Ahhoz, hogy ez a mennyiség véges pozitív szám legyen, teljesülnie kell a $b_1 + ... + b_q < 1$ relációnak. A variancia időbeli viselkedését leíró $g(t)$ függvényt volatilitásnak nevezik. A volatilitás az időhöz persze csak áttételesen kapcsolódik, valójában e_t múltbeli értékeinek nemlineáris függvénye. Ha ezek a múltbeli értékek jelentősen eltérnek nullától, akkor a volatilitás nagy és a folyamat viselkedése igen bizonytalan. Ekkor a folyamat tovább maradhat a zérustól (a várható értékétől) távol, de hirtelen megváltozhat menete az ellenkező irányba is. Kis volatilitás azonban a folyamat

finom változásait eredményezi mindaddig, míg nagy innováció nem lép fel, ami a folyamatnak a zérustól való nagy eltávolodását eredményezi. Mindezek eredményeképp a variancia időbeli klasztereződése az ARCH folyamat tipikus tulajdonsága.

Az ARCH modell paramétereinek becslése az AR modell paraméter becslésével analóg kérdés, csak most a rendelkezésre álló adatok négyzeteit használjuk. A legegyszerűbb eset az OLS technika, ami azonban most eléggé kevéssé hatékony, mert e_t^2 eloszlása igen távol van a gaussitól (Amano and Taniguchi, 2008). Ezért az ún. two-stage algoritmusok (Mousazadeh et al. 2007) vagy a normalizált legkisebb négyzetek módszere (Fryzlewicz et al., 2008) jóval előnyösebbek. Az ML módszer az innováció valószínűségi eloszlásának ismeretét feltételezi. Általánosan követett eljárás a pszeudo-maximum-likelihood, amikor az innovációt normális eloszlásúnak tekintik, jóllehet a valóságos eloszlás sosem lehet az (Fancq and Zakoian 2004).

Egy $y_1,...,y_n$ idősor általában nem tekinthető egy ARCH folyamat realizációjának, mert az adatsor elemei korreláltak. Az ilyen adatok AR-ARCH modellel kezelhetők, amikor is az autoregresszió e_t hibája egy ARCH modellt követ. Az $a_0,...,a_p$ autoregresszív paraméterek és az ARCH modell $b_0,...,b_q$ paraméterei Fryzlewicz et al. (2008) alapján becsülhetők. A p és q értéke AIC (Akaike, 1974) módosítása vagy inkább BIC (Liew and Chong, 2005; Hughes et al., 2004) alapján adható meg.

Az ARCH modellek korábbi meteorológiai alkalmazásáról nincs tudomásunk, jóval részletesebb matematikai ismeretek például Fan and Yao (2005) kötetéből nyerhetők.

4.2. ALKALMAZÁSOK

4.2.1. Napi parlagfű pollenkoncentráció

A 4.1.1 fejezetben ismertetett eljárást Szeged napi parlagfű pollenkoncentrációira alkalmaztuk az 1997-2006 időszakban (Matyasovszky and Makra, 2011).

A koncentrációk valószínűségi eloszlásának elemzésénél nehézséget okoz a koncentrációkban jelenlévő igen erős évi menet. Ezért a pollenszezon minden napja körül értelmeztünk egy időbeli környezetet, és az adott napra vonatkozó vizsgálatot e környezetbe eső adatokkal hajtottuk végre, vagyis minden napra ilyen módon végeztük el a khí-négyzet próbát és a Kolmogorov-Szmirnov-próbát. Az intervallumnak elég szélesnek kell lennie ahhoz, hogy elegendő adat jusson, de elég keskenynek is ahhoz, hogy az évi menet elhanyagolható legyen az intervallumokon belül. Végül úgy választottuk meg az intervallumot, hogy minden naphoz 90 adat tartozzék. Ahogy már említettük, a gamma-eloszlás illeszkedése egyáltalán nem volt megfelelő. Ellenben mindkét teszt esetében a napok nagy többségében az 5-20%-os szignifikancia-szinten fenntartható a lognormalitásra vonatkozó null-hipotézis. A pollenszezon kezdeténél és végénél a 0,1-1%-os szignifikancia-szint már cseppet sem látszik meggyőzőnek. Ennek azonban minden bizonnyal nem a ténylegesen rossz illeszkedés, hanem a szigorú évi menet az oka. A pollenszezon belsejében ugyanis a 90 adat az aktuális naphoz tartozó plusz-mínusz négy napos intervallumból kerül ki. Ezzel szemben a pollenszezon első vagy utolsó napján az intervallum már kilenc napos, ahol az évi menet sokkal kevésbé hanyagolható el.

Az erős évi menet azzal a további következménnyel is jár, hogy a (4.3)-ban szereplő paramétereknek időfüggőknek kell lenniük. Ezért (4.4) helyett

$$\sum_{k=2}^{n}\left(z_k-\left[\alpha_0+\beta_0 z_{k-1}+(\alpha_1+\beta_1 z_{k-1})(t_k-t_i)\right]\right)^2 K\left(\frac{t_k-t_i}{h}\right) \tag{4.17}$$

minimalizálandó $2 \le i \le n$ mellett, és $\hat{a}_0(t_i) = \hat{\alpha}_0$, $\hat{a}_1(t_i) = \hat{\beta}_0$. Az időfüggő σ^2 becslése: $\hat{\sigma}^2(t_i) = \hat{\lambda}_0$, ahol $\hat{\lambda}_0$ és $\hat{\lambda}_1$ minimalizálja a

$$\sum_{k=2}^{n} \left(\tilde{z}_k - [\lambda_0 + \lambda_1 (t_k - t_i)]\right)^2 K\left(\frac{t_k - t_i}{h}\right) \tag{4.18}$$

mennyiséget $2 \le i \le n$ mellett, továbbá

$$\tilde{z}_k = z_k - \hat{\mu}(t_k). \tag{4.19}$$

A (4.17) és (4.18) egyenlet a már többször alkalmazott nemparaméteres eljárásból adódik (Cai, 2007).

A napi parlagfű pollenkoncentrációk előrejelzése nyilván pontosítható, ha a 2.2 fejezetben említett meteorológiai változókat is figyelembe vesszük. Az ott ismertetett eljárás alapján pusztán a napi középhőmérséklet bevonása indokolt. Ekkor (4.17) helyett

$$\sum_{k=2}^{n} \left(z_k - [\alpha_0 + \beta_0 z_{k-1} + \gamma_0 x_{k-1} + (\alpha_1 + \beta_1 z_{k-1} + \gamma_1 x_{k-1})(t_k - t_i)]\right)^2 K\left(\frac{t_k - t_i}{h}\right) \tag{4.20}$$

minimalizálása szükséges, és

$$\hat{a}_0(t_i) = \hat{\alpha}_0, \hat{a}_1(t_i) = \hat{\beta}_0, \hat{b}(t_i) = \hat{\gamma}_0, \tag{4.21}$$

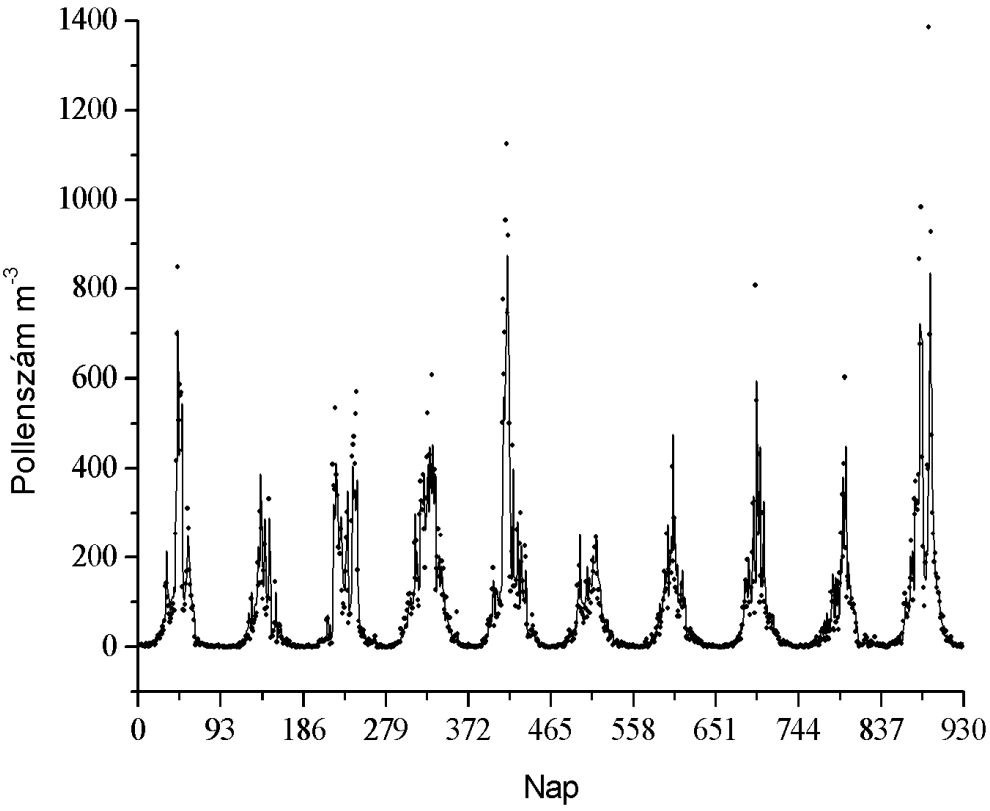
míg

$$\hat{\mu}(t_i) = \hat{a}_0(t_i) + \hat{a}_1(t_i) z_{i-1} + \hat{b}(t_i) x_{i-1}, \tag{4.22}$$

ahol $x_1, ..., x_n$ jelöli a napi középhőmérsékleti adatsort.

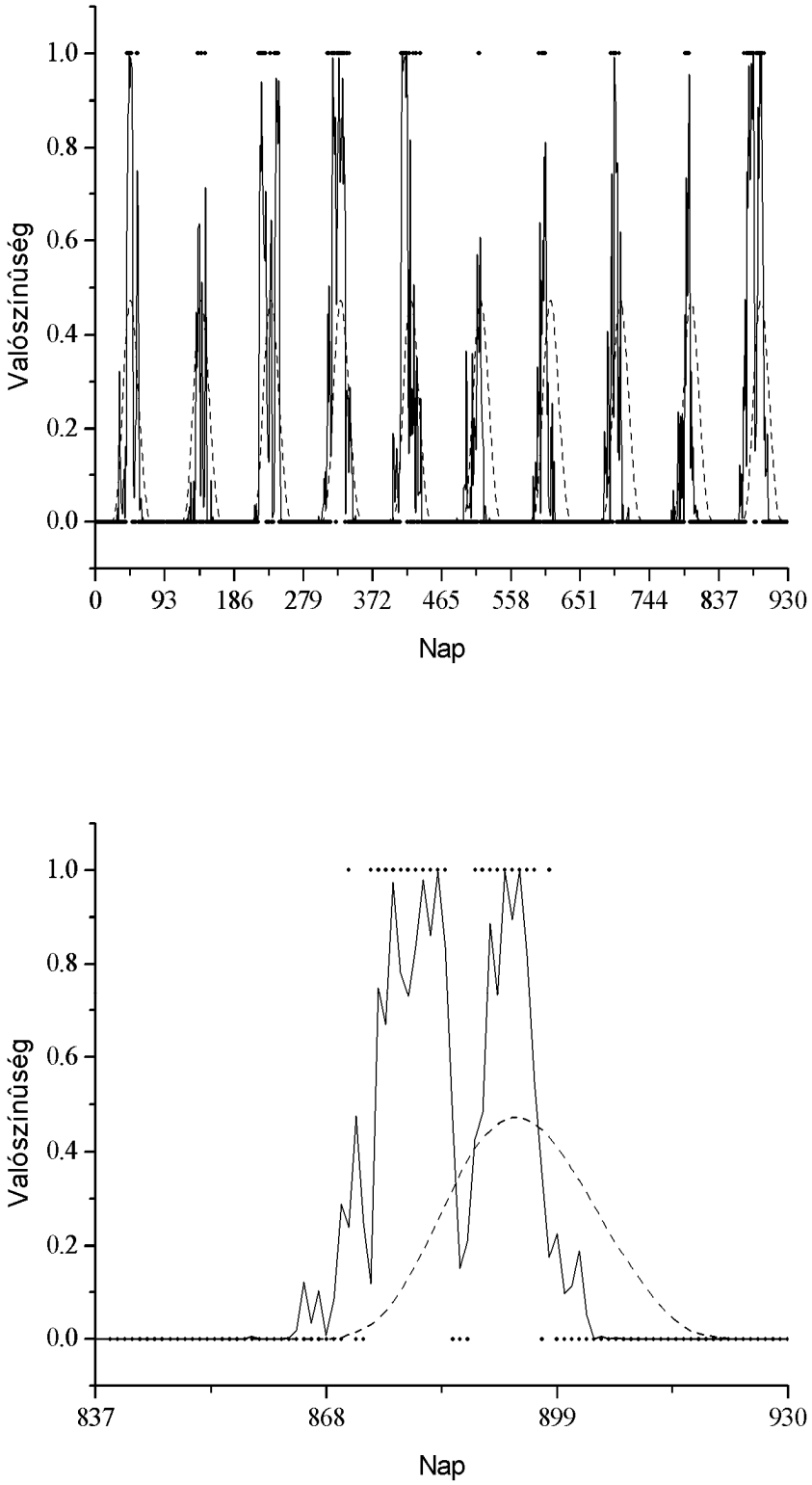
Az így értelmezett kiterjesztett AR(1) modell által megmagyarázott relatív variancia 53,5%, míg a megmagyarázott relatív varianciához hasonlóan értelmezett megmagyarázott relatív abszolút hibaátlag 40,3%. A 2.2 fejezetben más módszerrel nyert hasonló értékek 52,2%, illetve 37,4%. Ezek a számok azt mutatják, hogy a mostani eljárás nem hoz alapvető javulást az előzőekhez képest. E tény elismerése mellett azonban meg kell említeni, hogy a kiterjesztett AR(1) modell lényegesen gazdagabb információt szolgáltat. Nevezetesen, a (4.3)

sűrűségfüggvény, pontosabban annak időfüggő paraméterű és meteorológiai változót is tartalmazó változata nem csupán pontbecslést (például feltételes várható érték: lásd 16. ábra),



16. ábra
Napi pollenkoncentráció (pontok) becslése a kiterjesztett AR(1) modellel (folytonos vonal). A 93-as szám többszörösei az egyes évek pollenszezonainak végét jelöli.

hanem intervallumbecslést is lehetővé tesz. Könnyen kiszámolható ugyanis adott napi pollenkoncentráció és napi középhőmérséklet esetében a másnapi koncentráció adott intervallumba esésének becsült valószínűsége. A legérdekesebb kérdés talán az lehet, hogy milyen valószínűséggel számíthatunk valamilyen kritikus érték meghaladására. Az ilyen becslés jóságát úgy értékelhetjük, hogy összevetjük a becsült valószínűséget egy olyan indikátorváltozóval, ami egyet vagy nulla értéket vesz fel aszerint, hogy bekövetkezett-e vagy



17. ábra
A 200 pollenszám m^{-3}-es küszöb meghaladási valószínűségének becslése a kiterjesztett AR(1) modellel (folytonos vonal) és az éves trenddel (szaggatott vonal) és a küszöbmeghaladás indikátorváltozója (pontok) a teljes 10 éves periódusra (fenn), illetve az utolsó évre (lenn)

sem a küszöb meghaladása. Ezt illusztrálja a 17. ábra, illetve összegzi a 3. Táblázat. Úgy véljük, hogy különösen a 17. ábra alsó része igen meggyőző.

3. Táblázat
Átlagos négyzetes hiba gyöke (RMSE) és abszolút hibaátlag (MAE) a küszöbmeghaladás valószínűségi becslésére a kiterjesztett AR(1) modell és az éves trend esetén

Küszöb	RMSE		MAE	
Pollenszám m^{-3}	AR(1)	Trend	AR(1)	Trend
20	0.213	0.261	0.092	0.149
50	0.196	0.254	0.081	0.142
100	0.241	0.295	0.113	0.187
200	0.213	0.270	0.090	0.152

Sem a megmagyarázott relatív variancia ($1-RMSE^2_{AR(1)}/RMSE^2_{Trend}$), sem a megmagyarázott relatív abszolút hibaátlag ($1-MAE_{AR(1)}/MAE_{Trend}$) nem mutat világos függést a választott küszöbtől, sőt ez a függés nem is számottevő. Például a megmagyarázott relatív abszolút hibaátlag 38,3% és 43% között mozog a 20, illetve az 50 pollenszám m^{-3} küszöb mellett, míg a 100 és 200 pollenszám m^{-3} küszöbökhöz a közöttes 39,6% és 40,8% tartozik.

4.2.2. NGRIP és Vostok adatok együttes elemzése

A következőkben NGRIP (North Greenland Ice Core Project members 2004) O^{18}/O^{16} izotóparányával kapcsolatos $\delta^{18}O$ és Vostok deutérium tartalmának 122950 év hosszúságú adatsorának együttes vizsgálatáról számolunk be Matyasovszky (2010b) alapján. Mivel a Vostok adatok időben nem ekvidisztáns módon állnak rendelkezésre, ezért lineárisan interpoláltuk ugyanazon időpontokra, amelyekben az NGRIP adatok hozzáférhetők 50 éves időbeli felbontással. A jobb összehasonlítás érdekében mindkét idősort standardizáltuk, vagyis az eredeti adatok átlagát kivontuk az adatokból, majd e különbségeket osztottuk az adatok szórásával.

Grönland és Antarktisz éghajlatváltozásaiban meglévő alapvető egyezések, hasonlóságok mellett számottevő különbségek is találhatók. Szembetűnő, hogy Grönland éghajlata jóval gyorsabb, gyakran intenzívebb változásokat mutat, mint a fokozatosabban változó Antarktiszé. Ráadásul, amikor Grönland viszonylag lassú hűlési periódusban van, olyankor Antarktisz is inkább hűl, de legalábbis nem melegszik. Ilyenkor tehát a két félteke szinkronban van. Grönland melegedési időszakaival viszont gyakran antarktiszi hűlés társul. Ilyenkor tehát a két félteke aszinkronban van. Néhány jelentős grönlandi Dansgaard-Oeschger-esemény (DO esemény) mégis jelentkezett az Antarktiszon is, de csak a leghosszabbak és időben nem pontosan egybeesve (Steig and Alley, 2002). A két félteke klímaingadozásainak fázisrelációját (szinkronitását-aszinkronitását) magyarázó elmélet az ún. bipoláris mérleg, melyben komoly helyet foglalnak el a globális óceáni cirkuláció változásai (Severinghaus, 2009). Az nem teljesen tisztázott, hogy a mechanizmust az Északi vagy a Déli Félteke irányítja (Seidov et al., 2001), de Steig and Alley (2002) szerint ez pusztán statisztikai elemzéssel nem is deríthető ki. A következőkben cáfoljuk ezen utolsó megállapítást, és statisztikai úton válaszolunk arra a kérdésre, hogy a Grönland-Antarktisz éghajlatváltozásainak kapcsolatában melyik tag rendelkezik elsődleges szereppel.

Először mindkét adatsorhoz AR modelleket illesztettünk. AIC AR(5), illetve AR(3) modellt talált optimálisnak az NGRIP, illetve a Vostok adatsorra. Azonban mind az AR modell zajának varianciája, mind az ún. Portmanteau-próba (Ljung and Box, 1978) azt jelzi, hogy az NGRIP esetén az AR(3) modell csaknem olyan pontos, mint az AR(5). Ezért a két adatsorra egységesen az AR(3) modellt választottuk. A zaj varianciája 0,0444 és 0,0101 az NGRIP, illetve a Vostok adatsorra. Mindkét zaj igen alacsony szintű, vagyis viszonylag nagy pontossággal előrejelezhetők az adatsorok, ám mindez fokozottan igaz Vostokra.

A következőkben a két adatsor együttesét *p*-edrendű vektor autoregresszív (VAR(p)) folyamattal közelítjük. Legyen X_t és Y_t az NGRIP és a Vostok adatsort generáló sztochasztikus folyamat, ami a VAR(p) modellel kifejezve

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \sum_{j=1}^{p} \begin{bmatrix} a_j & b_j \\ c_j & d_j \end{bmatrix} \begin{bmatrix} X_{t-j} \\ Y_{t-j} \end{bmatrix} + \begin{bmatrix} e_{t,X} \\ e_{t,Y} \end{bmatrix}, \tag{4.23}$$

ahol $e_{t,X}$ és $e_{t,Y}$ külön-külön fehérzaj folyamat, továbbá $e_{t,X}$ és $e_{s,Y}$ korrelálatlan $t \neq s$ esetén. A rendelkezésünkre álló $\underline{z}_1 = (x_1, y_1)^T, ..., \underline{z}_n = (x_n, y_n)^T$ idősor alapján az $a_j, b_j, c_j, d_j, j = 1,..., p$ paraméterek az OLS módszerrel becsülhetők, vagyis

$$\sum_{i=p+1}^{n} (\underline{z}_i - \hat{\underline{z}}_i)^T (\underline{z}_i - \hat{\underline{z}}_i) \tag{4.24}$$

minimalizálásával, ahol

$$\hat{\underline{z}}_i = \sum_{j=1}^{p} \begin{bmatrix} a_j & b_j \\ c_j & d_j \end{bmatrix} \underline{z}_{i-j}. \tag{4.25}$$

Az eljárás az ismeretlen paraméterekre nézve egy lineáris egyenletrendszert szolgáltat. A feladat elvégzése nyomán az NGRIP és Vostok adatsorhoz tartozó zaj szórásnégyzete *p*=3 mellett 0,0435 és 0,0101, ami (az egyváltozós AR modellek zajának varianciájával való összehasonlítás révén) előrevetíti, hogy az NGRIP adatsor viselkedésére nézve Vostok rendelkezik némi információval, ám a Vostok adatsor viselkedésére nincs hatással az NGRIP adatsor. Hogy ez így van-e azt a Granger-féle okozatiság (Granger, 1969) elvének alkalmazásával deríthetjük ki. Nevezetesen ha (4.23)-ban minden c_j zérus, de nem minden b_j zérus, akkor X_t az Y_t Granger-okozata. Megfordítva, ha minden b_j zérus, de nem minden c_j zérus, akkor Y_t az X_t Granger-okozata. Továbbá ha nem minden b_j és nem minden c_j zérus, akkor a két folyamat kölcsönhatásban áll. Végül, ha minden b_j és minden c_j zérus (a szumma mögötti mátrixok diagonálisak (4.23)-ban), akkor a két folyamat egyáltalán nem áll

kapcsolatban. Ahhoz, hogy valamelyik értelemben Granger-okozatiságról beszélhessünk, meg kell vizsgálni, hogy a megfelelő AR paraméterek zérusnak tekinthetők-e. Több ezzel kapcsolatos statisztikai próba ismeretes, melyek közül a Wald-tesztet (Dolado and Lütkepohl, 1996) alkalmaztuk. Itt meg kell jegyezni, hogy a Granger-okozatiság elemzése során alkalmazott tesztek eléggé érzékenyek arra, hogy az adatsorokban fellép-e trend vagy sem, mert a tesztekhez tartozó próbastatisztikák valószínűségi eloszlása a trendmentes esetre érvényesek. Ezért a Wald-tesztet a trendtől megtisztított adatsorokra végeztük el oly módon, hogy mindkét adatsor értékeiből kivontuk a trend aktuális értékeit, majd előállítottuk a (4.23) VAR modellt (p=3 mellett) ezen új adatsorokra. A trendeket ezúttal is a WLR módszerrel becsültük, melynek részletei megtalálhatók Matyasovszky (2010b) tanulmányában. Ennek eredményeképp a 3%-os szignifikancia-szinten az NGRIP adatsor a Vostok adatsor Granger-okozata. Ez az eredmény látszólag eldönti a fejezet elején felvetett kérdést, nevezetesen úgy tűnik, hogy az Arktisz-Antarktisz éghajlat-ingadozásokat az Antarktisz vezérli. Erre a következtetésre azonban csak az adatsorok lineáris modellezésével jutunk, és mint majd látni fogjuk, a nemlineáris modellezés útján jelentősen eltérő álláspontra kell helyezkedni. A nemlineáris modellezés kétváltozós TAR modellel történik. E meglehetősen speciálisnak tűnő modell hasznosságának indoklásához egy kitérőt kell tennünk.

Például az NGRIP idősor nemlinearitása egy igen heuriszikus megfontolással is sejthető. Ismert ugyanis, hogy egy lineáris folyamat esetében az időskála megfordítható, vagyis a folyamat egy elemének az időben rákövetkező elemekkel való közelítése ugyanazt a modellt eredményezi, mint a megelőző elemekkel való közelítése. Nemlineáris folyamat esetében viszont az időskála nem megfordítható (Tong, 1990). Ez például úgy szemléltethető, hogy időben ábrázoljuk az idősort, és pici szögben balról nézzük az ábrát (az időskála az eredeti), majd pedig ugyanígy jobbról vesszük szemügyre (megfordítjuk az időskálát). Lineáris folyamat esetében a két irányból nagyon hasonlónak látjuk a görbét. Az NGRIP

adatsorra azonban ez semmiképp nem áll fenn; a két irányból nézve drasztikus különbség látszik, például a DO események erősen aszimmetrikus időbeli lefolyásának köszönhetően (lásd a későbbi 20. ábrát). A kérdést természetesen jóval egzaktabban is vizsgálták, például Braun (2009) az ún. surrogate data módszerrel igazolta az NGRIP adatsor nem lineáris voltát. Nemlineáris modellt legegyszerűbben úgy képezhetünk, hogy (4.1) autoregresszív együtthatóit függővé tesszük a folyamat múltbeli értékeitől, tehát az $a_0,...,a_p$ konstansok helyett $a_0(Y_{t-d}),...,a_p(Y_{t-d})$ függvények szerepelnek. Ez az ún. functional coefficient model (FCM). Az említett függvényeket lokálisan lineárisan (polinomiálisan) közelítve mód van a 2.1 fejezetben bemutatott nemparaméteres módszerhez igen hasonló eljárással történő becslésükre (Cai et al., 2000). Ezt a technikát alkalmaztuk a jelenlegi két adatsorra (NGRIP és Vostok), és arra a meglepő következtetésre jutottunk, hogy az idősor linearitására vonatkozó null-hipotézist még a null-hipotézis viszonylag szűk elfogadási tartományát biztosító 10%-os szignifikancia-szinten sem vethetjük el az FCM által nyújtott modell ellenében. A problémával kapcsolatos statisztikai próbát a Cai et al. (2000) által javasolt bootstrap eljárással végeztük el. E negatív eredmény egyik lehetséges oka az, hogy az idősorok lineárisak, ám az imént elmondottak szerint ez nehezen hihető. A másik lehetőség az, hogy az $a_0(Y_{t-d}),...,a_p(Y_{t-d})$ függvények nem kellően simák, tehát nem közelíthetők jól lokálisan lineárisan. Az ilyen nem simán, tehát hirtelen változó autoregresszív együtthatók viszont épp a TAR modell létjogosultságát jelzik.

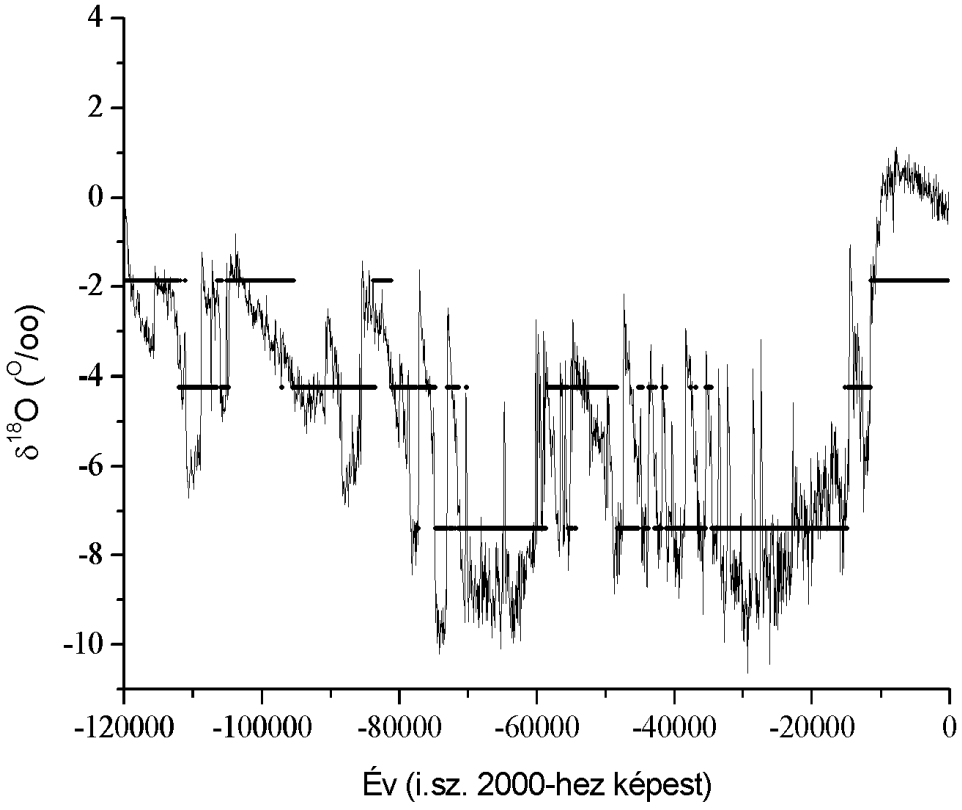
A (4.23) VAR modell jelöléseinek felhasználásával a kétváltozós TAR modell az

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} \alpha^{(k)} \\ \beta^{(k)} \end{bmatrix} + \sum_{j=1}^{p} \begin{bmatrix} a_j^{(k)} & b_j^{(k)} \\ c_j^{(k)} & d_j^{(k)} \end{bmatrix} \begin{bmatrix} X_{t-j} \\ Y_{t-j} \end{bmatrix} + \begin{bmatrix} e_{t,X}^{(k)} \\ e_{t,Y}^{(k)} \end{bmatrix}, \; k = 1,...,K \tag{4.26}$$

alakot ölti. Most tehát az autoregresszív együtthatók mátrixai, továbbá a zajtagok varianciája és a közöttük lévő kovariancia is rezsimfüggő. Küszöbváltozóként mind X_t (NGRIP), mind Y_t (Vostok) szóba jöhet, ám Tsay (1998) vektor TAR (VTAR) modellre vonatkozó módszerét

követve NGRIP a megfelelő küszöbváltozó, mert az ő választásával lényegesen pontosabb modell nyerhető. Hasonló megfontolásból a d=59 küszöbparaméter az optimális. Végül az autoregresszió rendjére és a rezsimek számára az AIC Tsay (1998) szerinti alakjával a p=3 és K=3 adódott. Mivel K=1 esetében a VTAR modell a lineáris VAR modellbe megy át, az optimálisnak adódott K=3 rezsim a nemlineáris VTAR modell létjogosultságát jelzi. Észrevehetjük, hogy NGRIP, mint küszöbváltozó az Arktisz fontosságát húzza alá, amit csak megerősít, hogy az 50 éves időbeli felbontás mellett a d=59 küszöbparaméter 2950 évnek felel meg. Ez gyakorlatilag éppen kétszerese a DO események előfordulásával kapcsolatos 1470 évnek (lásd 3.2.2 fejezet). A zaj varianciája 0,0410 és 0,0087 az NGRIP, illetve a Vostok adatsorra, ami azt jelzi, hogy a VTAR modell a VAR modellhez képest a Vostok adatsor esetében hozott nagyobb javulást. Mivel NGRIP a küszöbváltozó, a két adatsor együttes viselkedésében ő mindenképp szerepel. Ahhoz, hogy eldöntsük, hogy a két adatsor között kölcsönhatás van-e, vagy a Vostok az NGRIP Granger-okozata, meg kell vizsgálni, hogy az NGRIP adatsor kialakításában a Vostok adatsornak a VTAR modellben szereplő összes paramétere (mindegyik rezsimben) zérusnak tekinthető-e. A Wald-teszt (mely ezúttal is a trendtől megtisztított adatsorokra lett elvégezve) azt jelzi, hogy az említett paraméterek két rezsimben különböznek zérustól (az 1%-os szignifikancia-szinten), vagyis az Arktisz-Antarktisz éghajlat-ingadozások statisztikailag kölcsönhatásban állnak. Érdemes azonban megemlíteni, hogy ebben a kölcsönhatásban inkább az Arktisznak van elsődleges szerepe. Ez abból látható, hogy a küszöbváltozó az NGRIP, illetve, hogy a VTAR modellhez tartozó zaj varianciája nagyobb mértékben csökken Vostok, mint NGRIP esetén. Vagyis a nemlineáris modellben, szemben a lineáris modellel, NGRIP nagyobb plusz információt nyújt Vostokra nézve, mint megfordítva. Érdekes még, hogy a zaj varianciája a két idősorra az egyes rezsimekben eltérő tendenciát mutat. Az első rezsimben (a legalacsonyabb küszöbparaméterhez tartozó rezsimben) NGRIP zajának varianciája a legnagyobb, míg

Vostoké a legkisebb. A harmadik rezsimben (a legmagasabb küszöbparaméterhez tartozó rezsimben) NGRIP zajának varianciája a legkisebb, míg Vostoké a legnagyobb. Más szóval, NGRIP legpontosabb előrejelzése Vostok egyidejű legpontatlanabb előrejelzésével társul, illetve NGRIP legpontatlanabb előrejelzése Vostok egyidejű legpontosabb előrejelzésével jár együtt. Ez a két terület éghajlati aszinkronitásának egy újabb aspektusa. A három rezsim előfordulásának relatív gyakorisága rendre 20,9%, 70,2% és 8,9%.



18 .ábra
Az NGRIP $\delta^{18}O$ adatsor (folytonos vonal) (4.26) VTAR modelljének fixpontjai (pontok). A $\delta^{18}O$ értékek a jelenlegi (2000-es év) értéktől való eltérések.

Egy (4.1) stacionárius AR folyamat egyetlen fixponttal rendelkezik. Ez úgy értendő, hogy a 4.1.2. fejezetben szereplő *l*-lépéses $\hat{Y}_{t+l}$ előrejelzés a rendelkezésre álló idősor konkrét

értékeitől függetlenül $l \to \infty$ mellett ugyanazon számhoz, a várható értékhez konvergál. Egy TAR modell több fixponttal rendelkezhet, vagyis az, hogy $\hat{Y}_{t+l}$ mihez konvergál, függhet az $y_t, y_{t-1}, ..., y_{t-m}$, $m = \max\{p, d\}$ konkrét értékeitől. Ezen értékek olyan összességét, mely adott fixponthoz való konvergenciát eredményez, a fixpont vonzási tartományának nevezzük. Ezen kívül határciklus is felléphet, amikor $\hat{Y}_{t+l}$ nem egy számhoz, hanem egy periodikusan ismétlődő számsorozathoz tart. A jelenlegi kétváltozós VTAR modellünk három fixponttal és egy határciklussal rendelkezik. A határciklus periódusideje 3400 év, amely a DO események között eltelt átlagos időt reprezentálja. Ha ugyanis a következő fejezetben detektált 28 DO eseményt az első és utolsó esemény előfordulása közötti időszakban egyenletesen oszlatjuk el képzeletben, akkor a DO események 3400 évenkénti átlagos előfordulásához jutunk. A megfigyelt adatsor a teljes időszakban mindössze négyszer esett a határciklus vonzási tartományába, ám éppen akkor, amikor két egymást követő DO esemény között szinte pontosan 3400 év telt el. A három fixpontot a fixpontok vonzási tartományának előfordulásai időpontjai függvényében ábrázolva, az adatsort jellemző tendencia szakaszonként konstans közelítését kapjuk (18 ábra).

Végezetül hadd említsük meg, hogy többváltozós TAR modell meteorológiai alkalmazásával korábbi tanulmányokban nem találkoztunk.

4.2.3. Hirtelen éghajlatváltozás: Dansgaard-Oeschger-események

Az 1.2.2 fejezetben az Északi Hemiszféra hőmérsékletében beálló hirtelen éghajlatváltozások detektálását tanulmányoztuk először az évszázados, majd az évezredes időskálán. Megjegyezzük, hogy ez utóbbi skálán, pontosabban az elmúlt 1500 évre vonatkozóan 35, az Északi Hemiszféra különböző területeiről származó különböző típusú proxy adatsort is megvizsgáltunk (Matyasovszky and Ljungqvist, 2012). Végezetül a 10000 éves időskálán a Holocén időszak hirtelen éghajlatváltozásit tanulmányoztuk. Ezúttal tovább tágítjuk a kört és

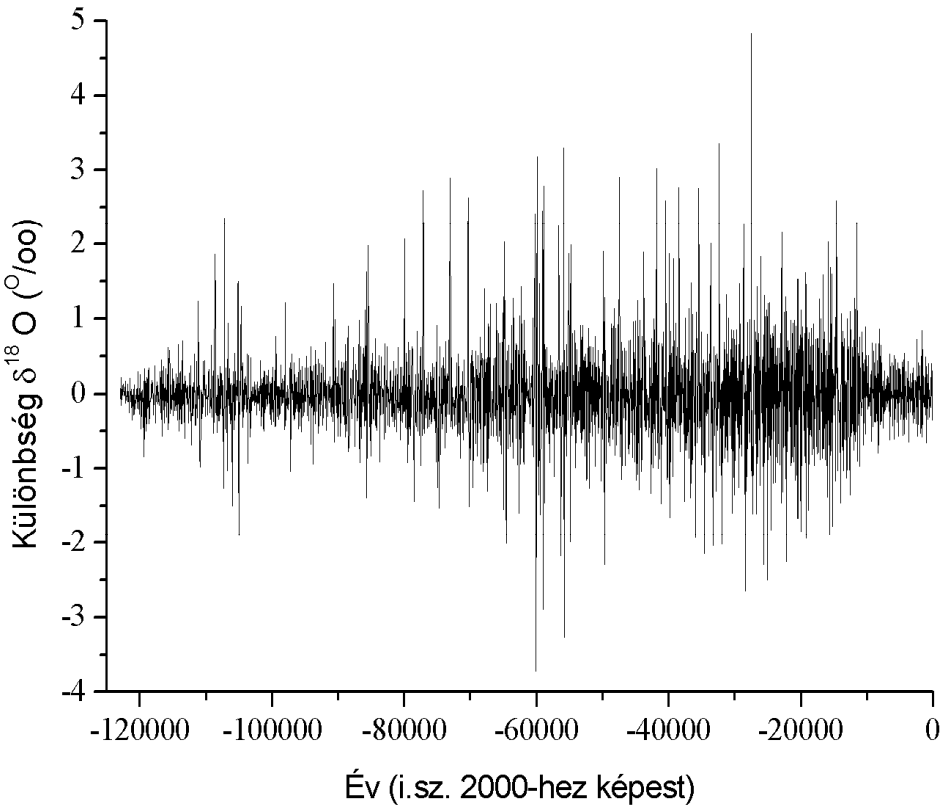
a 100000 éves időskálát tekintjük. Ismét az NGRIP (North Greenland Ice Core Project members 2004) jégfurat O^{18}/O^{16} izotóparányával kapcsolatos 122950 év hosszúságú $\delta^{18}O$ adatokkal dolgozunk, melyek 50 éves átlagokként állnak rendelkezésre. Az adatsor talán legjellemzőbb vonása a Dansgaard-Oeschger (DO) események megjelenése. Egy DO esemény egy gyors felmelegedést, majd lényegesen lassúbb lehűlést foglal magában. A felmelegedés időszaka néhány évtizedtől körülbelül egy évszázadig terjed, míg a lehűlés néhány évszázad alatt zajlik le. Az események azonosítása az iménti értelmezésnek megfelelően, de szubjektív úton történik. Bár például Ditlevsen et al. (2005) komoly erőfeszítést tett az objektivitás irányába, definíciójuk jelentős szubjektív elemeket is tartalmaz. Persze a DO események tökéletesen objektív azonosítása aligha megvalósítható, de a szubjektivitás mértékének csökkentése fontos szempont.

Matyasovszky (2010b) 7250 éves sávszélességet talált optimálisnak a $\delta^{18}O$ NGRIP adatsorban jelentkező trend nemparaméteres becsléséhez. Nyilvánvaló tehát, hogy az adatsor ilyen szélességű simítása nem képes reprezentálni a DO eseményeket, ezért ezek az események, mint hirtelen éghajlatváltozások detektálása nem történhet az 1.2.2 fejezetben leírt módon. Mivel egy DO eseményhez kapcsolódó megfigyelt értékek erősen eltérnek a trendtől, ezek a trendre rakódó zaj szórásának hirtelen megnövekedéseként jelentkeznek. Az ilyen módon változó szórás leírására az ARCH modell kínálkozik.

Úgy tűnik, hogy az 1.2.2 fejezet módszere és az ARCH modellezés, lévén két különböző eljárás, semmilyen módon nem köthető egymáshoz. Ez valójában nem így van. Vegyük ugyanis az eredeti adatsor helyett a $\Delta y_t = y_t - y_{t-1}, t = 2,...,n$ különbség adatsort. A különbségképzés ugyanis kiszűri az alacsony frekvenciákat, míg a ki nem szűrt frekvenciák a DO eseményekkel társíthatók. Nevezetesen, ha a trend sima, az $m(t) - m(t-1)$ tag a $\Delta y_t = m(t) - m(t-1) + e_t - e_{t-1}$ különbségben elhanyagolható, és Δy_t zérus körül ingadozik. Ha azonban a t=s időpontban a trend deriváltjának mondjuk a k-adik ugrása van, akkor

$\Delta y_s \approx c_k + e_s - e_{s-1}$, ami c_k körüli érték. Sok, sűrűn elhelyezkedő ilyen ugrás esetében logikus várakozás, hogy a Δy_t adatsor jól leírható ARCH modellel. Mivel a különbség idősorhoz illesztjük az ARCH modellt és a különbségképzés a derivált véges különbséges közelítése, a volatilitás szintje a trend deriváltja változásának mértékére utal. Pontosabban szólva, a nagy volatilitás nagy esélyt jelez arra nézve, hogy a trend deriváltja hirtelen változik, és ez utóbbi éppen a hirtelen éghajlatváltozás 1.2.2 fejezetben látott definíciójának lényegét jelenti (Matyasovszky, 2011).

Az említett várakozást erősíti meg a 19. ábra, ahol a különbség adatsor klaszteresződő szerkezete egy ARCH modell szükségességét sugallja. Az idősor azonban nem túl erős, de statisztikailag szignifikáns autokorrelációval is rendelkezik, ami AR-ARCH modell keresését



19. ábra
$\delta^{18}O$ NGRIP különbség adatsor az elmúlt 122900 évre

igényli. A BIC alapján negyedrendű AR tag adódott optimálisnak, ezért AR(4)-ARCH(q), q=0,1,…,8 modelleket illesztettünk az adatsorhoz. Most BIC szerint q értéke legalább 8 kell, legyen. Mivel ilyen magas rendű modell nem praktikus, a jóval gazdaságosabb GARCH modellel (Bollerslev, 1986) is próbálkoztunk. Az ARCH és GARCH modellek úgy viszonyulnak egymáshoz, ahogyan az AR és ARMA modellek lineáris folyamatok esetén, mivel egy GARCH modell az adatok négyzetére vonatkozó ARMA modell. A GARCH modellek ezért kevesebb becsülendő modellparaméterrel pontosabban közelíthetik a volatilitást, mint az ARCH modellek. BIC azonban még a GARCH modell fokára is irreálisan nagy értéket jelzett, ami nemlineáris ARCH folyamat szükségességét veti fel. E nemlinearitás leírására az aszimmetrikus volatilitást mutató, két ARCH rezsimből álló ún. GJR-ARCH modellt (Glosten et al., 1993), alkalmaztuk. Míg az ARCH modell az adatok négyzetére vonatkozóan egy AR modell, addig a GJR-ARCH modell az adatok négyzetére vonatkozó TAR modell. Mivel a TAR modell nem lineáris, így a nemlineáris ARCH modellhez további nemlinearitás csatlakozik az eredeti, tehát nem a négyzetes adatok esetében. A BIC alapján nyert optimális AR-GJR-ARCH modell

$$\Delta y_t = -0{,}0162\Delta y_{t-1} - 0{,}1198\Delta y_{t-2} - 0{,}0509\Delta y_{t-3} - 0{,}0444\Delta y_{t-4} + e_t\,, \tag{4.27}$$

ahol

$$e_t = \begin{cases} g_1(t)z_t\,, & e_{t-1} \geq 0{,}58 \\ g_2(t)z_t\,, & e_{t-1} < 0{,}58 \end{cases}, \tag{4.28}$$

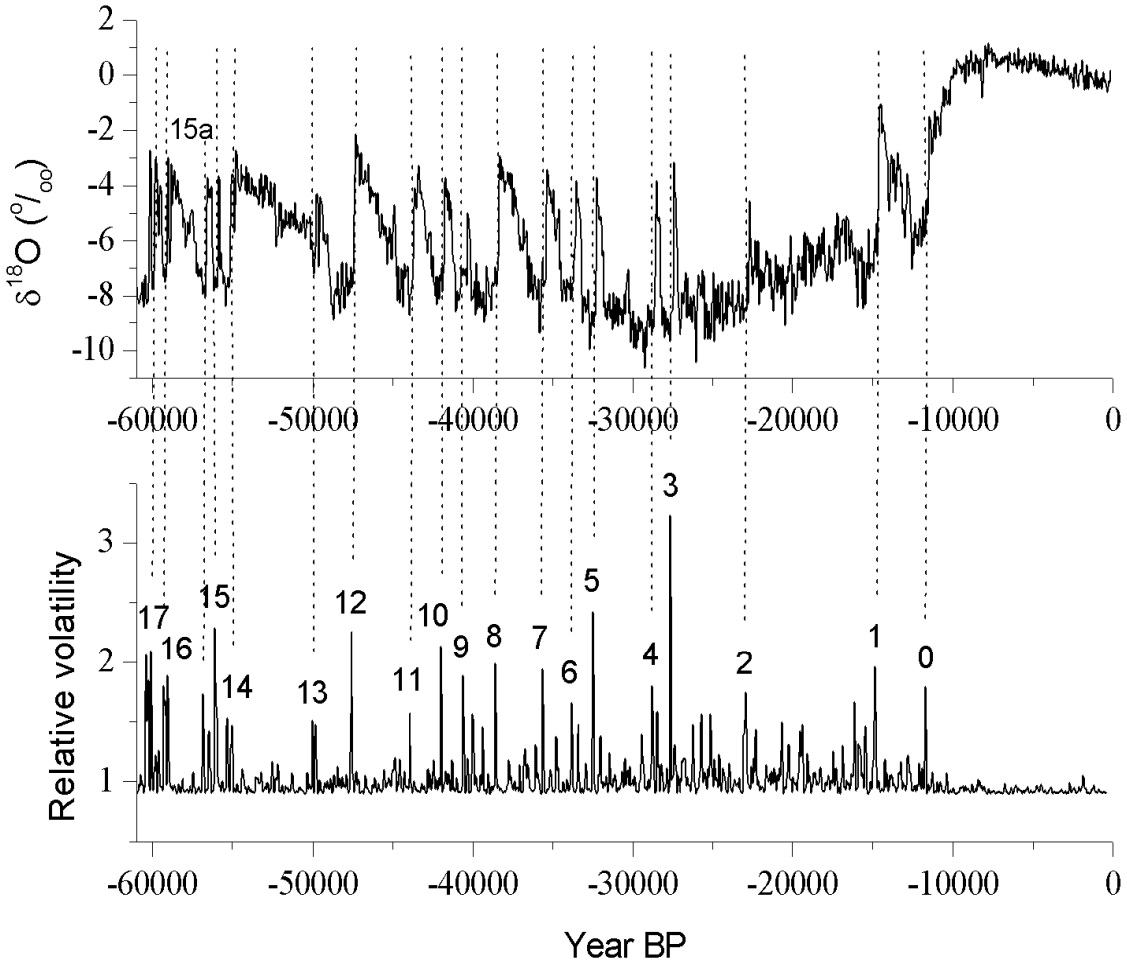
$$g_1(t) = \sqrt{0{,}3263 + 0{,}1750e_{t-1}^2 + 0{,}0758e_{t-2}^2}\,, \quad g_2(t) = \sqrt{0{,}3263 + 0{,}0903e_{t-1}^2 + 0{,}0092e_{t-2}^2}\,. \tag{4.29}$$

Megjegyezzük, hogy az eredeti GJR-ARCH modellben a küszöbparaméter definíció szerint zérus, ami most egy további becsült paraméter, nevezetesen 0,58‰ δ^{18}O.

Láthatjuk, hogy a (4.27) volatilitása akkor van az első rezsimjében, amikor az egymást követő adatok értéke erősen növekedő, mivel az e_t egylépéses előrejelzési hiba erősen pozitív

(lásd (4.28)). Ekkor a (4.29) volatilitás is nagyobb, ami az intenzívebb fluktuációk növekvő lehetőségét jelzi. Mivel a DO esemény hirtelen emelkedéssel indul, az első rezsim az ilyen eseménnyel hozható kapcsolatba. Tekintsük ezért az első rezsim $g_1(t)/\sigma$ relatív volatilitását (a feltételes szórás és szórás hányadosát). Azon események átlagos hossza, amikor a relatív volatilitás folyamatosan egynél nagyobb 1,3-nak adódik (50 éves egységben). Logikus ezért azokat az eseteket hirtelen éghajlatváltozásnak tekinteni, amikor az első rezsim egynél nagyobb relatív volatilitásának időbeli hossza meghaladja az 1,3-as átlagértéket. Ezen kívül a különösen gyors és így nagyon rövid felmelegedést is célszerű hirtelen éghajlatváltozásnak tekinteni. Konkrétabban, három olyan esetet találtunk, amikor az első rezsimben való egyfolytában tartózkodás hossza mindössze egy (az 50 éves időlépcsővel), ám a volatilitás különösen nagy. Ez utóbbit úgy értelmeztük, hogy a feltételes szórás legalább kétszerese a szórásnak. Mindezek alapján 28 hirtelen változást detektáltunk a korábban általánosan elfogadott 25 DO esemény (Dansgaard et al., 1993) helyett. Ezeket az eseteket például Ditlevsen et al. (2005) által objektívnak mondható módszerrel kimutatott DO eseményekkel volna célszerű összehasonlítani. Ők azonban GRIP adatokat használtak, ráadásul az alacsony frekvenciák kiküszöbölése céljából egy 10000 éves felülateresztő szűrő közbeiktatásával. Ezért a jobb összehasonlítás érdekében inkább Braun (2010) tanulmányában az elmúlt 60000 évre található, az NGRIP $\delta^{18}O$ adatokon alapuló szubjektíven meghatározott DO eseményeket tekintjük. Az általunk detektált DO események és Braun (2010) DO eseményei mind számban, mind időbeli elhelyezkedésükben meglehetősen jó egyezést mutatnak. Az egyetlen különbség, hogy a 15-ös esemény előtt egy további DO eseményt azonosítottunk (15a a 20. ábrán). E két időben igen közeli esemény szerintünk helyes különválasztását indokolja, hogy a Braun (2010) szerint is két külön eseménynek tekintett 16-os és 17-es hasonlóan közel jelentkeznek egymáshoz (20. ábra, felső rész). Megjegyezzük, hogy az általunk detektált időpontok a DO események kezdetét jelentik, mert módszerünk a kezdetet és nem az esemény

csúcsát azonosítja. Az összehasonlítás döntően a várt eredményt hozta. Arra ugyanis semmiképp nem számítottunk, hogy akár az események számában, akár időbeli elhelyezkedésükben gyökeres eltérés lesz. Ugyanakkor az is valószínűtlennek tűnt, hogy tökéletes lesz az egyezés. A különbség, döntően az, hogy eljárásunk hárommal több DO eseményt azonosított az általánosan elfogadott 25 (Dansgaard et al., 1993) DO eseményhez képest.



20. ábra
A relatív volatilitás alapján azonosított DO események (lenn) az NGRIP $\delta^{18}O$ adatsorban (fenn) az elmúlt 60000 évre. A $\delta^{18}O$ értékek a jelenlegi (2000-es év) értéktől való eltérések.

ÖSSZEFOGLALÁS

Az értekezésben modern matematikai statisztikai eszközöket mutattunk be és használtunk fel elméleti és alkalmazott klimatológiai vizsgálatokban. Célunk néhány széles körben hasznosítható, ám kevéssé elterjedt módszer megismertetése, majd egy-egy alkalmazásának tárgyalása volt. Az önmagukban is értékes eredményekkel egyben érzékeltetni kívántuk a bennük rejlő további gyakorlati lehetőségeket. A dolgozat legfontosabb részletei a következőkben foglalhatók össze.

Felhasználtunk (Cline, et al., 1995; Wishart, 2009), illetve részben kidolgoztunk (Matyasovszky, 2011; Matyasovszky and Ljungqvist, 2012) egy nemparaméteres regressziós technikán (Matyasovszky, 1992; 1998) alapuló módszert, mely alkalmas a hirtelen éghajlatváltozások (abrupt climate changes) detektálására amennyiben a rendelkezésre álló adatsor elemei kellő sűrűségben állnak rendelkezésre az adatsorban mutatkozó trend simaságához képest. A módszertani újítást az Északi Hemiszféra évi középhőmérsékleti anomáliáinak az 1850-2009 évek időszakára, az Északi Hemiszféra rekonstruált évi hőmérsékleti sorára az i.sz. 200-1995 évekre, továbbá az elmúlt 11700 évre vonatkozó oxigén izotóp adatokra, pontosabban az NGRIP (North Greenland Ice Core Project) jégfurat O^{18}/O^{16} izotóparányával kapcsolatos Holocén $\delta^{18}O$ adatokra alkalmaztuk.

Az első esetben az 1901, 1914, 1942, 1963 és 1975 éveket detektáltuk hirtelen változási időpontoknak. Az átfogó tendencia természetesen a melegedés, ám három hirtelen hűlési időpont (1901, 1942 és 1963) is található. Korábban nem találkoztunk olyan tanulmánnyal, mely a hirtelen változások ilyen finom szerkezetét tárta volna fel.

A hosszabb időskálát vizsgálva objektív definíciót nyújtottunk a Középkori Meleg Periódus (Medieval Warm Period: MWP) és a Kis Jégkorszak (Little Ice Age: LIA)

beazonosítására. Ezek a kifejezések széles körben elterjedtek, jóllehet nincs is egyértelműen elfogadott definíciójuk (Bradley et al., 2003). A hirtelen változások alapján a MWP a 795-1120 időszakra tehető, ami némiképp korábbi és hosszabb, mint Lamb (1977) időszaka, míg LIA az 1390-1880 évekre datálható, ami hosszabb, mint Grove (1988) időszaka. A legnyilvánvalóbb változás azonban a 19. század végén (1883) hirtelen meginduló intenzív melegedés.

A Holocén időszak legjellemzőbb éghajlati epizódja a Holocén Éghajlati Optimum. Mivel ez a meleg periódus fokozatos lehűléssel ért véget, behatárolása korábban eléggé bizonytalan volt. A detektált hirtelen változások szerint azonban 9900-3300 évvel ezelőttre datálható, ami jóval hosszabb, mint a korábban mások által definiált időszakok.

A hirtelen éghajlatváltozásokat a 100000 éves időskálán is vizsgáltuk (Matyasovszky, 2011). Az adatsor talán legjellemzőbb vonása a Dansgaard-Oeschger (DO) események jelentkezése. Egy DO esemény néhány évszázados időszakra terjed ki egy gyors felmelegedéssel és egy lényegesen lassúbb lehűléssel. Mivel az NGRIP jégfurat 122950 év hosszúságú $\delta^{18}O$ adatai 50 éves átlagokként állnak rendelkezésre, ezért az előzőekben említett eljárás ezúttal nem volt alkalmazható. A DO események karakterisztikus ideje és az adatok időbeli sűrűségének viszonyából fakadóan ezek az események most az adatsor varianciájának ingadozásaként jelennek meg. A variancia ilyen viselkedését egy AR-GJR-ARCH modell (Engle, 1982; Glosten et al., 1993) további finomításával írtuk le. Ennek segítségével a DO események korábbi döntően szubjektív beazonosítása helyett egy döntően objektív eljárással detektáltuk a DO eseményeket. Az általánosan elfogadott 25 (Dansgaard et al., 1993) DO eseményen túl 3 további esetet találtunk. Megjegyezzük, hogy ARCH (autoregressive conditional heteroescadistic) modell korábbi meteorológiai alkalmazására nem találtunk példát.

Az éghajlati idősorok spektrálanalízisének irodalma rendkívül gazdag. A folytonos spektrumot jellemző spektrális sűrűségfüggvény (a kontinuum-számosságú frekvenciák fontosságát jellemző függvény) becslése rendszerint a periodogram simításán (nemparaméteres regresszióján) alapszik. Az éghajlati adatsorok azonban általában nem mentesek a diszkrét periódusoktól (gondoljunk például az évi menetre), sőt a spektrálanalízis egyik fő feladata éppen a diszkrét periódusok detektálása. E frekvenciáknál a periodogram viselkedése alapvetően eltér a folytonos spektrumra jellemző periodogram elemek viselkedésétől. A cél tehát olyan, ún. robusztus eljárást értelmezni a spektrális sűrűségfüggvény becslésére, amely gyakorlatilag nem vesz tudomást az ilyen kiugró értékekről. A gondolat ugyan nem előzmény nélküli a meteorológiai irodalomban, ám Mann and Lee (1996) ún. medián szűrője pusztán csak segédeszköz egy elsőrendű autoregresszív (AR(1)) modellel nyert spektrális sűrűség pontosításához. Egy, a meteorológiai irodalomban korábban nem ismert robusztus becslés (Janas and von Sachs, 1995) finomítását alkalmaztuk (Matyasovszky, 2010a) például a NAO index havi adatsorának 1865-2002 időszakára. Az eredmények alapján elmondhatjuk, hogy a módszer kiállja az összehasonlítást, például a rendkívül hatékonynak tartott wavelet alapú spektrálanalízissel (Nicolay et al., 2008) is. Itt még hozzá kell tennünk, amit korábban nem említettünk, hogy a periodogram simítása helyett a jobb spektrális felbontást lehetővé tevő ún. tapered periodogram (Priestley, 1981) simítását végeztük el.

A spektrálanalízis során széles körben követett gyakorlat, hogy AR(1) modellt illesztenek az adatsorhoz, és ha a periodogram valamely frekvenciánál meghalad egy az AR(1) spektrumtól és a választott szignifikancia-szinttől függő küszöböt, akkor itt a spektrum különbözik az AR(1) spektrumtól. Az éghajlati adatsorok általában ekvidisztánsan elhelyezkedő időpontokban állnak rendelkezésre. Olykor azonban ez nem teljesül, amire jó példát szolgáltatnak a paleoklíma adatok. Nem ekvidisztáns időpontok esetében mind az

AR(1) modell illesztése, mind a periodogram definíciója módosításra szorul. Az első feladattal Mudelsee (2002) foglalkozott az autoregresszív együttható legkisebb négyzetes (ordinary least squares: OLS) becslésével. Kimutattuk azonban, és példákkal is alátámasztottuk, hogy ez a számos paleoklimatológiai vizsgálatban alkalmazott eljárás igen erősen túlbecsli a spektrális sűrűséget az alacsony frekvenciáknál. A probléma kiküszöbölésére az általunk javasolt súlyozott legkisebb négyzetes (weighted least squares: WLS) becslés alkalmas (Matyasovszky, 2012b). Az OLS módszeren alapuló, de az egyes frekvenciákat egyedileg kezelő Lomb-Scargle (L-S) periodogram (Lomb, 1976; Scargle, 1982) helyett az összes frekvenciát együttesen kezelő és ezért teljes legkisebb négyzetes (total least squares: TLS) becslésnek nevezett eljárást vezettünk be a periodogram előállítására (Matyasovszky, 2012b). A GISP2 Oxigén izotóp adatoknak a 15000 - 60000 évvel ezelőtti időszakára vonatkozó WLS-AR(1) modellünk és TLS periodogramunk Schulz and Mudelsee (2002) tanulmányának több furcsaságát tárja fel és magyarázza meg. Az ő OLS-AR(1) modelljükön és az L-S periodogramon alapuló, számos paleoklimatológiai vizsgálatban alkalmazott ún. REDFIT eljárásuk ugyanis komoly pontatlanságot rejt magában. Hasonló következtetésre jutottunk Vostok deuterium tartalom (az elmúlt 422766 évben) adatsorának elemzésekor is.

Az éghajlati adatsorokat jelentős részben vörös zaj jellemzi, vagyis az egyre kisebb frekvenciák egyre fontosabb szerepet játszanak az adatsor kialakításában. Másképp szólva, a spektrális sűrűség a zérus felé monoton növő. A vörös zaj spektrumot az adatsorhoz illesztett AR(1) modell spektrális sűrűségével szokás közelíteni, ami a legegyszerűbb, de egyáltalán nem biztos, hogy kielégítő eljárás. A vörös zaj becslésére ezért a meteorológiai irodalomban még nem alkalmazott ún. izoton regresszió (IR) módszerén alapuló eljárást javasoltuk (Matyasovszky, 2013b). Ekkor olyan görbét keresünk, mely a periodogramot négyzetes hibaösszegben optimálisan közelíti a vörös zajjal kapcsolatos monotonitási feltétel teljesülése

mellett (Zhao and Woodroofe, 2012), pontosabban, a már korábban említett okból kifolyólag, a feladat robusztus változatát tekintettük (Álvarez and Yohai, 2011). Az Északi Hemiszféra 200-1995 évek közötti hőmérsékleti sorára alkalmazva az eljárást az AR(1) és az IR spektrális sűrűségek között óriási különbség mutatkozott az alacsony frekvenciáknál. Jól példázza ez azt, hogy a vörös zaj spektrumnak AR(1) modellel történő közelítése esetenként rendkívül megbízhatatlan lehet.

NGRIP $\delta^{18}O$ és Vostok deutérium tartalmának 122950 év hosszúságú adatsorának együttesét vizsgáltuk egy vektorértékű nemlineáris idősor modellel (Matyasovszky, 2010c). Ez az ún. vektot TAR (VTAR) modell több rezsimből álló vektor AR (VAR) modell, ahol az adott időponthoz tartozó aktuális rezsimet egy küszöbváltozó múltbeli értéke határozza meg. Az elemzés célja az volt, hogy kiderítsük, hogy az Arktisz és Antarktisz éghajlat-ingadozásaiban megfigyelhető fázisreláció kialakításában elsődlegesen melyik tag tölti be a vezető szerepet. A két félteke klímaingadozásainak fázisrelációját magyarázó elméletben ugyanis nem teljesen tisztázott, hogy a mechanizmust az Északi vagy a Déli Félteke irányítja, de Steig and Alley (2002) szerint ez pusztán statisztikai elemzéssel nem is deríthető ki. Ez utóbbi megállapítás igaz lehet az adatsorok hagyományos lineáris elemzésére, ám a nemlineáris modellezéssel választ adtunk a kérdésre. Mivel a kétváltozós TAR modell illesztése során (Tsay, 1998) kiderült, hogy a küszöbváltozó az Arktiszt reprezentáló NGRIP, ezért NGRIP a két adatsor együttes viselkedésében alapvető szerepet játszik. Annak tisztázására, hogy NGRIP dominanciája egyértelmű-e megvizsgáltuk, hogy Vostok pusztán csak reagál-e NGRIP-re, vagy vissza is hat rá, azaz inkább kölcsönhatásban állnak. Ezt a Granger-féle okozatiság (Granger, 1969) elvének alkalmazásával derítettük fel. E szerint az Arktisz-Antarktisz éghajlat-ingadozások statisztikailag kölcsönhatásban állnak, ám a VTAR és VAR modell tulajdonságainak összehasonlításából kiderült, hogy ebben a kölcsönhatásban inkább az

Arktisznak van elsődleges szerepe. Megjegyezzük, hogy a TAR modell meteorológiai alkalmazásaival kapcsolatban csak az egyváltozós esetre találtunk példát, de saját korábbi tanulmányaink (Matyasovszky, 2001; 2003) kivételével itt is csupán Zwiers and von Storch (1990) egy igen egyszerű módszere említhető.

Feladatul tűztük ki a hazánkban nagyon elterjedt parlagfű erősen allergén pollenjének napi koncentráció becslését. Szeged, Legnano és Lyon napi parlagfű pollenkoncentrációit hoztuk kapcsolatba (Makra et al., 2011a) a megelőző napi koncentrációval és a megelőző napi átlaghőmérséklettel, csapadékösszeggel és átlagos szélsebességgel az 1997-2006 időszakban egy időfüggő nemparaméteres regressziós technika, az ún. time-varying coefficient model (Cai, 2007) segítségével. Itt ismét megemlíthetjük, hogy az eljárás korábbi meteorológiai alkalmazásáról nincsen tudomásunk. A Szegeden kívüli további két város bevonására azért került sor, mert a Kárpát-medencén kívül még a Pó-alföld (Legnano) és a Rajna völgye (Lyon) Európa erősen parlagfüves területei (Makra et al., 2011a). A becslés által megmagyarázott relatív variancia Szegedre a legnagyobb (52,2%) és Legnanora a legkisebb (22%), tehát a legpontosabban Szeged napi parlagfű pollenkoncentrációja becsülhető a három hely közül. A legfontosabb meteorológiai változónak a napi középhőmérséklet (Szeged és Legnano) és a napi csapadék (Lyon) bizonyult.

A parlagfű napi koncentráció becslését időfüggő medián és kvantilis regresszió (Koenker and Bassett, 1978) segítségével is elvégeztük (Makra and Matyasovszky, 2010). A medián regresszió azért hasznos, mert a napi koncentrációk valószínűségi eloszlása erősen aszimmetrikus és így a medián és a várható érték jelentősen különbözik. A gyakorlati feladatok során ugyanis nem annyira a minél kisebb négyzetes hiba, hanem a minél alacsony abszolút hiba biztosítása a cél, és a medián regresszió éppen az abszolút hibaátlagot minimalizálja. A medián regresszió természetesen számottevően kisebb abszolút hibaátlagot

nyújtott az előzőekben említett regresszióhoz képest (a megmagyarázott variancia csökkenése árán). A napi parlagfű pollenkoncentráció kvantilisei általában kisebbek az esős, mint a száraz napokon, továbbá a napi koncentráció valószínűségi eloszlása sokkal elnyújtottabb a magas koncentrációk felé a száraz napokon. Az esős napokhoz tartozó kvantilisek azt jelzik, hogy a pollenkoncentrációk jóval kisebb változékonyságúak a csapadékos napokon. Mindez világosan jelzi a csapadék koncentrációcsökkentő hatását. A hazai parlagfű pollenterhelés súlyosságára jól rávilágít a kvantilis regressziónak a nulla valószínűségi érték melletti alkalmazása, amivel a koncentrációk alsó határát lehet meghatározni, mert ez a kvantilis az a legnagyobb koncentráció, amelynél nagyobb koncentráció egy valószínűséggel fordul elő. Az egészségi kockázatot jelentő kritikus parlagfű koncentrációt a lehetséges legkisebb koncentráció csaknem 20 napon át bizonyosan meghaladja az említett küszöbértéket még úgy is, hogy ha pusztán az évi menettel foglalkozunk, vagyis a prediktorok értékét, tehát például egy előző napi esetlegesen magas koncentrációt, figyelembe sem vesszük.

Az előző vizsgálatok során kiderült, hogy a parlagfű pollenkoncentráció becsléséhez a leghasznosabb prediktor a megelőző nap koncentrációja. Célszerű ezért egy AR(1) folyamatot illeszteni a koncentrációk idősorához. Comtois (2000) munkájával szemben kimutattuk, hogy a napi parlagfű pollenkoncentrációk jó közelítéssel időben változó paraméterű lognormális eloszlásúnak tekinthetők (Matyasovszky and Makra, 2011). Ennek kihasználásával kidolgoztunk egy időfüggő lognormális AR(1) modellt, amely az aktuális nap koncentrációjának az előző napi koncentrációra vonatkozó időfüggő feltételes valószínűségi eloszlását adja meg. Előállítottuk a modell egy olyan kiterjesztését is, amelyben az említett eloszlás az előző napi középhőmérséklet értékeitől is függ. A modell által megmagyarázott variancia és a megmagyarázott abszolút hiba ugyan azt sugallja, hogy a mostani eljárás nem hoz gyökeres javulást az előzőekhez képest, de valójában lényegesen gazdagabb információt szolgáltat. A kiterjesztett modell által nyújtott feltételes eloszlásból ugyanis könnyen

kiszámolható adott napi pollenkoncentráció és napi középhőmérséklet esetében a másnapi koncentráció adott intervallumba esésének becsült valószínűsége, így például az is, hogy milyen valószínűséggel számíthatunk valamilyen kritikus koncentráció meghaladására.

Az elmúlt néhány évized során számos növényi faj pollenje által kiváltott allergiás tünetek és allergiás légúti betegségek számának erőteljes növekedése figyelhető meg világszerte (Damialis et al., 2007). Az ezzel párhuzamosan zajló globális éghajlatváltozás miatt logikus felvetés, hogy a pollenszezon fenológiai jellemzői (a pollenszezon kezdete, vége, tartama) és mennyiségi jellemzői (évi összes pollenszám, éves napi csúcspollen) is változást mutatnak. A kérdést vizsgálva 19 taxon napi pollenszámait elemeztük a rendelkezésünkre álló 1997-2007 közötti 11 éves időszakban Szegedre (Makra et al., 2011b). A trend létének igazolása mindössze 11 adat felhasználásával persze igen kevéssé ígérkezett sikeresnek. Valóban, a 19 taxon 5 karakterisztikájára elvégzett összesen 95 Mann-Kenndall-teszt (MK-teszet) (Önöz and Bayazit, 2003) 16, 10 és 3 esetben jelzett trendet a 10, 5 és 1 %-os szignifikancia-szinten, illetve pusztán az évi összes pollenszám esetében a 19 taxonra csupán 4, 1 és 0 esetben mutatkozott szignifikáns trend az említett szinteken. Ezért a napi pollenszámokra a 11 éves időszak 11 adata alapján az év összes napjára külön-külön elvégeztük az MK-tesztet. A trend létezéséről szóló döntés most azonos azzal a problémával, hogy a napi MK teszt értékek évi átlaga szignifikánsan különbözik-e nullától. Ekkor az 5%-os valószínűségi szinten már 11 taxon évi összes pollenszáma mutatott szignifikáns trendet, s e 11-ből 7 növekedést jelzett. Megtörténhet azonban, hogy a pollenszezon pozitív és negatív trendeket mutató időszakokból áll össze és az MK-teszt értékek évi átlaga emiatt nem ad átfogó (teljes évre számított) trendet. Ezért a napi MK-teszt értékek évi menetét nemparaméteres módszerrel becsültük, ami már az összes taxonra trendet jelezett. A nemparaméteres regresszió a napi bontású MK-statisztikákon keresztül tehát összehasonlíthatatlanul finomabb képet nyújt az allergén pollenek trendjéről, mint a mások

által követett szokványos eljárás. Meteorológiai változók napi értékeire hasonló vizsgálatot végeztünk, s ezek 11 éves trendjének évi menetét kapcsolatba hoztuk az egyes taxonok 11 éves trendjének évi menetével. A taxonok trendjének évi menete és a meteorológiai változók trendjeinek évi menete közötti többszörös korreláció meglepően nagynak adódott; a legnagyobb az *Artemesia* esetén 0,998, de az *Ambrosia* és *Urtica* esetén fellépő legalacsonyabb 0,827-es érték is igen magas. Az egyes meteorológiai változók trendjei tehát igen jól magyarázzák a pollenkoncentrációk trendjeit.

Irodalom

Akaike H, 1974: A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19: 716-723.

Álvarez EE, Yohai VJ, 2011: M-estimators for Isotonic Regression. arXiv: 1105.5065v1 stat.ME.

Amano T, Taniguchi M, 2008: Asymptotic efficiency of conditional least squares estimators for ARCH models. Statistics & Probability Letters, 78: 179-185.

Bellman RE, 1961: Adaptive control processes: a guided tour. Princeton University Press.

Benner TC, 1999: Central England temperature: long term variability and teleconnections. International Journal of Climatology, 19: 391–403.

Berk KN, (1974): Consistent autoregressive spectral estimates. Annals of Statistics, 2: 489-502.

Bollerslev T (1986): Generalized Autoregressive Conditional Heteroskedasticity. Journal of Econometrics, 31: 307-327.

Box JE, 2002: Survey of Greenland instrumental temperature records: 1873-2001. International Journal of Climatology, 22: 1829-1847.

Box GEP, Jenkins GM, 1970: Time Series Analysis, Forecasting and Control. San Francisco: Holden Day.

Bradley RS, Hughes MK, Diaz HF, 2003: Climate in medieval time. Science, 302: 404-405.

Braun H, 2009: Strong indications for nonlinear dynamics during Dansgaard-Oeschger events. Climate of the Past Discussions, 5, 1751-1762.

Braun H, 2010: Limitations of red noise in analysing Dansgaard-Oeschger events. Climate of the Past, 6: 85-92.

Cai Z, 2007: Trending time-varying coefficient time series models with serially correlated errors. Journal of Econometrics, 136: 163-188.

Cai Z, Fan J, Yao Q, 2000: Functional-Coefficient Regression Models for Nonlinear Time Series. Journal of the American Statistical Association, 95, 941-956.

Cai Z, Fan J, Yao Q, 2000: Functional-Coefficient Regression Models for Nonlinear Time Series. Journal of the American Statistical Association, 95, 941-956.

Chen ZG, Wu KH, Dahlhaus R, 2000: Hidden frequency estimation with data tapers. Journal of Time Series Analalysis, 21: 113-142.

Cheng H, Fleitmann D, Edwards LR, Wang X, Cruz FW, Auler AS, Mangini A, Wang Y, Kong X, Burns SJ, Matter A, 2009: Timing and structure of the 8.2 kyr B.P. event inferred from $\delta^{18}O$ records of stalagmites from China, Oman, and Brazil. Geology, 37: 1007-1010.

Chernozhukov V, 2005: Extremal quantile regression. Annals of Mathematical Statistics, 3: 806-839.

Cline DBH, Eubank RL, Speckman PL, 1995: Nonparametric estimation of regression curves with discontinuous derivatives. Journal of Statistical Research, 29: 17–30.

Clot B, 2003: Trends in airborne pollen: an overview of 21 years of data in Neuchâtel (Switzerland). Aerobiologia, 19: 227-234.

Comtois P, 2000: The gamma distribution as the true aerobiological probability density function (PDF). Aerobiologia, 16: 171-176.

Craven P, Wahba G, 1979: Smoothing noisy data with spine functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numerical Mathematics, 31: 377-403.

Cristofori A, Cristofolini F, Gottardini E, 2010: Twenty years of aerobiological monitoring in Trentino (Italy): assessment and evaluation of airborne pollen variability. Aerobiologia, 26: 253-261.

D'Amato G, Liccardi G, D'Amato M, Holgate S, 2005: Environmental risk factors and allergic bronchial asthma. Clinical and Experimental Allergy, 35: 1113-1124.

Damialis A, Halley JM, Gioulekas D, Vokou D, 2007: Long-term trends in atmospheric pollen levels in the city of Thessaloniki, Greece. Atmospheric Environment, 41: 7011-7021.

Damon PE, Sonnett CP, 1991: Solar and terrestrial components of the atmospheric ^{14}C variation spectrum. In: The Sun in Time. Edited by Sonnett CP, Gaimpapa MS, Matthews MS. Tucson: University of Arizona Press.

Dansgaard W, Johnsen SJ, Clausen HB, Dahl-Jensen D, Gundestrup NS, Hammer CU, Hvidberg CS, Steffensen JP, Sveinbjörnsdottir AE, Jouzel J, Bond G, 1993: Evidence for general instability of past climate from a 250-kyr ice-core record. Nature, 364: 218-220.

Deeming TJ, 1975: Fourier analysis with unequally-spaced data. Astrophysics and Space Science, 36: 137-158.

Dévényi D, Gulyás O, 1988: Matematikai statisztikai módszerek a meteorológiában. Budapest: Tankönykiadó.

Ditlevsen PD, Kristensen MS, Andersen KK, 2005: The recurrence time of Dansgaard-Oeschger events and limits on the possible periodic component. Journal of Climate, 18: 2594-2603

Dolado JJ, Lütkepohl H, 1996: Making Wald tests work for cointegrated VAR systems. Econometric Reviews, 15: 369-386.

Draper N, Smith H, 1981: Applied Regression Analysis. New York: John Wiley & Sons.

Engle R, 1982: Autoregressive Conditional Heteroscedasticity with Estimation of United Kingdom Inflation. Econometrics, 50: 987-1008.

F.-Fernandez M, V.-Fernandez J, 2004: Weighted Local Nonparametric Regression with Dependent Errors: Study of Real Private Residential Fixed Investment in the USA. Statistical Inference for Stochastic Processes, 7: 69–93.

Fan J, 1992: Design-adaptive nonparametric regression. Journal of American Statistical Association, 87: 998-1004.

Fan J, 1993: Local linear regression smoothers and their minimax efficiency. Annals of Statistics, 21: 196-216.

Fan J, Gijbels I, 1992: Variable bandwidth and local linear regression smoothers. Annals of Statistics, 20: 2008-2036.

Fan J, Jiang J, 1999: Variable bandwidth and one-step local M-Estimator. Science in China Series A, 29: 1-15, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.45.9261.

Fan J, Yao Q, 2003: Nonlinear Time Series: Nonparametric and Parametric Methods. Springer Series in Statistics. New York: Springer-Verlag.

Feng G, Gong Z, Zhi R, 2010: Latest Advances in Climate Change Detection Techniques. Acta Meteorolgica Sinica, 24: 1-16.

Fraedrich K, Jiang J, Gerstengarbe F-W, Werner PC, 1997: Multiscale detection of abrupt climate changes: application to River Nile flood levels. International Journal of Climatology, 17: 1301-1315.

Francq C, Zakoian J-M, 2004: Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. Bernoulli, 10: 605-637.

Fryzlewicz P, Sapatinas T, Rao SS, 2008: Normalized least-squares estimation in time-varying ARCH models. Annals of Statistics 36, 742–786.

Galbraith JV, Zinde-Walsh V, 2002: Autoregressive Approximation, with Econometric Applications, 401-421. In: Handbook of Applied Econometrics and Statistical Inference. Edited by Ullah A, Wan ATK, Chaturvedi A. New York: Marcel Dekker.

Glosten LR, Jagannathan R, Runkle DE, 1993: On the Relation between the Expected Value and the Volatility of the Nominal Excess Returns on Stocks. Journal of Finance, 48: 1779-1801.

Goodman J, 1998: Statistics of North Atlantic Oscillation variability. http://www.mit.edu/people/goodmanj/NAOI/index.html

Gourieroux C, Jasiak J, 2006: Autoregressive gamma processes. Journal of Forecasting, 25: 129-152.

Granger CWJ, 1969: Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37: 424-438.

Grenander U, Rosenblatt M, 1957: Statistical Analysis of Stationary Time Series. New York: John Wiley & Sons.

Groeneboom P, Wellner JA, 2001: Computing Chernoff's Distribution. Journal of Computational and Graphical Statistics, 10: 388-400.

Grootes PM, Stuiver M, 1997: Oxygen 18/16 variability in Greenland snow and ice with 10^{-3}– to 10^{5}-year time resolution. Journal of Geophysical Research, 102(C12): 26455–26470.

Grove JM, 1988: The Little Ice Age. New York: Methuen.

Grunwald GK, Hyndman RJ, Tedesco L, Tweedie RL, 2002: Non-Gaussian Conditional Linear AR(1) Models. Australian and New Zealand Journal of Statistics, 42: 479-495.

Hardle W, Müller M, 2000: Multivariate and Semiparametric Kernel Regression. In: Smoothing and Regression: Approaches, Computation, and Application. Edited by Schimek MG, pp. 357-391. Wiley Online Library, Doi: 10.1002/9781118150658.

Hays JD, Imbrie J, Shackleton NJ, 1976: Variations in the Earth's Orbit: Pacemaker of the Ice Ages. Nature, 194: 1121-1132.

Heng D, Leung Y, 2005: Cross-validation in non-parametric regression with outliers. Annals of Statistics, 33: 2291-2310.

Hirst JM, 1952: An automatic volumetric spore trap. Annals of Applied Biology, 39: 257-265.

Huber PJ, 1981: Robust statistics. New York: Wiley.

Hughes AW, King ML, Kwek KT, 2004: Selecting the order of an ARCH model. Economics Letters, 83: 269-275.

Hurrell JW, van Loon H, 1997: Decadal variations in climate associated with the North Atlantic Oscillation. Climatic Change, 36: 301–326.

Ivanov MA, Evtimov SN, 2010: 1963: The break point of the Northern Hemisphere temperature during the twentieth century. International Journal of Climatology, 30: 1738-1746.

Jäger S, 1998: Global aspect of ragweed in Europe. In: Satellite Symposium Proceedings: Ragweed in Europe. Edited by Spieksma FThM. Proceedings of the 6th International Congress on Aerobiology. Alk-Abelló, Perugia, Italy, pp. 6-10.

Janas D, von Sachs R, 1995: Consistency for non-linear functions of the periodogram of tapered data. Journal of Time Series Analysis, 16: 585-606.

Jiang J, Mendelssohn R, Schwing F, Fraedrich K, 2002: Coherency detection of multiscale abrupt changes in historic Nile flood levels. Geophysical Research Letters, 29: 1271, Doi: 10.1029/2002GL014826.

Johnsen SJ, Dahl-Jensen D, Gundestrup N, Steffensen JP, Clausen HB, Miller H, Masson-Delmotte V, Sveinbjörnsdottir AE, White J, 2001: Oxygen isotope and palaeotemperature records from six Greenland ice-core stations: Camp Century, Dye-3, GRIP, GISP2, Renland and NorthGRIP. Journal of Quaternary Science, 16: 299-307.

Jones PD, Mann ME, 2004: Climate Over Past Millennia. Reviews of Geophysics 42, RG2002, Doi: 10.1029/2003RG000143.

Jones PD, Parker DE, Osborn TJ, Briffa KR, 2010: Global and hemispheric temperature anomalies - land and marine instrumental records. In Trends: A Compendium of Data on Global Change. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A. Doi: 0.3334/CDIAC/cli.002

Kariya T, Kurata H. 2004: Generalized Least Squares. Chichester: John Wiley & Sons.

Kaufman DS, Ager TA, Anderson NJ, Anderson PM, Andrews JT, Bartlein PJ, Brubaker LB, Coats LL, Cwynar LC, Duvall ML, Dyke AS, Edwards ME, Eisner WR, Gajewski K, Geirsdottir A, Hu FS, Jennings AE, Kaplan MR, Kerwin MW, Lozhkin AV, MacDonald GM, Miller GH, Mock CJ, Oswald WW, Otto-Bliesner BL, Porinchu, DF, Ruhland K, Smol JP, Steig EJ, Wolfe BB, 2004: Holocene thermal maximum in the western Arctic (0–180°W). Quaternary Science Reviews, 23: 529–560.

Knight JR, Allan RJ, Folland CK, Vellinga M, Mann ME, 2005: A signature of persistent natural thermohalin circulation cycles in observed climate. Geophysical Research Letters 32: L20707. Doi:10.1029/2005GL024233.

Kobashi T, Severinghaus JP, Brook EJ, Barnola JM, Grachev AM, 2007: Precise timing and characterization of abrupt climate change 8200 years ago from air trapped in polar ice. Quaternaty Science Reviews, 26: 1212-1222.

Koenker R, Bassett GB, 1978: Regression quantiles. Econometrica, 46: 33-50.

Koenker R, 2005: Quantile regression. Cambridge: Cambridge University Press.

Kokoszka P, Mikosch T, 2000: The periodogram at Fourier frequencies. Stochastic Processes and their Applications, 86: 49-79.

Lamb HH, 1977: Climate History and the Future, Vol. 2, Climate: Present, Past and Future. New York: Methuen.

Lawrance AJ, 1982: The innovation distribution of a gamma distributed autoregressive process. Scandinavian Journal of Statistics, 9: 234-236.

Li Q, Racine J, 2004: Cross-validated local linear nonparametric regression. Statistica Sinica, 14: 485-512.

Liew VK-S, Chong T-L, 2005: Autoregressive Lag Length Selection Criteria in the Presence of ARCH errors. Economics Bulletin, 3: 1-5.

Ljung GM, Box GEP, 1978: On a measure of lack of fit in time series models. Biometrika, 65: 297–303.

Loehle C, Scafetta N, 2011: Climate Change Attribution Using Empirical Decomposition of Climatic Data. The Open Atmospheric Science Journal, 5: 74-86.

Lomb NR, 1976: Least-squares frequency analysis of unequally spaced data. Astrophysics and Space Science, 39: 447-462.

Luterbacher J, Schmutz C, Gyalistras D, Xoplaki E, Wanner H, 1999: Reconstruction of monthly NAO and EU indices back to AD 1675, Geophysical Research Letters, 26: 2745-2748.

Luterbacher J, Xoplaki E, Dietrich D, Jones PD, Davies TD, Portis D, Gonzalez-Rouco JF, von Storch H, Gyalistras D, Casty C, Wanner H, 2002: Extending North Atlantic Oscillation Reconstructions Back to 1500. Atmospheric Science Letters, Doi:10.1006/asle.2001.0044.

Makra L, Matyasovszky I, 2010: Assessment of daily ragweed pollen concentration with previous-day meteorological variables using regression and quantile regression analysis for Szeged, Hungary. Aerobiologia, 27: 247-259.

Makra L, Matyasovszky I, Deák JÁ, 2011a: Trends in the characteristics of allergenic pollen circulation in Central Europe based on the example of Szeged, Hungary. Atmospheric Environment, 45: 6010-6018.

Makra L, Matyasovszky I, Thibaudon M, Bonini M, 2011b: Forecasting ragweed pollen characteristics with nonparametric regression methods over the most polluted areas of Europe. International Journal of Biometeorology, 55: 361-371.

Mann ME, Lee JM, 1996: Robust estimation of background noise and signal detection in climatic time series. Climatic Change, 33: 409-445.

Mann HB, Wald A, 1943: On the statistical treatment of stochastic difference equations. Econometrica, 11: 173-220.

Matyasovszky I, 1992: Nonparametric Regression Methods for Trend Estimation of Climatological Time Series. 12th International Conference on Probability and Statistics, Toronto, Canada, 1992, pp. 5-10.

Matyasovszky I, 1998: Non-parametric estimation of climate trends. Időjárás, 102: 149-158.

Matyasovszky I, 2001: A nonlinear approach to modeling climatological time series. Theoretical and Applied Climatology, 69: 139-148.

Matyasovszky I, 2002: Statisztikus klimatológia. Budapest: Eötvös Kiadó.

Matyasovszky I, 2003: The relationship between NAO and temperature in Hungary and its nonlinear connection with ENSO. Theoretical and Applied Climatology, 74: 69-75.

Matyasovszky I, 2010a: Improving the methodology for spectral analysis of climatic time series. Theoretical and Applied Climatology, 101: 281-287.

Matyasovszky I, 2010b: Trends, time-dependent and nonlinear time series models for NGRIP and VOSTOK paleoclimate data. Theoretical and Applied Climatology, 101: 433-443.

Matyasovszky I, 2010c: Milankovitch forcing in paleoclimate data. Climate Research, 41: 151-156.

Matyasovszky I, 2011: Detecting abrupt climate changes on different time scales. Theoretical and Applied Climatology, 105: 445-454.

Matyasovszky I, Makra L, 2011: Autoregressive modelling of daily ragweed pollen concentrations for Szeged in Hungary. Theoretical and Applied Climatology, 104: 277-283.

Matyasovszky I, Ljungqvist FC, 2012: Abrupt temperature changes during the last 1,500 years. Theoretical and Applied Climatology, Doi: 10.1007/s00704-012-0725-8.

Matyasovszky I, 2013a: Spectral analysis of unevenly spaced climatological time series. Theoretical and Applied Climatology, 111: 371-378.

Matyasovszky I, 2013b: Estimating red noise spectra of climatological time series. Időjárás, accepted.

Mazzarella A, Scafetta N, 2012: Evidences for a quasi 60-year North Atlantic Oscillation since 1700 and its meaning for global climate change. Theoretical and Applied Climatology, Doi: 10.1007/s00704-011-0499-4.

Mousazadeh S, Karimi M, Farrokhrooz M, 2007: ARCH parameter estimation via constrained two-stage least squares method. Signal Processing and Application 2007, ISSPA 2007, 9th International Symposium 12-15 Feb. 2007, pp. 1-4.

Mudelsee M, 2002: TAUEST: a computer program for estimating persistence in unevenly spaced weather/climate time series. Computers & Geosciences, 28: 69-72.

Müller HG, 1991: Smooth optimal kernel estimators near endpoints. Biometrika, 78: 521-530.

National Research Council, 2002: Abrupt Climate Change: Inevitable Surprises. Washigton D.C.: National Academy Press.

Nie J, King JW, Fang X, 2008: Tibetian uplift intensified the 400 k.y. signal in paleoclimate records at 4 Ma. GSA Bulletin, 120: 1338-1344.

Nielsen AA, 2011: Least Squares Adjusments: Linear and Nonlinear Weighted Regression Analysis. Lecture Note, Lyngby, Denmark: IMM, DTU, http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=2804.

Nicolay S, Mabille G, Fettweis X, Erpicum M, 2008: 30 and 43 months period cycles found in air temperature time series using the Morlet wavelet method. Clymate Dynamics, Doi: 10.1007/s00382-008-0484-5.

North Greenland Ice Core Project members 2004: North Greenland Ice Core Project Oxygen Isotope Data. IGBP PAGES/World Data Center for Paleoclimatology Data Contribution Series # 2004-059. NOAA/NGDC Paleoclimatology Program, Boulder CO, USA. Contributor: Sigfús J. Johnsen, University of Copenhagen.

Önöz B, Bayazit M, 2003: The power of statistical tests for trend detection. Turkish Journal of Engineering and Environmental Sciences, 27: 247-251.

Priestley MB, 1981: Spectral Analysis and Time Series. New York: Academic Press.

Scargle JD, 1982: Studies in astronomical time series analysis II: statistical aspects of spectral analysis of unevenly spaced data. The Astrophysical Journal, 261: 835-853.

Schulz M, Mudelsee M, 2002: REDFIT: estimating red-noise spectra directly from unevenly spaced paleoclimatic time series. Computers & Geosciences, 28: 421–426.

Seidov D, Barron E, Haupt BJ, 2001: Meltwater and global ocean conveyor: northern versus southern connections. Global and Planetary Change, 30: 257-270.

Severinghaus JP, 2009: Climate change: southern see-saw seen. Nature, 457, Doi: 10.1038/4571093a.

Simonoff JS, 1996: Smoothing Methods in Statistics. Springer Series in Statistics. New York: Springer-Verlag.

Smadi MM, 2006: Observed Abrupt Changes in Minimum and Maximum Temperatures in Jordan in the 20th Century. American Journal of Environmental Sciences, 2: 114-120.

Steig EJ, Alley RB, 2002: Phase relationships between Antarctic and Greenland climate records. Annals of Glaciology, 35: 451-456.

Stoica P, Li J, He H, 2009: Spectral Analysis of Nonuniformly Sampled Data: A New approach Versus the Periodogram. IEEE Transactions on Signal Processing, 57: 1415-1425.

Szentimrey T, Farago T, Szalai, S, 1992: Window technique for climate trend analysis. Climate Dynamics, 6: 127-134.

Thomas ER, Wolf EW, Mulvaney R, Steffensen JP, Johnsen SI, Arrowsmith C, White JWC, Vaughn B, Popp T, 2007: The 8.2 ka event from Greenland ice cores. Quaternary Science Reviews, 26: 70-81.

Thompson DWJ, Wallace JM, Kennedy JJ, Jones PD, 2010: An abrupt drop in Northern Hemisphere sea surface temperature around 1970. Nature, 467: 444-447.

Tibshirani RJ, Hoefling H, Tibshirani R, 2011: Nearly-Isotonic Regression. Technometrics, 53: 54-61.

Tong H, 1990: Non-linear Time Series. Oxford: Calderon Press.

Tsay RS, 1989: Testing and Modeling Threshold Autoregressive Processes. Journal of American Statistical Association, 84: 231-240.

Tsay RS, 1998: Testing and Modeling Multivariate Threshold Models. Journal of the American Statistical Association, 93: 1188-1202.

Vio R, Adreani P, Biggs A, 2010: Unevenly-sampled signals: a general formalism of the Lomb-Scargle periodogram. Astronomy & Astrophysics, 519: A86, 12 p.

Wishart J, 2009: Kink estimation with correlated noise. Journal of Korean Statistical Society, 38: 131-143.

Zhao F, Xu Z, Huang J, 2007: Long-Term Trend and Abrupt Change for Major Climate Variables in the Upper Yellow River Basin. Acta Meteorologica Sinica, 21: 204-214.

Zhao O, Woodroofe M, 2012: Estimating a monotone trend. Statistica Sinica, 22: 359-378.

Zwiers F, von Storch H, 1990: Regime dependent auto-regressive time series modeling of the Southern Oscillation. Journal of Climate, 3: 1347-1360.