

Eötvös Loránd Tudományegyetem

Földrajz- és Földtudományi Intézet

Meteorológiai Tanszék

A surrogate data módszer éghajlati idősorok nemlinearitásának detektálására

DIPLOMAMUNKA



Készítette:

Kristóf Erzsébet

Meteorológus mesterszak,

Éghajlatkutató szakirány

Témavezető:

Dr. Matyasovszky István

ELTE TTK, Meteorológiai Tanszék

Budapest, 2015

Tartalomjegyzék

1. Bevezetés.....	3
2. Alapfogalmak	5
3. Szakirodalmi áttekintés	7
3.1. Pótadatsorok előállítási módjai	7
3.2. Hipotézisvizsgálat: a nullhipotézis és a választható próbastatisztikák	12
3.3. A pótadatsor-módszer meteorológiai alkalmazásai	15
4. A pótadatsor-módszer bemutatása és alkalmazása éghajlati idősorokon	19
4.1. A pótadatsor-módszer bemutatása	19
4.1.1. A pótadatsorok létrehozásához használt apparátus	20
4.1.2. A fázisok szimulációja	21
4.1.3. A korrelációk becslése	22
4.1.4. A próbastatisztika megválasztása.....	25
4.1.5. Hipotézisvizsgálat	26
4.2. Grönlandi jégfuratból származó oxigénizotóp idősor.....	27
4.2.1. Az NGRIP oxigénizotóp adatsor.....	27
4.2.2. A pótadatsor-módszer alkalmazása az NGRIP adatsorra.....	29
4.2.3. Az NGRIP adatsorra vonatkozó eredmények	31
4.3. Az Észak-atlanti Oszcilláció indexének idősora.....	33
4.3.1. A NAO-index adatsora.....	33
4.3.2. A pótadatsor-módszer alkalmazása a NAO-index idősorára	35
4.3.3. A NAO-index adatsorra vonatkozó eredmények	37
4.4. Az Északi Hemiszféra átlaghőmérséklete.....	40
4.4.1. Az NH átlagos felszínhőmérsékleti adatsor	40
4.4.2. A pótadatsor-módszer alkalmazása az NH idősorra	41
4.4.3. Az NH adatsorra vonatkozó eredmények.....	43
5. Összegzés	45
Köszönetnyilvánítás	46
Irodalomjegyzék	47

1. Bevezetés

A természetben számos folyamat nem tekinthető lineárisnak, idősoraikhoz nemlineáris idősormodellek illesztendők. A nemlinearitást azonosítani hivatott eljárások, azonban gyakran túl bonyolultak, vagy szubjektív elemeket tartalmaznak, amelyek következtében téves döntésre juthatnak az idősorok nemlinearitását illetően. A diplomamunkában az idősorok nemlinearitásának detektálására kidolgozott, körülbelül két évtizedes múlttal rendelkező eljárás, a *surrogate data* módszer (a továbbiakban: pótdatsor-módszer) kerül bemutatásra és alkalmazásra, saját vizsgálat keretében.

A nemlineáris megközelítés a lineárishoz képest sokkal több számítást igényel. A számítógépek fejlődésével azonban a nagy számításigény napjainkban egyre kisebb problémát jelent, amely a nemlinearitás vizsgálatának, így a pótdatsor-módszer alkalmazásának nagyobb teret biztosít a korábbi, analitikus módszerekkel szemben. Ennek keretében a nemlinearitás azonosítására alkalmas mérőszámok nemcsak az eredeti adatsorra, hanem számítógéppel szimulált pótdatsorokra is kiszámíthatók, azok összehasonlítása érdekében.

A pótdatsor-módszer az idősorok nemlinearitásának azonosítására használt matematikai statisztikai, hipotézisvizsgálati eljárás. Széleskörű meteorológiai alkalmazása azonban még nem terjedt el. A hipotézisvizsgálat során nullhipotézisként a vizsgált folyamat linearitása választandó. A pótdatsorok és a próbastatisztika ennek megfelelően konstruálandó meg. Először az „eredeti”, nemlinearitás szempontjából vizsgált idősorból állítandók elő a lineáris pótdatsorok, amelyre számos módszer létezik. Ezt követően kiszámítandó egy próbastatisztika az eredeti idősorra és a pótdatsorokra. A próbastatisztikát úgy kell megválasztani, hogy érzékeny legyen az idősor nemlinearitására. Ha a próba eredménye alapján a nullhipotézis fenntartható, akkor az idősor lineárisnak tekinthető a továbbiakban, a választott szignifikancia-szinten. Ellenkező esetben a linearitás elvethető, az eredetileg vizsgált idősor nemlineárisnak tekinthető adott szignifikancia-szinten.

Az idősorok nemlinearitásával kapcsolatos legfontosabb matematikai statisztikai alapfogalmak bevezetését a pótdatsor-módszer szakirodalmi áttekintése, vagyis a pótdatsorok néhány előállítási módjának és a választható próbastatisztikák, valamint az eljárás éghajlatkutatás területén való használatának bemutatása követi. A negyedik fejezetben a pótdatsor-módszer általunk alkalmazott módjának részletes leírására kerül sor, majd az eljárást három éghajlati idősor nemlinearitásának tesztelésére fogjuk felhasználni.

A pótatadsor-módszer használata során számos probléma merülhet föl, mind a pótatadsorok létrehozásával, mind az alkalmas próbastatisztika kiválasztásával kapcsolatban, amelyek megoldására gyakran nem térnek ki kellő részletességgel. Az eljárás negyedik fejezetben leírt alkalmazása során az előforduló problémák ismertetésével, azok áthidalásával is foglalkozunk, továbbá a pótatadsorok generálásának és a próbastatisztika kiszámításának részletes leírására törekszünk.

Az első vizsgálat egy ismert nemlineáris idősorra, egy grönlandi jégfuratból származó oxigénizotóp adatsorra vonatkozik. A második és a harmadik vizsgálat a nemlinearitás szempontjából kérdéses idősorokon, az Észak-atlanti Oszcilláció egyik indexének adatsorán, illetve az Északi Félgömb átlaghőmérsékleti adatsorán végzendő el.

A pótatadsor-módszer alkalmazásához szükséges programok megírásához a Fortran programnyelv 77-es verzióját, az adatsorok és az eredmények megjelenítéséhez, valamint a statisztikai próbák elvégzéséhez pedig az R statisztikai programnyelv 3.1.3 verziószámú felhasználói környezetét használtuk fel.

2. Alapfogalmak

Elsőként a diplomamunkában használt, valószínűségszámítással és matematikai statisztikával kapcsolatos alapfogalmak ismertetésére, majd a stacionárius sztochasztikus folyamatok linearitásának és nemlinearitásának értelmezésére kerül sor.

Legyen X valószínűségi vektorváltozó, T pedig végtelen sok paraméter, például időpontok összessége, ahol $t \in T$ természetes számok. Az X_t valószínűségi változók sokasága sztochasztikus folyamatot alkot, ha tetszőleges rögzített N és $t_1, t_2, \dots, t_N$ esetén az $X_{t_1}, X_{t_2}, \dots, X_{t_N}$ vektorváltozó alábbi együttes eloszlása létezik:

$$F(x_1, x_2, \dots, x_N) = P(X_{t_1} < x_1, X_{t_2} < x_2, \dots, X_{t_N} < x_N), \quad (2.1)$$

(Matyasovszky, 2002).

Egy sztochasztikus folyamatot stacionáriusnak nevezünk, ha az együttes eloszlásfüggvénye nem függ az időtől, azaz tetszőleges τ egész számra teljesül a következő:

$$P(X_{t_1} < x_1, \dots, X_{t_N} < x_N) = P(X_{t_1+\tau} < x_1, \dots, X_{t_N+\tau} < x_N), \quad (2.2)$$

(Matyasovszky, 2002).

Egy valószínűségi vektorváltozó abszolút folytonos, ha megszámlálhatatlanul sok állapotot vehet fel és az eloszlásfüggvénye integrális alakban a következőképpen állítható elő:

$$F(x_1, x_2, \dots, x_N) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_N} f(t_1, t_2, \dots, t_N) dt_1 dt_2 \dots dt_N, \quad (2.3)$$

ahol f az X valószínűségi vektorváltozó együttes sűrűségfüggvénye. Ha (2.1) normális eloszlás, akkor a folyamatot Gauss-folyamatnak nevezik.

Az E várható értékfüggvény és a D^2 szórásnégyzet abszolút folytonos valószínűségi változó esetén az alábbi módon jelölendő:

$$E[X_t] = m_t, \quad (2.4)$$

$$D^2[X_t] = \sigma_t^2. \quad (2.5)$$

A sztochasztikus folyamat kovariancia- és korrelációfüggvénye a következőképpen értelmezendő:

$$\text{kov}[X_{t_i}, X_{t_j}] = E[(X_{t_i} - m_i)(X_{t_j} - m_j)], \quad (2.6)$$

$$\text{kor}[X_{t_i}, X_{t_j}] = \frac{\text{kov}[X_{t_i}, X_{t_j}]}{D[X_{t_i}]D[X_{t_j}]} = \frac{E[(X_{t_i} - m_i)(X_{t_j} - m_j)]}{D[X_{t_i}]D[X_{t_j}]}, \quad (2.7)$$

ahol m_i a t_i időponthoz tartozó várható érték.

Stacionárius sztochasztikus folyamat várható értéke és szórása állandó, kétdimenziós eloszlás, azaz $N = 2$ esetén, pedig a kovariancia (jelölje: B) és a korreláció (jelölje: R) egyváltozósá alakul, azaz két tetszőleges t_1 és t_2 időponthoz tartozó kovariancia a két időpont egymáshoz képesti helyzetétől függ, ahol $k = t_2 - t_1$:

$$\text{kov}[X_{t_1}, X_{t_2}] = B(t_1, t_2) = B(k), \quad (2.8)$$

valamint:

$$\text{kor}[X_{t_1}, X_{t_2}] = \frac{B(t_1, t_2)}{\sigma(t_1)\sigma(t_2)} = \frac{B(k)}{\sigma^2} = R(t_1, t_2) = R(k), \quad (2.9)$$

(Matyasovszky, 2002).

Egy lineáris folyamat, amelynek pontos definícióját most nem ismertetjük, tetszőleges pontossággal közelíthető kellően nagy p -re $AR(p)$, vagyis p -ed rendű autoregresszív folyamattal, amely a következő alakban írható fel:

$$x_t = a_1 x_1 + \dots + a_p x_{t-p} + e_t, \quad (2.10)$$

ahol e_t független és azonos eloszlású változók sorozata (Matyasovszky, 2002).

Egy lineáris folyamatra lineáris becslés adható, vagyis olyan előrejelzés, amely a folyamatot reprezentáló idősor jövőbeli értékeit a megelőző elemek lineáris kombinációjaként adja meg. Egy $X_t, t = 1, 2, \dots$ sztochasztikus folyamat lineáris becslése az alábbi alakú, az időben p lépéssel, X_{t-p} -ig visszalépve:

$$\hat{X}_t = a_1 X_{t-1} + a_2 X_{t-2} + \dots + a_p X_{t-p}, \quad (2.11)$$

(Matyasovszky, 2002), ahol az $a_1, a_2, \dots, a_p$ az autoregresszív együtthatók.

Nemlineáris folyamat esetén X_t a megelőző időpontok nemlineáris függvényeként fejezhető ki, például a következőképpen:

$$X_t = f(X_{t-1}, X_{t-2}, \dots, X_{t-p}) + e_t, \quad (2.12)$$

(Matyasovszky, 2002).

3. Szakirodalmi áttekintés

A pótdadatsor-módszer körülbelül két évtizedes múltra tekint vissza. Az eljárást az elsők között *Theiler et al.* (1992) ismertette, amelynek lényege az alábbiak szerint foglalható össze.

A vizsgálatához kiválasztandó egy stacionárius sztochasztikus folyamatot reprezentáló idősor, amelynek linearitása kérdéses. A hipotézisvizsgálat során a nullhipotézis az idősor linearitása. Az eredeti idősor alapján olyan pótdadatsorok készíthetők, amelyek biztosan lineáris folyamatból származnak, és bizonyos tulajdonságaik megegyeznek az eredeti idősoréval. Nevezetesen, a kovarianciafüggvény és az eloszlásfüggvény. Az eredeti idősorra és mindegyik pótdadatsorra kiszámítandó egy próbastatisztika. Ez a statisztika annál jobb, minél inkább képes az idősor mögött meghúzódó folyamat valamilyen fontos tulajdonságának megragadására. A próbastatisztika valószínűségi eloszlása általában nem ismert, de jól közelíthető a nagyszámú pótdadatsorból számított próbastatisztika értékek empirikus eloszlásfüggvényével. Ez adott szignifikancia-szinthez tartozó kritikus érték meghatározása szempontjából fontos. Ha az idősorunkhoz tartozó próbastatisztika meghaladja a kiválasztott szignifikancia-szinthez tartozó kritikus értéket, akkor abból arra lehet következtetni, hogy a vizsgált idősor nemlineáris folyamatból származik. Egyébként a nullhipotézis fenntartható, azaz lineáris folyamattal jól modellezhető az eredeti idősor (*Theiler et al.*, 1992).

3.1. Pótdadatsorok előállítási módjai

A pótdadatsorok két fő típusa különböztethető meg, a *tipikus* és a *kényszerített* realizációk (*Theiler & Prichard*, 1996). A tipikus realizációk az eredeti idősorhoz lineáris modell, például autoregresszív mozgóátlag modell illesztése után, a modell iterálásával generálандók, többek között a zajtag véletlenszám-generátorral történő változtatásával. A pótdadatsorok tehát a választott modell tipikus realizációi. A pótdadatsorok másik típusa a kényszerített realizációk, amelyek közvetlenül az eredeti idősorból, annak valamilyen átalakításával, például Fourier-transzformációjával kerülnek létrehozásra (*Theiler & Prichard*, 1996). Az elnevezést az indokolja, hogy a pótdadatsorok megőrzik az eredeti idősor lineáris tulajdonságait. Utóbbi nemparaméteres módszer, hiszen a megfigyelésből származó, kérdéses idősorhoz nem szükséges a modellillesztés, így az eljárás nem függ semmilyen modellparamétertől. Ennek következtében előnyösebb a használata, mivel a tipikus realizációknál számos probléma merülhet fel a megfelelő típusú és rendű modell kiválasztásánál, a modell paramétereinek

becslésénél. Kényszerített realizációk esetén a megfigyelésből származó idősor tulajdonságai meghatározzák a folyamatot, az eredeti idősor rekonstruálására nincs szükség (Schreiber & Schmitz, 2000).

Pótadatsorok készítésére számos eljárás létezik. Az alábbiakban öt módszer kerül bemutatásra. A módszerek célja az eredeti idősorral azonos eloszlásfüggvényű és korrelációfüggvényű pótadatsorok létrehozása anélkül, hogy az átalakítások a nemlinearitás téves azonosításához vagy a linearitás hibás állításához vezetnének. Az eljárások többsége azonban vagy egyik feltételt sem, vagy csak az egyik kritériumot teljesíti. Az ismertetendő módszerek alapján megállapítható, hogy az eloszlásfüggvények egyezését könnyebb elérni, mint a korrelációfüggvényekét. Az egyik feltétel teljesítése általában csak a másik rovására történhet.

Az első módszer az eredeti, megfigyelésből származó $x_t, t = 1, \dots, n$ idősor elemeinek véletlenszerű felcserélésével (Random Shuffle) állítja elő a pótadatsorokat. Ezek átlagai és szórásnégyzetei megegyeznek a kiindulási idősoréval, viszont a korrelációfüggvényük megváltozik. A módszer tehát megőrzi az eredeti idősor eloszlásfüggvényét, de a korrelációfüggvényt nem, így ezzel az eljárással kizárólag az a nullhipotézis tesztelhető, hogy az eredeti idősor független és azonos eloszlású (Theiler et al., 1992).

A második eljárás a Fourier-transzformáció módszere (FT), amely a pótadatsorokat inverz Fourier-transzformációval állítja elő. Az eljárást a véletlen fázisok módszerének (Random Phases: RP) is nevezik (Lucio et al., 2012), ugyanis a Fourier-transzformációt alkalmazó pótadatsor-módszerek alapja a fázisok véletlen sorrendben való elrendezése. A nemlinearitás szempontjából kérdéses idősor Fourier-transzformálható, majd a $[-\pi, \pi]$ intervallumon egyenletes eloszlású, független fázisok szimulálандók véletlenszerűen. Ezután a pótadatsorok a szimulált fázisértékek felhasználásával, inverz Fourier-transzformációval hozhatóak létre (Theiler et al., 1992).

A módszer hátránya az, hogy nem megfelelően választott idősor esetén a pótadatsorokban a nullhipotézis téves elvetéséhez, azaz a nemlinearitás detektálásához vezető hamis alacsony és magas frekvenciák jelenhetnek meg. Utóbbit az eredeti idősor végpontjainak eltérő értéke okozhatja. A probléma elkerülhető például az idősor olyan részsorozatának kiválasztásával, amelynek kezdeti és végpontja hasonló értéket vesz fel (Theiler et al., 1992, Mammen et al., 2009).

A létrehozott pótadatsorok az eredetivel megegyező korrelációfüggvénnyel rendelkeznek, a Wiener-Hincsin-tétel alapján (Prichard & Theiler, 1994).

A Fourier-transzformációval létrehozott pótatadsorok normális eloszlásúak. Ennek következtében nullhipotézisként az tesztelhető, hogy az eredeti idősor lineáris Gauss-folyamat realizációja (*Theiler et al.*, 1992).

Ha az eredeti, nemlinearitás szempontjából vizsgált idősor nem normális eloszlású, akkor a Fourier-transzformáció módszerével létrehozott pótatadsorok az eredeti idősortól eltérő eloszlásfüggvénnyel rendelkeznek, amely a nullhipotézis hamis elvetéséhez vezethet. A nullhipotézis elutasításakor ugyanis nem tudni, hogy azt az eloszlásfüggvények különbözősége, vagy az eredeti idősor ténylegesen nemlineáris jellege okozza.

A pótatadsorokat létrehozó módszereket ezért úgy fejlesztették tovább, hogy azokkal olyan stacionárius sztochasztikus folyamatokat reprezentáló idősorok nemlinearitása is tesztelhető legyen, amelyek nem normális eloszlásúak. Ehhez a pótatadsorok megfigyelésből származó idősorral azonos eloszlásúvá transzformálása szükséges.

A harmadik eljárás, az amplitúdók eloszlásának módosításával végzett Fourier-transzformáció (Amplitude Adjusted Fourier Transform: AAFT) már alkalmas az eredeti idősorával megegyező eloszlás létrehozására. Az AAFT-módszerrel tesztelhető legáltalánosabb nullhipotézis a következő: az eredeti idősor egy lineáris Gauss-folyamat realizációja, amely egy monoton nemlineáris transzformáción ment keresztül. A nullhipotézis szerint tehát létezik olyan Gauss-folyamatból származó $y_t, t = 1, \dots, n$ idősor, amelyre teljesül az $x_t = h(y_t)$ egyenlőség, ahol h egy monoton nemlineáris függvény (*Theiler et al.*, 1992). A függvény monotonitása az invertálhatósághoz szükséges. Egy függvény ugyanis akkor invertálható, ha az értelmezési tartományának különböző pontjaiban különböző értékeket vesz fel, azaz injektív. Egy monoton függvény pedig injektív, tehát invertálható. A h függvény invertálhatóságának következtében az $y_t = h^{-1}(x_t)$ egyenlőség is teljesül.

Theiler et al. (1992) a h függvényt „statikusnak” nevezte, amellyel azt fejezte ki, hogy az x_t idősor nem függ a normális eloszlású idősor múltbeli értékeitől. A pótatadsorok a véletlenszerűen generált fázisok felhasználásával, a Gauss-folyamatot reprezentáló y_t idősor alapján állítandók elő inverz Fourier-transzformációval, majd ezek az x_t idősorral azonos eloszlásúvá alakítandók.

Az eredeti idősorhoz többféleképpen rendelhető normális eloszlású adatsor. Az egyik eljárás olyan véletlenszám-generátor alkalmazása, amely normális eloszlású, független számsorozatot állít elő. A normális eloszlásúvá transzformáláshoz az eredeti idősor elemei sorrendbe állítandók. A Gauss-folyamatot reprezentáló adatsor elemei szintén sorba rendezendők, az x_t idősor időbeli fejlődését követő módon. A normális eloszlású idősor

alapján a szimulált fázisok felhasználásával, inverz Fourier-transzformációval pótdat sorok állítandók elő, amelyek a megfigyelésből származó idő sorral azonos eloszlásfüggvényű adatsorokká transzformál andók vissza. Az eredeti idő sor eloszlása azonban nem pontosan ismert, ezért ehhez a h függvény becslése, az empirikus eloszlásfüggvény kerül alkalmazásra. Az így kapott adatsorok szolgálnak pótdat sorokként a hipotézisvizsgálat során (Theiler et al., 1992).

Ha azt feltételezzük, hogy az eredeti idő sorhoz rendelt normális eloszlású adatsor $AR(p)$ -folyamat realizációja, akkor *Autoregressive to Anything* (ARTA) modell is alkalmazható normális eloszlású idő sor korrelációfüggvényének szimulálására (Cario & Nelson, 1996).

Az AAFT-módszer eredményeként kapott pótdat sorokkal kapcsolatban több probléma is felmerülhet. Az eljárás csak végtelen hosszúságú sorozatokra pontos, véges idő sorok esetén az eredeti idő sor és a pótdat sorok korrelációfüggvényei nem egyeznek meg teljesen, ami elsőfajú hibához, azaz a nullhipotézis hibás elvetéséhez vezethet. A módszer alkalmazásánál a normális eloszlásúvá transzformálás jelent gondot, mert ez nem lineáris eljárás, ezért a próba eredménye szennyezett ezzel a transzformációval. A problémák forrása többek között az, hogy a normális eloszlásúvá transzformálás és az eredeti idő sor eloszlására való visszatranszformálás során felhasznált függvények nem pontosan egymás inverzei. Ennek következményeként a normális eloszlású pótdat sorok visszaalakításakor nem érhető el pontosan az eredeti, megfigyelésből származó idő sor eloszlása (Schreiber & Schmitz, 1996, Dolan & Spano, 2001).

Az AAFT-módszer hosszú idő sorokra alkalmazható sikeresebben, valamint akkor, ha az elemek közötti korreláció nem túl nagy (Schreiber & Schmitz, 2000).

Theiler et al. (1992) a fentiekben leírt eljárással egyváltozós pótdat sorokat hozott létre. A módszert Prichard & Theiler (1994) többváltozós pótdat sorok generálása is kiterjesztette.

A felmerülő problémák kezelésére Schreiber & Schmitz (1996) iteratív lépésekkel fejlesztette tovább a pótdat sor-módszert, a negyedik eljárás tehát az AAFT-módszer iteratív változata (Iterative Amplitude Adjusted Fourier Transform: IAAFT), noha már az eljárás kidolgozásakor egyértelművé tették, a véges hosszúságú idő sorok esetén az eloszlásfüggvények, illetve a korrelációfüggvények teljes egyezése egyidejűleg nem érhető el (Schreiber & Schmitz, 1996).

Az IAAFT-eljárás alkalmazásának nem feltétele a h függvény monotonitása. Kiindulópontja az eredeti, megfigyelésből származó idő sor véletlenszerű permutációja,

amelyet kétlépéses, nagy számításigényű iteráció-sorozat követ. Az iteráció addig folytatandó, amíg az eredeti idősor és a pótdatások az eloszlásfüggvényeik és a korrelációfüggvényeik tekintetében kellő pontossággal megközelítik egymást (Schreiber & Schmitz, 1996). Kugiumtzis (1999) az IAAFT-pótdatások két típusát különböztette meg. Az IAAFT-1 pótdatások az eredeti idősorral megegyező eloszlásfüggvénnyel rendelkeznek, azonban a korrelációfüggvényeik között jelentősebb különbségek figyelhetők meg. Az IAAFT-2 pótdatásoknál éppen fordított jelenség tapasztalható, vagyis a korrelációfüggvényeik a megfigyelésből származó idősoréval azonosak, viszont az eloszlásfüggvényeik nagyobb eltéréseket mutatnak. Az IAAFT-1 pótdatások esetén az iterációk számának növelésével egymáshoz nagyon hasonló pótdatások jönnek létre, korrelációfüggvényeik szinte azonos pontossággal közelítik meg az eredeti idősor korrelációfüggvényét, azaz kicsi a szórásuk. A hipotézisvizsgálat során, amennyiben a választott próbat statisztika érzékeny a korrelációra, ez elsőfajú hiba forrása lehet, hiszen tévesen szignifikáns eltérést jelezhet az eredeti idősor és a pótdatások között. A fentiek alapján az IAAFT-2 pótdatások alkalmazása előnyösebbnek tűnhet, azonban a feladattól függően változik, hogy az első vagy a második típushoz tartozó pótdatások használata jobb (Kugiumtzis, 1999).

Az IAAFT-eljárás korrigált változatát, a statikusan transzformált autoregresszív folyamatot (Statically Transformed Autoregressive Process: STAP) Kugiumtzis (2002) dolgozta ki. A módszer olyan pótdatásokat hoz létre, amelyek a tipikus realizációk közé tartoznak. A STAP-módszer kiindulópontja az, hogy bármilyen stacionárius folyamathoz létezik olyan lineáris autoregresszív folyamat, amelynek realizációi az eredetivel azonos eloszlásfüggvénnyel és korrelációfüggvénnyel rendelkező lineáris idősorok, azaz minden idősornak van egy lineáris „másolata”. Utóbbi pedig normális eloszlású folyamat transzformáltja. A transzformációt biztosító függvény monoton és a Theiler et al. (1992) által meghatározott módon statikus, amelynek eredeti alakja polinomokkal közelítendő. A korrelációfüggvény a közelítés alapján, analitikusan kerül kiszámításra. A módszer előnye az elsőfajú hiba valószínűségének csökkenése, hiszen a pótdatások korrelációfüggvényei jobban közelítik az eredeti idősorét (Kugiumtzis, 2002). Az eljárás hátránya az, hogy az előrejelzéshez használt modell illesztéséhez és a korrelációfüggvény analitikus meghatározásához paraméterek szükségesek, amelyek önkényes választás forrásai lehetnek. Az eljárás ezenkívül számításigényes.

A negyedik fejezetben bemutatásra kerülő pótdatás előállítási módszer az eredeti idősor és a pótdatások korrelációfüggvényeinek, illetve eloszlásfüggvényeinek egyezésére

törekszik, miközben az előbb ismertetett módszerekhez képest kevésbé számításigényes, továbbá mellőzi az önkényesen megválasztható paramétereket.

3.2. Hipotézisvizsgálat: a nullhipotézis és a választható próbastatisztikák

A pótdatumsor-módszer alkalmazása során a nullhipotézis az eredeti, megfigyelésből származó idősor linearitása, amely nem utasítható vagy tartható fenn teljes bizonyossággal, csak az elsőfajú hiba valószínűsége adható meg. A nullhipotézis lehet egyszerű vagy összetett. Az egyszerű nullhipotézis szerint az adatsort egy bizonyos, kiválasztott lineáris folyamat generálta. Az összetett nullhipotézis ezzel szemben nem egy kiválasztott folyamatra vonatkozik, hanem egy teljes folyamatsaládra, amely a lineáris adatsort létrehozhatta. Összetett nullhipotézis esetén a folyamatsalád minden tagjának jellemzésére alkalmas próbastatisztika választandó, noha ezáltal lényegtelenné válik, hogy az eredeti idősor a folyamatsalád pontosan melyik tagjának realizációja (*Theiler & Prichard, 1996*).

A pótdatumsor-módszer, mint hipotézisvizsgálati eljárás alkalmazásához bármilyen próbastatisztika alkalmazható, amely egy-egy számmal jellemzi az adatsorokat, valamint alkalmas a linearitás és a nemlinearitás közötti különbség észlelésére, vagyis az eloszlása különbözik az eredeti idősorra és a pótdatumsorokra vonatkozóan. A legoptimálisabb az, ha a próbastatisztika valamilyen, a folyamat szempontjából fizikailag releváns mennyiséget fejez ki. Többek között *Theiler et al. (1992)* és *Maiwald et al. (2008)* ismertetett alkalmas mérőszámokat, amelyek közül néhány az alábbiakban található.

Az egyszerűbb próbastatisztikák közé tartozik egy valószínűségi változó harmadik és negyedik momentum, azaz a *kurtosis* és a *skewness* (*Maiwald et al., 2008*). A két próbastatisztika a vizsgált valószínűségi változó eloszlását veti össze a normális eloszlással. Előbbi azt mutatja meg, hogy a valószínűségi változó sűrűségfüggvénye mennyire tér el a normális eloszlás sűrűségfüggvényétől, azaz mennyire lapult vagy csúcsos a normális eloszláshoz képest. Utóbbi a valószínűségi változó sűrűségfüggvényének ferdeségére utal. A normális eloszlás és bizonyos nem normális eloszlások azonban pontosan megegyező harmadik és negyedik momentummal rendelkeznek, amely azok próbastatisztikaként történő felhasználását problémássá teszi, hiszen másodfajú hibához, azaz a nullhipotézis téves fenntartásához vezethet (*Theiler & Prichard, 1996*).

A legelterjedtebb próbastatisztikák az idősorok kaotikus tulajdonságait jellemző nemlineáris mérőszámok, például a korrelációs dimenzió, a Ljapunov-exponens és a

nemlineáris előrejelzési hiba, vagyis az átlagos négyzetes hiba. A kaotikus, azaz hosszú távon előrejelezhetetlen viselkedés ugyanis nemlineáris folyamatok jelenlétére utalhat, noha a nemlinearitás nem jelent minden esetben kaotikus viselkedést (*Theiler et al.*, 1992).

A korrelációs dimenzió becslésére több eljárás létezik, többek között a Grassberger-Procaccia-Takens eljárás (*Grassberger & Procaccia*, 1983, *Takens*, 1985). A korrelációs dimenzió szignifikánsan különbözik nemlineáris és lineáris idősorok esetén: nemlineáris folyamatokra, kaotikus rendszerekre alacsony értéket vesz fel, ezért alkalmas próbastatisztika a hipotézisvizsgálat során (*Small et al.*, 2001). A hátrányaként említhető, hogy a becslésének bizonyos lépései szubjektív döntéseket igényelnek (*Theiler et al.*, 1992, *Small & Judd*, 1998), ezenkívül a korrelációs dimenziót alkalmazó eljárás rövid és zajos idősorokra megbízhatatlan, a klimatológiai adatsorok pedig gyakran ilyenek (*Grassberger*, 1986. In: *Paluš & Novotná*, 1994).

A Ljapunov-exponens azt fejezi ki, hogy egy kaotikus rendszerben két, egymáshoz közeli helyzetből indított pálya milyen mértékben távolodik el egymástól (*Wolff et al.*, 2004). Próbastatisztikaként azért alkalmas, mert nemlineáris folyamat esetén a Ljapunov-exponens pozitív értékeket vesz fel, a távolodás pedig exponenciális gyorsaságú, míg lineáris folyamatok során a mérőszám negatív, a távolság az exponenciálisnál lassabban nő. Hátránya a zajra való fokozott érzékenysége, amelynek következtében pozitív értéket vehet fel akkor is, ha a folyamat lineáris (*Theiler et al.*, 1992).

A harmadik, röviden ismertetésre kerülő nemlineáris statisztika az átlagos négyzetes hiba (mean squared error: MSE). A kiszámításához *Theiler et al.* (1992) a nemlinearitás tekintetében kérdéses idősort két részre bontotta. Az egyik részhez lineáris előrejelzési modellt illesztett a *k-nearest neighbor* (a továbbiakban: *k* darab legközelebbi szomszéd) módszer felhasználásával. A modellel ezután előrejelzést készítettek az idősor másik részére vonatkozóan. Az előrejelzési hibát a tényleges értékek és a becsült értékek eltérésnégyzetei összegének átlagát képezve számították ki (*Theiler et al.*, 1992).

A lineáris és a nemlineáris idősorokra kiszámított előrejelzési hibaérték szignifikánsan különbözik, ezért alkalmazható próbastatisztikaként. Az idősor két részsorozatának hossza, az előrejelzéshez szükséges időlépcső és a legközelebbi szomszédok számának megválasztása azonban önkényes lehet, gyakran szubjektív elemeken alapulhat.

Mikšovský & Raidl (2006) az átlagos négyzetes hiba négyzetgyökét (root mean squared error: RMSE) alkalmazta próbastatisztikaként. Az RMSE értéke nemlineáris idősorok esetén alacsonyabb, lineáris idősorokat tekintve pedig magasabb. A linearitás nullhipotézisének elutasítására így akkor kerülhet sor, ha a lineáris pótadatsorokra kiszámított RMSE értékek

meghaladják az eredeti idősorra meghatározott értéket (Mikšovsky & Raidl, 2006). Mikšovsky et al. (2008) az RMSE értékek alapján az előrejelzés pontosságát jellemző mérőszámot, a *skill score* értéket választotta próbastatisztikának, amely a Wilks (2006) által meghatározott módon került kiszámításra.

Paluš et al. (1993) próbastatisztikaként két, az információelméletben elterjedt mennyiséget használt fel, a redundanciát (a továbbiakban: általános redundancia) és a lineáris redundanciát. A lineáris redundancia definíciója a kovariancián alapul, így ebben az idősor elemei közötti lineáris függőségek nyilvánulnak meg. Az általános redundancia az eloszlásfüggvényből számítható ki, ezért általánosságban jellemzi az adatsor elemei közötti függőségeket. Az általános és a lineáris redundancia értéke normális eloszlás esetén elméletileg egyenlő. A vizsgálatok során mindkét redundancia értékét meghatározták a linearitása tekintetében kérdéses idősorra, majd az időeltolás függvényeiként ábrázolták azokat. Ezután a függvények alakjának összevetésével minőségi (kvalitatív) elemzést végeztek. A kérdéses idősort akkor tekintették lineárisnak, ha a függvények alakja megegyezett vagy nagy hasonlóságot mutatott. A vizsgálat következő lépése a mennyiségi (kvantitatív) analízis volt, a pótdatsor-módszer végrehajtásával. A hipotézisvizsgálat során a kérdéses idősor, valamint a pótdatsorok átlagos redundancia értékeinek szórását számították ki próbastatisztikaként (Paluš & Novotná, 1994).

Az előbbieken a legelterjedtebb próbastatisztikák kerültek röviden ismertetésre. Ezeken kívül számos egyéb próbastatisztika is létezik, amelyek felhasználhatók a pótdatsor-módszer alkalmazása során.

A diplomamunkában saját vizsgálat keretében bemutatott pótdatsor-módszerhez az átlagos négyzetes hibát választottuk próbastatisztikaként, a becslést az adatsor kettébontása nélkül végezve. A számítás menete a negyedik fejezetben kerül részletesebb ismertetésre.

3.3. A pótdatsor-módszer meteorológiai alkalmazásai

Noha a pótdatsor-módszert számos tudományterületen, például az orvostudományban és a csillagászatban sikeresen használják, addig széleskörű meteorológiai alkalmazása még nem terjedt el. Az 1990-es évektől többek között hőmérsékleti, geopotenciál és légnyomás adatsorokat elemeztek ezzel az eljárással, illetve éghajlati oszcillációs jelenségek, például az El Niño - Déli Oszcillációt reprezentáló Déli Oszcillációs Index (Southern Oscillation Index: SOI) vizsgálatára használták fel. Az alábbiakban olyan, éghajlatkutatással kapcsolatos példák kerülnek bemutatásra, amelyek során a pótdatsor-módszert alkalmazták.

Paluš & Novotná (1994) négy éghajlati adatsort vont vizsgálat alá az általános és a lineáris redundancia kiszámításán alapuló pótdatsor-módszerrel. Huszonkilenc darab pótdatsort képeztek gyors Fourier-transzformáció alkalmazásával, amelynek elvégzéséhez az idősorokat úgy kellett megválasztani, hogy elemszámuk kettő hatványa legyen. Az első két idősor tizennyolc és harminckét évre visszamenőleg tartalmazta az 500 hPa-os izobárszint geopotenciális magasságának prágai és krakkói napi értékeit, míg a másik két idősor prágai napi átlagos léghőmérsékleti és napi felszíni légnyomás értékekből állt, kétszáz évre vonatkozóan. Az elemzések során kvalitatív és kvantitatív vizsgálatot is elvégeztek. A mennyiségi vizsgálat során a linearitás nullhipotézise sem a geopotenciális magasság, sem a hőmérséklet esetében nem bizonyult elvethetőnek. A légnyomásra vonatkozó adatsor esetén, viszont az előzőektől eltérő eredményre jutottak mind a minőségi, mind a mennyiségi analízis keretében. A légnyomás idősorában nemlineáris komponenseket találtak, azaz nem lehetett lineáris stacionárius sztochasztikus folyamat realizációjának tekinteni (*Paluš & Novotná*, 1994).

Paluš (1996) a redundancia meghatározásán alapuló pótdatsor-módszert a prágai és krakkói állomásokról származó légnyomás és hőmérsékleti idősorokból képzett kétváltozós adatsorokra is elvégezte, azaz többváltozós pótdatsorokat, összesen harminc darab kétváltozós pótdatsort állított elő. A hipotézisvizsgálati eljárással kimutatta, hogy a hőmérsékleti idősor, amely egyváltozós vizsgálat során egyértelműen lineárisnak mutatkozott, kétváltozós vizsgálat esetén nem feltétlenül tekinthető annak. A légnyomás idősorát követően a hőmérséklet adatsorában is azonosítottak nemlineáris komponenst (*Paluš*, 1996).

Mikšovský & Raidl (2006) hat egyváltozós és három többváltozós idősorra alkalmazta a pótdatsor-módszert, amelyhez a vizsgálatonként negyvenkilenc darab pótdatsor IAAFT-módszerrel került generálásra. Az egyváltozós idősorok két méteres magasságban mért prágai napi maximum-, minimum- és átlaghőmérsékleti, valamint napi átlagos légnyomásértékeket

tartalmazó adatsorok voltak. A többváltozós pótdatsor-módszer elvégzéséhez az NCEP / NCAR (National Centers for Environmental Prediction: NCEP & National Center for Atmospheric Research: NCAR) tengerszinti átlagos légnyomás és 1000 hPa-os légnyomási szintre vonatkozó átlaghőmérsékleti reanalízis adatsorait használták fel. Az adatsorok harminc és tíz éves időszakra álltak rendelkezésre. A meteorológiai állapothatározókat tekintve a légnyomás adatsoraiban figyeltek meg erőteljesebb nemlinearitást, a hőmérsékleti idősorok esetén csak gyenge nemlinearitást észleltek. Az egyváltozós adatsorok vizsgálatakor csak a hosszabb, harminc éves idősorokban észleltek nemlineáris komponenst, a rövidebb, tíz éves idősorokban nem azonosítottak nemlineáris hatást. A többváltozós vizsgálatok során azonban az összes esetben detektáltak nemlinearitást. Ezt azzal magyarázták, hogy egy változó nem elégséges az éghajlati rendszer nemlineáris jellegének kimutatására, ahogyan arra *Paluš* (1996) vizsgálatai is utaltak. A rendszer jobban leírható több állapothatározó figyelembevételével, ugyanis több információt hordoznak a változók a teljes rendszerről együtt, mint külön-külön. A többváltozós idősorok alkalmazásával tehát a nemlineáris jelleg észlelhetősége növekszik. Összefoglalóan megállapítható, hogy a nemlineáris hatások a hosszabb és többváltozós idősorokban voltak jelentősebbek (*Mikšovský & Raidl*, 2006).

Mikšovský et al. (2008) a pótdatsor-módszert az NCEP / NCAR reanalízis adatsorokra és a Hadley Központ csatolt modellje (Hadley Centre Coupled Model, version 3: HadCM3) futtatásainak eredményeire alkalmazta. A vizsgálat célja a nemlinearitás térbeli és időbeli eloszlásának meghatározása volt. A 850 hPa és az 500 hPa légnyomási szintek geopotenciális magasságából relatív topográfiát számítottak negyven évre visszamenőleg, majd IAAFT-módszerrel pótdatsorokat szimuláltak. Megállapították, hogy a nemlinearitás gyengébb az Egyenlítő térségében, mint a közepes földrajzi szélességeken, továbbá a nemlineáris hatások általában markánsabbnak bizonyultak az Északi Félgömbön, ahol az évszakos különbségeket tekintve a téli félévben erőteljesebb nemlineáris hatások figyelhetők meg, mint nyáron. Ebből arra következtettek, hogy a közepes földrajzi szélességek nemlinearitása és a polárfront aktivitása kapcsolatban állhat egymással. A nemlinearitás legnagyobb értékei Európa, Észak-Amerika és a Csendes-óceán északi medencéje fölötti területekre estek. A szignifikáns nemlinearitásnak a különböző eredetű légtömegek interakciói lehettek a forrásai. A közepes déli földrajzi szélességeken a nemlinearitás maximális értékeit a kontinensek felett detektálták. Az óceánnal borított területek aránya itt sokkal nagyobb, mint a szárazföldre, így átlagosan kisebb mértékű nemlinearitást figyeltek meg az Északi Hemiszférához képest. A Déli Félgömbön a légkörben azonosítható nemlineáris komponensek gyengébb évszakos változékonysággal rendelkeztek. A vizsgálatok elvégzése

során a HadCM3 modellfuttatásokból származó idősorok erőteljesebb nemlinearitást mutattak, mint az NCEP / NCAR reanalízis adatsorai (*Mikšovský et al.*, 2008).

Patil et al. (2000) az előzőekhez képest eltérő módon elemezte a pótdatsor-módszer alkalmazásával a nemlinearitás térbeli eloszlását. A légkör nemlineáris jelenségeinek detektálására az NCEP / NCAR Ausztráliára vonatkozó, illetve globális reanalízis adatsorait, valamint ausztráliai mérőállomások idősorait vizsgálta, úgynevezett reziduális késleltetés diagramok (Residual Delay Map: RDM) szerkesztésével. Az eljárás alapja a következőképpen foglalható össze. Ha egy rövidtávú lineáris előrejelzési modell képes a dinamika, azaz a légköri jelenség időbeli fejlődésének leírására, akkor az adatok, illetve a modelleredmények hibája között nem észlelhető kapcsolat. Az RDM egy kétdimenziós ábra, amelynek egyik tengelyén a meteorológiai állapotváltozó, jelen esetben a tengerszinti légnyomásértékek normalizált értékei, a másik tengelyén a hibák, vagyis a reziduálisok találhatók. Az RDM alapján az idősor akkor tekinthető lineárisnak, ha nincs statisztikai kapcsolat az RDM változói között, tehát az ábrán nem azonosítható a pontpárok alapján semmilyen struktúra. Alacsony korrelációs dimenziójú, azaz nemlineáris folyamat jelenléte esetén azonban valamilyen struktúra, jellegzetesen „V” alakzat rajzolódik ki az RDM-en (*Patil et al.*, 2000).

Sugihara et al. (1999) az RDM elemzésén alapuló módszert magasabb földrajzi szélességeken sikeresen alkalmazta a nemlineáris folyamatok detektálására, azonban kvalitatív vizsgálatokon túl mennyiségi vizsgálatokat nem végzett, azaz pótdatsor-módszert sem használt. *Patil et al.* (2000) AAFT-módszerrel képzett lineáris pótdatsorokat, amelyek RDM ábrái strukturálatlanok. A nemlinearitás azonosításához az egyes adatsorok korrelációfüggvényeinek első két időeltoláshoz tartozó értékéből képeztek statisztikai mérőszámot, amely lineáris idősorok esetén alacsony értékeket, nemlineáris idősorokat tekintve magas értékeket vett fel. Megállapították, hogy nemlineáris jelenségek a közepes földrajzi szélességeken jellemzők, különösen Eurázsia fölött. A trópusok, az Atlanti- és a Csendes-óceán északi medencéje fölött gyengébb nemlinearitás jellemző, vagyis *Mikšovský et al.* (2008) vizsgálatával ellentétes eredményre jutottak. A horizontális térbeli elemzések után vertikális vizsgálatokat is folytattak. Kimutatták, hogy a horizontális vizsgálatokkal összevetve függőleges irányban eltérő földrajzi szélességeken tapasztalható erőteljesebb nemlinearitás. A nemlineáris komponensek térbeli eloszlása korrelált a jet stream átlagos elhelyezkedésével (*Patil et al.*, 2000).

Schreiber & Schmitz (2000) a Tahitin és az ausztráliai Darwinban mért tengerszinti légnyomásértékek különbségeiből képzett SOI-index 1866 és 1994 közötti, havi adatokat tartalmazó idősorának nemlinearitását vizsgálta az IAAFT-módszer alkalmazásával, a hipotézisvizsgálathoz próbastatisztikaként a nemlineáris előrejelzési hibát választva. A próbastatisztikát úgy konstruálták meg, hogy a linearitás nullhipotézise akkor legyen elvethető, ha az előállított kilencvenkilenc darab lineáris pótdatsorra kiszámított előrejelzési hiba értéke meghaladja az eredeti, megfigyelésből származó idősorra kiszámított értéket. A nullhipotézis, miszerint a SOI-index idősor lineáris stacionárius sztochasztikus folyamat realizációja, azonban nem bizonyult elvethetőnek (*Schreiber & Schmitz*, 2000).

A dolgozat negyedik fejezetében saját pótdatsor-módszerünket három klimatológiai adatsorra alkalmazzuk, vizsgálatonként ezer darab pótdatsort generálva.

4. A pótdadatsor-módszer bemutatása és alkalmazása éghajlati idősorokon

4.1. A pótdadatsor-módszer bemutatása

A szakirodalmi áttekintés alapján látható, hogy a nemlinearitás vizsgálatára alkalmas pótdadatsorok létrehozása bonyolult feladat, hiszen azok eloszlásfüggvényének és korrelációfüggvényének is egyeznie kell az eredeti, megfigyelésből származó idősoréval. A pótdadatsorok elkészítéséhez egyszerű módszert választva az eljárás nem lesz sikeres, vagyis téves döntéshez vezethet a hipotézisvizsgálat során.

A módszerek egyik része az eredeti idősorral azonos eloszlásfüggvényű pótdadatsorokat hoz létre, azonban az adatsor elemei függetlenné válnak egymástól. Erre szolgálnak példaképpen azok az eljárások, amelyekkel a pótdadatsorok az eredeti idősor elemeinek véletlenszerű felcserélésével generálандók. A módszerek másik részének alkalmazása esetén a létrehozott pótdadatsorok korrelációfüggvénye megegyezik az eredeti idősoréval, viszont ezek az eljárások Gauss-folyamatot állítanak elő. Erre példa a Fourier-transzformáció módszere. Ráadásul a bemutatott eljárások közül az IAAFT- és a STAP-módszerrel nagy számításigényű iteratív lépések sorozatával készíthetők pótdadatsorok.

További problémát jelent a megfelelő próbastatisztika kiválasztása a hipotézisvizsgálathoz, valamint annak alkalmazása, hiszen kiszámításához gyakran szubjektív, önkényesen meghatározható elemekre van szükség.

Az alábbiakban olyan, a korábbiaknál jóval kisebb számításigényű saját eljárást ismertetünk, amely mindkét kritériumot teljesítő pótdadatsort hoz létre, tehát az eredeti idősorral egyező eloszlásfüggvényt és korrelációfüggvényt állít elő.

A pótdadatsorok létrehozásának, az eljárással kapcsolatban felmerülő problémáknak és azok lehetséges megoldásainak ismertetését a módszer gyakorlati alkalmazása követi. Az eljárás három klimatológiai idősoron: egy grönlandi jégfuratból származó oxigénizotóp adatsoron, az Észak-atlanti Oszcilláció egyik indexének idősorán, illetve az Északi Hemiszféra átlaghőmérsékletére vonatkozó idősoron kerül bemutatásra.

4.1.1. A pótdatások létrehozásához használt apparátus

Az éghajlati idősorok nemlinearitásának vizsgálatára véletlenszerűen szimulált fázisok alkalmazásával létrehozott pótdatásokkal került sor.

Legyen X_t stacionárius sztochasztikus folyamat, amelyre rendelkezésre áll a megfigyelt $x_t, t = 1, \dots, n$ idősor. Az idősor az amplitúdói, fázisai és a körfrekvenciák alapján az alábbi összeggel rekonstruálható, ahol az összegzés az n elemű idősor esetén $\frac{n}{2}$ egész részéig végzendő el:

$$x_t = A_0 + \sum_{j=1}^{\left[\frac{n}{2}\right]} A_j \cos(\omega_j t + \varphi_j). \quad (4.1)$$

$$\text{Az összegben: } A_0 := \bar{x} := \frac{1}{n} \sum_{i=1}^n x_i \text{ az átlag,} \quad (4.2. a)$$

$$A_j := \sqrt{a_j^2 + b_j^2} \text{ az amplitúdó, } a_j, b_j \text{ a Fourier-együtthatók,} \quad (4.2. b)$$

$$\omega_j := \frac{2\pi j}{n} \text{ a körfrekvencia,} \quad (4.2. c)$$

$$\varphi_j \in [-\pi, \pi], \varphi_j := \arctan\left(-\frac{b_j}{a_j}\right) \text{ a fázis.} \quad (4.2. d)$$

Az a_j és b_j Fourier-együtthatók a következőképpen határozhatók meg:

$$a_j := \frac{2}{n} \sum_{t=1}^n x_t \cos(\omega_j t), \quad (4.3)$$

$$b_j := \frac{2}{n} \sum_{t=1}^n x_t \sin(\omega_j t), \quad (4.4)$$

(Matyasovszky, 2002).

Az eredeti, megfigyelésből származó x_t idősor egy y_t pótdatásorának szimulálására a stacionárius sztochasztikus folyamat diszkrét spektruma alábbiakban megadott alakjával kerül sor:

$$y_t = \sum_{j=1}^{\left[\frac{n}{2}\right]} A_j \cos(\omega_j t + \varphi_j), \quad (4.5)$$

(Matyasovszky, 2002).

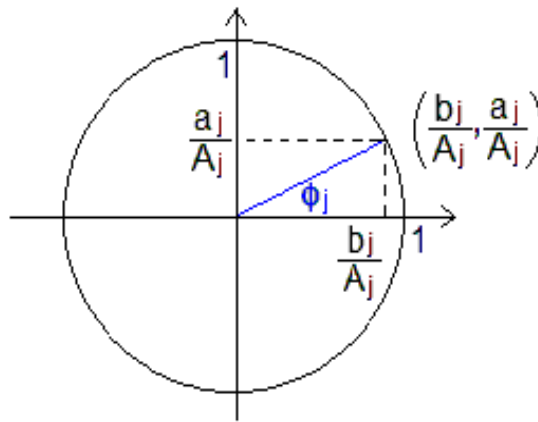
Az eredeti idősorral azonos elemszámú, azaz n elemű pótdatásor előállításához tehát meghatározandók a körfrekvenciák és az amplitúdók, továbbá $\left[\frac{n}{2}\right]$ darab egyenletes eloszlású fázis szimulálandó a $[-\pi, \pi]$ intervallumon.

4.1.2. A fázisok szimulációja

Az eredeti idősor fázisainak a (4.2. d) egyenlőség szerinti kiszámítása során problémát jelent, hogy a φ_j fázisszög értékei $-\pi$ és π között változnak, míg az arkusz tangens függvény $-\frac{\pi}{2}$ és $\frac{\pi}{2}$ közötti értékeket ad vissza. Tetszőleges tangens érték arkusz tangensét képezve, ezért az eredmény szűkebb intervallumon, csak $-\frac{\pi}{2}$ és $\frac{\pi}{2}$ között fog változni. Az arkusz koszinusz függvény, azonban 0 és π közötti értéket ad vissza, vagyis a fázis utóbbi függvény alkalmazásával számítandó ki a következő módon. A (4.2. b) egyenlőség négyzetét véve, majd azt $A_j^2 \neq 0$ -val osztva, az egységsugarú kör egyenletét kapjuk:

$$\left(\frac{a_j}{A_j}\right)^2 + \left(\frac{b_j}{A_j}\right)^2 = 1. \quad (4.6)$$

Ez azt jelenti, hogy a $\left(\frac{b_j}{A_j}, \frac{a_j}{A_j}\right)$ pont az egységsugarú köríven helyezkedik el, amely az 1. ábrán kerül szemléltetésre. A φ_j fázisszög koszinusza megadható b_j és A_j hányadosaként. Ennek következtében a φ_j fázisszög kifejezhető a hányados arkusz koszinuszaként, azaz $\varphi_j = \arccos\left(\frac{b_j}{A_j}\right)$.



1. ábra. A φ_j fázisszög szemléltetése, egységsugarú kör esetén.

A hipotézisvizsgálathoz 1000 darab, az eredeti idősorral megegyező hosszúságú pótadatsort hoztunk létre. A $[-\pi, \pi]$ intervallumon egyenletes eloszlású fázisokat Fortran 77-ben írt, véletlenszám-generátorral kerültek szimulálásra.

4.1.3. A korrelációk becslése

Az amplitúdók az I periodogramból számíthatók ki, amely a fizikai tartalommal nem rendelkező frekvenciák kihagyásával, a pozitív frekvenciákra a következőképpen írható fel:

$$I(\omega_j) = \frac{\hat{B}(0)}{\pi} + \frac{2}{\pi} \sum_{k=1}^{n-1} \hat{B}(k) \cos(\omega_j k), \quad (4.7)$$

ahol $\hat{B}(k)$ a kovariancia becsült értéke, amely az alábbiak szerint számítható ki:

$$\hat{B}(k) = \frac{1}{n} \sum_{i=1}^{n-k} (x_i - \bar{x})(x_{i+k} - \bar{x}) \quad 0 \leq k < n - 1. \quad (4.8)$$

A periodogram az alábbi alakban is megadható:

$$I(\omega_j) = \frac{n}{4\pi} A_j^2. \quad (4.9)$$

A periodogram a kovarianciák becslése után számítható ki, majd ez alapján az amplitúdók a (4.9) átrendezésével a következő módon kerülnek kiszámításra:

$$A_j = \sqrt{\frac{4\pi}{n} I(\omega_j)}, \quad (4.10)$$

(Matyasovszky, 2002).

Az eredeti, megfigyelésből származó x_t idősor és a pótdatsorok eloszlásfüggvényeinek, valamint kovarianciáinak egyezőeknek kell lenniük. A pótdatsorok azonban úgy állítandók elő, hogy lineáris folyamatból származzanak. Az eredeti idősor eloszlásához ezért Gauss-folyamat kerül hozzárendelésre, amelyet egyértelműen meghatároz a várható érték és a korreláció.

A korreláció és a kovariancia között a (2.9) szerinti összefüggés áll fenn, azaz a korreláció a kovarianciával az alábbi módon becsülhető:

$$\hat{R}(k) = \frac{\hat{B}(k)}{\sigma^2} = \frac{\hat{B}(k)}{\hat{B}(0)}. \quad (4.11)$$

Standard normális eloszlás esetén a k időkülönbséghez tartozó korreláció becslése egyenlő a k időeltoláshoz tartozó kovarianciával, tehát:

$$\hat{R}(k) = \hat{B}(k). \quad (4.12)$$

Az eredeti X_t stacionárius sztochasztikus folyamatot a megfigyelésből származó $x_t, t = 1, \dots, n$ idősor reprezentálja. A k időeltoláshoz tartozó korrelációt jelölje $r_X(k)$, azaz $\text{kor}[X_t, X_{t+k}] = r_X(k)$. Az idősorhoz definiálandó egy Φ eloszlásfüggvényű $(Y_t)_{1 \leq t \leq n} \in \mathbb{R}$ lineáris stacionárius Gauss-folyamat, amelynek várható értéke nulla és a szórása egy. A k időeltoláshoz tartozó korrelációját jelölje $r_Y(k)$, tehát $\text{kor}[Y_t, Y_{t+k}] = r_Y(k)$.

Az $F(X_t) = \Phi(Y_t)$ egyenlőségből kiindulva az X_t és az Y_t között a következő összefüggések állnak fenn:

$$X_t = F^{-1}(\Phi(Y_t)), \quad (4.13)$$

$$Y_t = \Phi^{-1}(F(X_t)), \quad (4.14)$$

ahol F az X_t eloszlásfüggvénye. Mindez azért igaz, mert minden valószínűségi változó tetszőleges eloszlásúvá alakítható át, amit az $F^{-1}(\Phi(.))$ transzformáció biztosít. Az $r_X(k)$ korrelációra ekkor az alábbi teljesül:

$$r_X(k) = \text{kor}[X_t, X_{t+k}] = \text{kor}[F^{-1}(\Phi(Y_t)), F^{-1}(\Phi(Y_{t+k}))], \quad (4.15)$$

(Cario & Nelson, 1996).

A kulcskérdés az, hogy mi lesz az Y_t korrelációfüggvénye, ami kielégíti a (4.15) egyenletet. A saját módszer alkalmazása során az Y_t korrelációfüggvényét a Spearman-féle rangkorreláció felhasználásával határozzuk meg.

A k időeltolás mellett az X_t folyamat Spearman-féle rangkorrelációja az alábbi mennyiség:

$$\varrho_X(k) = \text{kor}[F(X_t), F(X_{t+k})]. \quad (4.16)$$

A rangkorreláció abban tér el a korrelációtól, hogy előbbi a monoton kapcsolat szorosságát, míg utóbbi a lineáris kapcsolat szorosságát jellemzi. Az eloszlásfüggvény monoton, ezért az F^{-1} és a Φ^{-1} inverz függvények is monoton transzformációk. A Spearman-féle rangkorreláció monoton transzformációra nézve invariáns, ezért az X_t és az Y_t folyamatok Spearman-féle rangkorrelációja egyenlő, azaz:

$$\varrho_X(k) = \text{kor}[F(X_t), F(X_{t+k})] = \text{kor}[\Phi(Y_t), \Phi(Y_{t+k})] = \varrho_Y(k), \quad (4.17)$$

(Boudt et al., 2012). Az X_t folyamatra rendelkezünk adatsorral, így a rangkorreláció becsülhető X_t empirikus eloszlásfüggvénye ismeretében a következőképpen:

$$\varrho_Y(k) = \frac{\sum_{i=1}^{n-k} (F_{1,n-k}(x_i) - \overline{F_{1,n-k}})(F_{1+k,n}(x_{i+k}) - \overline{F_{1+k,n}})}{\sqrt{\sum_{i=1}^{n-k} (F_{1,n-k}(x_i) - \overline{F_{1,n-k}})^2 \sum_{i=1}^{n-k} (F_{1+k,n}(x_{i+k}) - \overline{F_{1+k,n}})^2}}, \quad (4.18)$$

ahol $F_{u,v}$ az $x_u, x_{u+1}, \dots, x_v$ empirikus eloszlásfüggvénye, $\overline{F_{u,v}}$ az $F_{u,v}(x_u), F_{u,v}(x_{u+1}), \dots, F_{u,v}(x_v)$ értékek átlaga (Boudt et al., 2012).

Az Y_t Gauss-folyamat, ezért az $r_Y(k)$ korreláció és a $\varrho_Y(k)$ Spearman-féle rangkorreláció között az alábbi összefüggés áll fenn:

$$r_Y(k) = 2 \sin\left(\frac{\pi}{6} \varrho_Y(k)\right). \quad (4.19)$$

Az $r_Y(k)$ korreláció kiszámításához tehát Y_t Spearman-féle rangkorrelációjának becslése szükséges. Az Y_t folyamat korrelációfüggvénye a (4.19) egyenlet módosított változatával, az alábbi módon határozandó meg:

$$r_Y(k) = \left(1 - \frac{k}{n}\right) 2 \sin\left(\frac{\pi}{6} \varrho_Y(k)\right), \quad k = 0, 1, \dots, n-1. \quad (4.20)$$

A kifejezésben az $\left(1 - \frac{k}{n}\right)$ tag azért szükséges, mert a (4.8) egyenletben az összegzés előtt $\left(\frac{1}{n}\right)$ szerepel osztóként, nem pedig $\left(\frac{1}{n-k}\right)$ (Matyasovszky, 2002).

Az eredeti $x_t, t = 1, \dots, n$ idősorhoz rendelt lineáris Gauss-folyamat periodogramja a (4.7) egyenlethez képest az alábbiak szerint módosul:

$$I(\omega_j) = \frac{1}{\pi} + \frac{2}{\pi} \sum_{k=1}^{n-1} r_Y(k) \cos(\omega_j k). \quad (4.21)$$

A korrelációk kiszámítása után egy korrekció szükséges, ugyanis a (4.21) periodogram felvehet negatív értéket, noha ennek nem volna szabad előfordulnia, hiszen a (4.10) egyenlőség miatt a periodogram nem lehet negatív. Ennek az az oka, hogy az (i, j) -ediként $r_Y(|i - j|)$ elemet tartalmazó $\underline{R}_Y$ mátrix általában nem pozitív definit, mert e korrelációk a (4.20) egyenletből számolódnak, hiszen az Y_t folyamatra nem áll rendelkezésünkre idősor. A probléma megoldása érdekében $\underline{R}_Y$ mátrixhoz hozzáadandó az $\varepsilon \underline{I}$ kifejezés, ahol ε kicsiny pozitív szám, míg $\underline{I}$ egységmátrix. Ez az egyszerű művelet biztosítja a módosított korreláció mátrix pozitív definit voltát. A korrekcióval a periodogram értéke is módosul, ezért ε kifejezetten kis értékűnek választandó. A gyakorlatban előforduló adatsorok többsége esetén ez teljesíthető.

A pótatadatok előállításához felhasználandó korrelációk, amelyeket jelöljön $r_Y'(k)$, tehát az alábbi formula alapján kerülnek kiszámításra:

$$r_Y'(k) = \frac{r_Y(k)}{1 + \varepsilon} \quad k = 1, \dots, n-1. \quad (4.22)$$

A 4.2. fejezetbeli példák esetén ε tízezred nagyságrendű volt, tehát a periodogram értékei csak minimálisan változtak.

A periodogram az $r_Y'(k)$ korrigált korrelációk becsült értékeinek felhasználásával, a (4.21) egyenlőség alapján számítandó ki, amelyben $r_Y(k)$ helyére $r_Y'(k)$ kerül.

A szimulált fázisok és a (4.10) egyenlőséggel meghatározott amplitúdók alapján a megfigyelt idősorok ezer darab pótdatsorát hoztuk létre, a (4.5) egyenlőség szerint.

Az előállított pótdatsorok normális eloszlásúak voltak, ezért vissza kellett transzformálni azokat az eredeti idősor eloszlására a (4.13) egyenlőség segítségével, amelyben F helyett az empirikus eloszlásfüggvény került alkalmazásra.

4.1.4. A próbastatisztika megválasztása

Próbastatisztikaként az átlagos négyzetes hibát választottuk, amelynek definíciója a következő:

$$MSE = \frac{1}{n-d} \sum_{i=d+1}^n (\hat{x}_i - x_i)^2, \quad (4.23)$$

ahol $\hat{x}_i$ az idősor i -edik tagjának becslése, míg a d beágyazási dimenzió lényege néhány sorral lejjebb fog kiderülni.

A számítás első lépése annak meghatározása, hogy az x_t idősorra adott $\hat{x}_t$ becslés milyen módszerrel történjék. A k darab legközelebbi szomszéd módszert alkalmazva a becslés a k darab legközelebbi társ átlaga alapján határozandó meg a következőképpen.

Rögzítsünk egy $d \geq 1$ természetes számot, a beágyazási dimenziót, és ezzel az alábbi vektort:

$$(x_{i-d}, \dots, x_{i-2}, x_{i-1}) \in \mathbb{R}^d. \quad (4.24)$$

Minden j -re elkészítendő $d + 1$ -től n -ig az alábbi vektor:

$$(x_{j-d}, \dots, x_{j-2}, x_{j-1}) \in \mathbb{R}^d. \quad (4.25)$$

Ezután kiszámítandó a (4.24) és a (4.25) vektorok távolsága, valamilyen távolságképző szabály szerint. A számítás során az egyes norma, vagyis az abszolút érték szerinti távolságot, és az euklidészi távolságnégyzetet választottuk, azaz:

$$\sum_{m=1}^d |x_{i-m} - x_{j-m}|, \quad (4.26)$$

amely az egyes norma szerinti távolság, illetve:

$$\sum_{m=1}^d (x_{i-m} - x_{j-m})^2, \quad (4.27)$$

amely az euklidészi távolságnégyzet, azonban a minimalizálás során azonos eredményt ad (4.27) négyzetgyökével, vagyis az euklidészi távolsággal, hiszen a két függvény minimumhelyei ugyanazok.

A távolságértékek növekvő sorba rendezése után vesszük az első k számút, persze a legkisebb elem kirekesztésével, hiszen az nulla, mert az adott elem önmagától vett távolságát jelenti.

A becslés a következő:

$$\hat{x}_i = \frac{1}{k} \sum_{m=1}^k x_{j(m)}, \quad (4.28)$$

ahol $x_{j(m)}$ a (4.24) vektorhoz tartozó m -edik legközelebbi (4.25) vektorhoz tartozó x_j .

A számítások során a k értékét ötnek, tíznek és tizenötnek választottuk. A fenti eljárással elkészített $\hat{x}_t$ becslés használandó fel az MSE értékek (4.23) képlet szerinti meghatározásakor.

4.1.5. Hipotézisvizsgálat

A hipotézisvizsgálat során nullhipotézisként az eredeti X_t folyamat linearitása választandó. Az X_t folyamat akkor tekinthető nemlineárisnak, ha a pótdatsorokra kiszámított MSE értékek meghaladják a megfigyelésből származó idősorra meghatározott MSE értéket. Ennek oka a következő: a (4.23) képletben a becslésre $k > 1$ darab legközelebbi szomszéd alapján átlagolva kerül sor. Ezzel az eljárással a nemlineáris folyamatok pontosabban becsülhetők, mint a lineárisak, hiszen a pontthalmazra jobban illeszkedik. Következmenyképpen, ha az eredeti folyamat nemlineáris, akkor az erre kiszámított MSE értéknek kisebbnek kell lennie a pótdatsorok MSE értékeinél, amelyeket lineáris folyamat realizációiként hoztunk létre.

Az MSE átlagos négyzetes hibaérték egy ismeretlen eloszlású valószínűségi változó. Nagyszámú pótdatsor azért szimulálandó, hogy empirikus eloszlásfüggvénnyel helyettesítjük a valószínű eloszlást. A kritikus érték ezek alapján történő meghatározása után dönthetünk a nullhipotézis fenntartásáról vagy elvetéséről. Egyoldalú, nemparaméteres próba hajtandó tehát végre, amely azt jelenti, hogy a nullhipotézis csak akkor utasítható el, ha az eredeti adatsorra kiszámított MSE érték egy irányban tér el a pótdatsorokra meghatározott MSE értékektől.

A szignifikancia-szint megállapítható az alapján, hogy hány pótdatsorunk van és azok közül hány esetben nagyobb a próbastatisztika értéke az eredeti idősorra kiszámított mérőszámnál. A kapott érték lesz az a szignifikancia-szint, amelyen a linearitás nullhipotézise elutasítható. A saját vizsgálatok során külön-külön 1000 darab pótdatsort szimuláltunk, ami azt jelenti, hogyha az összes pótdatsor MSE értéke meghaladja az eredeti idősorra kiszámított próbastatisztika értékét, akkor az elsőfajú hiba, vagyis a nullhipotézis téves elvetésének valószínűsége 0,001, azaz 0,01 százalék.

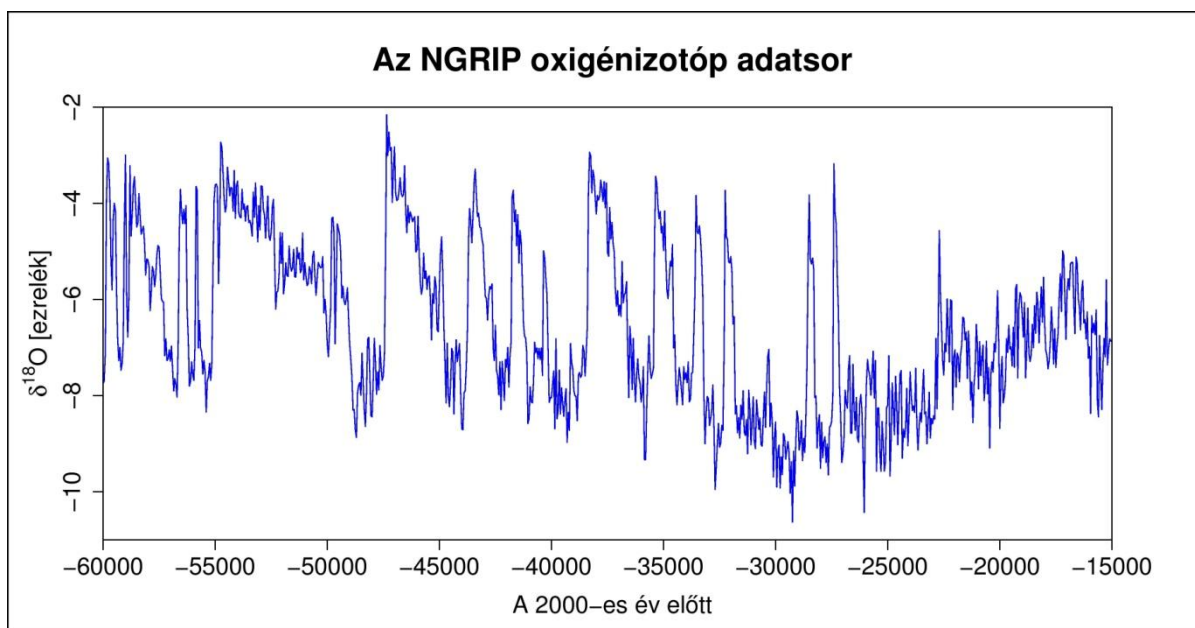
4.2. Grönlandi jégfuratból származó oxigénizotóp idősor

4.2.1. Az NGRIP oxigénizotóp adatsor

A saját módszer keretében elsőként vizsgált adatsor a Észak-grönlandi Jégfurat Projekt (North Greenland Ice Core Project: NGRIP) 18 és 16 tömegszámú stabil oxigénizotóp arányából képzett $\delta^{18}\text{O}$ idősora (a továbbiakban: NGRIP idősor), amelyet az 1990-es évek végén és a 2000-es évek elején mintavételezett Grönland északi, belső területéről származó jégfurat elemzése alapján állítottak össze ötven évenkénti átlagolással, több mint 120 ezer évre visszamenőlegesen (*North Greenland Ice Core Project members, 2004*).

Az ezrelékben kifejezett $\delta^{18}\text{O}$ arány a mintavételezett víz $^{18}\text{O}/^{16}\text{O}$ izotóparánya és az $^{18}\text{O}/^{16}\text{O}$ izotóparány előre meghatározott referenciaértéke különbségének, valamint az $^{18}\text{O}/^{16}\text{O}$ izotóparány referenciaértékének hányadosa (*Dansgaard, 1964*). Az NGRIP idősor esetében referenciaértékként a Nemzetközi Atomenergia-ügynökség által meghatározott szabványértéket (Vienna Standard Mean Ocean Water: VSMOW) alkalmazták (*North Greenland Ice Core Project members, 2004*).

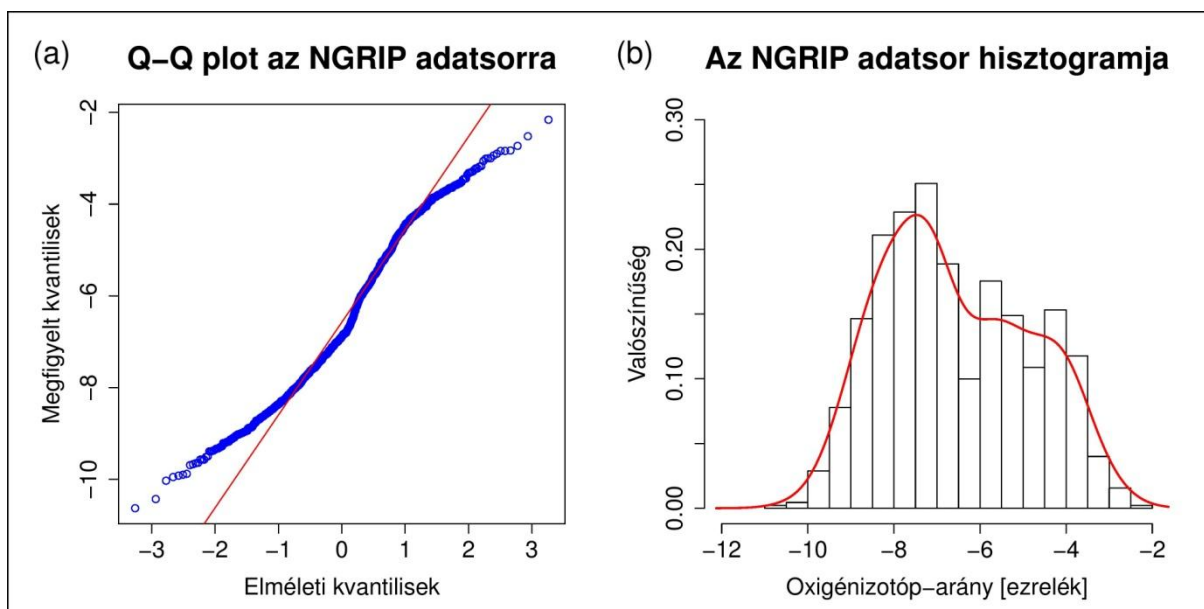
Az NGRIP adatsor 2000-hez képest 60 ezer és 15 ezer év közé eső időszakának nemlinearitása került vizsgálatra. Az idősor az 2. ábrán látható. A $\delta^{18}\text{O}$ izotóparány jó indikátora a hőmérséklet változásának, mivel az izotóparány és a hőmérséklet között szinte lineáris kapcsolat áll fenn. Az izotóparány növekedése a hőmérséklet emelkedését jelenti, míg ellenkező esetben a hőmérséklet csökkenésére utal (*Jouzel et al., 1994*).



2. ábra. Az NGRIP jégfuratból származó oxigénizotóp idősor a 60 ezer és 15 ezer évvel ezelőtti időszakra vonatkozóan (*North Greenland Ice Core Project members, 2004*).

Az adatsorban megfigyelhető a $\delta^{18}\text{O}$ arány nagymértékű fluktuációja körülbelül 60 ezer és 25 ezer év között. A változékonyság azonban a pleisztocén kor végére mérséklődött. Az adatsor alapján Dansgaard-Oeschger (DO) események azonosíthatók, vagyis gyors, pár évtized alatt lezajló felmelegedési, majd lassabb, néhány évszázad folyamán bekövetkező lehűlési periódusok (Dansgaard et al., 1993). A vizsgálatok kimutatták, hogy a DO-események jelentette hirtelen éghajlatváltozás nagy valószínűség szerint nemlineáris folyamat következménye, így az adatsor nemlineárisnak tekinthető (Braun, 2009). Az idősor tehát a nemlinearitásának ismeretében alkalmas a pótdatsor-módszerünk ellenőrzésére.

Az oxigénizotóp idősor normalitása először grafikusan kerül vizsgálatra, kétdimenziós kvantilis-kvantilis ábra (Q-Q plot) készítésével. A kvantilisok lényegében a gyakoriság vizsgálatához kijelölt osztópontokat jelentik. Például azt a kvantilist, amely két egyenlő gyakoriságú részre osztja a megfigyelt értékeket, mediánnak nevezik. A 3. ábra (a) részének x tengelyén a feltételezett normális eloszlás kvantilisei, tehát az elméleti kvantilisok, míg az y tengelyén az eredeti adatsor sorba rendezett elemei, azaz a megfigyelt kvantilisok találhatók. Ha az idősor normális eloszlású lenne, akkor az értékei az ábrán pirossal jelzett egyenesre illeszkednének. A 3. ábra (b) részén az oxigénizotóp idősor hisztogramja látható a sűrűségfüggvényével. A sűrűségfüggvény aszimmetrikus, a ferdeségét (skewness) kifejező együtttható értéke 0,256.



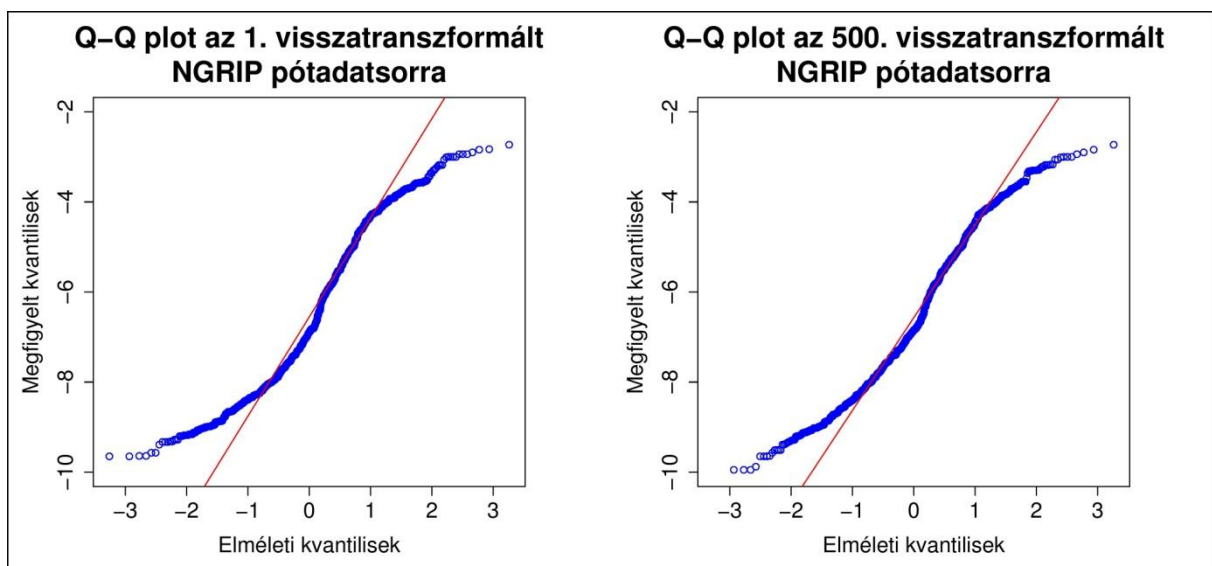
3. ábra. (a) Q-Q plot a grönlandi jégfuratból származó oxigénizotóp adatsor normalitásvizsgálatához, (b) az oxigénizotóp adatsor hisztogramja, a sűrűségfüggvényt piros görbe jelöli.

Az adatsor nem normális eloszlása statisztikai tesztel, Kolmogorov-Szmirnov-próbával, illetve Shapiro-Wilk próbával is ellenőrzésre került. A hipotézisvizsgálat során nullhipotézisként az adatsor normális eloszlása választandó. A Kolmogorov-Szmirnov próba esetén $p = 7,8 \cdot 10^{-6}$, a Shapiro-Wilk próba eredményeképpen, pedig $p = 2,2 \cdot 10^{-12}$ p -érték adódott. A nullhipotézis, tehát az $\alpha < 0,01$ százalékhöz tartozó szignifikancia-szint esetén is elvethető, hiszen $\alpha > p$, vagyis az NGRIP adatsor abszolút nem tekinthető normális eloszlásúnak.

4.2.2. A pótdatsor-módszer alkalmazása az NGRIP adatsorra

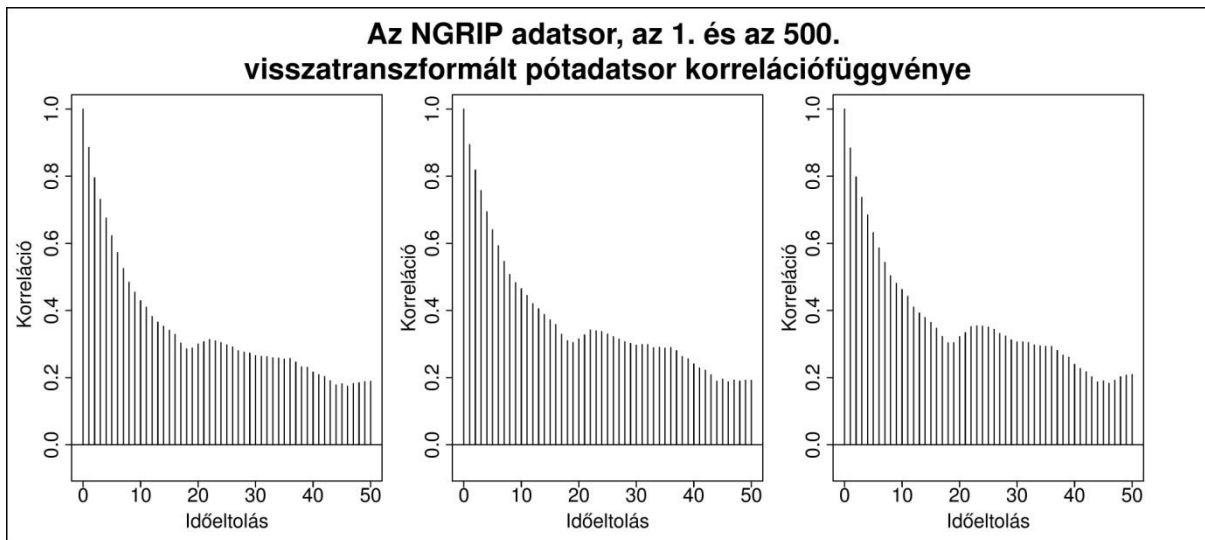
Az idősor nemlinearitásának vizsgálatához 1000 darab, az eredeti idősorral megegyezően 901 elemet tartalmazó pótdatsor került szimulálásra, amelyhez 450 500 darab egyenletes eloszlású fázist hoztunk létre véletlenszám-generátorral. Az idősor eredeti eloszlásához Gauss-folyamat került hozzárendelésre, amelyben a korrelációk becslése Spearman-rangkorrelációval történt. Az amplitúdókat tehát nem az eredeti idősor alapján, hanem az így meghatározott korrelációk felhasználásával, a periodogramból számítottuk ki.

A pótdatsorokat a körfrekvenciák, az amplitúdók és a szimulált fázisok alapján a (4.5) egyenlőség szerint állítottuk elő. A generált adatsorok normális eloszlásúak, azonban a pótdatsor-módszer alkalmazásához az eredeti idősor eloszlásával egyező eloszlásúaknak kell lenniük. Normális eloszlásuk miatt, tehát a pótdatsorok visszatranszformálandók az NGRIP idősor eredeti eloszlására. Szemléltetésére a 4. ábrán az első és az ötszázadik visszatranszformált pótdatsor Q-Q plotja található.



4. ábra. Q-Q plot az NGRIP adatsorral megegyező eloszlású első és ötszázadik visszatranszformált pótdatsorra.

Az eredeti idősor és a pótatadsorok korrelációfüggvényeinek egyezése szintén kvalitatív módon vizsgálendő. Az 5. ábrán található az eredeti idősor, illetve az első és az ötszázadik visszatranszformált pótatadsor első ötven időeltolásra becsült korrelációfüggvénye. Láthatóan a korrelációfüggvények hasonló menettel rendelkeznek.



5. ábra. Az NGRIP idősor (balra), az első visszatranszformált pótatadsor (középen) és az ötszázadik visszatranszformált pótatadsor (jobbra) korrelációfüggvényének becslése, az első ötven időeltolásra. A vizsgált pótatadsorok korrelációfüggvényei eredeti idősorétól való átlagos eltérése: 0,026.

A hipotézisvizsgálat utolsó előtti lépéseként a linearitásról való döntéshez kiszámításra került az 1001 darab adatsorra az MSE próbastatisztika, vagyis az összes transzformált pótatadsorra (jelölje: $MSE_{pót}$) és az eredeti idősorra (jelölje: MSE_{NGRIP}) is. A (4.23) képletben található becslésnél a beágyazási dimenzió értéke három, ugyanis ha igaz a nullhipotézis, miszerint az NGRIP idősor jól modellezhető lineáris autoregresszív folyamattal, akkor harmadrendű autoregresszív, azaz $AR(3)$ -modell adódott optimálisnak az Akaike-féle információs kritérium (Akaike information criterion: AIC) alapján (Akaike, 1974).

Az MSE értékek eloszlása azonban nem ismert, ezért empirikus eloszlásfüggvényük alapján került sor a szignifikancia-szint meghatározására, amelyen a linearitás nullhipotézise elvethető.

4.2.3. Az NGRIP adatsorra vonatkozó eredmények

A várakozásnak megfelelően az oxigénizotóp idősor linearitásának nullhipotézise nagy bizonyossággal vethető el.

Az MSE értékeket hat lehetséges módon határoztuk meg. A kiszámítására használt (4.23) képletben az egyes adatsorok becslését háromféleképpen, az öt, a tíz és a tizenöt legközelebbi társ értékének átlagát képezve készítettük el. A legközelebbi társakat pedig két eljárással kerestük: az abszolút érték, illetve az euklidészi távolság alapján. A legkedvezőbb eredmény a legközelebbi társak számát ötnek választva, az euklidészi távolság felhasználásával adódott. Ez alapján annak valószínűsége, hogy elvetjük a nullhipotézist, holott igaz: 0,3 százalék. Az idősor, tehát nem tekinthető lineárisnak. Az elsőfajú hiba valószínűsége, azaz a nullhipotézis helyes volta ellenére történő elvetése, azonban a legkedvezőtlenebb esetben, tizenöt legközelebbi társ és euklidészi távolság célfüggvény esetén is alig haladja meg a hét százalékot. A hipotézisvizsgálat eredményeit összefoglalóan az 1. táblázat tartalmazza.

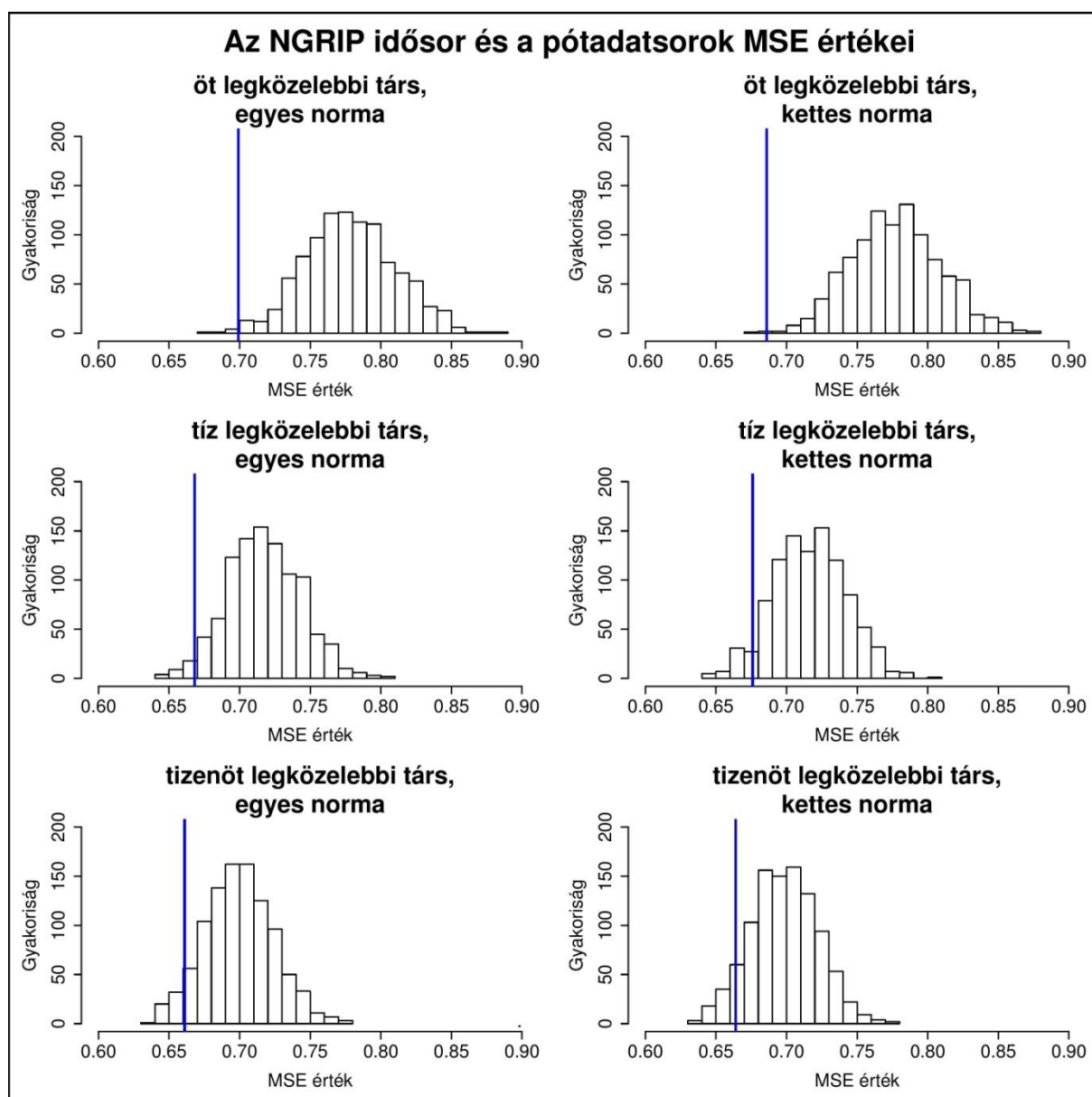
Az NGRIP idősor: eredmények				
legközelebbi társak száma (k)	esetszám: $MSE_{NGRIP} < MSE_{pót}$		az elsőfajú hiba valószínűsége (α)	
	célfüggvény: abszolút érték	célfüggvény: euklidészi távolság	célfüggvény: abszolút érték	célfüggvény: euklidészi távolság
5	995	997	0,5 %	0,3 %
10	977	940	2,3 %	6,0 %
15	944	928	5,6 %	7,2 %

1. táblázat. Az esetek száma, amikor az eredeti idősorra kiszámított MSE értéket meghaladta a pótdatsorokra kiszámított MSE érték, valamint az elsőfajú hiba valószínűsége az NGRIP adatsor nemlinearitásának pótdatsor-módszerrel való vizsgálata során.

A hipotézisvizsgálat alapján megállapítható, hogy a k darab legközelebbi társ számának növelésével romlott a szignifikancia-szint. Ennek lehetséges oka a következő: az MSE értékek kiszámítása során a feltételes várható érték kerül becslésre, amely a várható érték az idősor korábbi elemei ismeretében. A k legközelebbi társ a feltételes várható érték „durva” becslése, amelynek értékét növelve a feltételes várható érték egyre inkább megközelíti a várható értéket, ami lineáris és nemlineáris esetben nem különböztethető meg egymástól.

Az MSE értékek eloszlása a 6. ábrán került szemléltetésre. Az ábrán is látható, hogy a hipotézisvizsgálat során a hibás döntés hozatalának valószínűsége a második esetben a legkisebb, vagyis az MSE értéket az euklidészi távolság minimalizálásával, az öt legközelebbi társ alapján számítva, hiszen ekkor az eredeti idősorra kiszámított MSE értéket szinte minden alkalommal meghaladták a pótdatsorokra meghatározott MSE értékek. Megfigyelhető, hogy a k darab legközelebbi szomszéd számát növelve az elsőfajú hiba valószínűsége egyre nőtt.

A pótdatsor-módszer fentiekben leírt módon való alkalmazásával tehát igazolható az NGRIP idősor nemlinearitása.

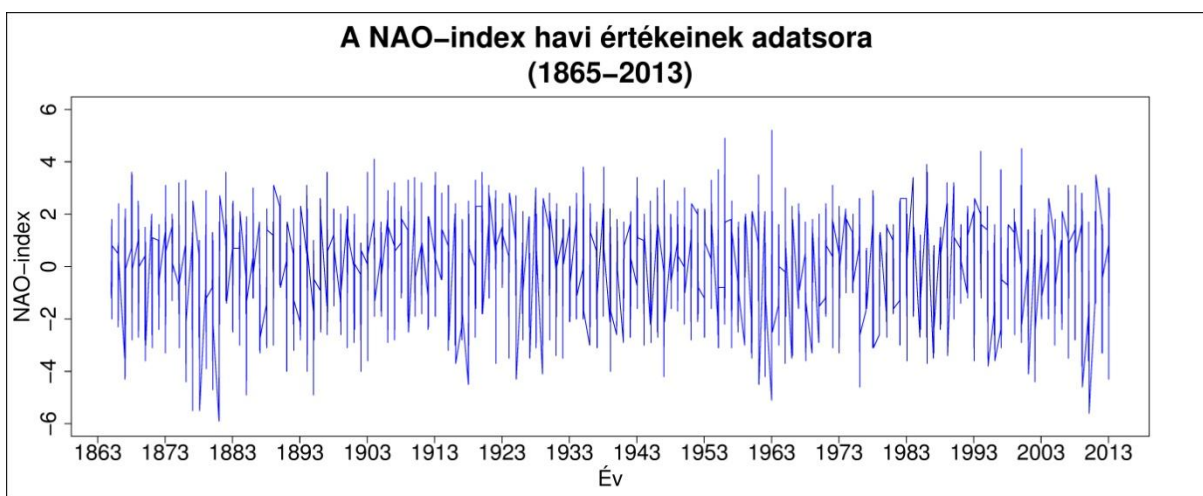


6. ábra. Az eredeti NGRIP idősor és az ezer darab pótdatsor MSE értékeinek eloszlása. Előbbit kék függőleges vonal jelöli.

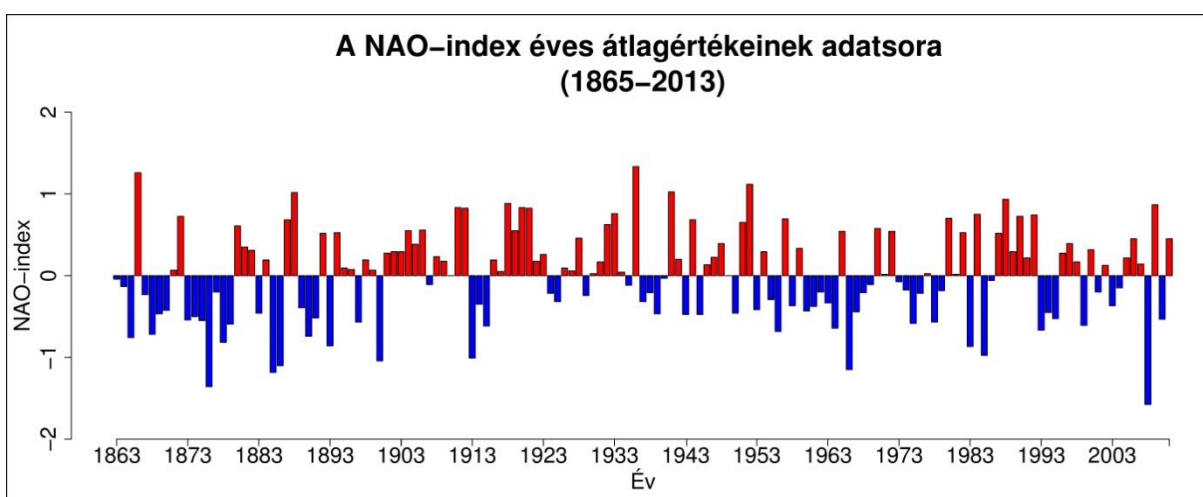
4.3. Az Észak-atlanti Oszcilláció indexének időszora

4.3.1. A NAO-index adatsora

A második, nemlinearitás szempontjából vizsgált időszor a Nemzeti Légkörkutató Központ (National Center for Atmospheric Research: NCAR) Észak-atlanti Oszcillációt (North Atlantic Oscillation: NAO) jellemző indexének 1788 elemű adatsora (a továbbiakban: NAO-index adatsor), amely 1865 januárja és 2013 decembere között havonta számított adatokat tartalmaz. Az időszor a 7. ábrán kerül bemutatásra. A 8. ábrán pedig a havi adatokból képzett éves átlagértékek találhatók.



7. ábra. A NAO-index időszora 1865 és 2013 között számított havi értékek alapján (Hurrell, James & National Center for Atmospheric Research Staff (Eds)., utolsó módosítás: 2014.09.05).

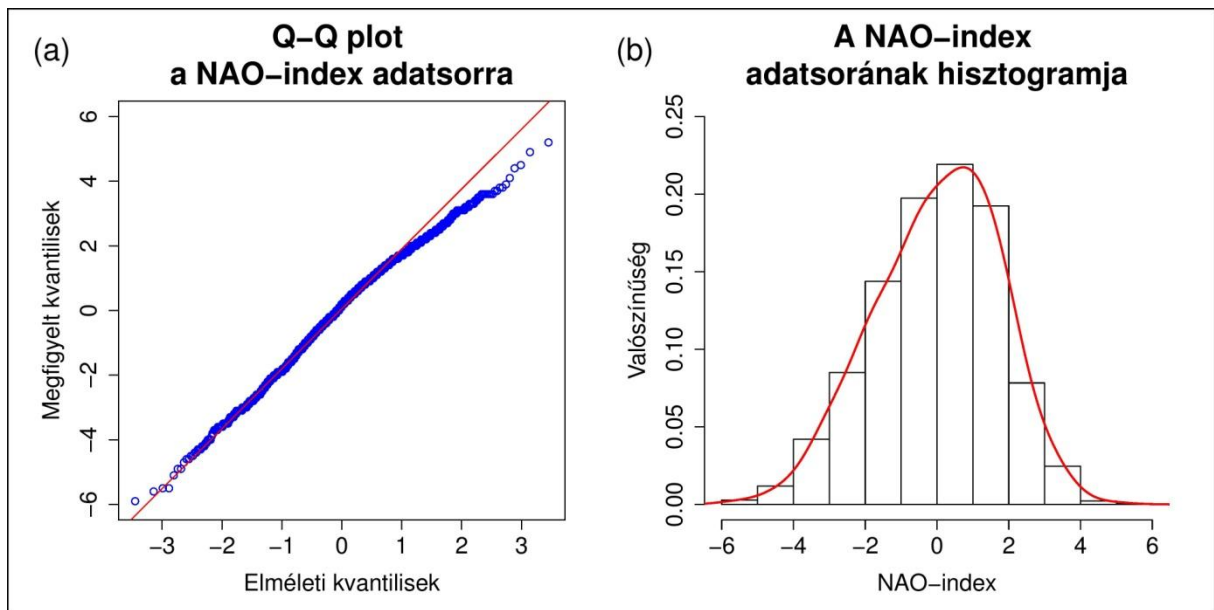


8. ábra. A NAO-index 1865 és 2013 közötti havi időszorából képzett éves átlagértékei. A pozitív értékek pirossal, a negatív értékek kékkel kerültek megjelenítésre (Hurrell, James & National Center for Atmospheric Research Staff (Eds)., utolsó módosítás: 2014.09.05).

A NAO-index az Atlanti-óceán medencéjének északi részén, Izland közelében lévő alacsony légnyomású és az Azori-szigetek térségében megfigyelhető magas légnyomású központ egyidejű változását kifejező mérőszám, amely nem rendelkezik általános definícióval, számítására több eljárás létezik. A NAO-indexek két alapvető típusa különböztethető meg, a főkomponens alapú (principal based) és az állomás alapú (station based) indexek. Az idősorok elérhetőek napi, havi, évszakos és éves változatban is. A pótdatumsor-módszer egy állomás alapú, havi adatokat tartalmazó indexen került alkalmazásra. Az indexet a portugáliai Lisszabonban, valamint az izlandi Stykkishólmurban és Reykjavíkban mért normalizált tengerszinti légnyomásértékek szerint határozták meg: a normalizált értékek előállításához kiszámították az 1864 és 1983 közötti időszak átlagát és szórását, majd az átlagot kivonták a teljes időszakra vonatkozó légnyomás adatsorból, végül a kapott értékeket osztották a szórással. Az izlandi és a portugáliai állomás a két légnyomásközpont legaktívabb területén helyezkedik el, ezért az itt mért értékek alapján a NAO jelenséget jól jellemző index határozható meg. A főkomponens és az állomás alapú indextípus között bizonyos időszakokban jelentősebb eltérések alakulhatnak ki, az 1899 és 2004 közötti periódusra, azonban 0,93 körüli korreláció mérhető közöttük. Az állomás alapú NAO-index térbeli reprezentativitása rosszabb, mint a főkomponens alapú indexé, azonban előbbi hosszabb időszakra áll rendelkezésre, mint utóbbi (*Hurrell, James & National Center for Atmospheric Research Staff (Eds), 2014*).

A NAO-index pozitív értéke az észak-atlanti alacsony légnyomású központ és az azori anticiklon erősen fejlett állapotát jelzi, tehát az észak-atlanti térségben alacsonyabb, a Mediterráneum nyugati területei fölött pedig magasabb légnyomásértékek mérhetők. Ebben az időszakban a nyugatias szelek intenzitásának növekedése figyelhető meg az Atlanti-óceán medencéjének északi része fölött, amely enyhébb és csapadékosabb időt okoz Nyugat-Európában. A negatív NAO-index arra utal, hogy egyik légnyomásközpont sem fejlett túlságosan, általában gyengébb nyugatias szelekkel jár együtt. A NAO jelentős hatást gyakorol az Északi Hemiszférán a szárazföld és a tengerfelszín hőmérsékletére is, különösen a téli félévben (*Hurrell & Deser, 2010*).

A NAO-index idősorának nem normális eloszlása először ismét Q-Q plottal kerül vizsgálatra, amely a 9. ábra (a) részén látható. Az adatsor nem illeszkedik az ábrán pirossal jelölt egyenesre, vagyis nem tekinthető normális eloszlásúnak. A NAO-index idősorának hisztogramja és sűrűségfüggvénye a 9. ábra (b) részén található. A sűrűségfüggvény ferdeségét kifejező együttható értéke -0,303.



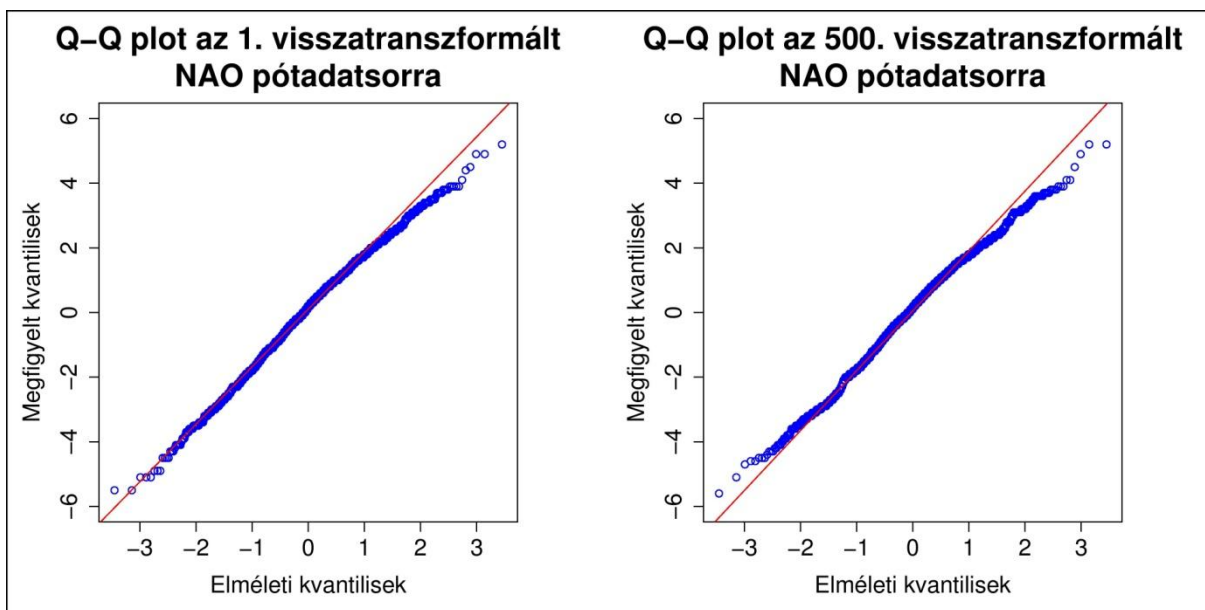
9. ábra. (a) Q-Q plot az Észak-atlanti Oszcillációt jellemző index nem normális eloszlásának vizsgálatához, (b) a NAO-index histogramja és sűrűségfüggvénye, utóbbi piros görbével került ábrázolásra.

Az adatsor nem normális eloszlása statisztikai teszttel is igazolható. Kolmogorov-Szmirnov próba alkalmazásával $p = 8,6 \cdot 10^{-5}$, a Shapiro-Wilk próba elvégzésével, pedig $p = 6,4 \cdot 10^{-9}$ értéket kaptunk eredményül, amelyek szerint a normális eloszlás nullhipotézise az $\alpha < 0,01$ százalék melletti szignifikancia-szinten is elutasítható. A NAO-index adatsora tehát nem kezelhető normális eloszlásúként.

4.3.2. A pótdatsor-módszer alkalmazása a NAO-index idősorára

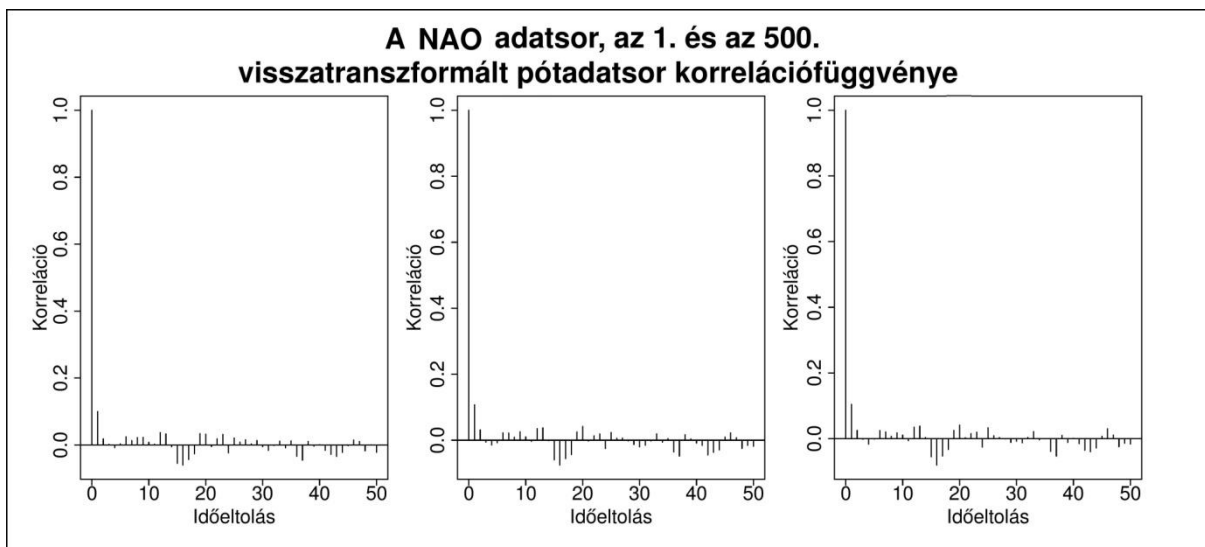
A nemlinearitás vizsgálatához 1000 darab, egyenként 1788 elemet tartalmazó, normális eloszlású pótdatsort állítottunk elő. A NAO-index adatsorának nem normális eloszlásához ismét Gauss-folyamat került hozzárendelésre. A korrelációk becslése után az amplitúdókat a periodogramból számítottuk ki, majd a körfrekvenciák, az amplitúdók és 894 ezer darab egyenletes eloszlású, véletlenszerűen generált fázis felhasználásával létrehoztuk a pótdatsorokat a (4.5) egyenlet alapján.

Az eredeti NAO-index idősor eloszlásával megegyező eloszlásúvá visszaalakított pótdatsorok nem normális eloszlása a 10. ábrán került szemléltetésre, az első és az ötszázadik visszatranszformált pótdatsor Q-Q plotjának bemutatásával.



10. ábra. Q-Q plot a NAO időssorral azonos eloszlású, első és ötszázadik visszatranszformált pótadatsorra.

Az oxigénizotóp időssorhoz hasonlóan, a NAO-index adatsora esetében is összevetettük az első és az ötszázadik visszatranszformált pótadatsor becsült korrelációfüggvényét az eredeti időssor becsült korrelációfüggvényével, amely a 11. ábrán található. Az első ötven időeltolásra a korrelációfüggvények szinte teljesen megegyeznek.



11. ábra. A NAO-index adatsora (balra), valamint az első visszatranszformált pótadatsor (középen) és az ötszázadik visszatranszformált pótadatsor (jobbra) korrelációfüggvényének becsült értéke, az első ötven időeltolásra. A vizsgált pótadatsorok korrelációfüggvényei eredeti idősorétól való átlagos eltérése: 0,007.

A pótatadatok transzformációja után a próbastatisztika, vagyis az eredeti NAO-index idősor és a pótatadatok MSE értékeinek kiszámítása következett a (4.23) képlet szerint. (Előbbi jelölje: MSE_{NAO} .) Meghatározásakor a beágyazási dimenzió értéke egy volt, mivel a nullhipotézis igaz volta esetében a NAO-index idősora a legoptimálisabban $AR(1)$ -folyamattal modellezhető az Akaike-féle információs kritérium alapján. A minimalizálási feladat során a célfüggvényként választott abszolút érték és euklidészi távolság függvény azonos eredményt adott, hiszen $d = 1$ esetén a két függvény minimumhelyei ugyanazok. Ennek következtében az MSE számításra három alkalommal került sor.

4.3.3. A NAO-index adatsorra vonatkozó eredmények

Az MSE értékeket háromféleképpen számítottuk ki, az öt, a tíz és a tizenöt legközelebbi társ értékének átlagolásával, minimalizálendő függvényként az abszolút értéket választva. A linearitás nullhipotézise azonban egyik esetben sem bizonyult elutasíthatónak. A nullhipotézis téves elvetésének valószínűsége még a legkedvezőbb esetben, azaz tizenöt legközelebbi társ esetén is majdnem elérte a negyven százalékot, vagyis a nemlinearitásra nincs szignifikáns bizonyíték. A NAO-index havi idősora tehát lineárisnak tekinthető, noha az NGRIP adatsorhoz képest jelentős különbségek tapasztalhatóak az MSE kiszámítása során a társak számának változtatásával. A hipotézisvizsgálat eredményei a 2. táblázatban találhatók összefoglalóan.

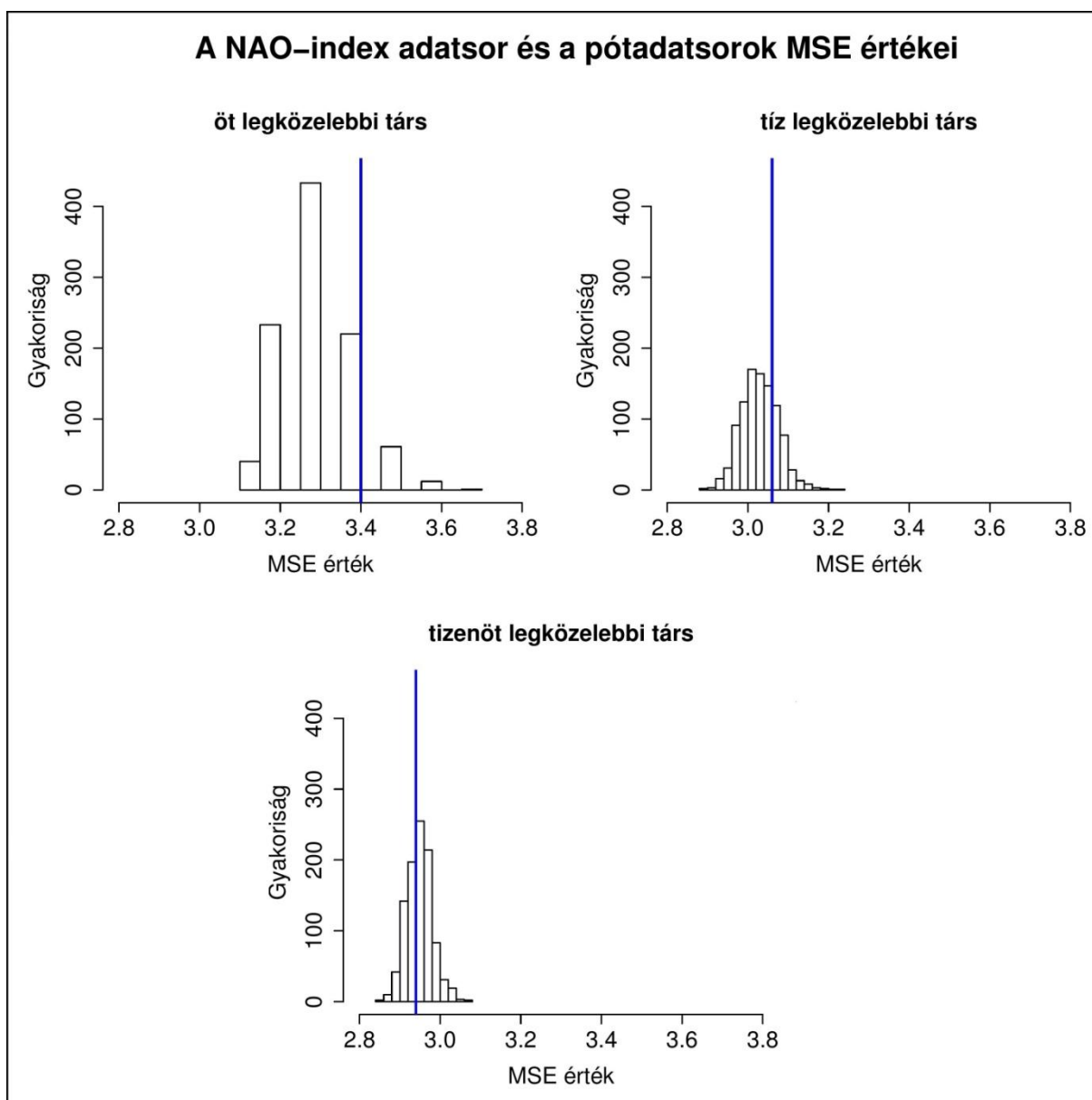
A NAO-index idősora: eredmények		
legközelebbi társak száma (k)	esetszám: $MSE_{NAO} < MSE_{pót}$	az elsőfajú hiba valószínűsége (α)
	célfüggvény: abszolút érték	célfüggvény: abszolút érték
5	99	90,1 %
10	263	73,7 %
15	613	38,7 %

2. táblázat. Esetszám arra vonatkozóan, hogy a pótatadatokra kiszámított MSE érték hány esetben haladja meg az eredeti idősorra kiszámított MSE értéket, illetve az elsőfajú hiba valószínűsége a NAO-index adatsora nemlinearitásának pótatadatok-módszerrel való vizsgálata során.

Az MSE értékek meghatározásakor a legközelebbi társak számának növelésével az elsőfajú hiba valószínűségének csökkenése figyelhető meg. Míg öt társ esetén a nullhipotézis téves elutasításának valószínűsége elérte a kilencven százalékot, addig tizenöt társ alapján végezve

az átlagolást, az elsőfajú hiba valószínűsége negyven százalék alá csökkent. Ez a feltételes várható érték véletlenszerű hibája csökkenésének következménye lehet, ugyanis ha k értéke kicsi, akkor a feltételes várható érték szisztematikusan jobban, véletlenszerűen rosszabbul közelíthető. A k értékének növelésével a feltételes várható érték szisztematikus hibája egyre nagyobbá válik, míg a véletlenszerű hibája csökken. A feltételes várható érték így kezdi megközelíteni a várható értéket.

Az MSE értékek eloszlása a 12. ábrán került megjelenítésre.



12. ábra. Az eredeti NAO-index idősor és az ezer darab pótdatsor MSE értékeinek eloszlása. Előbbit kék függőleges vonal jelöli.

A NAO jelenség rendelkezhet nemlineáris komponensekkel, amelyek nagymértékű bizonytalanság forrásai a kutatások során. Az eltérő eredményeket részben a különbözőképpen számított NAO-indexek okozhatják. A mérőszámok között jelentős különbségek figyelhetők meg bizonyos időszakokban, tehát a vizsgálat periódusától is függhet, hogy a folyamat lineárisként vagy nemlineárisként azonosítható inkább. Például az állomás és a főkomponens alapú NAO-index szignifikáns eltérést mutat az 1960-as és az 1990-es évek első felében.

A megfigyelések szerint a különböző módokon származtatott NAO-indexek térben is eltérhetnek egymástól, ugyanis a NAO-jelenség intenzitásának központi területei, azaz akciócentrumai időben változtatják helyzetüket, az állomás alapú NAO-index térbeli reprezentativitását lerontva, hiszen az állomások statikusak, nem követik az akciócentrumokat. A főkomponens alapú indexek alkalmazása során ez a probléma nem merülhet fel (*Hurrell, James & National Center for Atmospheric Research Staff (Eds), 2014*). Az állomás alapú NAO-indexek használatát mégis az teszi előnyössé a főkomponens alapúakkal szemben, hogy hosszú időre rendelkezésre állnak, míg utóbbi adatsorok rövidebbek, ezért statisztikai vizsgálatuk is nehezebb.

Az idő- és térbeli eloszlástól függően tehát eltérő következtetések vonhatók le az adatsor nemlinearitása tekintetében. Például *Pozo-Vázquez et al. (2001)* megállapította, hogy a téli NAO-index időszora nem tekinthető stacionáriusnak az 1875 és 1997 közötti időszakban; az időszor 1875 és 1939 között fehérzaj folyamat realizációjának tekinthető, míg 1940 után markáns, hat és tíz éves oszcilláció figyelhető meg az adatsorban.

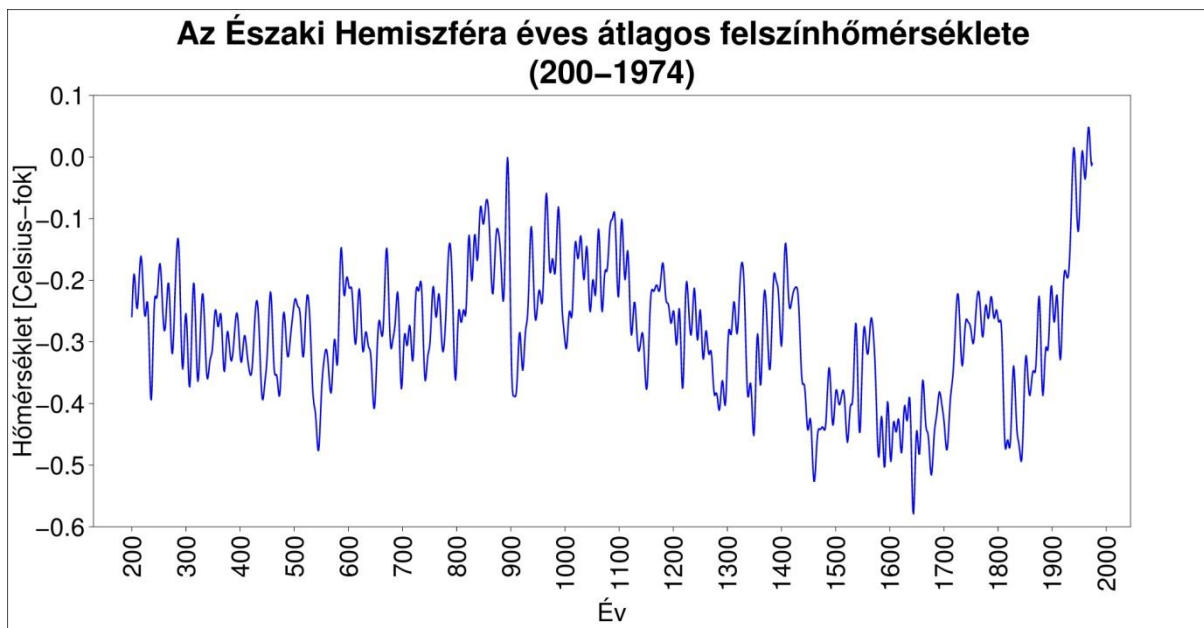
Masala (2014) pedig a vizsgálati periódustól függően eltérő karakterisztikák miatt azonosíthatta nemlineárisként a NAO-index 1950 és 2013 közötti időszakra vonatkozó, napi adatsorát.

Összefoglalásképpen megállapítható, habár a pótadatsor-módszer általunk alkalmazott változata szerint a NAO-index adatsora nem tekinthető nemlineáris folyamat realizációjának, azonban az eredmények arra utalnak, hogy a jelenség nemlineáris komponensekkel bírhat.

4.4. Az Északi Hemiszféra átlaghőmérséklete

4.4.1. Az NH átlagos felszínhőmérsékleti adatsor

A harmadik, nemlinearitás szempontjából vizsgált idősor az Északi Hemiszféra éves átlagos felszínhőmérséklete (a továbbiakban: NH idősor), 200 és 1974 között. Az 1775 elemet tartalmazó idősor a 13. ábrán került bemutatásra.



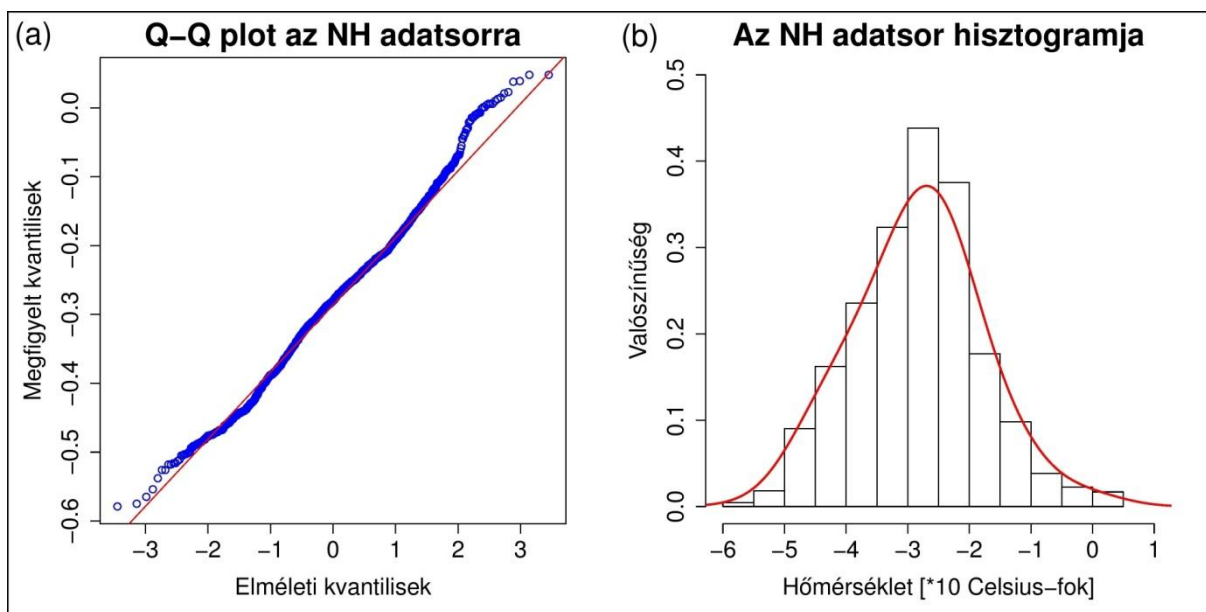
13. ábra. Az Északi Hemiszféra éves átlagos felszínhőmérséklete 200 és 1974 között (Jones & Mann, 2004).

A rendszeres, azonos elvek alapján végzett műszeres meteorológiai mérések körülbelül másfél évszázados múltra tekintenek vissza. A korábbi időszak hőmérséklete csak indirekt módon, úgynevezett *proxy* adatsorok készítésével elemezhető. Nagyfelbontású hőmérsékleti idősorok az 1850 előtti időszakra vonatkozóan például évgűrűk, jégfuratok, valamint függő cseppkövek és egyéb barlangi üledékek vizsgálata alapján rekonstruálhatók. Az NH idősor a műszeres méréseken alapuló HadCRUTv2 adatsor hőmérsékleti értékeit tartalmazza 1850-től kezdődően, amelyet a brit meteorológiai szolgálat keretei között működő Hadley Központ (Met Office Hadley Centre for Climate Change) és a Kelet-Angliai Egyetem Klímakutató Egysége (University of East Anglia, Climate Research Unit: CRU) állított össze (Jones & Mann, 2004).

Az NH idősorból az 1974 utáni adatok az elmúlt évtizedek erőteljes felmelegedése miatt eltávolításra kerültek. Hirtelen éghajlatváltozásra utaló események azonban a XX. század előtt

is megfigyelhetők, többek között a Középkori Meleg Periódus során 950 és 1250 között, illetve a Kis Jégkorszak idején 1350 és 1850 között (Matyasovszky, 2011).

Az NH idősor nem normális eloszlásának Q-Q plottal való ellenőrzése a 14. ábra (a) részén található. Az NH idősor nem illeszkedik az ábrán pirossal jelölt egyenesre, vagyis nem normális eloszlású. A 14. ábra (b) részén szemléltettük az adatsor hisztogramját és a sűrűségfüggvényét. A ferdeséget jellemző együttható értéke 0,128.



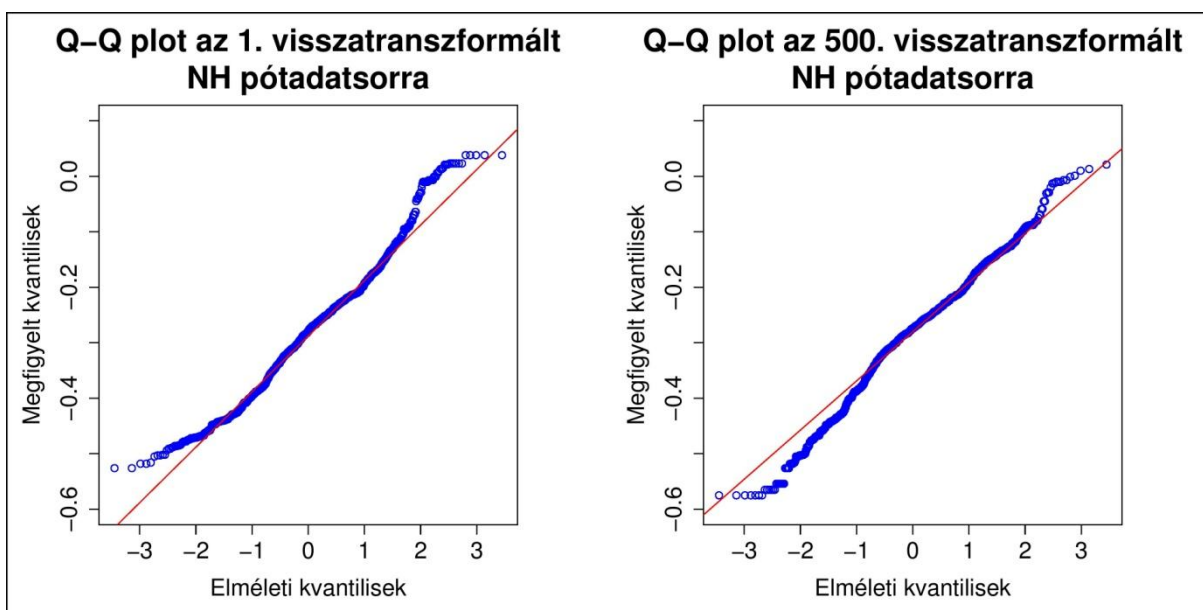
14. ábra. (a) Q-Q plot az Északi Hemiszféra éves átlagos felszínhőmérséklete nem normális eloszlásának vizsgálatához, (b) az NH adatsor hisztogramja és sűrűségfüggvénye, ahol utóbbit piros görbe jelöli; a hisztogram megfelelő megjelenítése érdekében az x tengelyen ábrázolt hőmérsékleti értékeket megszoroztuk tízzel.

Az idősor eloszlására vonatkozó hipotézisvizsgálat során a Kolmogorov-Szmirnov próba eredményeképpen $p = 10^{-2}$, a Shapiro-Wilk próba végrehajtásával, pedig $p = 3,1 \cdot 10^{-7}$ adódott p -értékként. A nullhipotézis, vagyis a normális eloszlás elvethető az $\alpha < 1$ százalék melletti szignifikancia-szint esetén, tehát az NH idősor nem tekinthető normális eloszlásúnak.

4.4.2. A pótatadatsor-módszer alkalmazása az NH idősorra

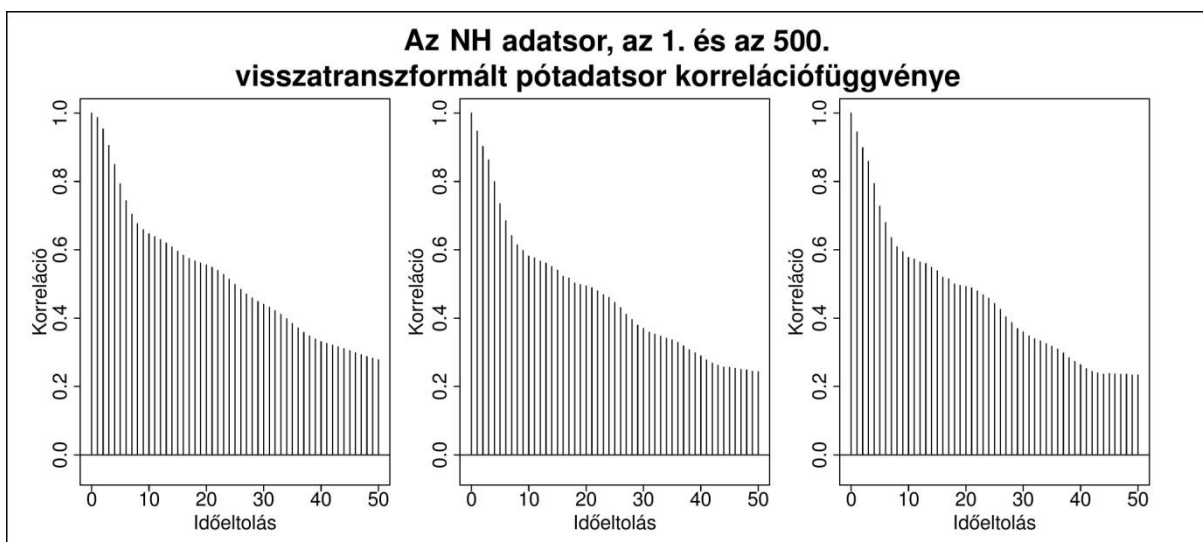
A pótatadatsor-módszer alkalmazásához első lépésként 887 500 darab egyenletes eloszlású fázist szimuláltunk, majd a periodogram felírásához szükséges korrelációk ismételtén a Spearman-féle rangkorrelációval kerültek becslésre. A körfrekvenciák és az amplitúdók periodogramból való kiszámítása után az 1000 darab, egyenként 1775 elemű, normális eloszlású pótatadatsort a 4.5 egyenlet alapján állítottuk elő.

A következő lépés a pótdatsorok eredeti, NH idősor eloszlásával megegyező eloszlásúra való visszatranszformálása volt. Szemléltetésképpen a 15. ábrán az első és az ötszázadik visszatranszformált pótdatsor kvantilis-kvantilis ábrája található.



15. ábra. Q-Q plot az NH idősorral megegyező eloszlású, első és ötszázadik visszatranszformált pótdatsorra.

Az NH idősor és a pótdatsorok korrelációfüggvényei ismét kvalitatív módon vizsgálandók. A 16. ábrán látható az eredeti idősor, illetve az első és az ötszázadik visszatranszformált pótdatsor első ötven időeltolásra becsült korrelációfüggvénye, amelyek hasonló menetűek.



16. ábra. Az NH idősor (balra), az első visszatranszformált pótdatsor (középen) és az ötszázadik visszatranszformált pótdatsor (jobbra) korrelációfüggvényének becslése, az első ötven időeltolásra. A vizsgált pótdatsorok korrelációfüggvényei eredeti idősorétól való átlagos eltérése: 0,059.

Az eredeti idősor és a pótatadsorok korrelációfüggvényeinek kismértékű eltérése annak a következménye, hogy a pótatadsorok (4.13) egyenlőség szerinti transzformációjához az empirikus eloszlásfüggvény került alkalmazásra, amely nem kellően sima, hiszen lépcsős függvény. A szimulált normális eloszlású idősorelem kis változtatása nagy különbséget jelenthet a visszatranszformált adatsorban, amennyiben annak környékén a valódi eloszlásfüggvénynek ugrása van.

A transzformációt az eredeti NH idősorra és a pótatadsorokra vonatkozó MSE érték kiszámítása követte, a (4.23) képlet szerint. A beágyazási dimenzió értéke ebben az esetben tizenöt volt, ugyanis a nullhipotézis teljesülése esetén az NH idősor jól modellezhető lineáris autoregresszív folyamattal, ekkor pedig a modellezéshez $AR(15)$ -folyamat adódott optimálisnak, a Akaike-féle információs kritérium alapján.

4.4.3. Az NH adatsorra vonatkozó eredmények

Az MSE értékeket – az NGRIP adatsorhoz hasonlóan – hatféleképpen határoztuk meg: az átlagolásra az öt, a tíz és a tizenöt legközelebbi társ értéke alapján került sor, minimalizálendő függvényként pedig az abszolút értéket és az euklidészi távolságot választottuk.

A linearitás nullhipotézise mindegyik esetben elutasításra került, még hozzá olyan szignifikancia-szinten, hogy az elsőfajú hiba, azaz a nullhipotézis téves elutasításának valószínűsége 0,1 százaléknál alacsonyabb.

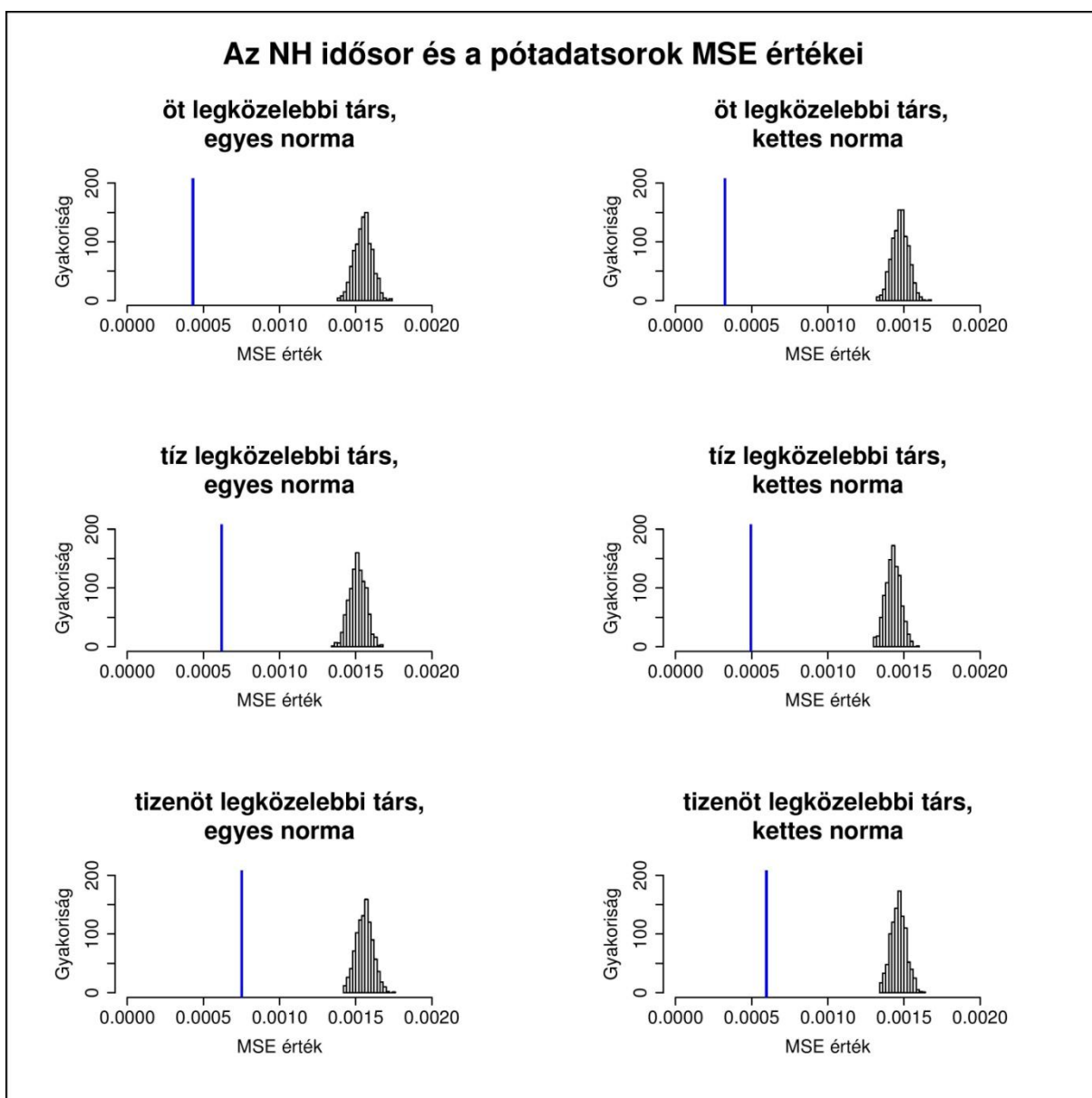
A hipotézisvizsgálat eredményeit összefoglalóan a 3. táblázatban tartalmazza.

Az NH idősor: eredmények		
legközelebbi társak száma (k)	az elsőfajú hiba valószínűsége (α)	
	célfüggvény: abszolút érték	célfüggvény: euklidészi távolság
5	< 0,1 %	< 0,1 %
10	< 0,1 %	< 0,1 %
15	< 0,1 %	< 0,1 %

3. táblázat. Az elsőfajú hiba valószínűsége az NH adatsor nemlinearitásának pótatadsor-módszerrel való vizsgálata során. A nullhipotézis: az idősor linearitása.

A pótatadsor-módszer általunk alkalmazott eljárása alapján tehát az Északi Hemiszféra éves átlagos felszínhőmérsékletének adatsora egyértelműen nemlineárisnak tekinthető.

Az MSE értékek eloszlását a 17. ábrán jelenítettük meg, amelyen jól megfigyelhető, hogy az eredeti idősorra kiszámított MSE értéket minden esetben meghaladta a pótdatsorokra kiszámított MSE érték.



17. ábra. Az eredeti NH idősor és az ezer darab pótdatsor MSE értékeinek eloszlása. Előbbit kék függőleges vonal jelöli.

5. Összegzés

A diplomamunkában az idősorok nemlinearitásának detektálására szolgáló pótatadsor-módszer új eljárását mutattuk be, amellyel a vizsgálat gyors és hatékony módon végezhető el. Az eljárással olyan pótatadsorok készíthetők, amelyek eloszlásfüggvényei, illetve korrelációfüggvényei egyezők az eredeti, megfigyelésen alapuló idősorével. A korábbi módszerek alkalmazásával kapcsolatban gyakori problémaként merült föl, hogy csak az egyik feltételt teljesítették, vagy egyiket sem. A saját módszer végrehajtására, így a pótatadsorok létrehozására önkényesen megválasztandó paraméterek bevezetése vagy bonyolult iteratív lépések sorozata nélkül került sor, a korábbiakhoz képest csekély számítási igényű programok felhasználásával.

A módszert három éghajlati idősorra alkalmaztuk. Az első az Észak-grönlandi Jégfurat Projekt 60 ezer évvel ezelőtt kezdődő, 45 ezer évre vonatkozó oxigénizotóp idősora, amelynek nemlinearitása ismert, így alkalmas volt az eljárás tesztelésére. A hipotézisvizsgálathoz az átlagos négyzetes hibát választottuk próbastatisztikaként, amit hatféleképpen számítottunk ki, az eredeti idősor linearitásának nullhipotézise pedig mindegyik esetben nagy bizonyossággal került elvetésre.

Az új eljárással vizsgált második idősor az Észak-atlanti Oszcilláció egyik indexének adatsora, amely 1865 januárja és 2013 decembere közötti havi adatokat tartalmaz. A hipotézisvizsgálat során az MSE próbastatisztikát háromféleképpen határoztuk meg, az eredeti idősor linearitásának nullhipotézise azonban nem bizonyult elutasíthatónak. Az elsőfajú hiba, vagyis a nullhipotézis téves elvetésének valószínűsége a legkedvezőbb esetben is negyven százalék körüli volt. Noha a vizsgálatunk alapján a NAO-index havi idősora lineárisnak tekinthető, a NAO jelenségnek lehetnek nemlineáris komponensei. A bizonytalanságot növeli az, hogy több NAO-index létezik, amelyek mind képzési módjaikban, mind térbeli és időbeli eloszlásuk tekintetében különbözőek, ezért statisztikailag eltérően viselkedhetnek.

A harmadik, nemlinearitás szempontjából elemzett adatsor az Északi Hemiszféra 200 és 1974 közötti évekre számított éves átlaghőmérsékletének idősora volt. Az MSE próbastatisztikát hatféleképpen számítottuk ki. Az adatsor linearitásának nullhipotézise mindegyik esetben nagy bizonyossággal került elutasításra, tehát az idősor nemlineárisnak tekinthető.

Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani témavezetőmnek, Dr. Matyasovszky Istvánnak, a dolgozat megírásához nyújtott szakmai segítségéért, a sok hasznos tanácsért, útmutatásért és türelméért.

Köszönettel tartozom Családomnak a támogatásért, akik nélkül ez a dolgozat nem jöhetett volna létre.

Irodalomjegyzék

- Akaike, H., 1974: A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19, 716-723.
- Boudt, K., Cornelissen, J., Croux, C., 2012: The Gaussian rank correlation estimator: robustness properties. *Statistics and Computing*, 22, 471-483.
- Braun, H., 2009: Strong indications for nonlinear dynamics during Dansgaard-Oeschger events. *Climate of the Past Discussion*, 5, 1751–1762.
- Cario, M. C., Nelson, B. L., 1996: Autoregressive to anything: Time-series input processes for simulation. *Operations Research Letters*, 19, 51-58.
- Dansgaard, W., 1964: Stable isotopes in precipitation. *Tellus*, 16, 436-468.
- Dansgaard, W., Johnsen, S. J., Clausen, H. B., Dahl-Jensen, D., Gundestrup, N. S., Hammer, C. U., Hvidberg, C. S., Steffensen, P., Sveinbjörnsdottir, A. E., Jouzel, J., Bond, G., 1993: Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature*, 364, 218-220.
- Dolan, K. T., Spano, M. L., 2001: Surrogate for nonlinear time series analysis. *Physical Review E*, 64, 046128.
- Grassberger, P., 1986: Do climatic attractors exist? *Nature*, 323, 609-612.
- Grassberger, P., Procaccia, I., 1983: Characterization of Strange Attractors. *Physical Review Letters*, 50, 346–349.
- Hurrell, J., Deser, C., 2010: North Atlantic climate variability: The role of the North Atlantic Oscillation. *Journal of Marine Systems*, 79, 231-244.
- Jouzel, J., Koster, R. D., Suozzo, R. J., Russell, G. L., 1994: Stable water isotope behavior during the last glacial maximum: A general circulation model analysis. *Journal of Geophysical Research*, 99, 27791-27801.
- Kugiumtzis, D., 1999: Test your surrogate data before you test for nonlinearity. *Physical Review E*, 60, 2808-2816.
- Kugiumtzis, D., 2002: Statically Transformed Autoregressive Process and Surrogate Data Test for Nonlinearity. *Physical Review E*, 66, 025201.

- Lucio, J. H., Valdés, R., Rodríguez, L. R., 2012: Improvements to surrogate data methods for nonstationary time series. *Physical Review E*, 85, 056202.
- Maiwald, T., Mammen, E., Nandi, S., Timmer, J., 2008: Surrogate Data – A Qualitative and Quantitative Analysis. 41-74. In: *Mathematical Methods in Signal Processing and Digital Image Analysis Understanding Complex Systems*, Springer-Verlag, Berlin.
- Mammen, E., Nandi, S., Maiwald, T., Timmer, J., 2009: Effect of Jump Discontinuity for Phase-Randomized Surrogate Data Testing. *International Journal of Bifurcation and Chaos*, 19, 403–408.
- Masala, G., 2014: North Atlantic Oscillation index stochastic modelling. *International Journal of Climatology*. DOI: 10.1002/joc.4236.
- Matyasovszky, I., 2002: Statisztikus klimatológia. Idősorok elemzése. *ELTE Eötvös Kiadó, Budapest*.
- Matyasovszky, I., 2011: Detecting abrupt climate changes on difference time scales. *Theoretical and Applied Climatology*, 105, 445-454.
- Mikšovský, J., Raidl, A., 2006: Testing for nonlinearity in European climatic time series by the method of surrogate data. *Theoretical and Applied Climatology*, 83, 21-33.
- Mikšovský, J., Pišoft, P., Raidl, A., 2008: Global Patterns of Nonlinearity in Real and GCM-simulated Atmospheric Data. 17-34. In: Donner, R. V., Barbosa, S. M.: *Nonlinear Time Series Analysis in the Geosciences - Applications in Climatology, Geodynamics, and Solar-Terrestrial Physics Geodynamics and Solar-terrestrial Physics, Lecture Notes in Earth Science*, 112, (2008), 390, Springer-Verlag, Berlin.
- Paluš, M., Albrecht, V., Dvůrák, I., 1993: Information theoretic test for nonlinearity in time series. *Physics Letters A*, 175, 203-209.
- Paluš, M., Novotná, D., 1994: Testing for Nonlinearity in Weather Records. *Physics Letters A*, 193, 67-74.
- Paluš, M., 1996: Detecting Nonlinearity in Multivariate Time Series. *Physics Letters A*, 213, 138-147.
- Patil, D. A. S., Hunt, B. R., Carton, J. A., 2000: Identifying Low-Dimensional Nonlinear Behavior in Atmospheric Data. *Monthly Weather Review*, 129, 2116-2125.

- Pozo-Vázquez, D., Esteban-Parra, M. J., Rodrigo, F. S., Castro-Díez, Y., 2001: A study of NAO variability and its possible non-linear influences on European surface temperature. *Climate Dynamics*, 17, 701-715.
- Prichard, D., Theiler, J., 1994: Generating Surrogate Data for Time Series with Several Simultaneously Measured Variables. *Physical Review Letters*, 73, 951-954.
- Schreiber, T., Schmitz, A., 1996: Improved surrogate data for nonlinearity tests. *Physical Review Letters*, 77, 635.
- Schreiber, T., Schmitz, A., 2000: Surrogate time series. *Physica D*, 142, 346-382.
- Small, M., Judd, K., 1998: Correlation dimension: A pivotal statistic for non-constrained realizations of composite hypotheses in surrogate data analysis. *Physica D*, 120, 386-400.
- Small M., Judd, K., Mees, A., 2001: Testing time series for nonlinearity. *Statistics and Computing*, 11, 257-268.
- Sugihara, G., Casdagli, M., Habjan, E., Hess, D., Dixon, P., Holland, G., 1999: Residual delay maps unveil global patterns of atmospheric nonlinearity and produce improved local forecasts. *Proceedings of the National Academy of Science*, 96, 14210-14215.
- Takens, F., 1985: On the numerical determination of the dimensions of an attractor. In: *Dynamical Systems and Bifurcations*, eds. Braaksma, B. L. J., Broer, H. W., Takens, F., *Lecture Notes in Mathematics*, 1125, 99-106.
- Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer, J. D., 1992: Testing for nonlinearity in time series: the method of surrogate data. *Physica D*, 58, 77-94.
- Theiler, J., Prichard, D., 1996: Constrained-Realization Monte-Carlo method for Hypothesis Testing. *Physica D*, 94, 221-235.
- Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*, 2nd edition. Elsevier, Amsterdam.
- Wolff, R. C., Yao, Q., Tong, H., 2004: Statistical tests for Lyapunov exponents of deterministic systems. *Studies in Nonlinear Dynamics and Econometrics*, 8, 174-193.

A felhasznált adatsorok forrásai:

Az Észak-grönlandi Jégfurat Projekt oxigénizotóp adatsora:

North Greenland Ice Core Project members. 2004. High-resolution record of Northern Hemisphere climate extending into the last interglacial period. *Nature*, v.431, No. 7005, pp. 147-151, 9 September 2004.

A NAO-index időszora:

Hurrell, James & National Center for Atmospheric Research Staff (Eds). Last modified 05 Sep 2014. "The Climate Data Guide: Hurrell North Atlantic Oscillation (NAO) Index (station-based)." Retrieved from <https://climatedataguide.ucar.edu/climate-data/hurrell-north-atlantic-oscillation-nao-index-station-based>.

Az 1800 éves hemiszférikus hőmérsékletre vonatkozó adatsor:

Jones, P. D., and Mann, M. E., (2004), Climate over past millennia, *Rev. Geophys.*, 42, RG2002, DOI:10.1029/2003RG000143.