

NÉHÁNY STATISZTIKUS MÓDSZER AZ ELMÉLETI ÉS ALKALMAZOTT KLIMATOLÓGIAI VIZSGÁLATOKBAN

Matyasovszky István

ELTE Meteorológiai Tanszék, 1117 Budapest, Pázmány Péter sétány 1/A
e-mail: matya@caesar.elte.hu

Bevezetés

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

Részben az általános megismerés, részben a jövő éghajlatának becslési lehetőségei megismerése szempontjából fontos a már lezajlott, illetve zajló éghajlatváltozás detektálása és becslése. Az éghajlatváltozást legegyszerűbben a várható érték időbeli változásával szokás leírni, vagyis a trend-elemzéssel. Manapság, alapvető fontosságú gyakorlati hatásai folytán, egyre inkább előtérbe kerülnek az extrém események, amelyek például a kvantilis regresszió segítségével vizsgálhatók. Az egyik legrégebbi statisztikus klimatológiai eszköz a spektrál-analízis, amikor egy szóban forgó idősor mögött meghúzódó sztochasztikus folyamatot véges sok (diszkrét spektrum) és megszámlálhatatlanul sok (folytonos spektrum) periodikus összetevő szuperpozíciójaként tekintjük. Adott idősor esetében a diszkrét spektrum trigonometrikus trendként jelentkezik. Végezetül meg kell említeni, hogy mivel az éghajlat nagyszámú nem-lineáris kölcsönhatás eredményeként jön létre, célszerű az éghajlati idősorokat a jól ismert lineáris idősor modellek helyett nem-lineáris idősor modellek segítségével elemezni.

Az éghajlat lényegének megismerése szempontjából vethető fel a különböző éghajlati elemek memóriája. Ilyenkor a rendelkezésre álló idősorok alapján próbáljuk becsülni, hogy az adott idősor aktuális értékeinek optimális statisztikai előrejelzéséhez a múltbeli értékek időben milyen hosszú szakaszára van szükség. Némiképp hasonló kérdés az ún. embedding dimenzió becslése. Ennek lényege, hogy az adott idősort egy zajjal terhelt dinamikai rendszer megfigyelésének tekintjük, és a feladat a dinamikai rendszer dimenziójának megadása. Számos statisztikai módszer igényli az embedding dimenzió ismeretét, így például az elmúlt két évtizedben rohamosan fejlődő szinguláris spektrum analízis (Vautard and Ghil, 1989), ám megbízható, jól definiált statisztikai tulajdonságokkal rendelkező becslése tudomásom szerint elmaradt a klimatológiai irodalomban. A megoldás iránt érdeklődők számára Fueda and Yaganawa (2001) munkáját ajánlom.

Trend-elemzés

A trendet ma is leggyakrabban az idő lineáris függvényének tekintik. Ezt a legkisebb négyzetek (least squares: LS) módszerével végzik, tehát az aktuális megfigyelések és becslésük négyzetes hibaösszegének minimalizálásával. A valóságos trend azonban rendszerint eltér a lineáristól. Lényegesen pontosabb becslés nyerhető az ún. súlyozott lokális regresszió (weighted local regression: WLR) segítségével. Mivel a trend az idő sima függvénye, minden időpont egy környezetében lokálisan illesztünk lineáris trendet úgy, hogy a négyzetes hibákat súlyozzuk. Az aktuális időponthoz tartozó hiba rendelkezik a legnagyobb súllyal, és az időponttól távolodva valamilyen ütem szerint csökkennek a súlyok. A súlyokat egy magfüggvény generálja, aminek megválasztása nincs túl komoly hatással a becslés pontosságára. A kör-

nyezet, melyet sávszélességnek nevezünk, azonban igen fontos szerepet játszik. Szerencsére ennek szubjektív megválasztása helyett van mód a becslésére is.

Ismeretes, hogy az Északi Hemiszféra átlaghőmérséklete az 1850–2010 közötti periódusban a lineáris trend szerint összességében $0,73\text{ }^{\circ}\text{C}$ emelkedést mutat, ami $0,045\text{ }^{\circ}\text{C}/10\text{ év}$ növekedési rátának felel meg. A WLR eljárás azonban az 1914–1942 időszakra összességében $0,133\text{ }^{\circ}\text{C}/10\text{ év}$, sőt az 1975–2010 periódusra $0,183\text{ }^{\circ}\text{C}/10\text{ év}$ emelkedő rátát nyújt úgy, hogy közben az 1942–1975 évek során $0,057\text{ }^{\circ}\text{C}/10\text{ év}$ ütemű csökkenést jelez (Matyasovszky, 2011).

Az említett évszámok, illetve időszakok nem véletlenül lettek kiválasztva. Az utóbbi időben ugyanis növekvő figyelem fordul az ún. hirtelen éghajlatváltozások (abrupt climate changes) felé. A hirtelen változást mutató időpontok azonosítása tradicionálisan azon alapszik, hogy a trend-függvényt szakaszonként konstansnak tekintik, mely szakaszokat a trend ugrásai határolják. Ez a megközelítés fizikailag tarthatatlan, hiszen nem képzelhető el az a helyzet, hogy az éghajlat valameddig változatlan, majd azonnal egy másik éghajlatba megy át. Ezért a trend-függvény deriváltjának ugrásait tekintem hirtelen változásnak (Matyasovszky, 2011). A feladat a WLR módszer általánosításával oldható meg. A definíció persze egy kompromisszum, mert a trend ugrásainak megengedése elfogadhatatlan, míg a második vagy magasabb deriváltjai ugrásának értelmezése túl sima trendet adna ahhoz, hogy hirtelen változásról beszélhessünk. A módszer az 1901, 1914, 1942, 1963 és 1975 éveket adta az Északi Hemiszféra hőmérsékletének hirtelen változásaira. A távolabbi múltba nyúló rekonstruált hőmérsékleti adatokra alkalmazva eljárásomat, be lehetett határolni egyebek mellett a Középkori Meleg Periódust (Viking-kor) és a Középkori Kis Jégkorszakot, vagy a Holocén Klíma Optimumot is.

Regresszió, kvantilis regresszió

A WLR azokban az esetekben is alkalmazható, amikor a független változó (prediktor) értékei nem időpontok vagy egyéb determinisztikus mennyiségek, hanem maguk is valószínűségi változó, sőt változók (prediktorok) realizációi. Ilyenkor a becslendő változó (prediktandusz) adott prediktorok melletti feltételes várható értékének becslése a cél. Példaként említhető a hazánkban nagyon elterjedt parlagfű erősen allergén pollenjének koncentráció becslése. Makra et al. (2011) Szeged, Legnano és Lyon napi parlagfű pollen-koncentrációit hozta kapcsolatba a megelőző napi koncentrációval és meteorológiai paraméterekkel, mivel a Kárpát-medencén kívül még a Pó-alföld és a Rhone völgye Európa erősen parlagfüves területei. A becslés által megmagyarázott relatív variancia Szegedre a legnagyobb és Legnanora a legkisebb, tehát a legpontosabban Szeged napi parlagfű pollen-koncentrációja becsülhető a három hely közül. A legfontosabb meteorológiai változónak a napi középhőmérséklet (Szeged és Legnano) és a napi csapadékösszeg (Lyon) bizonyult.

A hagyományos regressziós feladat helyett gyakran célszerű a medián regressziót elvégezni, amikor a prediktorok melletti feltételes medián becsléséről van szó. Ilyenkor a becslési hibák négyzetösszegének minimalizálása helyett a becslési hibák abszolút összegének minimalizálása történik. Különösen előnyös ez, ha a prediktandusz valószínűségi eloszlása erősen nem normális és esetenként nagyon kiugró értékek fordulhatnak elő. Jó példa erre ismét a szegedi napi parlagfű pollen-koncentráció. Makra and Matyasovszky (2010) ezúttal az esős és száraz napokat szétválasztva külön-külön oldotta meg a feladatot. Az előző napi koncentráción kívül az esős napokon a napi globálsugárzás, a száraz napokon a napi középhőmérséklet bizonyult fontos prediktornak. Részben a napok csapadék szerinti szétválasztásának, de főként a medián regressziónak köszönhetően az előző vizsgálathoz képest lényegesen kisebb átlagos abszolút hibákat, azaz lényegesen pontosabb becslést nyertünk.

Mivel a medián a $p=0,5$ valószínűségi értékhez tartozó kvantilis, a medián regresszió általánosításával bármely zérus és egy közötti p -re értelmezhető a kvantilis regresszió. Ez a prediktandusznak a prediktorok adott értéke melletti feltételes kvantilis becslését végzi. Megjegyezzük, hogy a p -kvantilis az a szám, amelynél kisebb értéket a szóban forgó valószínűségi változó p valószínűséggel vesz fel. A számítások szerint a napi parlagfű pollen-koncentráció kvantilisei általában kisebbek az esős, mint a száraz napokon, továbbá a napi koncentráció valószínűségi eloszlása sokkal elnyújtottabb a magas koncentrációk felé a száraz napokon. Az esős napokhoz tartozó kvantilisek azt jelzik, hogy a pollen-koncentrációk jóval kisebb változékonyságúak a csapadékos napokon.

Spektrálanalízis

A spektrálanalízis rendkívül kiterjedt módszertana döntően az ekvidisztáns időpontokban megfigyelt adatsorokra vonatkozik. Az éghajlati idősorok azonban olykor nem ekvidisztáns időpontokban állnak rendelkezésre. Tipikus példa erre a paleoklíma adatok. Ilyenkor a spektrálanalízis kiindulásaként szolgáló hagyományos periodogram helyett a Lomb-Scargle periodogramot (Lomb, 1976) szokás alkalmazni. Igen gyakran egy elsőrendű autoregresszív (AR(1)) folyamatot illesztnek az idősorhoz, és a feladat annak eldöntése, hogy az AR(1) modell által generált vörös zaj spektrumtól mely periódusidőknél különbözik szignifikánsan a periodogram. A Lomb-Scargle periodogramban fellépő pontatlanságok csökkentése érdekében egy új periodogramot vezettem be, mely az ún. totális legkisebb négyzetek módszerén alapul (Matyasovszky, 2013). A GISP2 oxigén izotóp adatok (i.e. 58000–13000 évek) és Vostok deutérium tartalom adatok (i.e. 420766–i.sz. 2000 évek) elemzése a korábbi eredményekhez képest több fontos különbséget hozott (Matyasovszky, 2013). Az összehasonlítás során az AR(1) spektrális sűrűségen kívül felhasználtam a periodogram robusztus simítását is (Matyasovszky, 2010a).

Ismeretes, hogy egy sztochasztikus folyamat spektruma a folyamat kovariancia-függvényének Fourier-transzformáltján alapszik. Az utóbbi időben a folyamattal kapcsolatos bizonyos karakterisztikák kovariancia-függvényének Fourier-transzformáltjával olyan új spektrum értelmezések bukkantak föl (Dette et al., 2015), amelyek segítségével mélyebben jellemezhetők idősorok statisztikai tulajdonságai.

Ezek meteorológiai alkalmazásával kapcsolatban diplomamunka témával, vagy akár doktori témával tudnék szolgálni.

Nem-lineáris idősor modellezés

Egy AR folyamat tipikus lineáris modell, mert a szóban forgó idősor aktuális értékét az idősor megelőző értékeinek lineáris kombinációja és egy zaj összege adja. Több idősor együttes modellezésekor vektor értékű AR folyamathoz jutunk. Ez nem-lineárisra például úgy tehető, hogy a lineáris kombinációt definiáló AR paraméterek függenek valamelyik idősor valahány lépéssel korábbi értékétől, vagyis a küszöbváltozó értékétől.

A grönlandi és az antarktiszi jégfuratok izotóp összetételének tanulmányozása során egyebek mellett fáziskülönbséget találtak a két terület éghajlatváltozásainak időbeli lefolyásában. Felvetődött ezért a kérdés, hogy a változások a két félteke közül vajon melyiken indulnak el és terjednek át a másikra. Mindkét lehetőségre születtek fizikailag alátámasztott elképzelések, amelyek közül azonban nem lehetett egyértelműen, statisztikailag igazoltan választani (Steig and Alley, 2002). Ennek oka minden bizonnyal az, hogy a statisztikai vizsgálatok megmaradtak a lineáris idősor elemzés keretein belül. Kidolgoztam ezért a grönlandi NGRIP és az antarktiszi Vostok adatsor együttesére egy egyszerű nem-lineáris AR modellt, egy ún. küszöbmodellt (Matyasovszky, 2010b). Mivel a küszöbváltozó - mely a nem-linearitást

beviszi a modellbe - a grönlandi, azt mondhatjuk, hogy Grönland képviseli az elsődleges szerepet a Grönland-Antarktisz éghajlatváltozási kapcsolatban. A küszöbváltozó választásának helyességét természetesen statisztikai teszttel igazoltam.

Embedding dimenzió becslése

Több módszer ismeretes az embedding dimenzió meghatározására, melyek közül a legismertebb eljárás valószínűleg az ún. false nearest neighbors technika (Rhodes and Morari, 1997). Ez jól példázza a módszereknek azt a problémáját, hogy végső soron mindegyikükben valamilyen szubjektív választás szükséges. Ennek következtében nem igazán ismeretesek a becsült dimenzió statisztikai tulajdonságai. Kivételt képez Fueda and Yaganawa (2001) eljárása, ami konzisztens becslést szolgáltat, vagyis egyre növekvő adathossz esetén a becsült dimenzió és a valóságos dimenzió egyezésének valószínűsége egyre közelebb kerül az egyhez. Rögzített adathossz mellett azonban az eljárásban szereplő fontos részfeladat megoldási pontossága a dimenzió növekedésével rohamosan romlik, így gyakorlati alkalmazása csak rendkívül hosszú adatsor esetén tanácsolható. A közeljövő feladataként a korreláció fogalmának általánosítása alapján a parciális távolság korreláció (Székely and Rizzo, 2014) segítségével kívánom az embedding dimenzió becslését a korábbi módszerek problémáitól mentesen becsülni.

Ezzel kapcsolatban szívesen várom az érdeklődő hallgatókat diplomamunka, vagy tágabban értelmezve a feladatot, doktori téma keretében.

Hivatkozások

- Detle, H., Hallin, M., Kley, T., Volgushev, S., 2015: Of copulas, quantiles, ranks and spectra: An L_1 -approach to spectral analysis. *Bernoulli*, 21: 781–831.
- Fueda, K., Yanagawa, T., 2001: Estimating the embedding dimension and delay time from chaotic time series with dynamic noise. *J. Japan Statist. Soc.*, 31: 27–38.
- Lomb, N.R., 1976: Least-squares frequency analysis of unequally spaced data. *Astrophys. Space Sci.*, 39: 447–462.
- Makra, L., Matyasovszky, I., 2010: Assessment of daily ragweed pollen concentration with previous-day meteorological variables using regression and quantile regression analysis for Szeged, Hungary. *Aerobiologia*, 27: 247–259.
- Makra, L., Matyasovszky, I., Thibaudon, M., Bonini, M., 2011: Forecasting ragweed pollen characteristics with nonparametric regression methods over the most polluted areas of Europe. *Int. J. Biometeorol.*, 55: 361–371.
- Matyasovszky, I., 2010a: Improving the methodology for spectral analysis of climatic time series. *Theor. Appl. Climatol.*, 101: 281–287.
- Matyasovszky, I., 2010b: Trends, time-dependent and nonlinear time series models for NGRIP and VOSTOK paleoclimate data. *Theor. Appl. Climatol.*, 101: 433–443.
- Matyasovszky, I., 2011: Detecting abrupt climate changes on different time scales. *Theor. Appl. Climatol.*, 105: 445–454.
- Matyasovszky, I., 2013: Spectral analysis of unevenly spaced climatological time series. *Theor. Appl. Climatol.*, 111: 371–378.
- Rhodes, C., Morari, M., 1997: The fast nearest neighbors algorithm: An overview. *Compute. Chem. Eng.*, 21: S1149–S1154.
- Steig, E.J., Alley, R.B., 2002: Phase relationships between Antarctic and Greenland climate records. *Ann. Glaciol.*, 35: 451–456.
- Székely, G., Rizzo, M., 2014: Partial distance correlation with methods for dissimilarities. *Ann. Statist.*, 6: 2382–2412.
- Vautard, R., Ghil, M., 1989: Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Physica D*, 35: 395–424.