Opponensi vélemény

Matyasovszky István: „Néhány statisztikus módszer az elméleti és alkalmazott klimatológiai vizsgálatokban” című MTA doktori disszertációjáról

A dolgozat négy fejezetre tagolódik, melyeket összefoglalás zár le. Mindegyik fejezet szerkezete hasonló elveket követ, a matematikai statisztikai eljárás rövid ismertetését a konkrét meteorológiai, klimatológiai alkalmazás bemutatása és annak elemzése követi. A szerző a bevezető részben leszögezi, hogy „Mivel nem matematikai műről van szó, a matematikai eszközök tárgyalása csak olyan mértékben történik, ami feltétlenül szükséges a problémák és az alkalmazások megértéséhez”. Majd így folytatja: „Az elmélet és alkalmazás remélt egyensúlyának megtalálásában talán leginkább a nagyszámban felhasznált statisztikai próbák bemutatásának a mélysége jelentette a legnagyobb gondot.” Valóban, a kiterjedt irodalomjegyzék is mutatja, hogy a szerző jártas a releváns modern statisztikai irodalomban, és a disszertáció számos különböző módszer klimatológiai alkalmazását tartalmazza. Ezek között több olyan statisztikai eljárás is található, melyeket a klimatológiában először a szerző alkalmazott.

Azonban ezen a ponton az opponensnek is le kell szögeznie, hogy – mivel nem jártas a klimatológiában, és feltételezi, hogy nem ezért kapta meg a megtisztelő felkérést, hogy bírálója legyen a jelen MTA doktori disszertációnak – feladatául csak azt tűzheti ki, hogy az bemutatott statisztikai eljárások jogosultsága, a belőlük levont következtetések helyessége alapján formálja meg véleményét.

Elismerve ugyan azt, hogy szerteágazó területeket érintő, nagy ívű munkáról van szó, azonban nem osztom a szerző fent bemutatott véleményét. Azzal egyetértek, hogy a bonyolultabb statisztikai eljárások részletes bemutatása, ismertetése szétfeszítette volna a dolgozat kereteit, de azzal a választott megoldással már nem, hogy a szerző lényegében teljességgel elhagyta annak bemutatását, indoklását, hogy az egyes statisztikai eljárások alkalmazásának mik a feltételei, és hogy a konkrét esetekben azok – legalábbis közelítőleg – teljesülnek-e.

Az 1. fejezet a trend becslésével foglalkozik. A felhasznált statisztikai eljárás Fernandez and Fernandez (2004) összefoglaló munkájára épít. Azonban ez utóbbi munka is pontosabban definiálja az alkalmazott modellben a hibasorozatra tett feltevést. Álláspontom szerint lényeges lett volna pontosan meghatározni, hogy milyen feltételeknek kell teljesülniük. Továbbá szerintem a módszer leírásához mindenképpen hozzá kell, hogy tartozzék az is, hogy a trend becslése milyen statisztikai – esetleg aszimptotikus – tulajdonságokkal rendelkezik. Enélkül az alkalmazásokban kapott eredmények, az azok alapján megfogalmazott állítások pontossága, elfogadhatósága kérdéses lesz. Bántó pontatlanság a 8. oldal első bekezdésének végén megfogalmazott állítás, miszerint: „Ekkor viszont a lineáris trend jelenléte ellenőrizhető a fejezet elején említett t-próbával.” Hiszen

ehhez feltétlen szükség van a normális eloszlás feltételezésére, de erről ezen a ponton hallgat a disszertáció. (A teljesség kedvéért meg kell jegyezni, hogy később, más fejezetekben tartalmazza a dolgozat, hogy szükség van a Gauss-eloszlás feltételezésére, azon a fent idézett helyen ez elmarad.)

A trend becslésére három konkrét példát tartalmaz a dolgozat. Ezek

Északi Hemiszféra átlaghőmérséklete,
Hirtelen éghajlatváltozások,
Allergén pollenek.

A „Hirtelen éghajlatváltozások" alkalmazásban a trend különböző rezsimjeit kívánja a szerző statisztikai eljárás segítségével kideríteni. Ekkor fontos kérdés a feltételezett rezsimek darabszámának meghatározása. A 14. oldal 4. sora ad erre vonatkozóan támpontot. Eszerint akkor kell növelni a rezsimek számát, ha az új rezsim együtthatója „szignifikánsan különbözik zérustól". A dolgozat értelmezésében a szignifikánsan különbözik kifejezés azt jelenti, hogy jobb közelítést ad. Számomra a „szignifikáns" jelző statisztikai fogalom, melyhez szignifikancia-szint is tartozik. Amely viszont nem adható meg az eloszlásokra tett feltételezés nélkül.

A 14. oldal alján kezdődő mondat szerint: „A sávszélesség becslési tulajdonságainak ismeretében viszont nyilvánvaló, hogy a deriváltjában ugrásokkal rendelkező trendre vonatkozó sávszélesség nem lesz nagyobb, mint a sima trendre kapott becsült sávszélesség. Ezért ismét alkalmazható Fernandez and Fernandez (2004) módszere akár van ugrás, akár nincs." Nem meggyőző számomra ez a gondolatmenet.

Az „allergén pollenek" statisztikai elemzése szakasz átfogó képet akar nyerni a regionális pollenflóráról. Ehhez 11 év napi pollenszámai adják az alapstatisztikai adatokat 19 taxon esetében. Az adott taxonra az egyes évek ugyanazon napján észlelt pollenszámok 11 hosszúságú idősort szolgáltatnak. A nem-paraméteres Mann-Kendall próba segítségével ellenőrizhető a trend megléte. A trend hiánya esetén a próbastatisztika aszimptotikusan normális eloszlású. Így a dolgozat abból a feltevésből indul ki, hogy a nullhipotézis fennállása esetén (nincsen trend) – pollenszezon idejében – az év egyes napjaihoz – a vonatkozó Mann-Kendall próbastatiszika értékei által – hozzárendelhető nulla várható értékű normális eloszlású, de korrelált sorozat. Erre a sorozatra AR(1) folyamatot illesztve becsülhető a szórás és így t-próba végezhető. Azonban a 11 hosszúságú idősorra számolt Mann-Kendall próbastatisztika eloszlása szerintem még igen távol lehet az aszimptotikus – normális – eloszlástól. Továbbá mi indokolja az AR(1) folyamat alkalmazását? Hogyan lehet beállítani az fenti „két lépcsős" próba esetén az egyes szignifikancia szintekhez alkalmazandó kritikus értékeket? Végezetül mi indokolja azt, hogy a kiinduló idősorokat az egyes évek ugyanazon naptári napjain mért pollenszámok adják? Nem lenne-e helyesebb a pollenszezon kezdete alapján, vagy esetleg a pollenszemek nagyságának éves lefutása alapján megfeleltetni egymásnak az egyes évek napjait?

A fejezet záró mondata utal Makra et al. (2011a) tanulmányára, de sajnos a statisztikai eljárás vonatkozásában az nem tartalmaz többet a dolgozatban közöltnél.

A 2. fejezet tárgya regresszió, kvantilis regresszió, illetve ennek alkalmazása a napi parlagfű koncentrációra. Az kapott eredmények igen érdekesek.

A 3. fejezet tárgya a Spektrálanalízis alkalmazása a klimatológiában. Nem túl lényeges, de vitatkoznom kell a 34. oldalon tett azon megállapítással, miszerint tetszőleges stacionárius idősor véges diszkrét spektrumú és „folytonos spektrumú, tehát megszámlálhatatlan periodikus tag összege". Praktikus szempontból valóban csak a véges diszkrét spektrum, illetve az abszolút folytonos spektrum az érdekes, azok statisztikai elemzése végezhető el, azonban elméleti szempontból a kijelentés pontatlan. A 35. oldal alján szerepel az a kijelentés, hogy a „spektrális sűrűségfüggvény becslése analóg a trendfüggvény becslésével". Ennek alapján a szerző közöl is egy képletet, amelyben az idő a körfrekvencia helyettesíti. Ha jól látom, a szerző később nem használja ezt a spektrális sűrűségfüggvény becslési képletet, amely teljességgel eltér az irodalomban szokásosaktól.

Az elméleti összefoglaló három kérdéskört tárgyal. Ezek: robusztus becslés, nem ekvidisztáns időpontokban rendelkezésre álló adatsor, vörös zaj becslése. Kiemelendő, hogy itt a módszerek leírása, az felhasznált irodalomra vonatkozó hivatkozások pontosabbak, mint az előző fejezetekben, így jobban nyomon követhetőek a statisztikai módszerek részletei, azok alkalmazhatóságának feltételei. Ugyanakkor a 43. oldal közepén a szerző kijelenti, hogy a becsült Fourier-együttható kovarianciamátrixa – $D^{-1}Z^{T}BZD^{-1}$ – aszimptotikusan a $\pi\, k(\lambda)D^{-1}$ értékhez tart. A részleteket illetően utalva Matyasovszky (2012b) cikkére. (Valószínűleg a 2013a lenne itt a jó hivatkozás.) Ugyanakkor az idézett cikk nem tartalmaz erre az állításra vonatkozó matematikai levezetést.

Igen érdekes gondolat a vörös zaj alkalmazása a klimatológiai idősorokban. A módszer alapja az izoton regresszió. A hivatkozott Zhao and Woodroofe (2012) cikk gyengén függő hibasorozat esetére közöl aszimptotikus eredményeket, azonban nem a spektrum becslése esetére, hanem valamely rögzített n hosszúságú idősor trendjének becslése esetére. Ugyanígy Álvarez and Yohai (2011) a „klasszikus" regresszióra vonatkozó robusztus becsléssel foglalkozik. Milyen matematikai eredmények érhetőek el a szakirodalomban a izoton regressziónak a spektrál-sűrűségfüggvényre való átvitelére vonatkozóan? Az elméleti bemutatás után a szerző számos érdekes alkalmazást tárgyal. Ezek: a NAO index, a GISP2 Oxigén izotóp adatok a 15 ezer – 60 ezer évvel ezelőtti időszakra, Vostok deuterium tartalom adatsora az elmúlt 422 766 évben, az Északi Hemiszféra hőmérséklete a 200-1995 évekre.

A kapott eredmények érdekesek, számos esetben jó megfelelést mutatnak a szakirodalomban korábban közöltekkel, illetve fontos új szempontokhoz szolgáltatnak támpontot.

A 4. fejezet az autoregresszív idősor modellezés lehetséges általánosításával foglalkozik. Az áttekintett modellek: gamma eloszlású AR(1) modell, nemlineáris AR modell, treshold autoregresszív modell, ARCH modell. A leírás itt is matematikailag jobban kiérlelt, mint a korábbi fejezetekben. Egyetlen ponttal kapcsolatban merült fel hiányérzet. A 65. oldalon közölt szerint az „autoregresszió rendje és a rezsimek száma Akaike (1974) nyomán becsülhető.". Nem világos számomra, hogy miért? Mik azok a feltételek, amelyek

ahhoz kellenek, hogy az AIC kritérium alkalmazható legyen és a jelen esetben hogyan teljesülnek ezek? Ugyancsak hiányoltam pontosabb irodalmi hivatkozást a 66. oldal alján tett kijelentéssel kapcsolatban, miszerint a (4.15) „próbastatisztika aszimptotikusan χ^2-eloszlású.

Véleményem szerint az ebben a fejezetben szereplő alkalmazások során kapott eredmények a legjelentősebbek. Ezek: napi parlagfű koncentráció Szeged környékén, NGRIP és Vostok adatok együttes elemzése, hirtelen éghajlatváltozás, Dansgaard.Oeschger-események.

Az eredmények jelentőségét nem csökkenti az a tény, hogy szerintem a 76. oldal alján – egy idősor linearitásával kapcsolatban – tett megállapítás pontatlan. Eszerint: „Ismert ugyanis, hogy egy lineáris folyamat esetében az időskála megfordítható, vagyis a folyamat egy elemének az időben rákövetkező elemekkel való közelítése ugyanazt a modellt eredményezi, mint a megelőző elemekkel való közelítése." Az adódó modell nyilvánvalóan nem lesz ugyanaz, hiszen például a hibafolyamat az egyik esetben a múltbeli folyamatértékek függvénye, az idő megfordítása esetén pedig a jövőbelieké.

A disszertáció 100 számozott lapot tartalmaz, irodalomjegyzéke 132 tételt sorol fel. Elírás csak igen ritkán fordul elő benne, stílusa körültekintő, gondosan szerkesztett munka.

Összefoglalva, véleményem szerint a dolgozat két, igen eltérő szempont alapján értékelhető. Az egyik, az alkalmazott matematikai statisztikai módszerek megválasztása, a módszerek bemutatása, alkalmazhatósági feltételeinek ellenőrzése. A másik, a bemutatott módszerek alkalmazásával elért eredmények közlése, azok klimatológiai jelentősége, a korábbi, a szakirodalomban megtalálható eredményekkel egyezés, azok továbbfejlesztése.

Amint azt fentebb részletesen kifejtettem, az első szempont vonatkozásában sok hiányérzetem van, és nem értek egyet a szerzőnek a bevezetőben lerögzített, a matematikai statisztikai próbák tárgyalásával kapcsolatban kifejtett elvével.

Ha azonban elfogadjuk azt, hogy a közölt statisztikai módszerek alkalmazhatóságának feltételei teljesülnek az egyes konkrét statisztikai adatsorok esetén, tehát a második szempont alapján értékeljük a disszertációt, akkor megállapítható, hogy a szerző úttörő munkát végzett a modern matematikai statisztikai módszereknek a klimatológiában történő meghonosítása területén.

Ez utóbbit jelentősebbnek tartom, ezért javaslom a dolgozat nyilvános vitára bocsátását, és a fokozat odaítélését.

Budapest, 2014. március 14.

Michaletzky György