

VÉLEMÉNY

Dr.Kovács József „Néhány adatelemző módszer alkalmazása földtudományi feladatok megoldására, különös tekintettel csoportosító eljárásokra” című MTA doktori értekezéséről

1. ÁLTALÁNOS MEGJEGYZÉSEK

A 146 számozott oldal terjedelmű disszertáció egy rövid Bevezetés után az 1. fejezetben a földtudományok területén használatos adatelemző módszerek és szoftverek ismertségének helyzetét tárgyalja, a 2. fejezet összefoglalást ad a statisztikai modellezési feladatok megközelítési problémáiról. A dolgozat tartalmaz még 4 érdemi fejezetet, az elért eredmények tézisszerű összefoglalását (7. fejezet), továbbá egy közel 250 tételből álló hivatkozások jegyzékét.

A dolgozatban vizsgált kutatási feladatok jellegüknél fogva csak csoportmunkával oldhatók meg, ezért a publikációk ezen a területen jellemzően többszerzősek. A disszertáció témakörében felsorolt 32 saját publikáció – néhány kivételtől eltekintve – az elmúlt másfél évtizedben jelent meg, közülük Jelölt 16-ban elsőhelyes szerző. Azt is fontos megemlíteni itt, hogy társszerzői között több doktorandusza és hallgatója volt.

Az eredmények ismertetésénél a szerző tisztázza a szerepét a dolgozatban felhasznált publikációk vonatkozásában. Az 1-7. tézisek esetében csak elsőszerzős publikációkra hivatkozik, a 8-9. tézisben megfogalmazott eredményeinek interpretálását paleontológus és fizikus kollégák segítették. Ezek alapján megállapítható, hogy a tézisekben megfogalmazott eredmények elérésében Jelölt hozzájárulása meghatározó volt.

Az értekezés szép kivitelű, megfelelően strukturált, az ábrák és diagramok jól és szépen megrajzoltak, jelentős segítséget nyújtanak a sokirányú összetevőből álló anyag követéséhez. Az értekezés formai szempontból megfelel az általános követelményeknek.

Kritikai észrevételként megjegyzem, hogy a tézispontok sorrendje messze nem követi a disszertáció fejezeteiben kifejtett eredmények sorrendjét, ami megnehezíti az opponens számára az egyes fejezetekben foglalt eredmények és a tézisek összevetését. Pl. az 1. tézis tartalmazza a dolgozat egyik fontos módszertani eredményének a leírását, a kombinált klaszter- és diszkriminanciaanalízis (CCDA) módszerét, ami pedig a 6. fejezet első pontja. Megjegyzendő az is, hogy néhány helyen apróbb megfogalmazási pontatlanságok fordulnak elő, amelyek azonban nincsenek befolyással a disszertáció eredményeire és ezekre részletesen nem térek ki.

2. A TÉMAVÁLASZTÁS, AZ ÉRTEKEZÉS ÉRTÉKELÉSE

Témaválasztás

Jelölt doktori értekezése a témakör aktuális és mind elméleti, módszertani, mind pedig gyakorlati modellezési szempontból jelentős feladatokkal foglalkozik. A föld- és környezettudományi folyamatok modellezése számos kihívással párosul, amely minden vonatkozásában különös figyelmet és szakértelmet követel. Ezen a tudományterületen

viszonylag kevésbé elterjedtek a többváltozós statisztikai analízis egyes módszerei, ezért a dolgozatban szereplő elemzési eljárások használata és új adatelemző módszerek kidolgozása, valamint különböző feladatokon való részletes bemutatása jelentősen hozzájárulhat a kutatók számára e tudományterületen az adatokban rejlő összefüggések feltárásához és mélyebb elemzéséhez.

A vizsgálatok jellege

A vizsgálatok alapkérdése a modellek, vagyis adott paraméterek együttesével megadott matematikai konstrukciók megválasztása, amelyek meghatározóan valószínűségelméleti, matematikai statisztikai és a paraméterek időben történő változásának modellezése esetében az idősor-analízisbeli (vagy a 4. fejezetben kiegészítésképpen más matematikai) eszközökre támaszkodnak és amelyeket a vizsgálat céljának megfelelően kell megválasztani. Fontos része a vizsgálatoknak a kapott modell kiválasztása objektív kritériumok alapján és különösen, az eredmények interpretálása. Meg kell említeni azonban azt is, hogy a modellalkotás egy összetett folyamat, amelynek jelentős része a vizsgált objektumokat jellemző paraméterek mérési eredményinek előzetes elemzése, ahol sok esetben hibás adatok pótlására, illetve az ún. „outlier” adatok szűrésére is megfelelő eljárással sort kell keríteni, figyelembe véve a vizsgált paraméterek jellegét és a vizsgálatban elfoglalt szerepét.

Értelemszerűen az értekezésben nem találhatók új elméleti matematikai megállapítások, ugyanakkor új vizsgálati módszerek és a használatukat lehetővé tevő számítógépes programok kerültek kidolgozásra benne. Ezen felül ismert matematikai apparátusok alkalmazása a modellezési technikával kombinálva és konkrét feladatra adaptálva újszerű megoldásokat eredményeztek. A dolgozatban alkalmazott módszerek és eljárások használata a föld- és környezettudományok területén jelentősen hozzájárulhat az megfigyelési eredmények sokrétű elemzéséhez és megértéséhez.

Vizsgálati módszerek

A többváltozós statisztikai módszerek közül a dolgozatban alkalmazott legfontosabb klasszikus vizsgálati módszerek a dimenziószám csökkentő és csoportosító eljárások, amelyek további statisztikai elemzések kiindulópontjaiként is szolgálhatnak: főkomponens analízis (PCA), lineáris diszkriminancia-analízis (LDA), klaszteranalízis (CA), hierarchikus klaszteranalízis (HCA), k-középpontú klaszteranalízis és többdimenziós skálázás (MDS). Ezek célja részben adattömörítés a további vizsgálatokhoz, részben pedig a megfigyelt mérési eredmények olyan csoportosítása, amely a kialakított csoportokat a lehető legjobban megkülönbözteti, illetve az, hogy az egyes csoportokon belüli adatok bizonyos értelemben legjobban hasonlítanak egymáshoz. Mindezek a megfigyelési adatokban rejlő lényegkiemelést szolgálják.

Az egyes statisztikai eljárások más-más matematikai megközelítésre támaszkodnak. A PCA a változók közötti kovarianciamátrix spektrális felbontását felhasználva a mintatér lineáris transzformációjával egy olyan új alacsonyabb dimenziójú, főkomponensekből álló vektorváltozót definiálhat, amely az eredeti mintában rejlő variancia jelentős részét magyarázza. Az LDA módszer a független változók olyan lineáris kombinációját keresi, amely a függő változó alapján a mintatéren a lehető legjobban csoportosítja a megfigyelési eredményeket. A kapott lineáris kombináció felhasználható a további osztályozás előtt a dimenzió csökkentésére is. A klaszteranalízis különböző módszerei (HCA, k-középpontú) a megfigyelési eredményeket a közöttük levő távolság vagy hasonlóság alapján sorolja különböző csoportokba a megfigyelt adatok felhasználásával. Az eredményezett csoportok

számát és az odartartozó elemeket az alkalmazott eljárás hozza létre. A lineáris modellektől eltérően (PCA, LDA) az MDS eljárás célja az, hogy az adott feladatra értelmezett távolság-, vagy hasonlóság-mérték alapján a megfigyelési adatokat olyan alacsonyabb dimenziójú térben pontfelhőként reprezentálja úgy, hogy a kapott pontok minél jobban visszatükrözzék az eredeti adatok közötti távolsági- vagy hasonlósági mértékeket. Ez az eljárás 3-nál nem nagyobb dimenzióban vizuális elemzést is lehetővé teszi.

Az alkalmazott statisztikai eljárások között kiemelt szerepet tölt be a kidolgozott CCDA (1. tétel), a klaszter- és diszkriminanciaanalízis (CA és LDA) kombinálásából eredő statisztikai módszer, amely többdimenziós vektorváltozó megfigyelt eredményei alapján a mérőhelyek bizonyos szempont szerinti optimális csoportosítását valósítja meg úgy, hogy a csoportok száma optimális legyen amellett, hogy a legnagyobb homogén csoportok alakuljanak ki. Jellemzően ilyen helyzettel találkozhatunk disszertációban vizsgált vízminták esetében (Balaton, Kis-Balaton, Fertő tó, Duna, Budai kasztvizek), ahol a többdimenziós vektorváltozó egyes koordinátái a vízösszetétel különböző mért paramétereit jelentik. A CCDA módszer részletes bemutatására a 6. fejezet első pontjában kerül sor, az egyes lépéseit a 6.1 folyamatábra jeleníti meg.

Külön értéket képvisel az, hogy nem csak maga a CCDA módszer lett kidolgozva, hanem elkészültek a megvalósításához nélkülözhetetlen és nyilvánosan hozzáférhető program kódok R programozási nyelven, valamint program dokumentációja is, melyek lehetővé teszik más kutatók számára is a CCDA módszer eredményes alkalmazását.

A többváltozós statisztikai analízis módszereitől elütő megközelítés vezetett eredményre a 4. fejezetben a permafroszt és a talaj hőmérsékleti folyamatai közötti kapcsolat vizsgálatánál, ami a modell fizikai hátterét a hőterjedést időben leíró parciális differenciálegyenletre, a statisztikai elemzés pedig idősoranalízis módszereire támaszkodott.

Eredmények

Az 1. fejezet, amely az adatelemző módszerek ismertségének és használatának kérdéseivel foglalkozik, kitérve a fontosabb adatelemző szoftverek kérdésére is, és ad egy általános képet erről. Ennek alapján az látszik, hogy a többváltozós statisztikai analízis módszerei és adaptált változatai kevésbé elterjedtek, azonban fontos szerephez juthatnak a kutatott tudományterületen. Ugyanakkor ez a fejezet a maga nemében kicsit szokatlan, de teljes mértékben hozzátartozik a dolgozat jelentőségének megítéléséhez.

A 2. fejezet a föld- és környezettudományok különböző területein előforduló adatok jellegére vonatkozóan megfogalmazza annak szükségességét, hogy az adatok értelmezésénél és vizsgálatainál figyelembe kell venni a minták véletlen jellegét. Helyénvalónak tartom a szerző által kifejtett adatelemzési módszertani megközelítést (lásd 2.3. ábra), külön kiemelve az egyes lépésekben rejlő problematikát és egyben utalva a későbbi fejezetekben elfoglalt helyére.

3-6. fejezetek a föld- és környezettudományok különböző problémáit vizsgálják; az egyes fejezetek, illetve azok részei ugyanakkor önmagukban is kerek egész tudományos elemzés eredményét mutatják. Értékelésükkel kapcsolatban nem térek ki azok szaktudománybeli interpretációs kérdéseire.

A 3. fejezetben (8. tétel) a felső kréta korból származó, viszonylag nemrégén feltárt (1999-2011) rovarpete-maradványok morfológiai adatainak vizsgálata kerül bemutatásra. A standardizált változókra alkalmazott és Ward eljárással (négyzetes távolság mellett) végrehajtott HCA eljárás után az LDA módszer használatával sikerült megadni a 97 leletből

álló együttes morfometriai leírását. Az előzetes elemzések alapján a szélsőértékeket mutató és a HCA által elkülönített 5 tételből álló csoport elhagyásával a HCA 3 csoportot határoz meg, melyek a PCA alkalmazásával jól magyarázhatók a pete érési folyamatainak fázisaival.

A 4. fejezet (9. tézis) az alapvető feladat egy dél-amerikai hegy különböző magasságaiban és a talaj különböző mélységű mintavételi pontjaiban mért hosszú idejű hőmérsékleti idősorok alapján a talaj hőmérsékleti folyamatainak változása és a permafroszt közötti összefüggés vizsgálata. A fejezet új eredményeket tartalmaz a hőmérsékleti folyamatok és a permafroszt változásainak jellegére nézve, melyek a hőterjedést leíró fizikai modell és az idősorokra alkalmazott wavelet transzformáció felhasználásával születtek.

Az 5. fejezet (6. tézis) a Balaton vízminőségét térben és időben leíró paraméterek vizsgálatával foglalkozik. A feladat itt a hasonlóságot mutató adatok alapján a mérőhelyek csoportosítása. Ward klaszterező (HCA) módszerének alkalmazása négyzetes euklideszi távolság mellett új eredményekhez vezetett a mérőhelyek csoportosításával kapcsolatban, illetve a Balatonra nézve az optimális mérőhely-szám és mérőhelyek kijelölésére.

A 6. fejezet 6.1. pontja (1. és 5. tézis) a CCDA – a gyakorlati alkalmazhatóságot is megvilágító – módszer részletes leírását tartalmazza (erre korábban, a vizsgálati módszerek leírásánál kitértem). A 6.1. pont részben, továbbá a 6.2 és 6.3 pontok a CCDA módszer felhasználásával készült komplex kutatási eredményeket foglalják össze, amelyek a Balaton (6. tézis), Kis-Balaton (7. tézis), Fertő tó és a Duna magyarországi monitoringhálózata (4. tézis) vízminőségi paramétereire vonatkozó mérési eredmények alapján születtek. A Budapest területén található 27 termálkút, illetve forrás vízösszetételi és hőmérsékleti idősoraira (1960–2009) vonatkozó eredményeket tartalmazzák a 2. és 3. tézisek.

3. KÉRDÉSEK A JELÖLTHÖZ:

1. A permafroszt vizsgálatánál a ψ wavelet függvény megválasztására mennyire érzékeny a spektrum becslés (42. o.)? Vannak-e olyan eredmények, amelyek a hőmérsékleti folyamatokat általánosított ARMA folyamatokkal modellezik?
2. A disszertációbeli feladatokra alkalmazott klaszterező eljárásoknál felmerült-e más, a négyzetes távolságtól eltérő távolságfogalmak használata, pl. a Mahalanobis-, vagy Manhattan-távolság?
3. Vannak-e eredmények a Velencei tó vízösszetételére nézve és azok mennyire különböz(het)nek a disszertációban vizsgált tavakétól?

4. ÖSSZEFOGLALÓ ÉRTÉKELÉS

Az értekezés egységes rendszerben, megfelelő formában mutatja be Jelöltnek azon vizsgálati módszereit és tudományos eredményeit, amelyek a föld- és környezettudomány különböző problémáira adnak választ. Az értekezésben felhasznált többváltozós statisztikai és idősoranalízis módszerekkel és különböző számítógépes eljárásokkal vizsgált kutatási problémák elméleti és gyakorlati szempontból is egyaránt fontosak az alkalmazott módszertannal együtt. Elmondható, hogy a Jelölt a matematikai eszközöket, vizsgálati módszereket jól választotta meg és megfelelően alkalmazta, ezért a levont statisztikai következtetések korrektek. A tézisekben megfogalmazott tudományos eredmények megalapozottak, eléérésükben Jelöltnek meghatározó szerepe volt, továbbá megfelelő módon publikálásra kerültek, melyeket elfogadom új tudományos eredményekként.

Összefoglalóan megállapítom, hogy Kovács József értekezése mind formai, mind tartalmi vonatkozásban kielégíti az MTA doktori szabályzatában előírt követelményeket. Jelölt a PhD fokozat megszerzése óta jelentős tudományos eredményekkel és gyakorlati alkalmazásokkal gyarapította a földtudomány tudományterületét. Mindezek alapján javaslom a nyilvános vita kitűzését és Kovács József részére az MTA doktora cím odaítélését.

Budapest, 2020. március 19.



Dr.Szeidl László

a mat. tud. (MTA) doktora