BÍRÁLAT
Dr. KOVÁCS JÓZSEF
NÉHÁNY ADATELEMZŐ MÓDSZER ALKALMAZÁSA FÖLDTUDOMÁNYI FELADATOK MEGOLDÁSÁRA, KÜLÖNÖS TEKINTETTEL A CSOPORTOSÍTÓ ELJÁRÁSOKRA
című
MTA doktori disszertációjáról

1 Vélemény a dolgozat felépítéséről, stílusáról, arányairól, eredetiségéről

A Magyar Tudományos Akadémia (MTA) előírása szerint az "MTA doktora cím megszerzése iránti kérelemhez olyan doktori művet kell mellékelni, amely önmagában véve is alkalmas a kérelmező eredeti tudományos teljesítményének értékelésére, megítélésére, valamely tudományos kérdés megoldásának bemutatása alapján." Az eredeti tudományos teljesítmény egyértelmű értékelése a nemzetközi impakt faktoros szaklapokban történő tudományos közlemények rendszeres és magas szintű publikálása, melyet az MTA doktori cím megszerzése iránti kérelem habilitációs dokumentációja számszerűen kér is. Ilyen módon természetes, hogy egy doktori értekezés akár vezető tudományos szaklapokban megjelent közleményeken alapulhat. A jelölt értekezése is ilyen. Az értekezés mindenhol pontosan jelöli az adott szövegrész forrását és külön jelöli, ahol még nem publikált szöveg található. Ilyen módon az értekezés a legpontosabban tesz eleget az MTA előírásnak és teljes összhangban van a habilitációs követelményekkel is.

„Az értekezés célja, hogy bemutassa egy-egy szakmai kérdés eldöntése esetén az adatelemző módszerek használhatóságát, és érzékeltesse jelentőségüket tudományterületünkön…” (4.oldal). A dolgozat első fejezete lényegében a cél hasznosságát mutatja be egy kérdőíves felmérés eredményeinek elemzésével. Ennek a felmérésnek egyik eredménye szerint „a megkérdezett szakemberek igénylik és támogatják egy olyan, segédanyagként is használható folyamatábra elkészítését, ami rávilágít arra, hogy ismert és általánosan elterjedt adatelemző programcsomagok által kínált módszerek közül adott típus adathalmazra milyen adatelemző módszereket milyen sorrendben célszerű alkalmazni.”

A második fejezetben a szerző egy általános ajánlású folyamat ábrát mutat be, amelyben összefoglalja a „földtudományokban leggyakrabban előforduló több paraméterrel jellemezhető, térben és időben levő megfigyelések elemzésének sorrendjét”. A harmadik fejezetben a többváltozós eljárásokon alapuló morfometriai elemzésekre olvashatunk egy példát. A fejezet egy (a dolgozat írásakor) megjelenés alatt álló dolgozat magyar nyelvű változata. Több idősor mérési eredményeinek feldolgozására példa a negyedik fejezet, amely az Atacama sivatag permafrost folyamatainak waveleth analízissel történő elemezési eredményeit mutatja be. A fejezet a jelölt társszerzőségében készült öt (közülük a dolgozat benyújtásának időpontjában három publikálásra elfogadott) dolgozat alapján. Ugyancsak egy már megjelent dolgozat magyar nyelvű változata az ötödik fejezet is, amely a Balaton több ponton mért vízminőségi adatsorainak többváltozós statisztikai feldolgozását mutatja be.

A hatodik fejezet a szerző saját fejlesztésű elemzési rendszerének, a Kombinált Klaszter és Diszkriminancia Analízisnek (CCDA) három alkalmazási területéről számol be ugyancsak publikált dolgozatok alapján.

A disszertáció egy kilenc pontot tartalmazó összefoglalóval zárul, amely a tézisfüzet tanúsága szerint a jelölt kilenc tézisének felel meg.

A dolgozat a bevezetőben megjelölt célt 132 oldalon, 62 ábra és 16 táblázat segítségével dolgozza fel. Irodalomjegyzéke 140 hivatkozott tétel. A dolgozat szövegtörzsében minden irodalmi tételre történt hivatkozás. A dolgozat stílusa jó, még a kívülálló számára is olvasmányos, a fogalmazás szabatos, tömör, de mindvégig érthető. A dolgozat egyik nagy formai pozitívuma, hogy elütéseket nem tartalmaz.

A dolgozat megközelítésének újszerűségét, a témaválasztás mellett, a vizsgálatokban használt módszerek komplexitása adja. Külön érdemes hangsúlyozni, hogy a munkában alkalmazott statisztikai eljárások végrehajtása precíz, a protokollnak megfelelő. Ahol nem vagy „nem annyira", azok a „Tételes megjegyzésekben" képezik a kritika tárgyát.

A dolgozatban kifejtett gondolatok és feldolgozott adatok, a szerző irodalmi munkásságának hivatkozott publikációi alapján, teljesen nyilvánvalóan a szerző sajátjai.

2 Tételes észrevételek

2.1 Az adatelemző eljárások ismertségének és használatának kérdései

A fejezet egy kérdőíves felmérés eredményét ismerteti. Lényegében arra keres választ, hogy „mennyire elfogadottak ezek a módszerek (t.i. a statisztikai módszerek) a vízminőségi adatokkal dolgozó kollégák körében". A kérdőív hazai kitöltői a vízügyi igazgatóságok 39 szakembere volt, míg a külföldi 19 válaszadó jobbára kutatóintézetek egyetemek dolgozói voltak. A válaszokat szép diagramok szemléltetik. Az eredmények a hazai szakemberek statisztikai ismereteire nézve nem kifejezetten hízelgők.

Ennek kapcsán a kérdésem az, hogy vajon a hazai válaszadók összetétele (hatósági feladatokat ellátó szakemberekről van szó), vajon alkalmas-e az általánosításra. Egyik oldalról a hatósági feladatok ellátása nem feltétlenül jelent adatfeldolgozást is. Másfelől a hazai hidrológia és hidrogeológia olyan szervezeteiben, mint a pl. GOLDER Kft, Smaragd Kft, a HYDROInform Kft vagy a Magyar Bányászati és Földtani Hivatal hidrológiai szervezetei a sztochasztikus modellek és a gépi tanuló algoritmusok széles körét alkalmazzák, amelyek pontosan olyan statisztikai elővizsgálatokat igényelnek, amelyekre a kérdőív vonatkozott.

2.2 Az adattípus fogalom használata

A fejezetben szerző először áttekinti a föld-és környezettudományok területén előforduló „adattípusokat". Meglátása szerint az adattípusok rendszerét a 2.2.ábrán közölt négydimenziós adathalmaz foglalja össze.

Úgy érzem, hogy adattípus fogalom használata ebben a fejezetben nem kifejezetten szerencsés, hiszen adattípus alatt rendszerint a Bool-típusra, vagy intervallum típusra, vagy skalárra stb. gondolunk. A szerző által használt „adattípus" fogalom sokkal inkább megfelel ***Bárdossy és Fodor (2004)*** *által bevezetett „elemzési rendszerek" fogalomnak. Az idézett szerzők szerint a* ***skaláris értékelések*** *során nem eltekintünk a minták tér és időbeli helyzetétől, csak az adatok numerikus értékére koncentrálunk. Ebbe a körbe tartoznak a földtani objektumok kémiai, mineralógiai vagy fizikai tulajdonságainak statisztikai elemzései.* ***A minták térbeli értékelésekor*** *elfogadjuk, hogy minden adathoz tartozik egy térbeli koordináta és a változók kapcsolatát térben értékeljük.* ***A minták tér-és időbeli értékelésekor*** *a térbeli koordináták mellett a mérési eredmények időben is változnak.*

2.3 Adatelemzési „protokolok”

A második fejezet további részében a szerző három alternatívában mutat be egy-egy adatelemzési „protokolt”.

Ezek a ’protokolok’ gyakorlati ajánlások, ám *sok szempontból nem harmonizálnak a statisztikai feldolgozások EDA (Exploratory Data Analysis= Feltáró adat elemzés) szemléletével. Az EDA sokkal inkább szemlélet, mintsem módszertani gyűjtemény. Tukey 1961-ben definiálta ezt a megközelítést. Ez tartalmaz grafikus eljárásokat (hisztogram, box-plot, Pareto-chart, SPC chart stb.) és dimenzió csökkentő eljárásokat (pl. többdimenziós skálázás, PCA, FA, klaszteranalízisek stb.). Az EDA szemlélete szerint minden egyes diszkrét eljárásnak van egy „protokolja” ám a diszkrét eljárások protokol-ajánlása helyett, napjaink feldolgozásában az adatok „tulajdonsága” és a vizsgálat célja határozza a feldolgozás menetét. Pl. a permeabilitás vagy szivárgási tényező esetében a szakmailag megalapozott, de az adatok eloszlása szempontjából extrémnek minősülő értékek határozzák meg az áramlási rendet. Ilyen esetekben a klasszikus átlagok helyett célszerű a maximum likelihood becslésű középértékek valamelyikének alkalmazása és a feldolgozás is a maximum likelihood elv alapján történik.*

Az első „protokol” a több mintavételi pontban rendelkezésre állóadathalmazok feldolgozásával, a második az egy adott időpotban (időintervallumban) rendelkezésre álló több mintavételi pont elemzésével, míg a harmadik „protokol” kifejezetten az idősorok elemzésének áttekintésével foglalkozik. Nagyon logikus áttekintés ez. A javasolt eljárásrenddel lényegében egyet is értek az első és a harmadik „protokol” esetében. Azonban a második „protokol” szellemét ebben a formában már problémásnak vélem.

Ez a „protokol” (egy adott időpont/intervallum térbeli adatainak kezelése) lényegében a geostatisztikai feldolgozások témaköre. Úgy tűnik, hogy az ehhez ajánlott feldolgozási rendszer kissé távol áll a geostatisztika paradigmáitól. *Meglepő módon a szerző nem foglalkozik azzal a kérdéskörrel, ami a geostatisztika és a statisztika együttes alkalmazását erősen korlátozza. A statisztikai és geostatisztikai szemlélet ugyanis egészen másként viszonyul a térbeli adatokhoz. A statisztikai megközelítésben egy attribútum több adatponti (térbeli) értékét úgy tekintjük, hogy az adott valószínűségi változót vizsgáljuk az adatponti értékeken keresztül. Ugyanezt a rendszert a geostatisztika egy többváltozós valószínűségi függvény egy realizációjának tekinti. Ennek a többváltozós valószínűségi függvénynek annyi változója van, ahány adatpontból ismert a jelenség (pl. Journel 1986, Olea 1999, Goovaerts 1974 stb.). Vagyis ha pl. a Na-tartalmat 10 megfigyelő pontban mértük, akkor a statisztika oldaláról egy 10 elemű mintánk van a Na-tartalomra, a geostatisztika oldaláról pedig 10 valószínűségi változó egy véletlen realizációja áll rendelkezésünkre. Ez a modell elfogadja és alkalmazza azt a tényt, hogy a mérési eredmények között sztochasztikus kapcsolat van. Ezt a kapcsolatot a sztochasztikus modell választásától függően jellemezhetjük autokorrelációval, a gyenge stacionaritás elfogadása mellett, vagy félvariogrammal vagy autokovarianciával, a belső hipotézis elfogadása esetén. Ha a térbeli kapcsolatot a félvariogrammal jellemezzük, mint ahogy a szerző is ajánlja, akkor elfogadjuk, hogy a szórás egy trend/drift mentén változhat. Ez viszont kizárja mind a főkomponens analízis, mind a diszkriminancia analízis alkalmazhatóságát (pl. Olea 1999, Goovaerts 1997, Caers 2011). Emiatt a geostatisztikai modellekben soha nem alkalmaznak sem statisztikai próbákat, sem olyan statisztikai megoldásokat, amelyek a szórás stacionaritását igénylik. A statisztikai próbákat pl. a box-plot technika vagy a Q-Q és P-P diagramok helyettesítik, míg a „többváltozós eljárások” szerepét a koszimulációk, vagy a multiple-point szimulációk veszik át. Ugyancsak zavaró, ahogy a szerző a geostatisztikai feldolgozást két kimeneti eredményre szűkíti: a félvariogramra és a kontúrtérképre. Napjaink geostatisztikájának célja a vizsgált/mintázott jelenség kapcsán a méréshez/kiterjesztéshez kapcsolódó*

bizonytalanság megjelenítése és jellemzése. E cél elérésében a geostatisztikai sztochasztikus szimulációk szerepe a meghatározó, a „kontúrtérkép" csak egy köztes eredmény.

2.4 Főkomponens analízis kontra faktor analízis problémája.

A szerző az első protokolban mind a faktor mind a főkomponens analízis alkalmazását egyaránt javasolja. Ugyanakkor a 3.fejezetben közölt esettanulmányban a főkomponens elemzést használja. Jóllehet a KMO statisztika azt is mutatta, hogy az adatok bizonytalansággal terheltek, a szerző és társai mégis a teljes variancia felbontását végezték el a bizonytalanság figyelmen kívül hagyásával. Faktor analízis választása mellett ugyanakkor a mérések bizonytalansága meg tudott volna jelenni az értelmezésben is.

Valóban, hiszen a főkomponens analízis során bármely adott változó varianciát teljes egészében leírjuk a főkomponensekkel. Ezzel elfogadjuk, hogy minden mérésünk tökéletes, a bizonytalanságnak nem adunk teret. A faktor analízisek során bármely változó teljes varianciáját három tag összegére bontjuk: közös variancia, egyedi variancia, hiba variancia. A „közös variancia" (=kommunlitás) az adott változó varianciájának az a része, amelyet a faktorokkal leírunk. Ez egy adott változóra vonatkozó mérési adatok varianciájának az a része, amely a többi változóval leírható. A másik két variancia azokhoz a mérési bizonytalanságokhoz, reprezentativitási problémákhoz és a mérési hibákhoz kapcsolódik, amely minden természeti folyamat/jelenség teljesen természetes része. A fentiekből lényegében az következik, hogy a főkomponens analízisben nem vesszük figyelembe azt a változékonyságot, amely akár a mérési bizonytalanságra, akár a változó saját bizonytalanságára utal. A faktor analízis során viszont komolyan számba vesszük a varianciának ezt a többi változóval nem magyarázható részét. Röviden a főkomponens analízis olyan kapcsolatokat is megjelenít, amelyek háttere mérési/mintázási bizonytalanság, míg a faktor analízis csak a „tiszta" kapcsolatokat mutatja. Persze a faktor analízis ezen előnyének ára is van: attól függően, hogy a „kummunalitásokat" hogyan határozzuk meg, a korrelációs mátrix a megoldás során szinguláris is lehet. Emiatt a faktor analízisnek több algoritmusa létezik, amelyek természetesen „kicsit" más eredményeket adnak ugyanarra a mintára.

Érdekes módon a szerző az adatok bizonytalanságát az 5.fejezet esettanulmányában is figyelmen kívül hagyta. A 74.oldalon ugyanis a következő olvasható: „..más szerzők a főkomponensek számát a sajátértékek alapján határozzák meg….Azonban ez az adattömörítés az adatok varianciájának csak a 65-85%-át magyarázza, és a fennmaradó változékonyság…a későbbiekben nem kerül felhasználásra." A szerző ezt a problémának érzi, míg az opponens az információkban rejlő bizonytalanság eredményének. A sajátérték=1 feltétel alapja ugyanis az, hogy nem engedjük meg, hogy a transzformált főkomponens térben bármely főkomponenst szórása (azaz sajátértéke) kisebb legyen, mint az eredeti változók standardizált formáinak szórásnégyzete.

A fentiek kapcsán kérdésem: a szerző miért nem vette figyelembe a mértékekhez kapcsolódó fentiekben körvonalazott bizonytalanságot?

2.5 A klaszteranalízissel kapcsolatos észrevételek

A klaszteranalízisek az automatikus csoportosító eljárások nagyon népszerű algoritmusai. Tény, hogy ezek a módszerek a mintatér pontjainak (Q-típusú eljárások) és a mintákon mért változók csoportjainak (R-típusú eljárások) felderítésére egyaránt alkalmasak. A szerző vizsgálatait a hierarchikus csoportosító eljárásokra alapozta.

Az ilyen eljárások eredménye lényegében két hasonlósági mértéken alapul: Ez a hasonlósági mutató és a redukciós mérték. A hasonlósági mutató/mérték határozza meg azokat a „magokat", amelyek köré a csoportok (klaszterek) kiépülnek. A redukciós mérték dönt a minta és klaszterek valamint a klaszterek és klaszterek összevonásáról. A szerző által alkalmazott megközelítésben ez utóbbi a Ward-algoritmus csoport homogenizálása volt.

A klaszter eljárásoknak nagyon sokféle implementációja van, amely egy adott mintára nézve nem feltétlenül (sőt rendszerint nem is) azonos csoportosításokat ad.

Történt e vizsgálat arra nézve, hogy más algoritmus választása mennyire változtatta volna meg a Ward-algoritmussal kapott csoportokat?

Az 5. és 6. fejezetek olvasva az a benyomásom, hogy a szerző a klaszteranalízis eredményét úgy tekinti, hogy „A" csoportokat, és nem úgy, hogy a sok csoportképzési lehetőség egyikét kapta meg. A diszkriminancia analízissel élesen fogalmazva csak az látható be, hogy a „választott csoportosító algoritmus, a választott diszkriminancia implementáció alapján szignifikánsan létezik". Ez viszont nem jelenti azt, hogy a kapott csoportosítás olyan mintázat, amely a kérdésre a „legjobb" választ adja.

3 A tézisekről

Sajnos a tézisek megfogalmazása nem szerencsés. A tézis rendszerint egy olyan, néhány erős mondatban megfogalmazott állítás, amely a szerző eredményét összegzi a vizsgált témakörben. Ezt a formai elvárást a 3. és a 7. tézis tökéletesen kielégíti, a többit nagyon erőteljesen le kellett volna rövidíteni. Az 1., 2., 8. és 9. tézisek kb. 2-2 oldat, a 4., 5. és 6. tézisek egyenként egy-egy oldalt jelentenek a tézisfüzetben.

Tartalmilag elfogadom az 1-6. téziseket. Nem tudom megítélni a 8. és 9. tételek tézisértékét.

4 Összegző megállapítás

Kijelentem, hogy a disszertációt nyilvános védésre alkalmasnak tartom.

Szeged, 2019. augusztus 28.

……

Geiger János, PhD

A bírálatban hivatkozott irodalmak jegyzéke

1. Bárdossy, Gy., Fodor, J. (2004): Evaluation of Uncertainties and Risks in Geology. Sringer, 221 pp
2. *Olea, R.A. (1999):* Geostatistics for Engineers and Earth Scientists. Springer Science+Business Media. 303 pp
3. *Caers, J.* (2011): Modeling Uncertainty in the Earth Sciences. Wiley-Blackwell. 225 pp
4. *Goovaerts, P (1997):* Geostatistics for Natural Resources Evaluation. Oxford Univ.Press. 483 pp.
5. *Journel, A.*G. (1986): Models and tools for the earth sciences. Mathematical Geology. v.18, no.1, p.119-140