HUN REN Csillagászati és Földtudományi Kutatóközpont CSFK

HUN-REN Magyar Kutatási Hálózat

HUN-REN Csillagászati és Földtudományi Kutatóközpont

1121 Budapest, Konkoly-Thege Miklós út 15-17. Tel: (+36) 1 391-9322 www.csfk.org

Válaszok Dr. Geiger János opponensi véleményére

Köszönöm Dr. Geiger Jánosnak, hogy igen alaposan átnézte és véleményezte az MTA doktori disszertációmat. Hálás vagyok a dolgozattal kapcsolatos elismerő szavaiért és kritikai észrevételeiért. Az alábbi oldalakon szeretnék kitérni pontról-pontra a megjegyzéseire és konkrét kérdéseire. A szóhasználattal, vagy pontosabb megfogalmazással kapcsolatos megjegyzéseire röviden, míg a többi kérdésre részletesebben kívánok válaszolni. A Tisztelt Bíráló kérdéseit félkövérrel szedtem, míg az idézeteket és esetleges kiemeléseket dőlttel.

1. Szóhasználati illetve a dolgozatban pontosabb megfogalmazásra ösztönző megjegyzések.

1.1. Bíráló a dolgozat szöveges értékelésében megjegyzi, hogy „Laterális eloszlások kialakítására a krigelést (vélhetően ordinary kriging) használja”.

Válasz: Köszönöm a megjegyzést, a disszertáció törzsszövegében valóban nem tértem ki rá, de *oridinary point kriging* eljárást alkalmaztam; ezt a Hatvani et al. (2017): 2.5. fejezete tartalmazza.

1.2. **(#1) Az 1.1 alfejezet címe: Környezeti adatok tér- és időbeli változékonysága, hozzáférhetősége – gyakorlati szempontok**

Nekem úgy tűnik, hogy ez a cím meglehetősen pontatlanul utal az alfejezet tényleges mondanivalójára. Ebben az alfejezetben a szerző, a címmel ellentétben, kizárólag a környezeti adatok feldolgozhatóságának kritériumaival foglalkozik.

Válasz: Igen, ahogy a Tisztelt Bíráló is megjegyzi a fókusz valóban a környezeti adatok feldolgozhatóságának kritériumain van, ugyanakkor az adott fejezetben (1.1. fejezet, 3. bekezdés) ezen kritériumok tér- és időbeli aspektusaira helyeztem a hangsúlyt. Célszerűbb lett volna pontosabb címválasztással élnem, például: *A környezeti adatok feldolgozhatóságához szükséges tér- és időbeli kritériumok.*

1.3.* (#2) A szerző által ismertetett térbeli kritérium szerint *„Egy adott változóra térben legyen annyi adatpont, hogy azok között legyen megfelelő erősségű a térbeli autokorreláció, i.e. kapcsolatiságot lehessen vizsgálni közöttük..”

Ez a kritérium lényegében a térbeli változásokat leíró monitoring rendszerek tervezésének alapja. Az autokorreláció létére adott kritériumot nagyon szűkítőnek látom azokban az esetekben, amikor nem tervezett, hanem már meglevő információszerző pontok mérési eredményeit kívánjuk monitoringként alkalmazni . Az autokorreláció léte a gyenge vagy erős stacionaritáshoz kötődik. A gázokban, folyadékokban lebegő szuszpendált, pl. agyag, finom

aleurit méretű szemcsék keveredésük során Brown-mozgást mutatnak. Az ilyen Brown-típusú keveredési folyamatokra jellemző, hogy nincs véges szórásuk. Emiatt az autokorreláció nem számolható. Talán, ha autokorreláció helyett a jelenség átlagos félvariogramjának véges hatástávolsága lenne a kritérium, akkor ez a feltétel már lényegesen kevésbé lenne szűkítő hatású.

A geostatisztika oldaláról szemléve azt mondhatjuk, hogy csak a tisztán röghatás típusú févariogram modell esetében jelenthető ki, hogy a vizsgált jelenség nem mutat a térben lineáris folytonosságot, vagy legalábbis a lineáris folytonosság megítélése az adatpontok száma és térbeli helyzete alapján nem lehetséges. Ha a tapsztalati félvariogram bármely nem-röghatás típusú modellel leírható, akkor vannak hatástávolságon belüli pontpárok, így a jelenség térbeli folytonosságának elemzése akár valamely krigelési becsléssel, akár valamely sztochasztikus szimulációval leírható, a grid-rendszer geometriájának kellő megválasztását követően.

Válasz: A „*kapcsolatiságot lehessen vizsgálni közöttük*” megfogalmazással valójában arra utaltam, hogy az adatpárokból számított félvariogram(ok)nak véges hatástávolsággal kell rendelkezniük.

1.4. **(#3) Ugyanebben az alfejezetben szerepel a következő állítás: *„Mindazonáltal fontos leszögezni, hogy bármilyen transzformáció, amely „hozzáad” adatot az eredeti idősorhoz vagy „elvesz” belőle, elkerülhetetlenül megváltoztatja annak spektrális tulajdonságait.”***

A fenti mondatban a „transzformáció” jelentését kellene tisztázni. A statisztikában és geostatisztikában alkalmazott numerikus transzformációk ugyanis függvények, amelyekkel csak az adatértékeket kezeljük. Az eredeti idősor elemeinek számát csökkentő beavatkozás lehet pl. a szűrés, de ez nem transzformáció. Az adatsor számosságát növelő beavatkozást sem lehet matematikai értelemben transzformációnak nevezni.

Emellett az ilyen jellegű beavatkozások is legfeljebb megváltoztathatják, de nem feltétlenül változtatják is meg a spektrális tulajdonságokat.

Válasz: Köszönöm szépen a megjegyzést, pontatlan volt a szóhasználat, adatpótlásra szerettem volna utalni.

1.5. (#4) Ugyanebben az alfejezetben: *„E fenti szempontok mind meghatározzák azon geomatematikai módszerek körét, amelyek alkalmazhatóak egy adott környezeti probléma kezelésében.”*

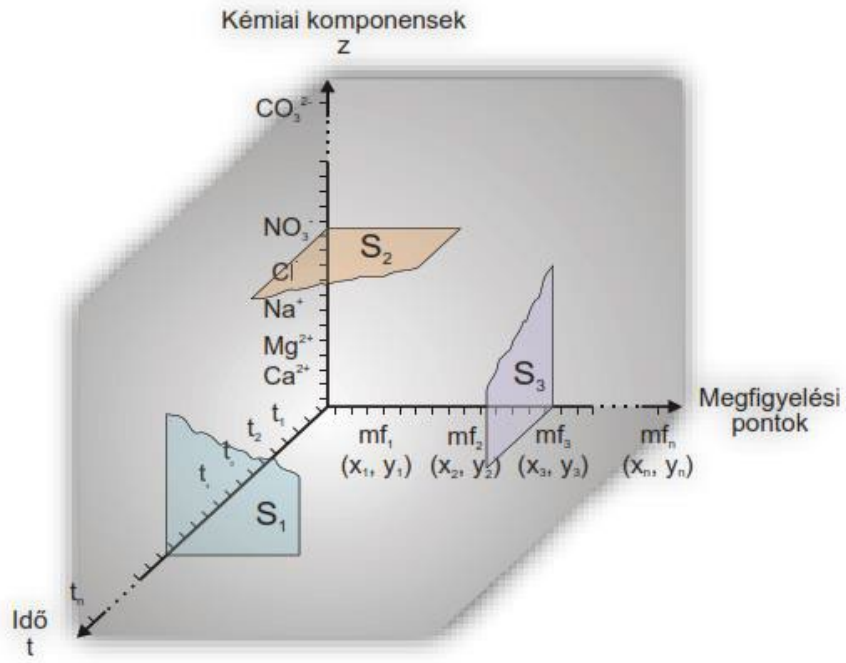
Ezt az állítást is talán érdemes lett volna árnyalni, pl. a „jelenlegi ismereteink mellett alkalmazhatóak” változattal. Érdemes belegondolni abba, hogy a hatványfüggvény félvariogram modeljének alkalmazása előtt pl. a repedésrendszerek geostatisztikai leképezése szinte megoldhatatlan volt.

Válasz: Jobban kellett volna árnyalnom a megfogalmazást, ahogy a Tisztelt Bíráló is javasolja.

1.6. (#5): Mit jelent pontosan az „egységes rendszerbe foglalt matematikai eszköztár”?

Számomra kissé heurisztikus ez a megfogalmazás. Lehet ez attól egységes, mert minden tanulmány lényegében ugyanabból, az egyébként nagyon sokrétű, módszerhalmazból használ fel vizsgálati elemeket. Lehet attól is egységes, hogy az input-output utak egymáshoz kapcsolódnak. Szellemében azonban nem lehet egységes, mert a statisztikai és geostatisztikai szemlélet egészen másként viszonyul a térbeli adatokhoz. A statisztikai megközelítésben egy attribútum több adatponti (térbeli) értékét úgy tekintjük, hogy az adott attributumot mint valószínűségi változót vizsgáljuk az adatponti értékeken keresztül. Ugyanezt a rendszert a geostatisztika egy többváltozós valószínűségi függvény egy realizációjának tekinti. Azaz annyi valószínűségi változónk van, ahány adatpont, és ezek együttes eloszlása definiálja az attributumot. Ez a különbség elméletileg nagymértékben szűkíti azon klasszikus statisztikai módszerek számát amelyek a geostatisztikai szemléleten belül is alkalmazhatók. Ez utóbbi következmény pedig jelenleg nem támogat egy szellemében egységes statisztikai és geostatisztikai rendszert.

#6: A 3.2 alfejezetben szerepel a következő gondolat: *„A bemutatott tanulmányokban vizsgált adatokat egy egységes rendszerben, négydimenziós térben lehet elképzelni (3.2-1. ábra), ahol a tengelyeket az idő (t), a vizsgált kémiai komponensek és a megfigyelési pontok adják.*” A hivatkozott ábra és annak címe az alábbi:



3.2-1. ábra. Adathalmaz ábrázolása négy dimenzióban Kovács et al. (2012c) alapján.

Ennek az ábrának a címe Kovács József MTA doktori művében teljesen korrekt: *„Egy négydimenziós adathalmaz modellje*”. Talán érdemes lett volna átvenni az ábra feliratát is Kovács József eredeti publikációjából.

A másik problémám az ábra célja. Kovács József ezt az ábrát a föld és környezettudományokban használt adatmodellek és bizonyos feldolgozási protokolok általános

kategorizálása használta nagyon egyedi és nagyon előremutató módon. Jelen munka szerzőjének vélhetően az volt a célja, hogy a rövid értekezésben felhasznált elemző módszereket elhelyezze a Kovács-féle rendszerben. Ugyanakkor ez a cél talán néhány soros magyarázattal sokkal egyértelműbben elérhető lett volna, hiszen a két műben alkalmazott módszerek, a gépi tanuló algoritmusoktól eltekintve, nagyon hasonlóak.

Válasz: Szeretnék egyszerre válaszolni a fenti két kérdésre (#5 és #6), mivel meglátásom szerint ezek szorosan összefüggenek. A fejezetet azzal a mondattal kezdem: *„A bemutatott tanulmányokban vizsgált adatokat egy egységes rendszerben, négydimenziós térben lehet elképzelni (3.2-1. ábra)”,* majd bemutatom magát az ábrát és a következő bekezdésekben kifejtem, hogy a disszertációban és az ahhoz kapcsolódó tanulmányokban alkalmazott módszerek hol helyezkednek el a Kovács által meghatározott diagramon.

A megfogalmazással és magával az alfejezettel a holisztikus szemléltetés volt a célom. Ez valóban félreérthető volt egy nagy szakmai jártassággal rendelkező kolléga számára, ugyanakkor fontosnak tartottam, hogy azok az olvasók, akik nem feltétlenül jártasak az összes alkalmazott eszközben, kapjanak egy átfogó képet arról, mely módszereket lehet hasonló típusú adatokon alkalmazni. Ezzel egyúttal lehetőségem nyílt a dolgozat sokszínű módszertani megközelítésének bemutatására is.

Fel szeretném hívni a Tisztelt Bíráló figyelmét, hogy Mádlné Szőnyi Judit bíráló 2.1-es kérdésére adott válaszomban (2. ábra) egy közvetlen példát mutatok be, hogy az alkalmazott módszerek milyen döntések nyomán és milyen sorrendben kerültek alkalmazásra a Balatonra vonatkozó célok fényében. Ez a válaszom a Tisztelt Bíráló kérdéséhez is kapcsolódik.

2. Mélyebb szakmai kérdések

2.1. (#7): A 3.2.1 alfejezetben (Néhány alkalmazott módszer részletesebb ismertetése) a Kombinált klaszter és diszkriminancia analízis (CCDA) nevű R-csomag alkalmazása kapcsán a jelölt a következőket írja: *”A két módszer összevonásának esszenciája a hagyományos csoportosító eljárásokkal ellentétben, hogy objektív mérőszámot rendel a HCA eredményéhez, így segítségével homogén és optimális csoportok is meghatározhatók.”*

A szobánforgó R-csomag a cluster analízis hierarchikus megoldását használja. Az ilyen eljárások eredménye lényegében két nagyon is objektív mértéken alapul: Az egyik a hasonlósági mutató, a másik 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 klaszter és klaszter összevonásáról. A CCDA R-csomag a Ward-algoritmust alkalmazza redukciós mértékként. A klaszter eljárásoknak nagyon sokféle implementációja van, amely egy adott mintára nézve nem feltétlenül (sőt rendszerint nem is) azonos csoportosításokat ad.

Kérdéseim:

- **(i) Történt e vizsgálat arra nézve, hogy más redukciós mérték választása mennyire változtatta volna meg a Ward-algoritmussal kapott csoportokat?**
- **(ii) Milyen hasonlósági mutatót, vagy mutatókat használ a CCDA csomag?**

Az a benyomásom, hogy a jelölt 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 algoritmus alapján szignifikánsan létezik". Ez viszont nem jelenti azt, hogy a kapott csoportosítás az egyetlen optimális, vagy szignifikáns megoldás. Csak arról beszélhetünk, hogy szóbanforgó csoportosítás a sok optimális mintázat egyike.

Kérdésem:

- **(iii) Történt-e érzékenység vizsgálat vagy alternatívák kialakítása a kapott csoportokra?**

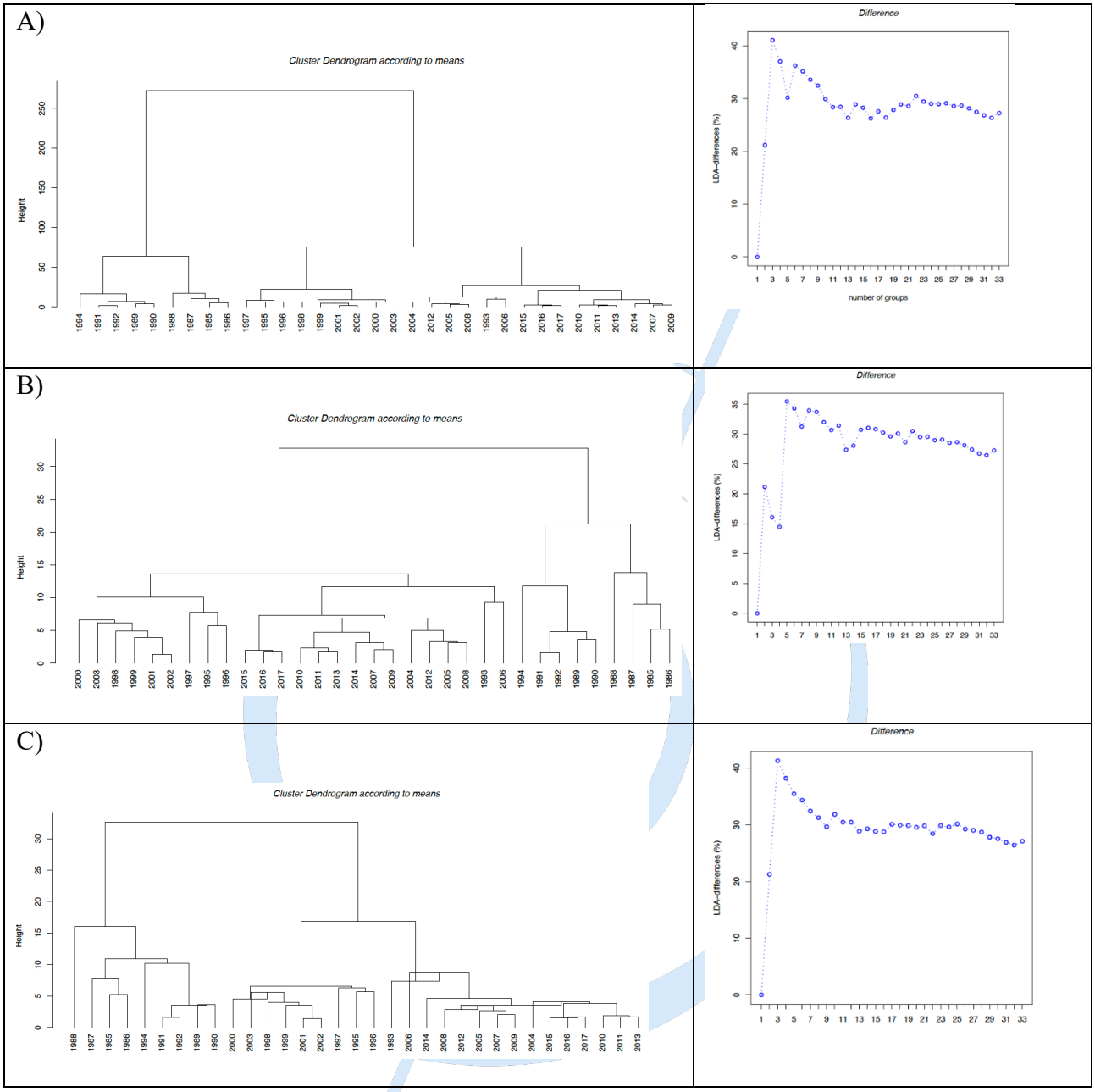
Válasz: Szeretnék a fenti három kérdésre együtt, de tételesen, válaszolni, mert ezek szorosan összefüggenek.

Ahogy a Tisztelt Bíráló is írja, a `ccda`, (v. 1.1.1.) ahogy Kovács S. et al. (2022) https://cran.r-project.org/web/packages/ccda/ccda.pdf és Kovács J. et al. (2014) is újra Ward módszerét használja redukciós mértékként. (i) Nincs beépített lehetőség ezt megváltoztatni. Ennek ellenére belenyúltam a forráskódba – aminek fejlesztésében nem, de tesztelésében részt vettem - és noha eddig (ii) nem vizsgáltam, hogy más redukciós mérték esetén mennyire kaptam volna más csoportokat Hatvani et al. (2020)-ban, ezt most a bírálat által motiválva megtettem. Köszönöm a hasznos felvetést!

(iii) a `ccda` a `hclust` függvényt használja a `stats` csomagból a hierarchikus klaszterek kiszámítására. A Bíráló javaslatára a Ward módszerén kívül két másik módszerrel (`average` és `median`) is megvizsgáltam a lehetséges csoportosításokat, hogy milyen egyéb optimális klaszteredményt kaphatunk alkalmazott algoritmustól függően. (A kód amivel dolgoztam válaszom végén található). A Bíráló kérdései által motiválva frissítjük a `ccda` csomagot, hogy ne csak Ward módszerével lehessen csoportokat kialakítani.

Az eredmények alapján a medián módszer (1C ábra) pontosan ugyanannyi és ugyanolyan összetételű csoportokat eredményezett, mint a Ward-módszer (1A ábra és Hatvani et al., 2020: Fig. 3). Az átlag módszer viszont a dolgozatban bemutatott három csoportot további két alcsoportra bontotta, így összesen öt klasztert kaptam. Fontos hangsúlyozni, hogy ezek az alcsoportok az eredeti felosztás *hierarchikus* tagolásai, nem pedig új, „összekeveredett" összetételű csoportok.

Ezek fényében kijelenthetem, hogy a publikált csoportosítás, noha csak egyike az optimális csoportoknak, de az robusztusnak tartom az alább bemutatott eredmények fényében. Ez természetesen nem zárja ki annak lehetőségét, hogy más módszerrel vagy távolságmetrikával eltérő klasztereket kapjunk, ugyanakkor a választott megközelítés összevethetővé teszi az eredményeket korábbi vizsgálataimmal (pl. Hatvani et al., 2011; Kovács et al., 2012, 2015) így szakmailag elfogadható és megalapozott.



1. ábra Dendrogramok (bal) és ccda output ábrák (jobb), ahol a maximális érték adja meg az optimális csoportszámot Ward módszere (A) átlag (B) és medián (C) módszer választása esetében.

Hatvani, I. G., Kovács, J., Kovácsné Székely, I., Jakusch, P., Korponai, J. (2011). Analysis of long-term water quality changes in the Kis-Balaton Water Protection System with time series, cluster analysis and Wilks' lambda distribution. *Ecological Engineering, 37*, 629–635.

Kovács, J., Kovács, S., Hatvani, I. G., Zsuga, K., Korponai, J., Gribovszki, Z., et al. (2015). Spatial optimization of monitoring networks on the examples of a river, a lake-wetland system and a sub-surface water system. *Water Resources Management, 29*, 5275–5294.

Kovács, J., Nagy, M., Czauner, B., Kovácsné Székely, I., Borsodi, A. K., Hatvani, I. G. (2012). Delimiting sub-areas in water bodies using multivariate data analysis on the example of Lake Balaton (W Hungary). *Journal of Environmental Management, 110*, 151–158.

Ezeken felül a Bíráló javaslatai által motiválva és általam írt – és alább bemutatott - kód alapján a csomag fejlesztői frissítették a `ccda` R package-et 1.1.2-es verzióra, ahol már `ward.D` az alapértelmezett beállítás, de minden, a `hclust` függvény számára natív módszert is elfogad a program anélkül, hogy a forráskód módosítására lenne szükség. A forráskódban végzett módosításaimat sárga háttérrel emeltem ki alább.

```
# Function to run CCDA with different linkage methods
ccda_with_custom_linkage <- function(dataset, names_vector, nr, nameslist,
linkage_method = "ward.D") {
  length_dataset = length(dataset[, 1])
  dataset = cbind(names_vector, dataset)
  st = 2
  dim_dataset = length(dataset[1, ])
  ngroups = sum(table(dataset[, 1]) != 0)

  if (length(names_vector) != length_dataset) {
    stop("length of names vector does not equal to the length of dataset")
  }
  if (length(nameslist) != ngroups) {
    stop("number of names does not equal the number of various names in names
vector")
  }
  if (ngroups <= 1) {
    stop("all data have the same origin")
  }
  if (prior != "proportions" & prior != "equal") {
    stop("prior has to be set to proportions or equal")
  }

  mean_mtx = matrix(rep(0, ngroups * (-st + dim_dataset + 1)), nrow = ngroups)
  for (i in 1:ngroups) {
    mean_mtx[i, ] = colMeans(dataset[(dataset[, 1] == nameslist[i]),
st:dim_dataset])
  }

 # Apply custom linkage method
  cluster = hclust(dist(scale(mean_mtx)) * dist(scale(mean_mtx)), method =
linkage_method)
  clusterings = cutree(cluster, k = c(1:ngroups))

  grouping_mtx = matrix(rep(1, times = ngroups * length_dataset), ncol =
ngroups)
  for (x in 1:ngroups) {
    for (z in 1:ngroups) {
      for (i in 1:length_dataset) {
        if (dataset[i, 1] == nameslist[z]) {
          grouping_mtx[i, x] = clusterings[z, x]
        }
      }
    }
  }

  grouping_percentage = c(rep(1, times = ngroups))
  for (x in 2:ngroups) {
    grouping_percentage[x] = percentage(dataset[, st:dim_dataset],
grouping_mtx[, x], prior)
  }
```

```
  random_grouping_percentage = matrix(rep(1, times = ngroups * nr), ncol = nr)
  for (y in 1:nr) {
    random_grouping_mtx = grouping_mtx[sample(1:length_dataset), ]
    for (x in 2:ngroups) {
      random_grouping_percentage[x, y] = percentage(dataset[, st:dim_dataset],
random_grouping_mtx[, x], prior)
    }
  }

  q95 = rep(1, ngroups)
  for (i in 1:ngroups) {
    q95[i] = quantile(random_grouping_percentage[i, ], prob = 0.95)
  }

  ratio = grouping_percentage
  D = ratio - q95
  k = min(which(D == max(D)))
  groups = matrix(paste("sub-group", clusterings[, k]), ncol = 1, nrow =
ngroups)
  rownames(groups) = nameslist

  cat(paste("
", "Number of optimal groups: ", k, "
", sep = ""))
  cat(paste("
", "Maximal difference between q95 and ratio ", round(max(D), digits = 3), "
", "
", sep = ""))
  cat(paste("further investigation of the following ", k, "sub-groups
recommended:"))
  print(groups)

  groups = matrix(paste("sub-group", clusterings[, k]), ncol = 1, nrow =
ngroups, dimnames = list(nameslist, paste("further investigation of the
following ", k, "sub-groups  recommended:")))

  return(list(nameslist = nameslist, q95 = q95, ratio = ratio, difference = D,
sub_groups = groups, cluster = cluster))
}

##############################################################################################

# Define linkage methods to compare
linkage_methods <- c("ward.D", "average", "median")
cluster_results <- list()

ggplot_list <- list()

# Loop through linkage methods
for (method in linkage_methods) {
  cluster_results[[method]] <- ccda_with_custom_linkage(
    dataset = rbind(KB, SZIB, SZEB, SB)[complete.cases(rbind(KB, SZIB, SZEB,
SB)[, param2]), param2],
    names_vector = rbind(KB, SZIB, SZEB, SB)[complete.cases(rbind(KB, SZIB,
SZEB, SB)[, param2]),]$year,
    nr = 200,
    nameslist = unique(rbind(KB, SZIB, SZEB, SB)[complete.cases(rbind(KB, SZIB,
SZEB, SB)[, param2]),]$year),
    linkage_method = method
```

```
  )
}

# Generate and save plots
pdf("res_KB_SZIB_SZEB_SB_comparison.pdf")
par(mfrow=c(1,2))
plotccda.cluster(cluster_results[["ward.D"]])
plotccda.results(cluster_results[["ward.D"]])
plotccda.cluster(cluster_results[["average"]])
plotccda.results(cluster_results[["average"]])
plotccda.cluster(cluster_results[["median"]])
plotccda.results(cluster_results[["median"]])
dev.off()
```

2.2. **(#8): Ugyancsak a 3.2.1 alfejeztben olvasható a PCA (Főkomponens analízis) rövid összefogalója: *„…(a PCA) a vizsgált változók lineáris transzformációjával új ortogonális főkomponenseket hoz létre, amelyeket az egymással korreláló eredeti változók határoznak meg. Ezen főkomponensek száma azonos az eredeti változók számával, és az eredeti teljes variancia egy adott százalékát magyarázzák, rendre szigorúan monoton csökkenő formában (Tabachnick és Fidell, 2014).”***

Az összefoglaló, rövidsége ellenére, lényegében helyes: a főkomponens analízis során bármely adott változó varianciát teljes egészében leírjuk a főkomponensekkel. Meglátásom szerint a PCA-t földtudományi alkalmazásának problémája pontosan ebben van. Ugyanis azzal, hogy minden változó teljes varianciáját bontjuk fel, elfogadjuk, hogy minden mérésünk tökéletes. Ilyen mérések neglehetősen ritkán fordulnak elő földtudományi feldolgozásokban. Márpedig a PCA alkalmazásával sem mérési problémának, sem az adott tujadonság saját meghatározhatósági bizonytalanságának nem adunk teret.

Ezzel szemben a faktor analízisek során bármely változó teljes varianciáját három tag összegére bontjuk: közös variancia, egyedi variancia, hiba variancia. A „közös variancia” (=kommunlitás) adatok varianciájának az a része, amelyet az összes változó egymásrahatása alakított ki. Ez az amelyet szakmailag értelmezni kell. A másik két variancia azokhoz a mérési bizonytalanságokhoz, reprezentativitási problémákhoz és a mérési hibákhoz kapcsolódik, amelyek minden összetett természeti folyamat/jelenség mérésének teljesen természetes részei.

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.

Másként fogalmazva, a főkomponens analízis akár olyan kapcsolatokat is megjeleníthet, amelyek háttere mérési/mintázási bizonytalanság. A faktoranalízis ezzel szemben csak a „tiszta” kapcsolatokat mutatja. Hogy milyen „tiszta kapcsolatot” kívánunk megjeleníteni, az a kommunalitások meghatározásán múlik. Emiatt (is) a FA-nek több implementációja van. Mindez felveti, hogy talán az FA jobban illeszkedett volna mérési adatok természetes bizonytalanságához, mint a PCA.

Vannak olyan esetek, amikor a FA és PCA „hasonló” eredményt ad. Ám Fabrigar et al. (1999) bemutatta, hogy ha a kommulitások viszonylag alacsonyak (pl. a (−0.4,0.4) intervallumban vannak), akkor a két eljárás nagyon különböző eredményeket mutathat.

Válasz: Köszönöm a Bíráló megjegyzéseit és noha matematikai érvelésével egyetértek, jelen alkalmazások esetében úgy gondolom megfelelő volt PCA használata is. A Balaton esetében elvégeztem a vizsgálatot Maximum likelihood FA-val is. Jól látszik, hogy Keszthelyen, Szigligeten, Siófokon majdnem teljesen azonos a magyarázott variancia mértéke a FA és a PCA esetében. A Szemesi-medencében volt egy elhanyagolható eltérés a 2. és 3. faktorban (1. Táblázat).

1. Táblázat. A Balaton négy medencéjének vízminőségi változóira elvégzett PCA és FA eredményei

Medence	Faktor	PCA	FA
		Var. %	Var. %
Keszthely	FAC1_1	49,3%	49,3%
	FAC2_1	19,0%	19,0%
	FAC3_1	10,5%	10,5%
Szigliget	FAC1_2	46,5%	46,5%
	FAC2_2	18,0%	18,0%
Szemes	FAC1_3	42,7%	42,0%
	FAC2_3	19,1%	17,6%
	FAC3_3	11,3%	13,5%
Siófok	FAC1_4	38,5%	38,5%
	FAC2_4	19,0%	19,0%
	FAC3_4	12,7%	12,7%

Meg szeretném jegyezni, hogy megközelítés szempontjából, az általam feldolgozott adatok különböző mérési technikákkal meghatározott változók (disszertáció 3.1-1. táblázat) akár évtizedeket átívelő adatait. Ebben az esetben számos lehetséges „mérési hiba problémája” kiküszöbölhető PCA alkalmazásával, mivel az adatok -A Balaton esetében - diakronikusak, és ezek átlagaival dolgozom, így feltételeztem, hogy a véletlenszerű hatásokból adódó mérési hibák hosszú távon kiegyenlítődnek (Di Franco és Marradi, 2013). A dolgozatban leírt célok számára a rendelkezésre álló adatokkal a PCA alkalmazás véleményem szerint megfelelő volt, hiszen arra voltam kíváncsi, hogy a változóimhoz milyen külső magyarázó változó lefutása hasonlít a legjobban.

Di Franco, G., Marradi, A. (2013). Factor analysis and principal component analysis. https://www.torrossa.com/en/resources/an/3112303

2.3. **(#9): A jelölt a félvariogramot használta a „mintavételezési gyakoriság reprezentativitásának" vizsgálatára mind az antarktiszi jégfuratok, mind a Fertő-tó, mind a Mediterrán térség vizsgálatakor.**

Ha a tapasztalati félvariogram bármely nem-röghatás típusú modellel leírható, ott van/vannak hatástávolságon belüli pontpárok. Emiatt a síkbeli kiterjesztés (kontúrtérkép) sikere lényegében két dolgon múlik: (1) az pontcsoportosulások hisztogram torzító hatását sikerült-e kiszűrni valamely csoportbontó algoritmussal; (2) a grid méret (geometria) meghatározásakor tekintettel voltunk-e a pontok távolságára. Ha a csoportbontó algoritmus megfelelő, akkor az adat értékek gyakorisági eloszlása megfelelően reprezentálja a rendelkezésre álló információ mennyiséget. Ha a grid geometria meghatározása helyes, és a variogram model korrekt, akkor a kapott kontúrtérkép megfelelően reprezentálja a rendelkezésre álló adatok térbelisége által megengedett kiterjeszthetőséget.

Egészen más kérdés, hogy a kapott eredmény mennyire jellemző az adatok által visszatükrözött folyamatra. Ennek eldöntéséhez az adatrendszert kezdő állapotú monitoringnak tekintjük, melyet a vizsgált jelenségre kell kalibrálnunk. Ebben egyebek mellett felhasználható pl. a regressziós krigelés krigelési varianciája (Szatmári, 2017) a variogram modelljének hatástávolsága és a röghatás értéke (pl. Füst és Geiger, 2010; Füst, 2011) vagy például a sztochasztikus realizációk szórástérképe (Goovaerts, 1997;). A fentiek miatt a félvariogramot használini a térbeli minta-reprezentativitás jellemzésére nem a legjobb megoldás.

Válasz: Köszönöm a problémafelvetést. Jóllehet nem a legjobb megoldás a félvariogram alkalmazása, az Antarktisz régió esetében az izotóphidrológiában bevett módszertant követtem (pl. Bowen és Wilkinson, 2002; Kern et al., 2014). A Fertőn kapott eredményekre szeretnék a következő kérdésre adott válaszomban kitérni.

G.J. Bowen, B. Wilkinson (2002). Spatial distribution of $\delta^{18}O$ in meteoric precipitation. *Geology*, *30,* 315-318

Kern, Z., Kohán, B., Leuenberger, M. (2014). Precipitation isoscape of high reliefs: interpolation scheme designed and tested for monthly resolved precipitation oxygen isotope records of an Alpine domain. *Atmospheric chemistry and physics, 14(4),* 1897-1907.

2.4. *(#10): A szerző értelmezése szerint,* ***ha a mintavételi pontok távolsága nagyobb, mint az átlagos félvariogram hatástávolsága, akkor egy eseteleges interpoláció bizonytalan értékeket ad (ökörszem struktúrák jelennek meg a térképen)****.*

Ez az interpretáció, jóllehet erősen leegyszerűsített, alapvetően elfogadható, ha a variogram modell nem az eredeti, hanem a standardizát adatokra készült, és ha a standardizált adatok félvariogram modeljából származik a hatástávolság. A szerző vizsgálataiban azonban a standardizálásra semmilyen utalás nem olvasható.

Válasz: Noha a tanulmányban nem a standardizált adatokból készült a variogram, a Tisztelt Bíráló javaslatára kiszámoltam azokat is és *azonos hatástávolságot kaptam mint az eredeti tanulmányban*. Hatvani et al. (2018)-ban a=1110 és 1180 m volt *Enterococcus* és *E-Coli* paraméterekre rendre, míg most ez 1120 m és 1090 m volt.

***(#11):* A szerző a Fertő-tó esetében kijelenti, hogy „..with an approximately 1 km spatial range, interpolation was insufficient in the case of distant sites with the current sampling grid, and that in general the sites are mostly representative of local phenomena." Sajnos a sampling grid méretéről nem ad információt a dolgozat. A mintavételi pontok nem grid-mintázásra utalnak. Ilyen esetekben a monitoring-geostatisztika azt javasolja, hogy a síkbeli kiterjesztés olyan grid geometriára történjen, amelyet az adatok megengednek. Az ilyen grid-méretezési problémákat Hengl (2006) dolgozata tekinti át remek ajánlásokkal. Ennek, a vélhetően nagy-léptékű, gridnek „down-scaling"-jével kapható meg egy olyan megjelenítés, ami az éppen rendelkezésre álló információknak a leginkább megfelel.**

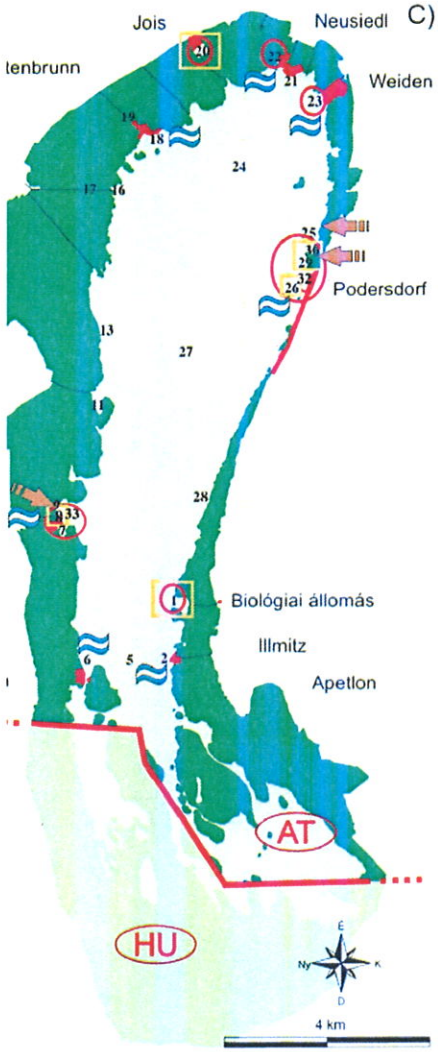
Úgy vélem, hogy az ilyen elsődleges-monitoring adatok térbeli heterogenitása leképezésének legjobb eszközei a sztochasztikus szimulációk, amelyek egyúttal lehetőséget adnak a bizonytalanság számszerűsítésére is.

5.tézis A tézist csak az alábbi mellék mondat kihagyásával tudom elfogadni: „…geostatisztikai vizsgálatokkal megállapítottam, hogy a jelenlegi mintavételi hálózattal nem ajánlott interpolálni SFIB-re". A problémát a geostatisztikai alkalmazásokhoz kapcsolódó megjegyzéseim mutatják.

Mejegyzem, a tézis a fenti kifogásolt mellékmondat kihagyásával is érthető marad.

Válasz: Ebben a válaszban szeretnék egyszerre kitérni az 5. tézist érintő kritikákra, valamint a fenti #11 kérdésre. A megfogalmazásom pontatlan volt. Mind a Hatvani et al. (2018) 'Spatial sampling frequnecy estimation' fejezetében, mind a 'Geostatistical assessment of SFIB variance' c. fejezetben „*current sampling grid*"-et írok, miközben a mintavételi hálózatra kívántam utalni '*sampling network'*.

A tézisben és az említett mondatban is azt kívántam hangsúlyozni, hogy a Fertő jelenlegi mintavételi hálózatán (2. ábra) mért fekálindikátor-baktériumok térbeli varianciája alapján meghatározott hatástávolság (1 km) olyan csekély, hogy amennyiben csak az eddig működő mintavételi pontokon történne a tó állapotfelmérése - amit Herzig et al. (2019)-ben is javasoltunk - , az nagy valószínűséggel nem nyújtana reprezentatív képet, sem a tó egészéről, sem annak vízminőség szempontjából elkülönülő élőhelyeiről (Magyar et al., 2013).



A Fertőre vonatkozó tanulmány (Hatvani et al., 2018) konklúziójában már pontosan fogalmaztam: "*Based on the geostatistical analysis, it is suggested that an intense campaign be initiated, both in time and space, in order to be able to map the different areas/habitats of the lake more representatively in terms of SFIB*". nem utalva a mintavételi grid-re, hanem a Fertő mintavételi hálózatára (2. ábra). Bízom benne, hogy ez a pontos idézet segít feloldani az esetleges félreértést a Tisztelt Bírálóval.

Herzig, A., Hatvani, I. G., Tanos, P., et al. (2019). Mikrobiologisch-hygienische Untersuchungen am Neusiedler See – von der Einzeluntersuchung zum Gesamtkonzept. *Österreichische Wasser- und Abfallwirtschaft*, *71*, 537–555.

Magyar, N., Hatvani, I. G., Székely, I. K., Herzig, A., Dinka, M., & Kovács, J. (2013). Application of multivariate statistical methods in determining spatial changes in water quality in the Austrian part of Neusiedler See. *Ecological Engineering*, *55*, 82-92.

2. ábra. A Fertő ausztriai mintavételi hálózata Hatvani et al. 2018 szerint.

Összegezve le szeretném szögezni, hogy a Bíráló kritikáiban leírt matematikai érveléssel nem kívánok vitatkozni, azokkal egyetértek. Mindazonáltal bízok benne, hogy a megközelítésbeli különbségekből adódó kritikákat fel tudtam oldani.

Még egyszer köszönöm a bírálatot!

Budapest, 2025. 04. 14.

Dr. Hatvani István Gábor