Házi feladatok megoldása 7.

Hierarchikus klaszterelemzés SPSS-ben és R-ben

Smahajcsik-Szabó Tamás, M9IJYM

1. Végezz HKA-t SPSS-ben a minimális távolság módszerrel, SED távolsággal a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel, standardizálással! Hány klaszteres megoldás tűnik a legjobbnak az Agglomeration Schedule Coefficients oszlopának grafikus ábrázolása alapján? És a jégcsapdiagram alapján? És a dendrogram alapján?

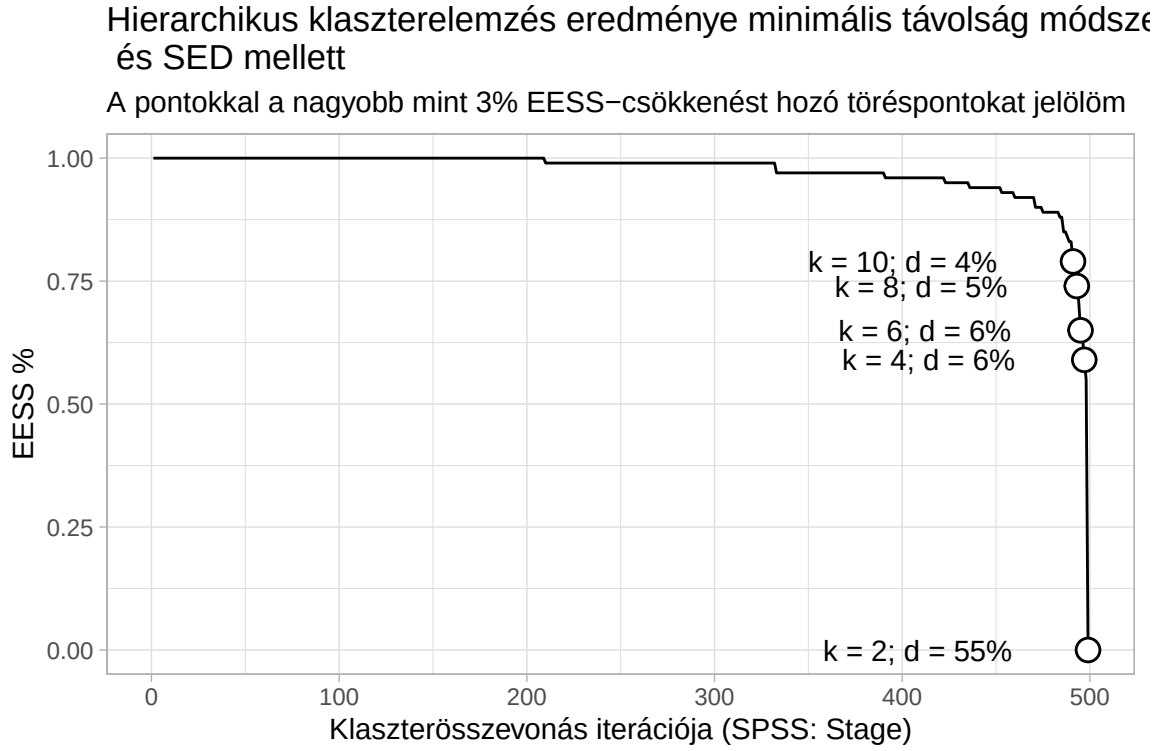
Az SPSS *Agglomeration Schedule* táblázatának utolsó sorait az alábbi táblázatba foglaltam.

k	EESS	d
16	0.88	0.00
15	0.85	0.03
14	0.85	0.00
13	0.84	0.01
12	0.83	0.01
11	0.83	0.00
10	0.79	0.04
9	0.79	0.00
8	0.74	0.05
7	0.71	0.03
6	0.65	0.06
5	0.64	0.01
4	0.59	0.06
3	0.55	0.03
2	0.00	0.55

k=klaszterszám, EESS%, d=EESS% csökkenés

Az eredményekről az alábbi könyökábra is tájékoztat. K=10 megoldás esetén az EESS % értékének csökkenése elsőként éri el a 4%-ot, de ez értéke egészen k=2-ig nem ér el 5-6%-nál nagyobb csökkenést. Végül a d (itt a különbözetet jelöltem így) értéke 55% k = 3 után.

Noha az *icile* diagram áttekinthetősége csökkent a nagy változószám mellett, a dendogram már inkább volt a segítségemre. A nagyobb klaszterösszevonásokra viszonylag az iterációk végén került sor, a dendogram nem mutatja az elvárt összevonási mintázatot, feltételezem, a minimális távolság módszere szuboptimális az adott adatokon.

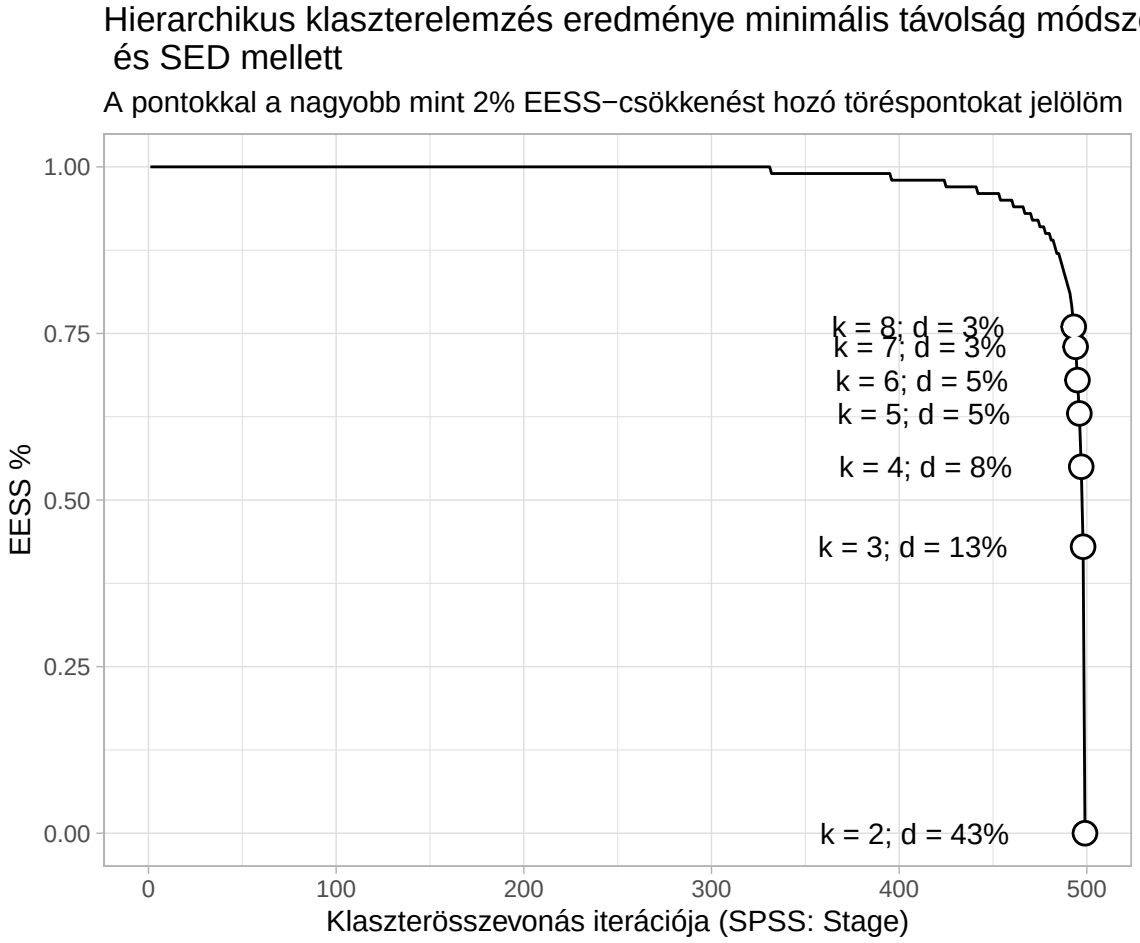


2. Végezz HKA-t SPSS-ben a Ward módszerrel, SED távolsággal a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel, standardizálással! Hány klaszteres megoldás tűnik a legjobbnak az Agglomeration Schedule Coefficients oszlopának grafikus ábrázolása alapján? És a jégcsapdiagram alapján? És a dendrogram alapján?

Hasonló grafikát képeztem Ward módszerének kiértékelésére, mint fentebb. Elsőként az SPSS *Agglomeration Schedule* táblázatának utolsó sorait közlöm.

k	EESS	d
16	0.87	0.01
15	0.86	0.01
14	0.85	0.01
13	0.84	0.01
12	0.83	0.01
11	0.82	0.01
10	0.81	0.01
9	0.79	0.02
8	0.76	0.03
7	0.73	0.03
6	0.68	0.05
5	0.63	0.05
4	0.55	0.08
3	0.43	0.13
2	0.00	0.43

k=klaszterszám, EESS%, d=EESS% csökkenés

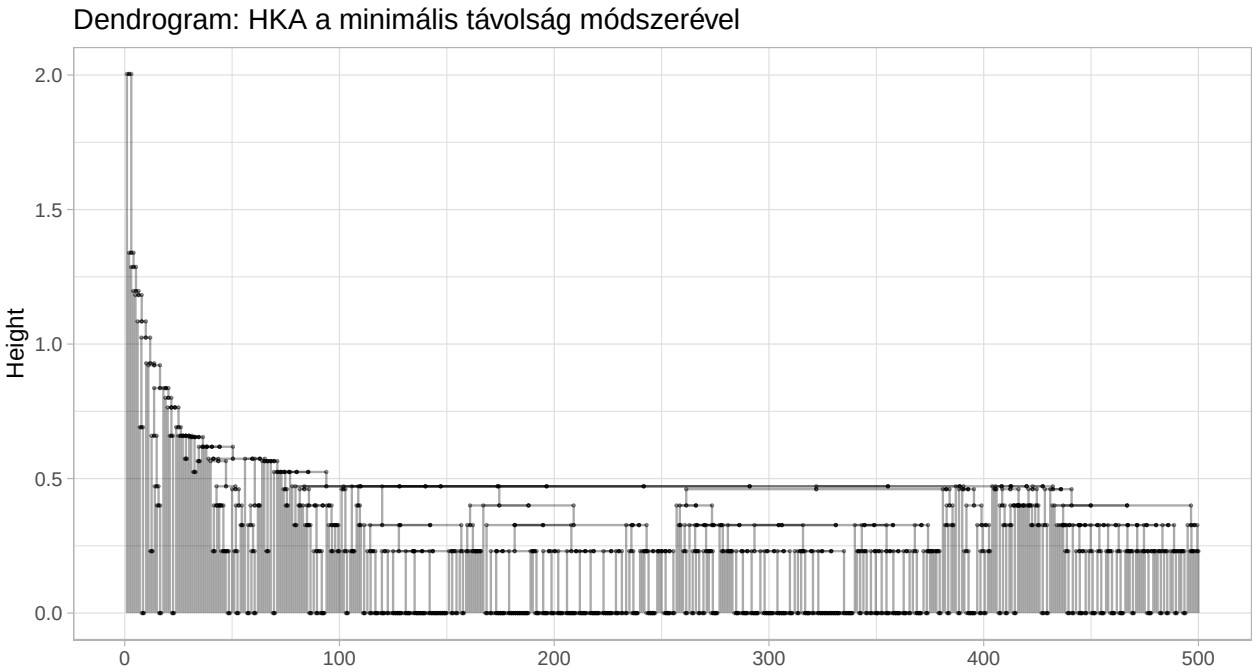


3. Mekkora az előző feladat 5-klaszteres megoldásában EESS% értéke? Vesd össze ezt a ROPstat hasonló elemzésében kapott értékkel (6. óra/7. feladat)!

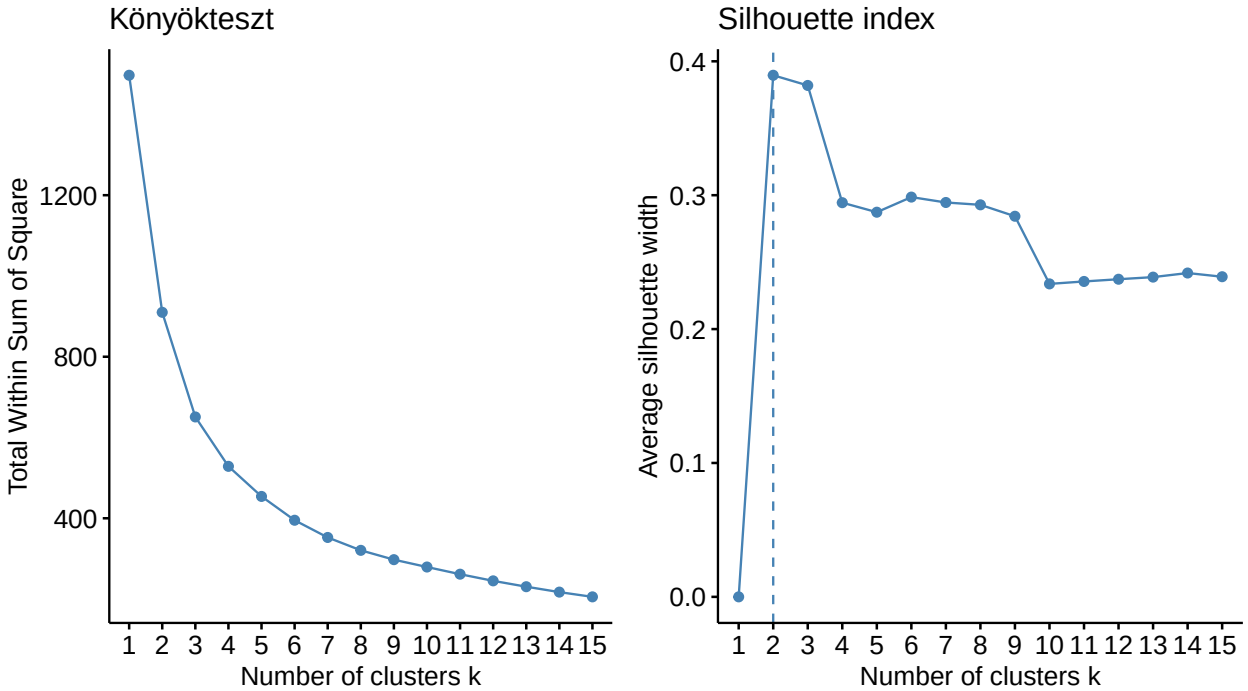
5 klaszteres megoldás esetén az EESS% 63%. A ROPStat megoldása EESS% = 69.89% volt.

4. Végezz HKA-t R-ben a minimális távolság módszerrel, SED távolsággal a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel, standardizálással! Hány klaszteres megoldás tűnik a legjobbnak a dendrogram alapján? És az 5.22. ábra szerint elvégzett elemzés grafikonja alapján?

A dendogram minimális távolság módszere mellett nagyon hasonló eredményt hoz mint az SPSS megoldása. Jól láthatóan a klaszterképzés eseti, azaz nem nagyobb homogénebb csoportok kialakítása zajlik. A minimális távolság módszerének megfelelősége vetődik fel.



A könyök teszt hasonlóképp nem egyértelmű k=15 alatt, a Silhouett Index k=2-t valószínűsít.

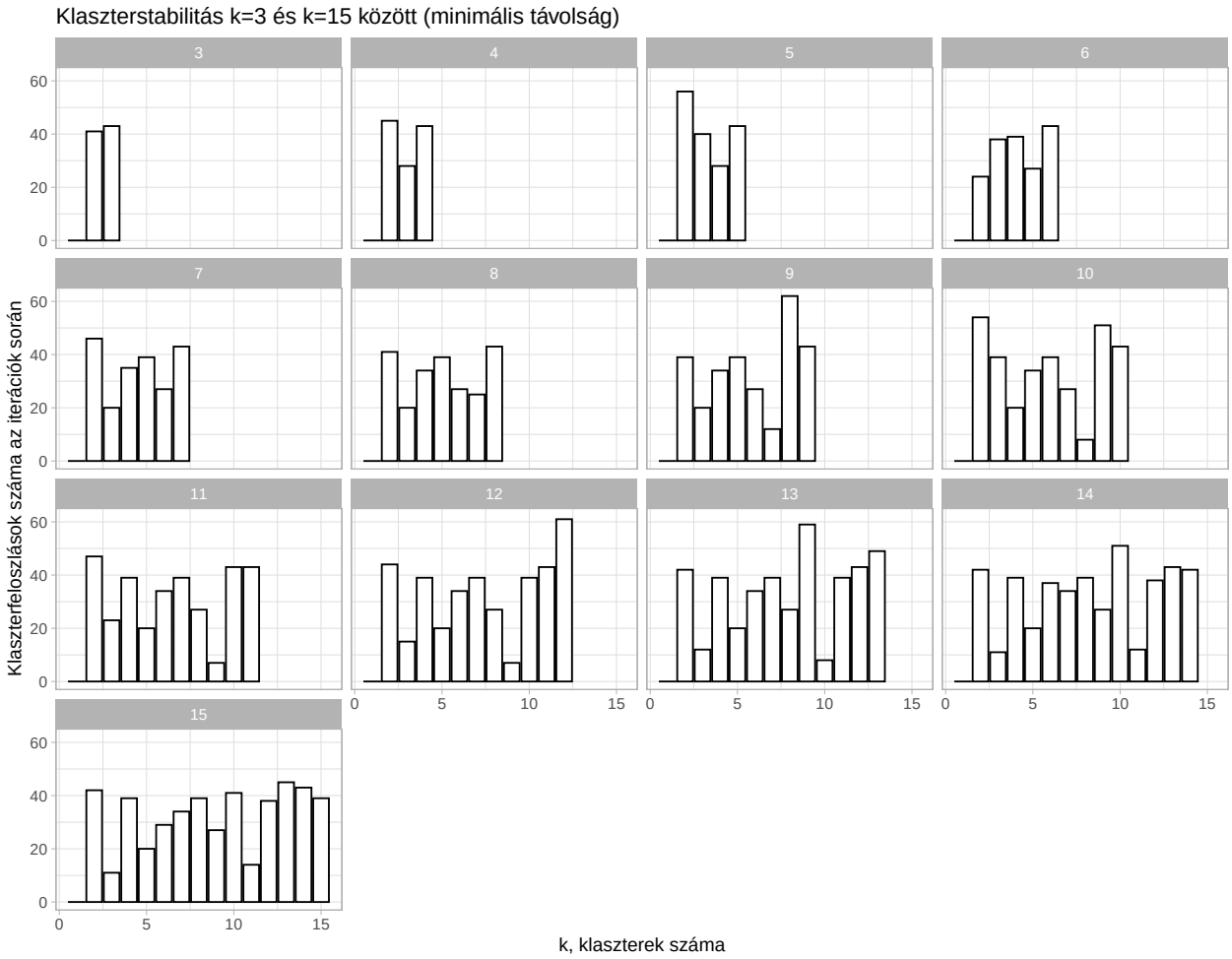


A kettőtől tizenötig terjedő klaszteres megoldások további tesztelésére az *fpc* R csomag *clusterboost* függvényét használtam. Ez a funkció alapesetben B=100 bootstrap újramintázással teszteli az adott klaszterstruktúra stabilitását, azaz számon tartja, a korábbi iterációban képzett klaszter megmarad-e, vagy feloszlik-e az újabb iteráció során. Az alábbiakban a stabilitási indexre [0 és 1 között], illetve a feloszlások számára közlöm az eredményeket.

A Silhouette index alapján kínált k=2 megoldás elemzése során azt találjuk 100 esetből 40 alkalommal nem volt megismételhető egy stabil második klaszter. Mint az első feladat megoldásánál, itt is a minimális távolság nem megfelelőségében látom a klaszterstruktúra nem illeszedésének okát.

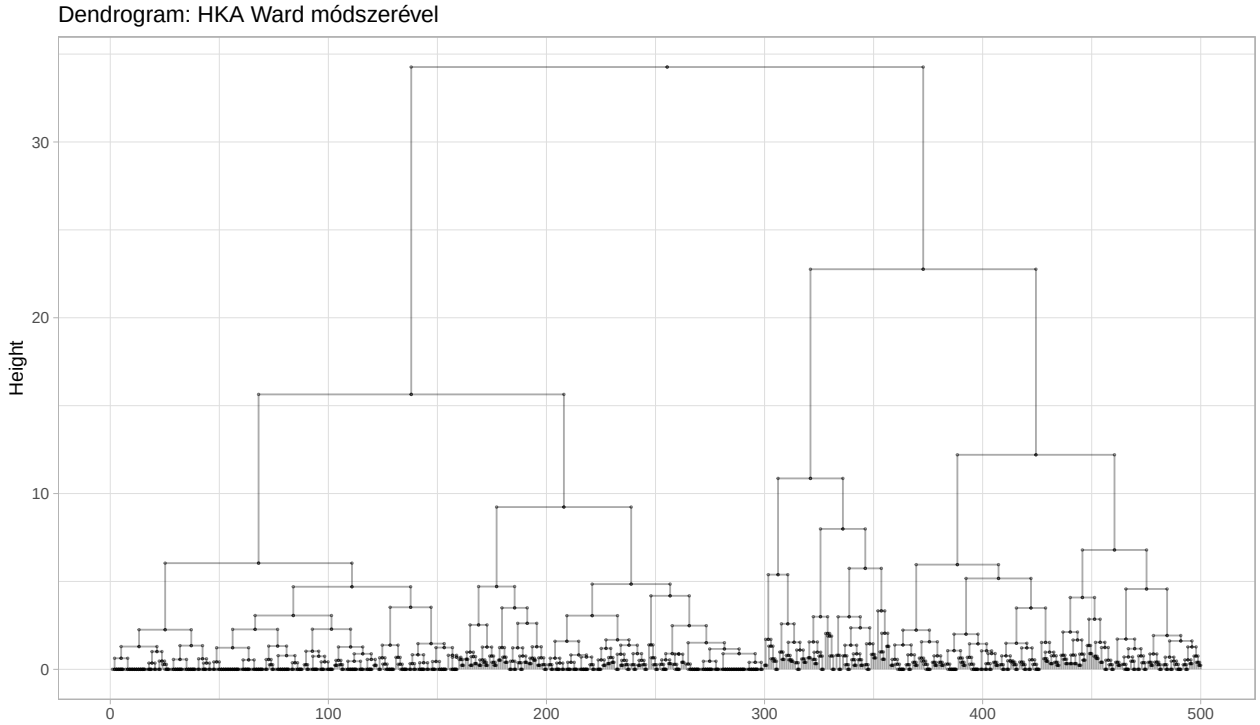
k	stabilitás	feloszlások száma
1	0.9987	0
2	0.6000	40

Az alábbi ábrán a k=3 és k=15 közti klaszterstruktúra stabilitását mutatom be. Az első klaszter teljesen stabil, apró vonalka jelzi, hogy 0 a feloszlások száma. A többi esetben 30 és 60 között ingadozik a feloszlások száma 100 futásból, a k értéke mentén váltzakozva. Ennek alapján a minimális távolság egy algoritmustól függő megoldást hoz, amely arra tereli a figyelmet, hogy más módszer lenne megfelelő valós struktúrák kinyerésére a kérdéses három változó esetén.



5. Végezz HKA-t R-ben a Ward módszerrel, SED távolsággal a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel, standardizálással! Hány klaszteres megoldás tűnik a legjobbnak a dendrogram alapján? És az 5.22. ábra szerint elvégzett elemzés grafikonja alapján?

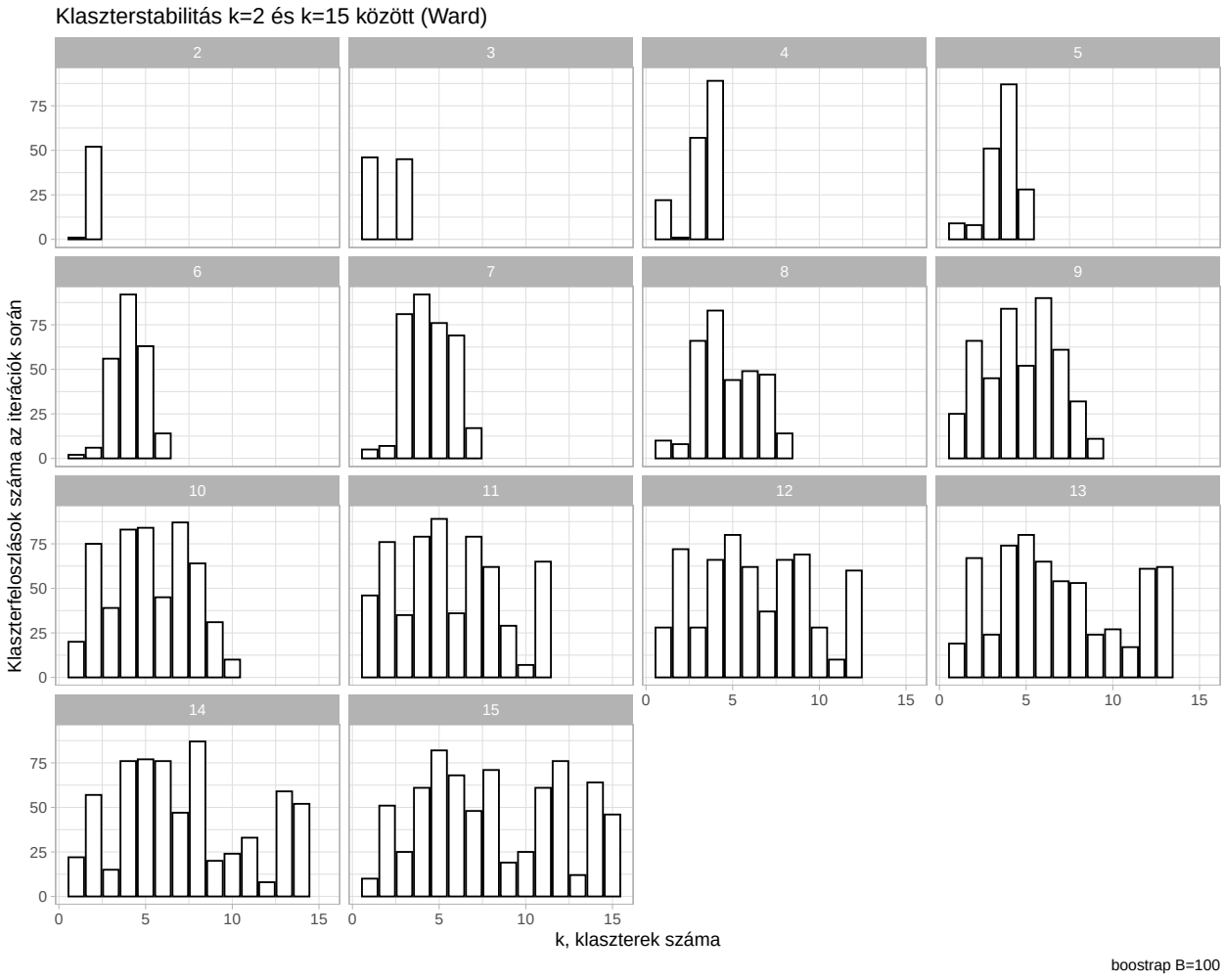
Az alábbi dendogram alalpján a k=7, illetve k=8 megoldások tűnnek megfelelőnek. Ez a Silhouette index adta k=2 megoldással nem fér össze.



Alább táblázatosan összegzem a stabilitást Ward módszerével.

k	stabil klaszterek száma	átlagos klaszterstabilitás	minimum klaszterstabilits	maximum klaszterstabilitás
2	0.500	0.672	0.569	0.774
3	0.333	0.624	0.518	0.788
4	0.250	0.574	0.350	0.750
5	0.200	0.582	0.373	0.739
6	0.167	0.557	0.362	0.756
7	0.286	0.535	0.374	0.744
8	0.250	0.563	0.347	0.762
9	0.111	0.537	0.345	0.742
10	0.100	0.523	0.380	0.726
11	0.091	0.527	0.367	0.737
12	0.083	0.541	0.390	0.714
13	0.077	0.554	0.404	0.700
14	0.214	0.563	0.345	0.751
15	0.133	0.551	0.403	0.756

Ha grafikusan tekintjük a stabilitást (lásd az alábbi diagrammot), noha a feloszlások száma magasabb is mint 60 nem egy esetben, több olyan klaszterstruktúra is képződött, ahol nem egy klaszter stabilitása magasabb volt (<30). Mindez, illetve a dendogram alapján a Ward egy megfelelőbb módszernek tűnik a klaszterstruktúra feltárására az adott adatokon.



6. Mentsd el az előző feladat 5-klaszteres megoldását, másold be ROPstatba és ott a Validálás modul segítségével számítsd ki a főbb QC mutatókat! Vesd össze ezt a ROPstat hasonló elemzésében kapott értékekkel (6. óra/7. feladat)!

Az alábbi táblázat tartalmazza a főbb mutatókat a ROPStat és az R megoldásaiból. A két rendszer nagyon hasonló melgoldásokra jutott. A legnagyobb eltérést a módosított Xien-Beni indexben találtam.

QC	ROPStat	R
EESS%	69.89	69.66
Pontbisz	0.361	0.36
Xbmod	0.394	0.449
Sil.eh.	0.595	0.587
Hcatlag	0.609	0.613
HCmin-HCmax	0.24-1.44	0.25-1.90