Házi feladatok megoldása 7.

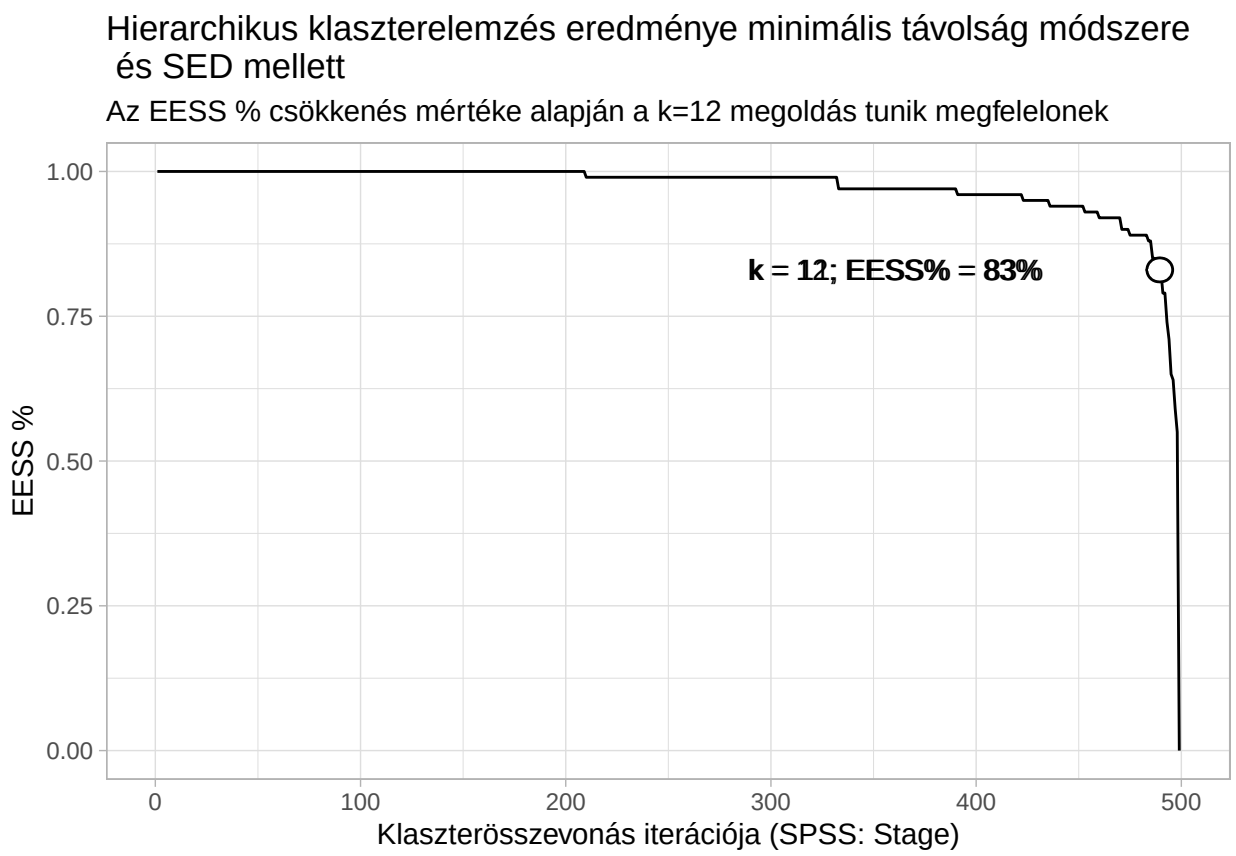
Hierarchikus klaszterelemzés SPSS-ben és R-ben

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?

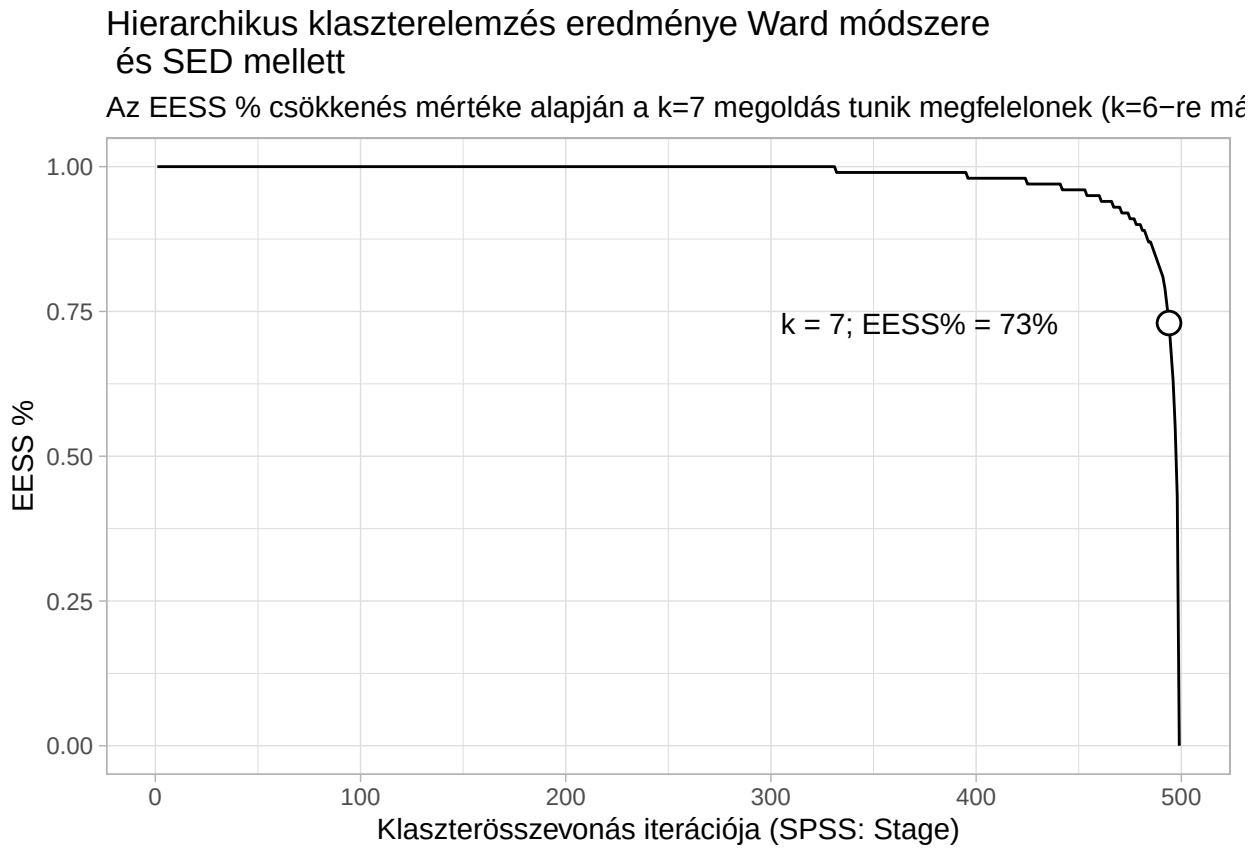
Az eredményekről az alábbi könyökábra tájékoztat. K=11 megoldás esetén az EESS % értékének csökkenése elsőként éri el a 4%-ot, így k=12-t fogadtam el. Noha később ez az érték 55% k = 3 után, ezt tekintettem egy első nagyobb törésnek a magyarázott variancia arányában, így ezt a struktúrát fogadtam el.

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, és a fentebbi okokból elogadom a k=12 megoldást, de 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, mely alapján a k=7 megoldást fogadom el az EESS% csökkenésének első nagyobb esete alapján. A dendogram is ezt erősítette meg.

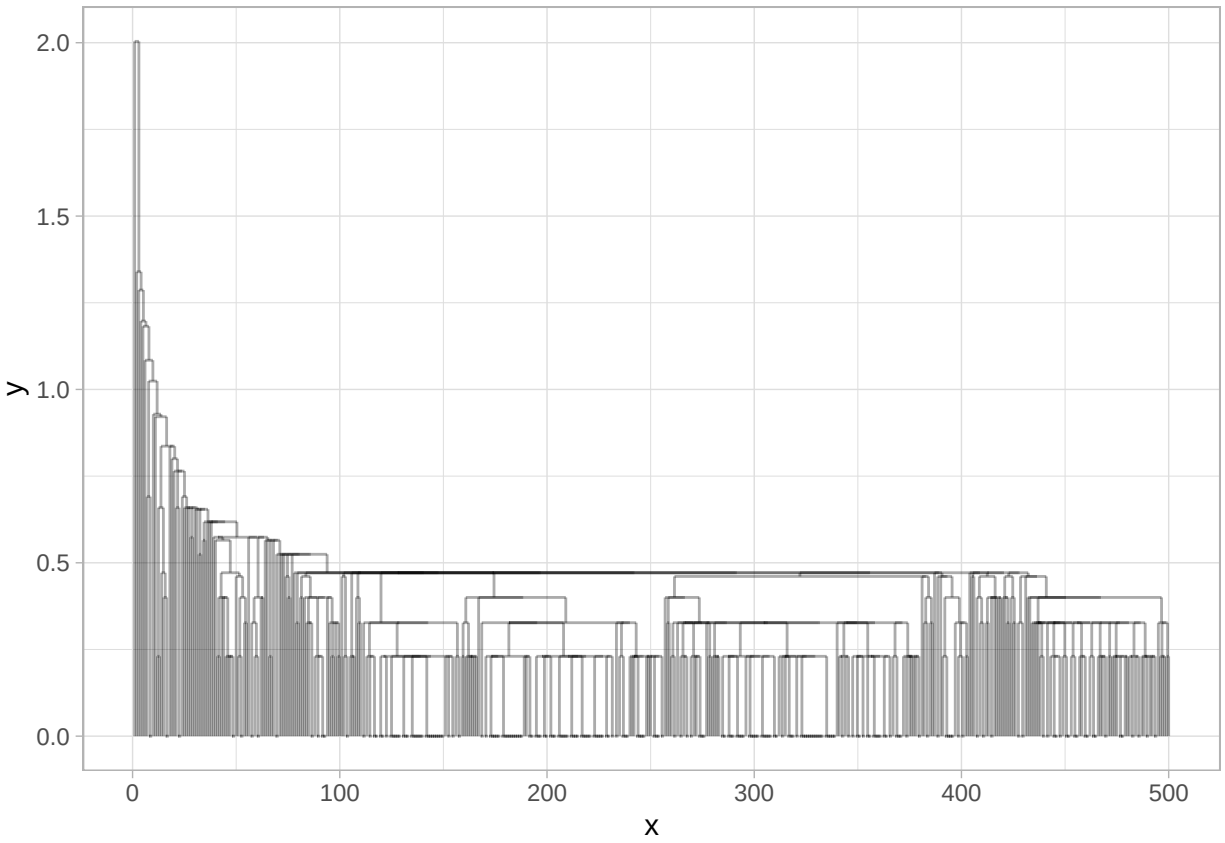


3. Mekkora az előző feladat 5-klaszteres megoldásában EESS% értéke? Vesd össze ezt a ROPstat hasonló elemzésében kapott értékkel (6. óra/7. feladat)!

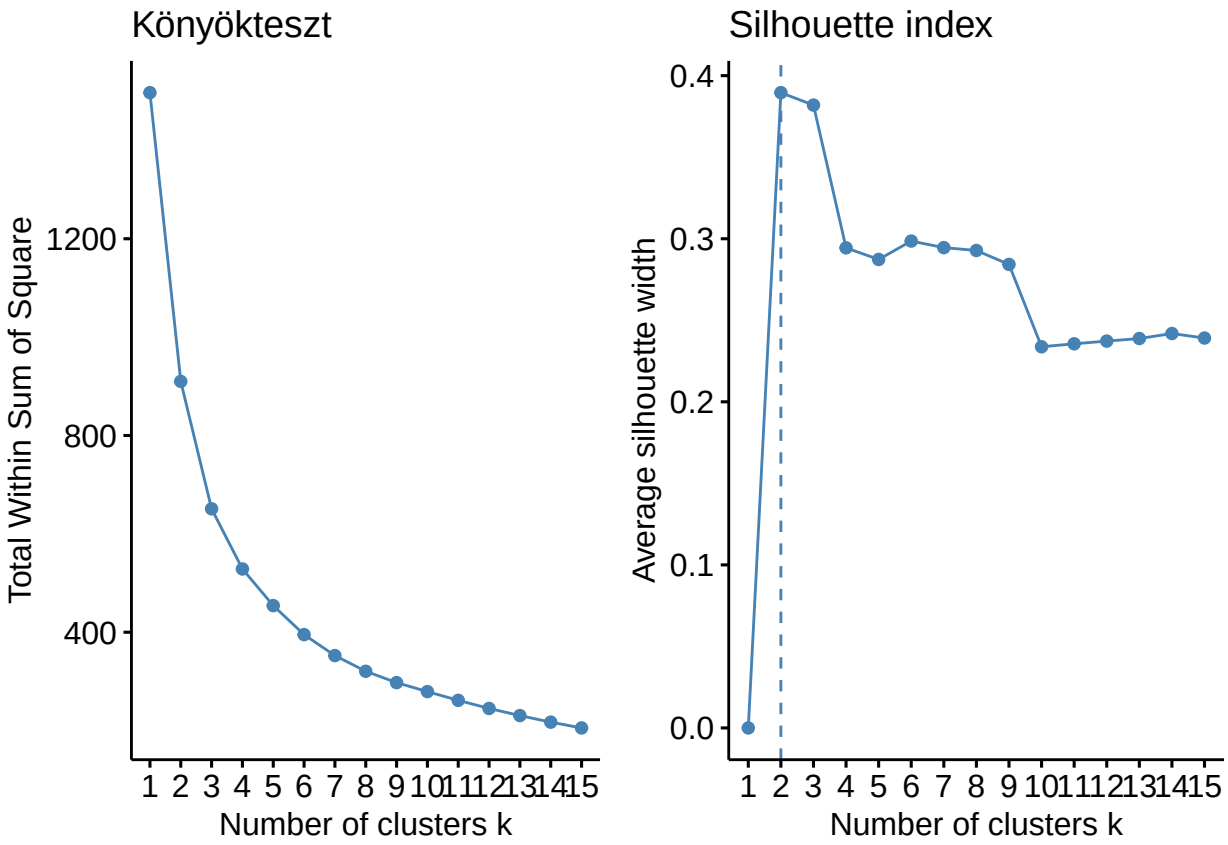
5 klaszteres megoldás esetén az EESS% 63%. A ROPStat megoldása EESS% = 69.89% volt.

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 nagyo hasonló eredményt hoz mint az SPSS megoldása. Jól láthatóan a klaszterképzés eseti, nem nagyobb homogénebb csoportok kialakítása zajlik.



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



A két és a tizenkét 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

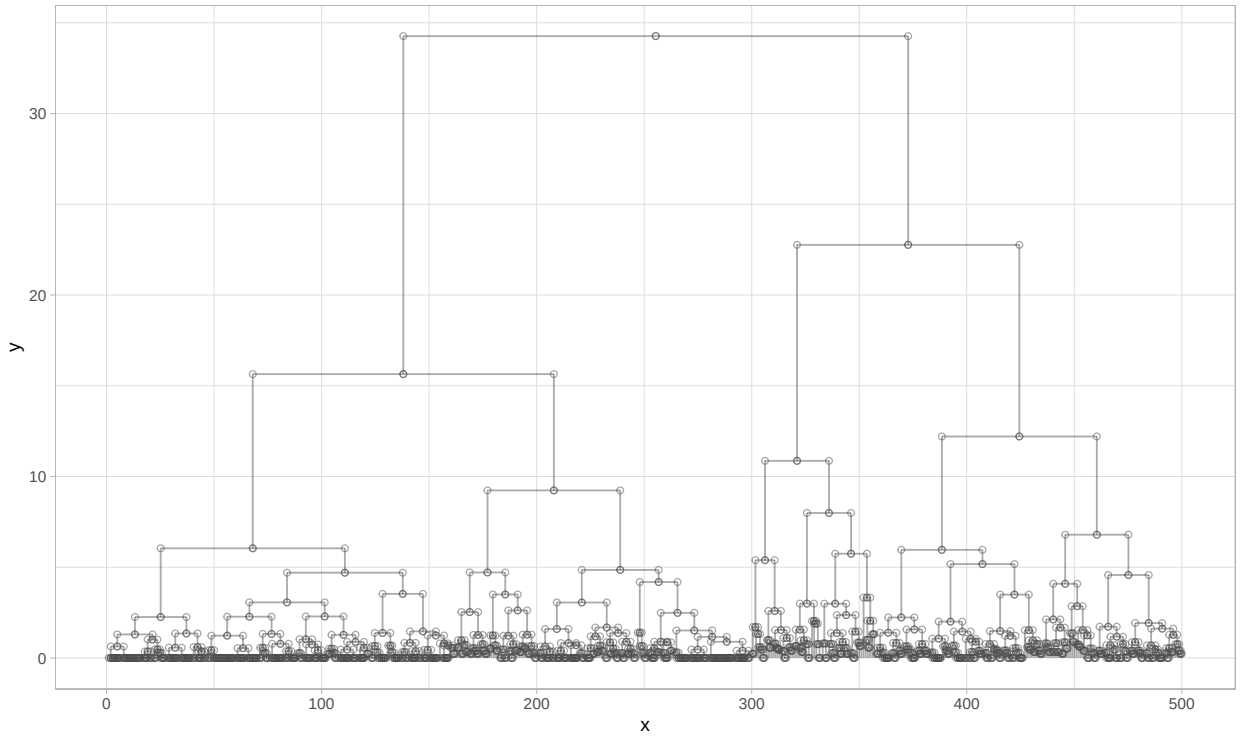
Tizenkét klasztert tesztelve feltűnik, hogy több esetben is a k=2-höz képest is stabilabb struktrúrát tárunk fel, például a 3. vagy a 9. klaszter, de fontos, hogy a k=12 megoldás is csak egy EESS% mutató alapján került korábban kiválasztásra. Ennek alapján a minimális távolság egy algoritmustól függő megoldást hoz, amely arra tereli a figyelmet, hogy más módszer lenne megfelelő valós struktúrák kinyerésére a kérdéses három változó esetén. Az eredményeket továbbá úgy értelmezném, hogy a minimális távolság esetén adott egy erős, első klaszter k=2 és k=12 esetén is, de minden más struktúra instabil, azaz nem tudunk egy stabil csoportosítást leképezni, ha ezt a módszert követjük.

k	stabilitás	feloszlások száma
1	0.9796	0
2	0.5801	42

k	stabilitás	feloszlások száma
3	0.8112	19
4	0.5900	41
5	0.7400	26
6	0.6900	31
7	0.6200	38
8	0.5700	43
9	0.9450	6
10	0.5802	42
11	0.6500	35
12	0.4099	61

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.



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

k	stabil klaszterek száma	átlagos klaszterstabilitás	minimum klaszterstabilits	maximum klaszterstabilitás
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

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
HCtlag	0.609	0.613
HCmin-HCmax	0.24-1.44	0.25-1.90