Házi feladatok megoldása 8.

k-középpontú klaszteranalízis (KKA) ROPstatban és SPSS-ben

Smahajcsik-Szabó Tamás, M9IJYM

1. Végezz KKA-t ROPstatban a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel, standardizálással k = 6 és 10 között és mentsd el minden megoldás klaszterváltozóját! Melyik megoldás tűnik a legjobbnak szerinted? Miért?

A ROPStatban elvégzett k-központú elemzések adekvációs mutatóit az alábbi táblázatban foglaltam össze. Kedvező pontbiszeriális együtthatója (0.37), magas módosított Xien-Beni mutatója (0.534 értékkel a többi klaszterstruktúra felett áll), a legjobb Silhouette indexe (0.61), és ugyancsak a legjobb GDI24 mutatója (0.484) a **k=7** struktúrát fogadtam el a legkedvezőbb megoldásnak. Ugyan nem a legjobb a megmagyarázott variancia ebben a megoldásban (76.83%), a többi mutatóban mutatott jobb teljesítménye ezt ellensúlyozza.

k	EESS%	Pontbisz	XBmod	Sil.eh.	HCatlag	CLdelta	GDI24	HCmin-HCmax
6	74.70	0.369	0.491	0.614	0.513	0.888	0.345	0.24-1.44
7	76.83	0.370	0.534	0.610	0.471	0.895	0.484	0.24-1.03
8	78.72	0.355	0.456	0.598	0.434	0.893	0.381	0.24-1.03
9	80.26	0.351	0.496	0.590	0.404	0.898	0.381	0.24-1.03
10	81.48	0.342	0.473	0.592	0.379	0.896	0.342	0.18-1.03

2. Készíts ábrát az előző feladat legjobb megoldásának centroidjai alapján a sima és a standardizált átlagokra! Hogyan tudnád értelmezni a klasztereket?

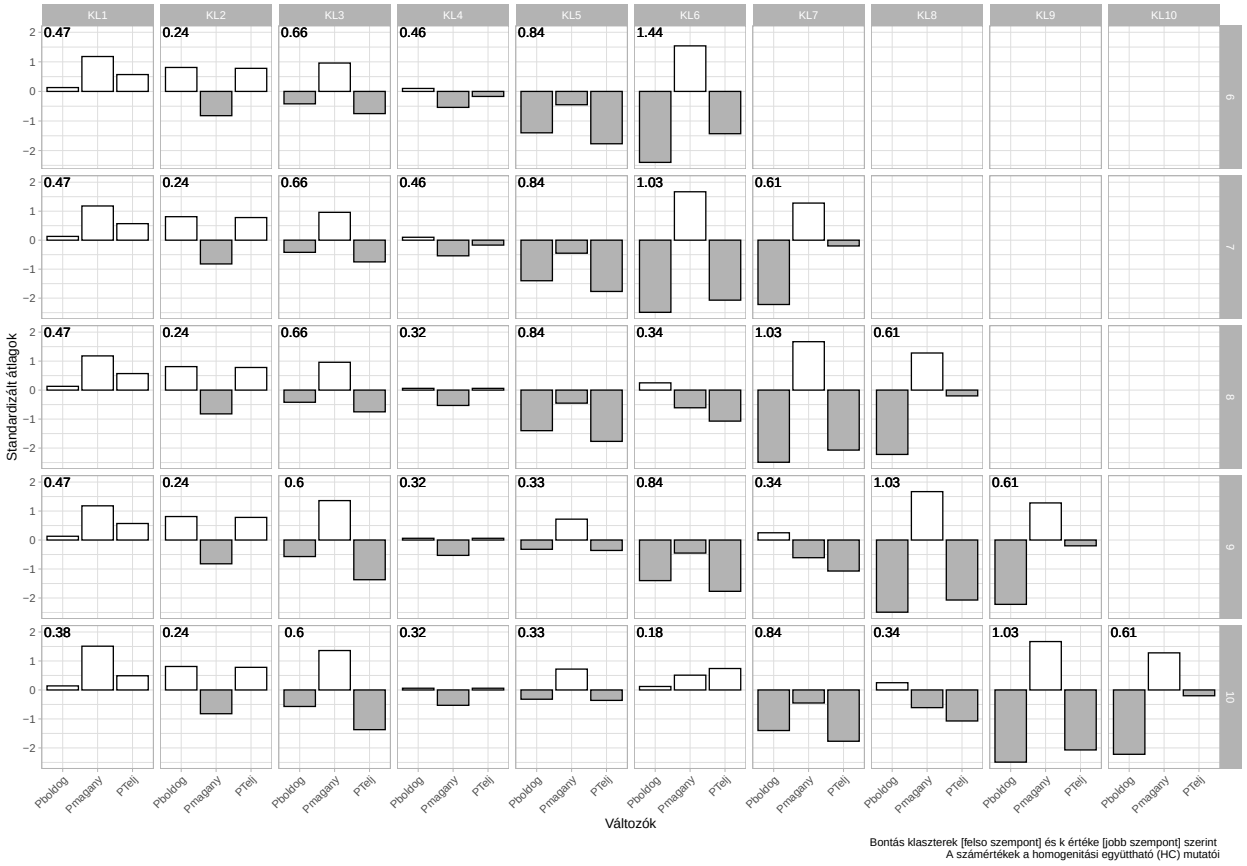
Tovább elemezve a kapott struktúrákat, a standardizált átlagok és homogenitási együtthatók alapján az alábbi ábrát képeztem klaszterstruktúránként. A k=6 megoldáshoz képest jól látható, hogy a k=7 struktúrában adódik egy KL7 klaszter, mely mintázatában új struktúrát tár fel kedvező homogenitással (HC=0.61), miközben csökken a KL6 klaszter heterogenitása (1.44-ről 1.03-ra). Ehhez képest a k >= 8 megoldások esetén a KL7 struktúra továbbra is megmarad, minden esetben az utolsó klaszterként (KL8, KL9 és KL10 formájában, azonos homogenitással), ebből arra következtetek, hogy a 8, 9 és 10 klaszteres megoldások a minta egyéb klasztereit tagolják tovább.

k=8 esetén a korábban már azonosított KL4 alakult tovább, homogenitása kedvezőbb lett (0.32), képződött egy KL6 klaszter szintén jó homogenitással. A KL4 (k=8) egy átlagosan boldog és átlagosan telesítő, viszonylag alacsony szinten magányos klaszter, noha homogén, nem látok benne különösebben szakmailag érdekes csoportot.

k=9 esetén pedig további gondot jelent, hogy a korábban is már hasonló mintázatot mutató klaszterek (pl. k=7-nél a KL3 és KL6) mellett megjelent egy KL5 klaszter, mely a redundancia irányába mutat: A rosszul teljesítő, szomorú és magányos személyeket sürítő KL3 mellett adódott már 3 további klaszter is hasonló mintázattal (KL5, KL8 és KL9 is), melyek a változókban mutatott áltagos értékeik arányában mutatnak változatosságot, de döntően ugyanazt a jelenségkört mutatják be.

A 10 klaszteres megoldás eseténe hasonló redundanciát látok a KL3, KL5, KL9 és KL10 miatt, noha szakmailag érdekesnek tartom a KL6 homogén klasztert (jól teljesítő, magányos, átlagosan boldog - talán

introvertált) klasztert.



3. Ennél a megoldásnál hány százalékponttal nő meg EESS% értéke a relokáció hatására a HKA-hoz képest?

7 klaszteres megoldás esetén Ward-módszerével végzett hierarchikus agglomeratív klaszterelemzésnél ugyanaezen a mintán az EESS% = 76.83%, tehát a két megoldás ugyanannyi varianciát magyaráz.

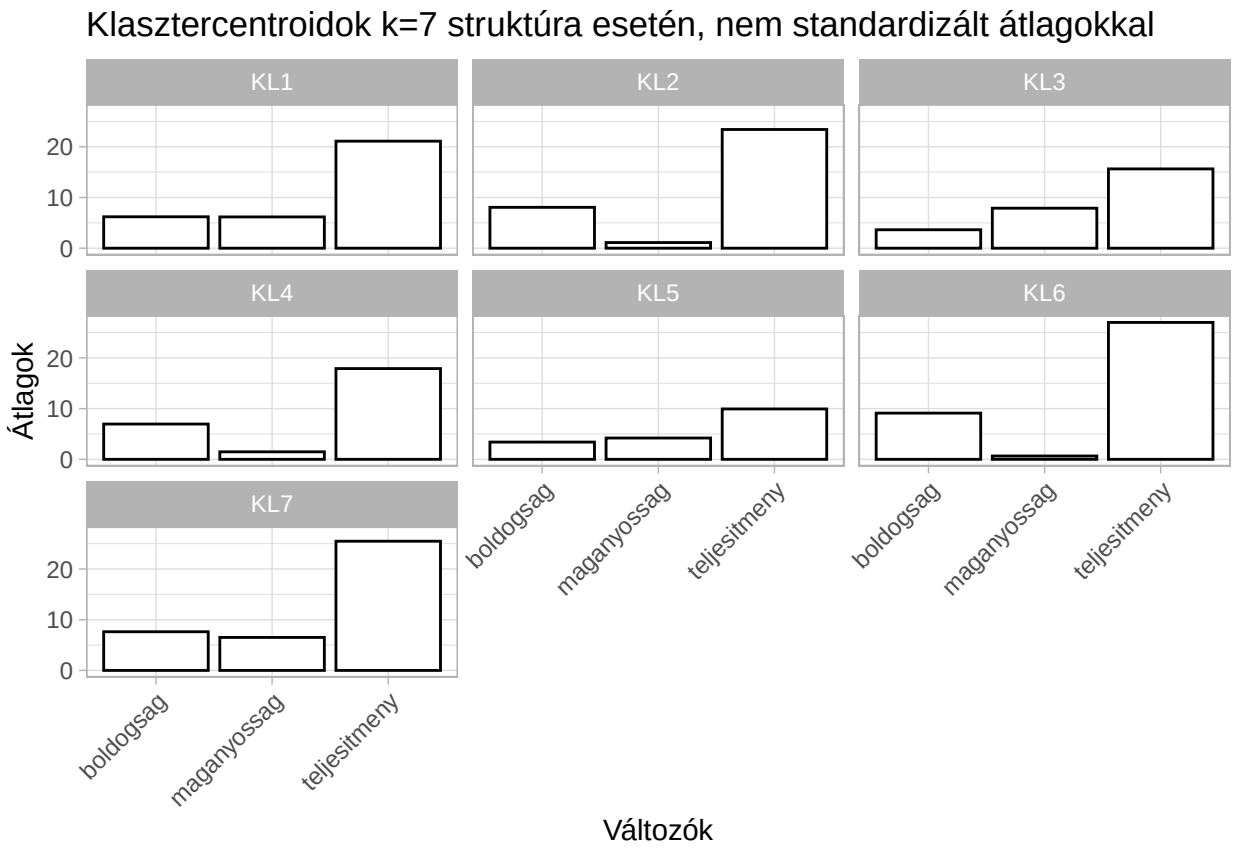
4. Végezz KKA-t SPSS-ben a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel, standardizálással a ROPstatban kapott legjobbnak tűnő klaszterszámra! EESS% értéke mekkora? Készíts ábrát a kapott megoldás centroidjairól! Hogyan tudnád értelmezni a klasztereket?

A hét klaszteres megoldásra az SPSS-ben 25-szörös iteráció mellett az EESS% 81.51%.

Valtozo	Cluster MS	df	Error MS	df_1	F	Sig.	SS_hatas	SS_hiba
Teljesitmeny	1381.3586	6	2.270833	493	608.3048	0	8288.151	1119.5206
Boldogsag	210.6344	6	1.993074	493	105.6832	0	1263.806	982.5856
Maganyossag	624.5612	6	1.854991	493	336.6923	0	3747.367	914.5107

SS_hatas	SS_hiba	SS_Total	EESS%
13299.33	3016.617	16315.94	81.51123

A klasztercentroidokat az alábbi táblázat mutatja be.



A struktúra szerint adott több olyan klaszter, ahol a teljesítmény magasabb, miközben a magányosság alacsonyabb, a boldogság pedig emelkedett (KL2, KL4, KL6). Adott két klaszter (KL1 és KL7), ahol a magányosság szintje is emelkedett, noha eme reziliens csoportokban megtartott a teljesítmény. KL3 és a hozzá hasonlóbb mintázatot mutató KL5 esetében a boldogság szinjte alacsopnyabb, miközben a magányosság és a teljesítmény emelkedett.

5. Mentsd el az előző feladat klaszterváltozóját, másold be ROPstatba és ott a Validálás modul segítségével számítsd ki a főbb QC mutatókat! Vesd össze ezeket a ROPstat hasonló elemzésében kapott értékekkel (lásd 1. feladat)!

A Validálás modul segítségével kapott eredményeket az alábbi táblázat összegzi 7 klaszteres struktúrára.

Software	EESS%	Pontbisz	XBmod	Sil.eh.	HCatlag	CLdelta	GDI24	HCmin-HCmax
ROPStat	76.83	0.370	0.534	0.610	0.471	0.895	0.484	0.24-1.03
SPSS	74.91	0.334	0.209	0.565	0.511	0.855	0.204	0.23-1.55

Az SPSS által használt KKA algoritmus (esetünkben a McQueen-féle) gyengébb teljesítménye valószínűsíthető volt a fentebbi feladat megoldásában is, a klaszterstruktúrákat tekintve.