A klaszteranalízis, mint sertéstelepeket minősítő eljárás[1]

Kovács Sándor – Balogh Péter
Debreceni Egyetem Agrár- és Műszaki Tudományok Centruma,
Agrárgazdasági és Vidékfejlesztési Kar,
Gazdaságelemzési és Statisztikai Tanszék, Debrecen
kovacss@agr.unideb.hu

ÖSSZEFOGLALÁS

A klaszteranalízis a többváltozós statisztikai módszerek egyik legismertebb eljárása, amely tulajdonképpen egy csoportosító eljárás. A megfigyelési egységekhez rendelt változók mentén kívánjuk a megfigyeléseket csoportosítani. Célunk olyan klaszterek létrehozása, amelyeknek elemei a lehető legszorosabban kapcsolódnak egymáshoz, és viszonylag jobban eltérnek a többi klaszter elemeitől. Ezek alapján például telepek egyféle minőségi besorolása adható meg a klaszterek mentén. A klaszteranalízisen belül is számos módszer létezik, valamint az alkalmazható távolságmértékekből is széles a választék. Mindezekkel részletesen foglalkozunk a cikkben. Az elméleti háttér bemutatása után konkrét esettanulmány keretében bemutatjuk a módszer gyakorlati alkalmazhatóságát sertéstelepek szakértői minősítésének ellenőrzésére. Az esettanulmányban mind a hierarchikus módszert, mind a nem hierarchikus módszert kipróbáltuk, és összehasonlítottuk. Külön szeretnénk felhívni a figyelmet arra, hogy a klasztermódszer egyik legfontosabb problémája az optimális klaszterszám meghatározása.

Kulcsszavak: *klaszteranalízis, sertés, sertéstelepek, minősítés*

SUMMARY

Cluster Analysis is one of the most favorite multivariable statistical methods, which is actually a special type of aggregating method. Observations are clustered by variables belonged to the observations. Our purpose is to create such clusters, in which the elements are the most similar, and between the clusters they are the most variant. For example these clusters could be the qualitative classifications of farms.

There have been several methods in Cluster Analysis as well as numerous distance measures, which could be used. In this article, we study all of these methods and measures. After we show the theoretical background, we apply the method in a given casestudy to control the qualitative classifications of experts. In this study, we use both the hierarchical and the non-hierarchical method, and also compare them. We would like to attract the attention that the most important problem of the analysis is to determine the optimal of clusters.

Keywords: *clusteranalysis, swine, swine-farms, qualifying*

BEVEZETÉS

Amennyiben egy *N* elemű sokaságot (pl. 62 sertéstelepet) egy osztályozó változó szerint (pl. fajlagos takarmányhányad nagysága szerint) kívánunk csoportosítani, akkor a sokaság elemeinek egy adott ismérv szerinti sorba rendezéséről van szó. Ha egy második osztályozó változót (pl. az 1 dolgozóra jutó sertések számát) is figyelembe kell venni, akkor az előbbi sorba rendezett sokaságot általában már nem tarthatjuk meg, hanem különböző elemeket is tartalmazó csoportokat kell alkotnunk. További harmadik osztályozó változóra is tekintettel akarunk lenni (pl. az 1 sertésre jutó súlygyarapodásra is), s ekkor az újabb csoportosítás az előző csoportokat tovább bontja. Így egy negyedik, egy ötödik osztályozó változó szerinti további csoportosítás a hagyományos módon történő osztályozás szerint már alig, vagy egyáltalán nem végezhető el.

Ezen probléma megoldására egyszerre több osztályozó változó szerinti csoportosításra alkalmas osztályozó eljárást alakítottak ki klaszteranalízis néven. Az elnevezés az angol Cluster szóból ered, ami ebben a vonatkozásban csoportot jelent. A klaszteranalízis két részből áll, egy csoportképzési eljárásból és a csoportok elemzéséből. A hazai szakirodalomban a csoportképzést automatikus osztályozásnak is nevezik. A külföldi szakirodalomban a klaszteranalízis elnevezésen kívül numerikus taxonómia és taxometria elnevezéssel is találkozunk (Sokal és Sneath, 1963). A csoportképzés alapja a sokaság elemeinek elhelyezkedése a két, a három, általában az *n*-dimenzisós térben, amikor is a sokaság egy-egy eleme a két, a három, illetve az *n*-dimenziós tér egy-egy pontja, majd e pontok egymástól számított távolsága.

MÓDSZERTAN

A klaszterelemzés a többváltozós statisztikai eljárások egyik kedvelt módszere. Mielőtt az elméleti háttér bemutatására térnénk rá, szeretnénk összehasonlítani a módszert a többi igen kedvelt többváltozós technikával, s rávilágítani az esetleges hasonlóságokra, illetve különbségekre. A klaszterelemzés a faktorelemzéshez hasonló módszer, mely az összefüggések halmazát vizsgálja. A klaszterelemzés sem tesz különbséget függő és független változók között. Sokkal inkább a változók halmazán belüli kölcsönös összefüggéseket vizsgálja.

A klaszterelemzés fő célja, hogy a megfigyelési egyedeket viszonylag homogén csoportokba sorolja a kiválasztott változók alapján úgy, hogy az adott csoportba tartozó megfigyelési egységek hasonlítsanak egymásra, de különbözzenek más csoportok tagjaitól (Malhotra, 2001).

Ily módon a klaszterelemzés a faktorelemzés olyan kiegészítő módszere, amikor a megfigyelési

[1]OTKA F 62949/2006 támogatásával

egyedeket sokkal kisebb számú klaszterekbe soroljuk. A klaszterelemzés csakúgy, mint a diszkriminanciaanalízis, csoportosítással foglalkozik. A két módszer közötti különbség azonban az, hogy míg a diszkriminanciaanalízis megköveteli a klaszterbe tartozás előzetes ismeretét, s ez alapján mindegyik megfigyelési egységre vagy esetre csoportosító szabályt alakít ki, addig a klaszterelemzésnél nem rendelkezünk előzetes információval egyetlen megfigyelési egyed csoportba tartozásáról sem. A csoportok nem adottak előre, hanem az adatok alapján alakítja ki a módszer azokat. A klaszteranalízis segítségével nemcsak a megfigyelési egyedeket tudjuk csoportosítani, hanem változócsoportok kialakítására is lehetőség van.

A klaszterelemzés menete (Malhotra, 2001)**:**

- A probléma megfogalmazása
- A távolsági mérték kiválasztása
- A klasztermódszer kiválasztása
- Döntés a klaszterek számáról
- A klaszterek értelmezése és jellemzése
- A klaszterelemzés érvényességének ellenőrzése

1. A probléma megfogalmazása

Ennek a folyamatnak a során adjuk meg, hogy a csoportok kialakítása mely változók szerint történjen. Ez azért nagyon lényeges pontja az elemzésnek, mert egyetlen nem megfelelő változó bevonása is ronthat a bevonása nélkül egyébként megfelelő csoportosításon. A változók kiválasztásában segítségünkre lehet korábbi kutatásaink eredménye, egyéb elméleti megfontolások, de dönthetünk a saját intuíciónkra támaszkodva is.

2. A távolsági mérték kiválasztása

Az egyedek csoportosítása előtt definiálni kell azt, hogy hogyan mérjük a köztük lévő hasonlóságot, vagy különbséget. A hasonlóságok illetve különbségek mérésére a megfigyelési egyedek páronként vett távolságát alkalmazzák legelterjedtebben. Az egymáshoz hasonlóbb megfigyelési egyedek között kisebb a távolság, mint azok között, melyek kevésbé hasonlók. Két megfigyelési egyed távolságát sokféleképpen számíthatjuk ki. A leggyakoribb mértékek a következők (Szűcs, 2002):

1. Az euklideszi távolság
$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$
2. Négyzetes euklideszi távolság
$$d_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$
3. City-Block (Manhattan) távolság
$$d_{ij} = \sum_{k=1}^{p} | x_{ik} - x_{jk} |$$
4. Csebisev-távolság
$$d_{ij} = \max_{1 \le k \le p} | x_{ik} - x_{jk} |$$
5. Pearson-féle távolság
$$d_{ij} = \sqrt{\sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{S_k{}^2}}$$
6. Négyzetes Pearson-féle
$$d_{ij} = \sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{S_k{}^2}$$

Mivel általános esetben az elemzésbe vont változók különböző mértékegységűek, ezért a mértékegység befolyásoló hatását az elemzés előtt ki kell küszöbölnünk, hogy ezáltal is növeljük a jobb besorolás esélyét. A változókból kivonjuk az átlagukat, majd osztjuk a szórásukkal, ezáltal egy 0 átlagú, 1 szórású változóvá transzformáljuk őket (standardizálás). Az elemzést célszerű különböző távolsági mértékek használatával újra elvégezni, majd a kapott eredményeket összehasonlítani.

3. A klasztermódszer kiválasztása

A távolsági mértékek bemutatásánál kitértünk arra, hogy két elem közötti távolságot hogyan definiálhatunk, de azt nem említettük, hogyan definiáljuk két klaszternek a távolságát. A továbbiakban ennek a lehetőségét mutatjuk be. A klaszterejárások lehetnek hierarchikusak és nem hierarchikusak.

3.1 Hierarchikus klaszterező módszerek

A leggyakrabban alkalmazott eljárások tartoznak ebbe a csoportba. A hierarchikus módszereket két csoportra bonthatjuk, mint agglomeratív és divizív eljárások, amelyekről később részletesebben lesz szó. A hierarchikus eljárások a korábbi fázisokban létrehozott csoportosításon már nem változtatnak, így az elemek átsorolása másik klaszterbe semmiképpen nem lehetséges. Mindez azt is jelenti, hogy optimális megoldást ezen eljárások nem biztosítanak. A hierarchikus klaszterelemzés nem csak egyedek összehasonlítására használható, hanem a változók számának csökkentésére, és azok csoportosítására is. Kiszűrhetünk általa olyan egyedi változókat is, amelyek jelentős mértékben magyarázzák az adatokat. Ez gyakorlatilag hasonlít arra, amit a faktoranalízis is megvalósít. A változók összevonása a korábban bemutatott módszerek szerint történik azzal a különbséggel, hogy a két változó hasonlóságának mérésére más távolsági mértékeket alkalmaznak. Szűcs (2002) munkájában részletesebben kitér ezekre a mértékekre, jelen tanulmányban azonban ezekkel nem foglalkozunk.

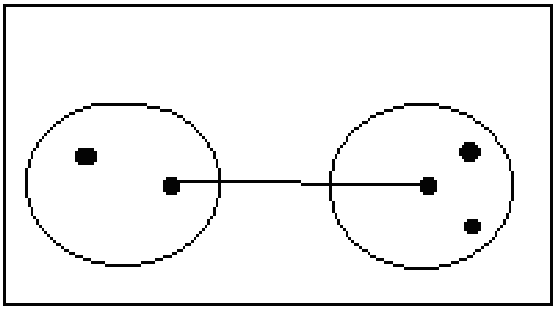
3.1.1 Agglomeratív eljárások

Ezen eljárástípus esetén n db egyelemű klaszterből indulunk ki az elemzés kezdetén. Ezután megkeressük a távolság mátrix minimális (maximális) elemét, vagyis a két leghasonlóbb (legkülönbözőbb) klasztert. Ezt a két klasztert összevonjuk, majd a klaszterszámot csökkentjük. A további összevonásokat addig végezzük, míg minden elem egy klaszterbe nem kerül.

3.1.1.1 Egyszerű lánc módszer

A módszer másik elnevezése a legközelebbi szomszéd módszer, mely utal a távolságképzés technikájára. Két klaszter távolságát a két csoport legközelebbi tagjai közötti távolságként definiálja (*1. ábra*).

1. ábra: **A legközelebbi szomszéd módszer elve**



Forrás: Malhotra (2001)(1)

Figure 1: Nearest neighbour method
Source(1)

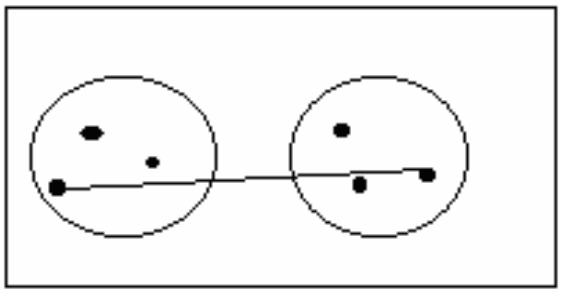
D_1 (I, J) = min d ($\underline{x}_i$, $\underline{x}_j$), ahol I és J klaszterek, n_i és n_j a két klaszter elemszáma. i=1,2,….., n_i ; j =1,2,….., n_j

Az egyszerű lánc módszer egyik igen nagy előnye, hogy képes a szabálytalan, azaz nem ellipszis alakú csoportok felismerésére is. Ez egyben hátrány is lehet, hiszen összefűzhetünk általa különböző tulajdonságú csoportokat, így a kialakított klaszterek nagyon heterogének lehetnek. Ezen módszer alkalmazása akkor javasolható, ha a klaszterek összekötése a cél, mintsem a klaszterek homogenitása. Előnyösen alkalmazható nagy elemszámú minták részekre osztásához.

3.1.1.2 Teljes lánc módszer

Legtávolabbi módszer néven is említik, mivel a távolságképzés éppen az ellentéte az egyszerű lánc módszernek. A két klaszter távolságát a két csoport legtávolabbi tagjai közötti távolságként definiálja (*2. ábra*).

2. ábra: **A legtávolabbi szomszéd módszer elve**



Forrás: Malhotra (2001)(1)

Figure 2: Furthest neighbour method
Source(1)

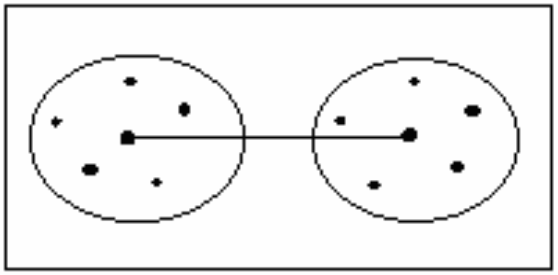
D_2 (I, J) = max d ($\underline{x}_i$, $\underline{x}_j$), ahol I és J klaszterek, n_i és n_j a két klaszter elemszáma. i=1,2,….., n_i ; j=1,2,….., n_j

A teljes lánc módszer kis méretű klasztereket hoz létre, és ezáltal a klaszterek elemeinek kisebb változtatása lényegesen átrendezheti a csoportosítást.

3.1.1.3 Centroid módszer

A klaszterhez tartozó egyedek az n-dimenziós tér pontjai, az egyedekre jellemző változóértékek pedig a pontok koordinátái. A klaszter centoridja az átlagos klaszteregyedet jelöli, amelynek kordinátái az egyedek koordinátáinak (változóértékeinek) átlaga. A módszer lényege pedig, hogy két klaszter távolságát a centroidjaik távolságával definiálja *(3. ábra)*.

3. ábra: **Centroid módszer**



Forrás: Malhotra (2001)(1)

Figure 3: Centroid method
Source(1)

I klaszter centroidja $\vec{x} = \frac{1}{n_i} \sum_{i=1}^{n_i} x_i$,

J klaszter centroidja $\vec{y} = \frac{1}{n_j} \sum_{j=1}^{n_j} y_j$,

az I és J klaszter centroidjának távolsága:

$$D_3(I,J) = d(\vec{x},\vec{y})$$

A centroid módszer hátránya, hogy ha a két összevonandó klaszter elemszáma lényegesen eltér, akkor a kisebb méretű klaszter jellege elvész az egyesítés során, azaz az új klaszter centroidja a nagyobb klaszter centroidjához lesz közelebb.

3.1.1.4 Medián módszer

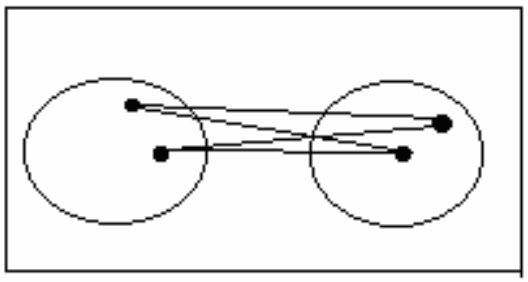
A centroid módszernek az erősen eltérő elemszámú klaszterek összevonásából adódó problémáját igyekszik feloldani a medián módszer. Az I és J klaszterek egyesítése után kapott új klaszter és a K klaszter távolsága (Füstös és Meszéna, 1983):

$$D_4 = \frac{1}{2}(d_{IK} + d_{JK}) - \frac{1}{4} d_{IJ}$$

3.1.1.5 Csoportátlag módszer

Átlagoljuk az egyik csoport minden elemének távolságát a másik csoport elemitől, majd azt a két csoportot vonjuk össze, amelyek esetén az objektumok közötti átlagos távolság minimális (*4. ábra*).

4. ábra: **Az átlagos távolság módszer**



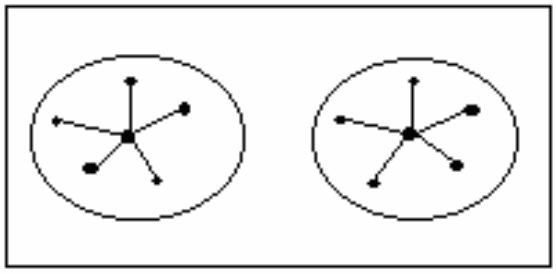
Forrás: Malhotra (2001)(1)

Figure 4: Average distance method
Source(1)

3.1.1.6 Ward módszer

A csoportok összevonásánál információveszteség keletkezik. A Ward módszer arra törekszik, hogy a csoportosítást minimális információveszteséggel hajtsa végre. Az információ-veszteség alatt az elemek csoportátlaguktól való eltéréseinek négyzetösszegét, azaz a csoporton belüli varianciát kell érteni. A teljes minta varianciája felbontható a csoportokon belüli, és a csoportok közötti variancia összegére. A cél olyan csoportosítás kialakítása, mely során a csoportokon belüli varianciák összege minimális (*5. ábra*).

5. ábra: **A Ward-féle módszer**



Forrás: Malhotra (2001)(1)

Figure 5: Ward's method
Source(1)

A csoporton belüli varianciát az alábbi kritérium szerint minimalizáljuk (Füstös és Meszéna, 1983):

$$D_6(I,J) = \frac{n_i n_j}{n_i + n_j}(\underline{\bar{x}} - \underline{\bar{y}})^T(\underline{\bar{x}} - \underline{\bar{y}}),$$

ahol $\underline{x}$ és $\underline{y}$ a két klaszter átlagvektora

3.1.1.7 Lance és Williams flexibilis módszere

Lance és Williams (1967) általános formulát ad a klaszterek közötti D távolság értelmezésére, mely speciális esetekben az előző módszerek valamelyikét adja. A formula az I és J klaszter összevonásával kapott klaszter és a K klaszter távolságát a következőképpen definiálja:

$$D_7(K, IJ) = \alpha_i d_{KI} + \alpha_j d_{KJ} + \beta d_{IJ} - \gamma \left| d_{KI} - d_{KJ} \right|$$

Lance és Williams javasolta, hogy az együtthatókat a következők szerint válasszák:

$$\alpha_i + \alpha_j + \beta = 1 \; ; \; \alpha_i = \alpha_j \; ; \; \beta < 1 \; ; \; \gamma = 0$$

A ß–ra kis negatív érték megadását javasolták, pl. -0.25. Az általános formula az együtthatók következő választásával az eddigi módszereket adja meg (Lance és Williams, 1967):

Egyszerű lánc

$$\alpha_i = \alpha_j = \frac{1}{2} \; ; \; \beta = 0 \; ; \; \gamma = -\frac{1}{2}$$

Teljes lánc

$$\alpha_i = \alpha_j = \frac{1}{2} \; ; \; \beta = 0 \; ; \; \gamma = \frac{1}{2}$$

Centroid

$$\alpha_i = \frac{n_i}{n_i + n_j} \; ; \; \alpha_j = \frac{n_j}{n_i + n_j} \; ; \beta = \alpha_i \alpha_j \; ; \; \gamma = 0$$

Medián

$$\alpha_i = \alpha_j = \frac{1}{2} \; ; \; \beta = -\frac{1}{4} \; ; \; \gamma = 0$$

Csoportátlag

$$\alpha_i = \frac{n_i}{n_i + n_j} \; ; \; \alpha_j = \frac{n_j}{n_i + n_j} \; ; \; \beta = 0 \; ; \; \gamma = 0$$

Ward módszer

$$\alpha_i = \frac{n_k + n_i}{n_k + n_i + n_j} \; ; \; \alpha_j = \frac{n_k + n_j}{n_k + n_i + n_j} \; ;$$

$$\beta = \frac{-n_k}{n_k + n_i + n_j} \; ; \; \gamma = 0$$

3.1.1.8 A távolsági mértékek és az algoritmusok kapcsolata

Lance és Williams (1967) foglalkozott azzal a kérdéssel, hogy mely algoritmusokhoz milyen távolsági mértékek alkalmazhatók célszerűen (*1. táblázat*).

Nem-Euklideszi távolsági mértékre példa a City-Block távolság, a Szemi metrikákra az ún. cosinus mérték. Ez utóbbi kettő alkalmazhatósága centroid módszer esetében kérdéses, Ward módszer esetében pedig biztosan nem alkalmazható.

1. táblázat

Távolsági mértékek alkalmazhatósága az egyes klaszter módszereknél

Algoritmus(1)	Euklideszi(2)	Nem-Euklideszi(3)	Szemi metrikák(4)
Egyszerű lánc(5)	+	+	+
Teljes lánc(6)	+	+	+
Csoportátlag(7)	+	+	+
Median(8)	+	+	+
Centroid(9)	+	?	?
Ward módszer(10)	+	-	-

Forrás: Lance és Williams (1967)(11)

Table 1: Applicable distance measures and cluster methods

Algorithm, method(1), Euclidian(2), Non-euclidian(3), Semi-metrics(4), Nearest neighbour(5), Furthest neighbour(6), Average distance(7), Median(8), Centroid(9), Ward' method(10), Source(11)

3.1.2 Divizív eljárások

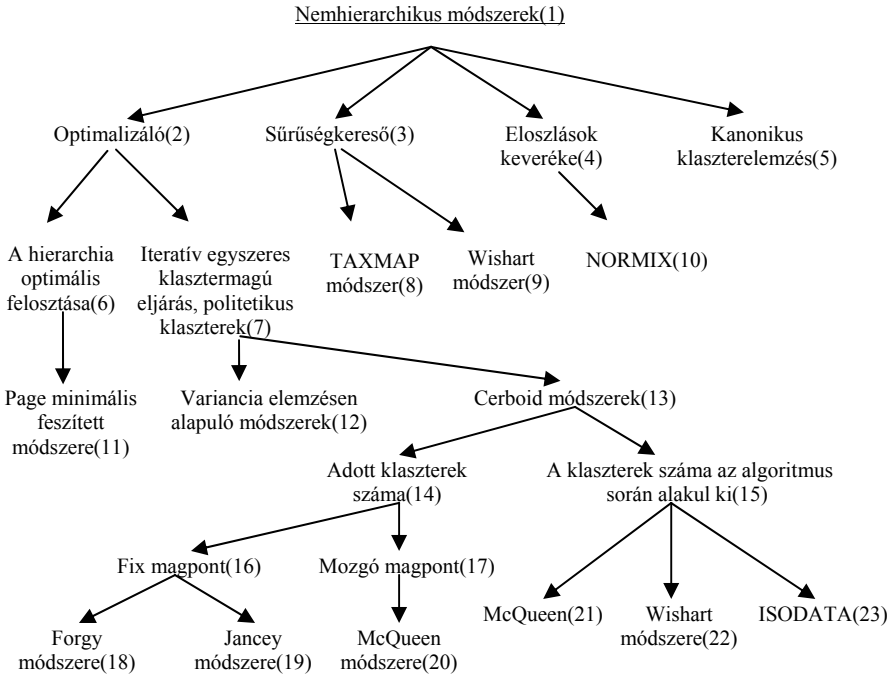
A hierarchikus módszereknek ez a típusa *n* elemet oszt egymás után kisebb elemszámú csoportokba egészen addig, amíg minden elem külön klasztert alkot. A kezdő lépésben *n* objektumot kell kétfelé bontani. Ezt többféle módon végezhetjük el, így még kis adathalmaz esetén is nagy számítógépkapacitásra van szükség az összes lehetőség vizsgálatára.

3.2 A nemhierarchikus módszerek

A nemhierarchikus módszerek nagy előnye, hogy a nagyméretű problémák kezelésére jól alkalmazhatóak. Az ilyen típusú módszerek lényege, hogy ha két elem egyszer egy csoportba került, akkor a továbbiakban nem biztos, hogy együtt is marad, később a két elem külön csoportba is kerülhet. Van olyan módszer, ahol a klaszterek száma a csoportosítás során alakul ki, más módszereknél előre meg kell adni paraméterként.

A nemhierarchikus módszerek általános menete a kezdő klaszterek kialakításával kezdődik. Ezután az objektumok elhelyezése történik a kezdő klaszterekbe, majd az objektumok átrendezése a klaszterek között valamilyen optimalizáló kritérium alapján. Az egyes felosztásokat jól definiált döntésfüggvénnyel értékelhetjük. Az algoritmus az iterációs felosztássorozat révén a döntésfüggvény lokális optimumát keresi (Füstös és Meszéna, 1983). Mivel a hierarchikus módszereknek igen sok altípusa létezik (*6. ábra*), ezért a továbbiakban csak a legelterjedtebb MacQueen-féle k-középpontú módszerről szólunk részletesebben.

6. ábra: **A nemhierarchikus módszerek csoportosítása**



Forrás: Füstös és Meszéna (1983)(24)

Figure 6: Non-hierarchic clustermethods

Non-hierarchical methods(1), Optimalizing methods(2), Methods based on density(3), Admixture of distributions(4), Canonical clustering(5), Optimalizing the hierarchical structure(6), Politetical, iterative clustering(7), TAXMAP method(8), Wishart's method(9), NORMIX method(10), Page's minimal graph method(11), Methods based on variance analysis(12), Centroid methods(13), Cluster numbers are given ahead(14), Cluster numbers are changing during the algorithm(15), Fixed cluster center(16), Moving cluster center(17), Forgy's method(18), Jancey's method(19), McQueen's K-means clustering(20), McQueen's method(21), Wishart's method(22), ISODATA(23), Source(24)

3.2.1 MacQueen-féle k-középpontú módszer

A módszer állandó klaszterszámmal dolgozik, amelyet az eljárás elején adunk meg paraméterként. Az algoritmus a klaszterek összevonásakor a legközelebbi centroid kritériummal dolgozik. Az eljárás során további feltétel, hogy minden objektum egyszerre egy, és csak egy klaszterbe kerülhet. Először kiindulunk az első k elemből, mint ún. magpontból. Az elemeket a hozzá legközelebbi magközéppontú klaszterhez soroljuk. Minden egyes besorolása után a klaszterek új centroidjait kiszámítjuk, és az új magpontoknak megfeleltetjük a centroidokat. Az adatokat újra hasonlítjuk a már megváltozott magpontokhoz. Ezt a folyamatot addig ismételjük, míg a klaszterek állandósulnak.

4. Döntés a klaszterek számáról

A nem hierarchikus klaszterelemzésnél igen lényeges a klaszterek számának helyes megválasztása. Ennek érdekében képezni kell a csoportok belső és a külső varianciájának hányadosát a klaszterek számának függvényében. Az a pont utal a megfelelő klaszterszámra, amely után éles ugrás következik a belső-külső variancia arányában. E ponton túl már nem érdemes növelni a klaszterszámot, mert a belső változékonyság nagyobb mértékben növekedett, mint a külső. Hierarchikus klaszterelemzés esetén a klaszterek összevonásának történetéből, az összevonási sémából (dendogram) következtethetünk a klaszterek számára. Azon klaszterek összevonása már nem szükséges, melyek közti távolság feltűnően nagy. Célszerű figyelembe venni a klaszterek relatív méretét is. Például ha egy 3 klaszteres megoldásban a klaszterek elemszáma 8, 7, 6 lenne, a 4 klaszteres megoldásban pedig 8, 6, 6, 1 lenne, akkor a 3 klaszteres megoldás az elfogadhatóbb, hiszen egy elemből álló klaszternek nem sok értelme van.

5. A klaszterek értelmezése, jellemzése

A klaszterek értelmezését, jellemzését a centroidjaik elemzésével végezzük el (*4. táblázat*). A centroidok lehetővé teszik, hogy minden klaszterhez egy nevet illesszünk. Gyakran segít a jellemzésben olyan változók bevonása is, amelyeket nem használtunk fel az elemzésben. A klasztereket jól elkülönítő változók diszkriminancia vagy szóráselemzéssel tárhatók fel.

6. A klaszterelemzés érvényességének ellenőrzése

Természetesen addig nem fogadhatunk el egy csoportosítást, amíg nem ellenőriztük annak érvényességét. Erre a célra is számos lehetőségünk adódik. A klaszterelemzést elvégezzük ugyanazokkal az adatokkal, de más távolságmértéket alkalmazva. Az így kapott eredményeket összehasonlítjuk, és megállapítjuk, mennyire stabil a megoldásunk. Különböző klaszterelőrásokkal dolgozunk, s összehasonlítjuk az eredményeket. Az adatokat véletlenszerűen két almintára bontjuk, s mindkettőre elvégezzük az elemzést. Összehasonlítjuk az alminták klaszterátlagait. Véletlenszerűen elhagyunk változókat, és csökkentett változószámmal végezzük el újra az elemzést. Összehasonlítjuk az eredményeket azon eredményekkel, melyeket az elhagyás előtt kaptunk. Nemhierarchikus elemzéseknél a megoldás függhet az elemek adatbázisban elfoglalt sorrendjétől is. Ennek érdekében az elemzést az esetek különböző sorrendjével kell kipróbálni addig, amíg nem stabilizálódik a megoldás.

7. A dendogram

Az eljárások által létrehozott klasztereket követhetjük nyomon az ún. dendogramon, amelyről leolvashatjuk azt is, hogy két objektum, illetve két kialakított csoport melyik lépésben került összevonásra. Ebből az is kiderül, hogy milyen a két csoport közötti hasonlóság mértéke (Falus és Ollé, 2004). A klaszteranalízis egyik leglényegesebb pontja azt megérteni, hogy készül a dendogram, és hogyan értelmezhető. Mindezt pár egyszerű példán keresztül mutatjuk be figyelmen kívül hagyva az apróbb részleteket. A számolás, s a végleges dendogram attól függően változik, hogy hogyan választottunk távolsági mértéket és klaszterező algoritmust. Az adathalmaz 5 elemből és két változóból (v_1 és v_2) áll (*7. és 8. ábra*).

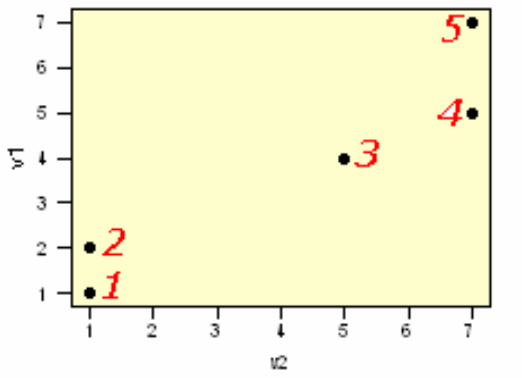
7. ábra: **Az elemek koordinátái**

Elemek(1)	v_1	v_2
1	1	1
2	2	1
3	4	5
4	7	7
5	5	7

Forrás: Anonymous (2008)(2)

Figure 7: Co-ordinates of the elements
Elements(1), Source(2)

8. ábra: **Az elemek elrendeződése a koordinátarendszerben**



Forrás: Anonymous (2008)(1)

Figure 8: Position of the elements in the co-ordinate system
Source(1)

Ezen adatok alapján elkészítjük az Euklideszi távolság mátrixot *(9. ábra)*. Természetes módon csak az alsó háromszöget adjuk meg, mivel például a 2. és 4. elem távolsága megegyezik a 4. és 2. elem távolságával. A távolságok az alábbiak szerint alakultak:

9. ábra: **Az euklideszi távolságmátrix**

	1	2	3	4	5
1	0.0				
2	1.0	0.0			
3	5.0	4.5	0.0		
4	8.5	7.8	3.6	0.0	
5	7.2	6.7	2.2	2.0	0.0

Forrás: Anonymous (2008)(1)

Figure 9: The Euclidian distance-matrix
Source(1)

Látható, hogy a két leghasonlóbb elem az 1. és a 2. elem *(10. ábra)*.

10. ábra: **A leghasonlóbb elemek a mátrixban**

	1	2	3	4	5
1	0.0				
2	1.0	0.0			
3	5.0	4.5	0.0		
4	8.5	7.8	3.6	0.0	
5	7.2	6.7	2.2	2.0	0.0

Forrás: Anonymous (2008)(1)

Figure 10: The most similar elements of the Euclidian distance-matrix
Source(1)

Ez a két elem alkotja az első klasztert, a további távolságokat ezen klaszter és a többi 3 elem között számoljuk ki *(11. ábra)*.

A klaszterező algoritmusnak a csoportátlag módszert választjuk, az első klaszterünk esetében a centroid koordinátái (v_1 átlaga, v_2 átlaga): v_1=1.5, v_2=1.0. A javított távolság mátrix a következő (A jelöli 1. és 2. elem által képezett klasztert):

11. ábra: **A két leghasonlóbb elem egy csoportba kerül**

	A	3	4	5
A	0.0			
3	4.7	0.0		
4	8.1	3.6	0.0	
5	6.9	2.2	2.0	0.0

Forrás: Anonymous (2008)(1)

Figure 11: The most similar cases are clustered in Group A
Source(1)

A legrövidebb távolság a 4. és 5. elem között van (távolság=2.0), így ezen elemek egy újabb B klaszterbe tömörülnek (a centroid koordinátái: v_1=6, v_2=7). Ezen értékekkel, klaszterekkel újraszámolt távolság mátrix a *12. ábrán* látható.

12. ábra: **További két elem került összevonásra a B csoportba**

	A	B	3
A	0.0		
B	7.5	0.0	
3	4.7	2.8	0.0

Forrás: Anonymous (2008)(1)

Figure 12: Two other cases are aggregated in Group B
Source(1)

Ezek után a legkisebb távolság a B klaszter és a 3. elem között van (2.8). Így a 3. elem bekerül a B klaszterbe, ekkor az már 3 elemet tartalmaz. A csoport centroidjának koordinátái v_1=(4+5+7)/3=5.3, v_2=(5+7+7)/3=6.3). A *13. ábra* alapján nyílvánvaló módon most már csak a két klaszter maradt, amit összevonhatnánk (távolságuk=6.4).

13. ábra: **A kialakult 2 csoport**

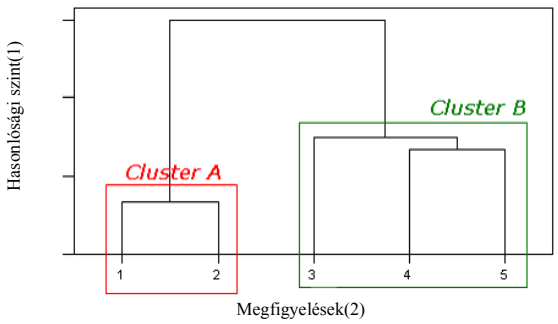
	A	B
A	0.0	
B	6.4	0.0

Forrás: Anonymous (2008)(1)

Figure 13: The final cluster solution (Group A and Group B)
Source(1)

Az egész összevonási folyamat a *14. ábrán* látható dendogramba foglalható össze.

14. ábra: **Az összevonás folyamata**

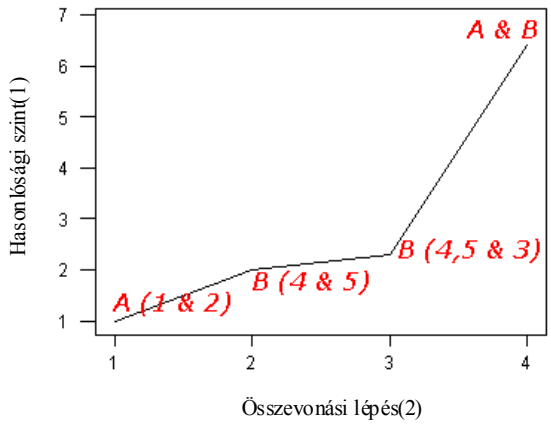


Forrás: Anonymous (2008)(3)

Figure 14: Aggregation schedule
Similarity level(1), Observations(2), Source(3)

Úgy tűnik a dendogramból, hogy az eredeti adathalmazt két klaszter megfelelően szétválasztja. Az esetek számának növelésével azonban ez nem adódik ilyen nyílvánvaló módon. A legfőbb probléma a klaszterelemzés alkalmazásakor az optimális klaszterszám megállapítása. Ahogy az egyesítés folyamata előrehalad, összevonhatunk olyan klasztereket is, melyek kevésbé hasonlóak, így a klaszterezés mesterkéltté válhat. Eléggé szubjektív dolog meghatározni a megfelelő klaszterszámot, de ebben lehet segítségünkre az összevonási séma. A különböző hasonlósági szintekhez (Similarity) különböző klaszterszám tartozik. A hirtelen ugrások a hasonlósági szintben azt mutatják, hogy kevésbé hasonló klasztereket vontunk össze nem célszerű módon *(15. ábra)*.

15. ábra: **Az optimális klaszterszám megállapítása**



Forrás: Anonymous (2008)(3)

Figure 15: Determining the optimal number of clusters
Similarity level(1), Aggregation level(2), Source(3)

A példabeli dendogram egy egyszerű adathalmaz alapján készült. Jegyezzük meg, hogy az A és B klaszter összevonása után már nagy ugrást tapasztalunk a hasonlósági szinten. Ez erősíti meg azt a feltételezésünket, hogy adathalmazunkat két klaszterrel jó reprezentálhatjuk.

SAJÁT VIZSGÁLATOK

Magyarországi sertéstelepek adatai (2004 évben) alapján vizsgáltuk, hogy az egyes telepek szakértők által előzetesen kialakított minőségi besorolása mennyire megbízható. A klaszteranalízist (Bíró et al., 2008) reprezentatív felmérésből származó 62 sertéstelep adatai alapján végeztük SPSS 13 program használatával. A klaszterelemzés részletes számítógépes kidolgozására ezt a programot használják leginkább a szakirodalomban (Székelyi és Barna, 2002).

Sertéstelepeket soroltunk csoportokba öt arányskála mérési szintű változó alapján (választott malacszám/**valmalac**, 1 sertésre jutó dolgozó/ **dolgsert**, sertések súlygyarapodása/**sulygyar**, 1 kocára jutó hízó száma/**kocahízó**, fajlagos takarmányhányad/**fajltakh)**. Ezen változók alapján hierarchikus klaszterelemzést végeztünk az esetekre, a kívánt klaszterszámot 3-ra állítottuk előzetes kutatások alapján. Azért, hogy igazoljuk, hogy a 3 klaszteres megoldás megfelelő-e, megvizsgáltuk a változó eloszlását, melyet az *2. táblázat* tartalmaz:

2. táblázat

A klaszterek elemszámának eloszlása

Csoport(1)	Gyakoriság(2)	Megoszlás, %(3)	Kumulált relatív gyakoriság, %(4)
1	16	25,8	25,8
2	14	22,6	48,4
3	32	51,6	100,0
Összesen(5)	62	100,0	

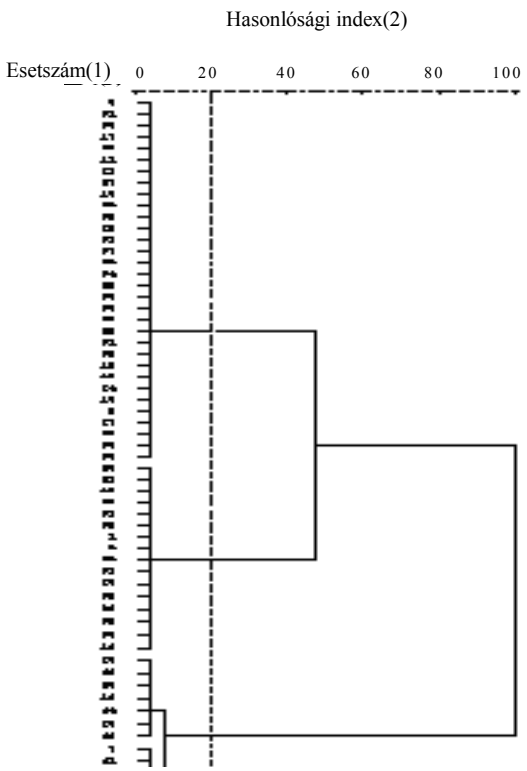
Forrás: Saját számítás SPSS 13.0 programmal(6)

Table 2: Distribution of the numbers of the clusterelements
Group, Cluster(1), Frequency(2), Relative frequency(3), Cumulative relative frequency(4), Total(5), Source: Own calculation by using SPSS 13.0 program(6)

Az eredményből látszik, hogy a változónk megfelelő eloszlású az egyes kategóriákban, mivel egyik klaszter sem tartalmaz kevés elemet, ezért elfogadható a 3 klaszteres megoldás.

A klaszterelemzés végrehajtása során a Ward-féle módszerrel, valamint a klasszikus Euklideszi távolsági mértékkel dolgoztunk. Azért a Ward módszert használtuk, mert ez az a módszer, amely a létrehozandó klaszterek belső heterogenitásának minimalizálására törekszik. Egy klaszter egy elemmel való bővítését, illetve két klaszter összevonását akkor hajtja végre a program, ha ettől az új klaszter belső heterogenitásának növekedése kisebb, mint miden más lehetséges klaszterstruktúra esetében. A program az összes beállítás elvégzése után a *16. ábrán* látható dendogramot készítette el:

16. ábra: **A Ward módszer dendogramja**



Forrás: Saját számítás SPSS 13.0 programmal(3)

Figure 16: The Dendogram of Ward's method
Case number(1), Similarity index(2), Source: Own calculation by using SPSS 13.0 program(3)

A dendogramból azt olvashatjuk le, hogy a megfelelő szinten 3 jól elkülönülő csoport alakult ki, ezzel együtt létrejött egy új változó, mely tartalmazza a telepek klaszterbesorolását.
A *3. táblázat* az eredeti minősítés és a Ward módszer szerinti minősítés összehasonlítását mutatja:

3. táblázat

Az eredeti és a Ward módszer szerinti besorolás összehasonlítása

Eredeti szakértői besorolás(1)	Ward módszer szerinti besorolás(2)			Összesen(3)
Csoport(4)	1	2	3	
1		**13**	2	15
2		1	**29**	30
3	**16**		1	17
Összesen(3)	16	14	32	62

Forrás: Saját számítás SPSS 13.0 programmal(5)

Table 3: Comparing the professional classification and Ward's classification
Original professional classification(1), Classification of the Ward's method(2), Total(3), Group(4), Source: Own calculation by using SPSS 13.0 program(5)

A táblázatból látható, hogy a két besorolás minimális eltéréssel tökéletesen fedi egymást.

Végül elemeztük a csoportátlagokat is a megfelelő változók tekintetében, hogy a csoportokat el tudjuk nevezni, s felfedjük, mely csoportok milyen telepeket tartalmaznak. A kapott csoportok változók szerinti átlagait tartalmazza a *4. táblázat*.

4. táblázat

A klaszterelemzés változói szerinti csoportátlagok

Csoport(1)	VALMALAC(2)	KOCAHIZO(3)	FAJLTAKH(4)	SULYGYAR(5)	DOLGSERT(6)
1	-- 15,1143	-- 11,5415	++ 6,1200	-- 498,1429	-- 252,1429
2	**0** 17,7147	**0** 14,0792	+ 5,0493	- 583,5244	**0** 462,1563
3	++20,6900	++ 17,4226	- 3,8794	+ 727,6875	+ 587,0000
Összesen(7)	17,8953	14,3690	4,9892	601,4481	446,9516

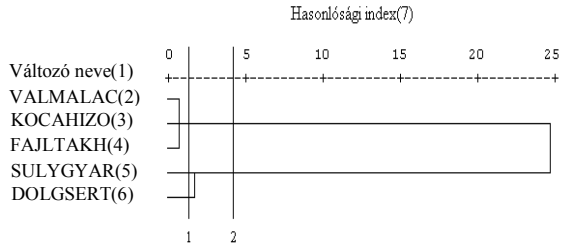
0: átlag körüli; -- nagyon átlag alatti; - kicsit átlag alatti; ++ nagyon átlag feletti; + kicsit átlag feletti(8)
Forrás: Saját számítás SPSS 13.0 programmal(9)

Table 4: Average group values of variables
Group(1), Chosen pigs(2), Store-pig per sow(3), Specific nutritive quotient(4), Average weight gain(5), Number of swine per worker(6), Total(7), Special signs: 0: average; -,--: subaverage; + ,++: above the average(8), Source: Own calculation by using SPSS 13.0 program(9)

Az eredeti outputot kiegészítettük saját jelöléseinkkel a könnyebb értelmezhetőség érdekében. A jelölést a táblázat alján adtuk meg, az adott változó átlagához viszonyítottuk a csoportokon belüli átlagokat.
Ez alapján megállapítható, hogy az 1-es csoportba sorolt telepek 4 változó szerint nagyon elmaradnak az átlagtól, míg a **fajltakh** változó (jelentése fajlagos takarmányhasznosítás) esetében a többi telep átlagos mutatóját jóval meghaladják.
A 2-es csoportról elmondható, hogy az átlag körül alakulnak a különböző mutatók átlagai. Ezzel szemben a 3-as csoportba sorolt telepek a legjobb paraméterekkel rendelkeznek.

A hierarchikus klaszterelemzést Ward módszerrel végrehajtottuk a klaszterelemzésben alkalmazott változókra is. Az eredmény a *17. ábráról* olvasható le.

17. ábra: **A változók összevonásának dendogrammja**



Forrás: Saját számítás SPSS 13.0 programmal(8)

Figure 17: The Dendogramm of the variables

Variable name(1), Chosen pigs(2), Store-pig per sow(3), Specific nutritive quotient(4), Average weight gain(5), Number of swine per worker(6), Similarity index(7), Source: Own calculation by using SPSS 13.0 program(8)

Ha az 1-es szintvonalon vágjuk el az ábrát, akkor 3 változócsoport alakul ki, míg a 2-es szintvonal mentén elvágva 2 csoport jön létre. Érdemesebb az 1-es szintvonalon (minél közelebb a 0 esethez) elvágni az ábrát, mert az így kialakult csoportok elemei jobban hasonlítanak egymásra. A klaszterelemzést a nem hierarchikus McQueen féle K-középpont módszerrel is elvégeztük, majd elmentettük az így keletkezett besorolást. Az elkészült kereszttábla *(5. táblázat)* szemlélteti, milyen különbségeket eredményezett a két módszer alkalmazása.

5. táblázat

A McQueen-féle módszer és a Ward módszer besorolásának összehasonlítása

McQueen-féle módszer szerinti besorolás(1)	Ward Módszer szerinti besorolás(2)			Összesen(3)
Csoport(4)	1	2	3	
1			**16**	16
2	**14**			14
3		**32**		32
Összesen(3)	14	32	16	62

Forrás: Saját számítás SPSS 13.0 programmal(5)

Table 5: Compering the classification of McQueen's and Ward's method

Classification by McQueen's method(1), Classification by Ward's method(2), Total(3), Group(4), Source: Own calculation by using SPSS 13.0 program(5)

Az *5. táblázatból* látható, hogy a két módszer egyformán sorolta be telepeinket a különböző csoportokba. Ilyen módon eredményünket, illetve a klaszterek helytállóságát két módszer kipróbálásával is érvényesítettük.

IRODALOM

Anonymous (2008): http://www.biochemistry.ucla.edu/biochem/Faculty/Mallick/260/multivar/dend.htm

Bíró O.-Ózsvári L.-Lakner Z. (2008): Az állat-egészségügyi menedzsment hatása a sertéstenyésztő telepek teljesítményére – Egy módszertani kísérlet és tanulságai. Magyar Állatorvosok Lapja közlésre elfogadva 2008. februárban

Falus I.-Ollé J. (2004): Statisztikai módszerek pedagógusok számára, Okker Kiadó, 271.

Füstös L.-Meszéna Gy. (1983): Bevezetés az adatelemzés sokváltozós módszereibe, Tankönyvkiadó, Budapest, 91-137.

Lance, G. H.-Williams, W. T. (1967): A general theory of classificatory sorting strategies, Computer Journal 9, 373-380.

Malhotra N. K. (2001): Marketingkutatás, Műszaki Kiadó, Budapest, 698-721.

Sokal, P. P.-Sneath, P. M. (1963): Principles of numerical taxonomy, Freeman, San Francisco, London

Székelyi M.-Barna I. (2002): Túlélőkészlet az SPSS-hez, Typotex Kiadó, Budapest

Szűcs I. (2002): Alkalmazott Statisztika, Agroinform Kiadó, Budapest, 501-507.