Többváltozós statisztika, ANOVA és adatredukciós módszerek (BMNPS07700M)

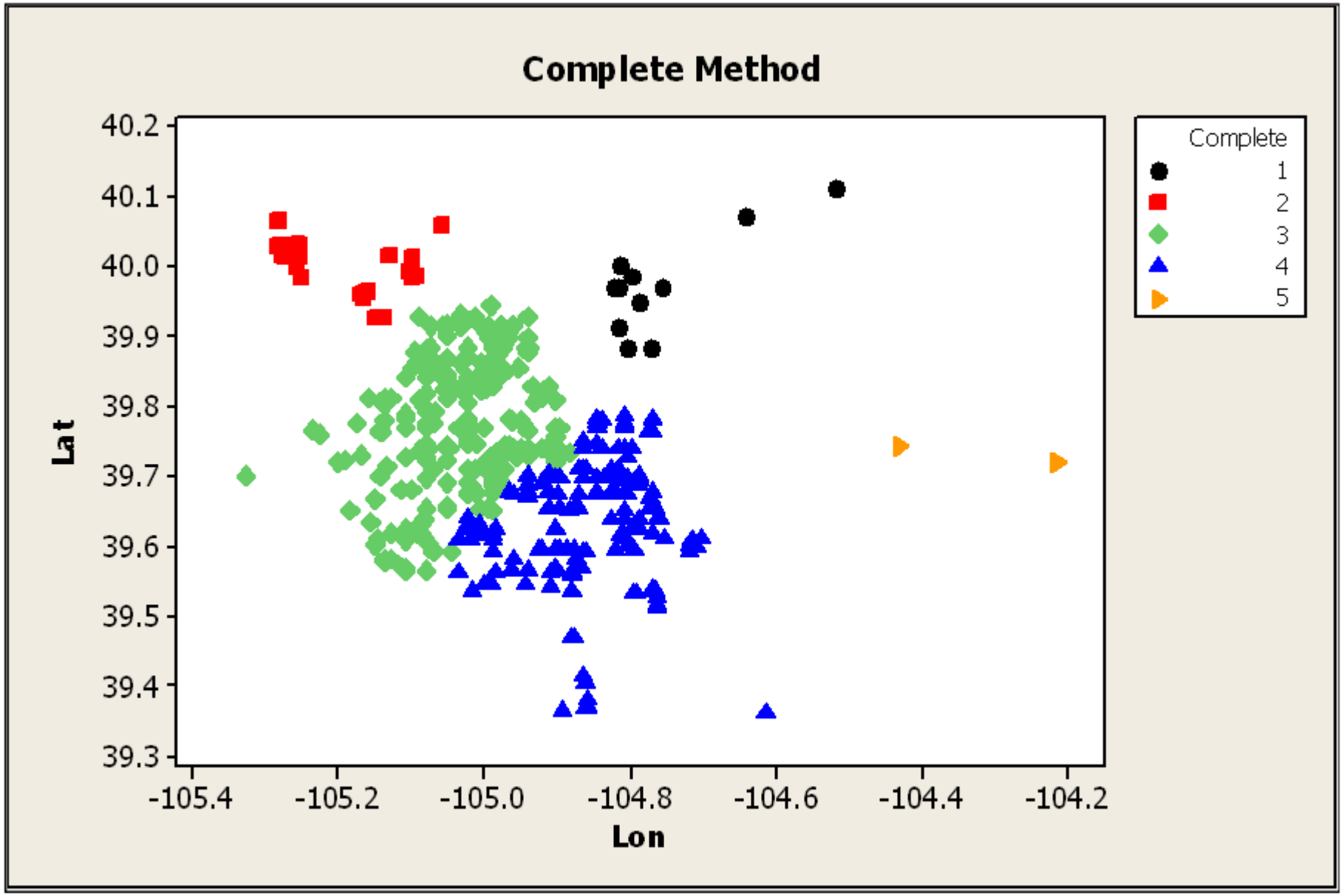
Készítette: Soltész-Várhelyi Klára

Klaszterelemzés

Klaszteranalízis

- Adatredukció
  - Nagy adathalmaz kezelhető méretű információhalmazokba rendezése
  - Adatok csoportosítása kisszámú csoportba
  - Példa: potenciális vásárlók csoportosítása, minden csoportnak más típusú reklám/termék
  - Csoportosítás módja: csoporton belüli esetek hasonlóságának maximalizálása, csoportok közötti hasonlóság minimalizálása
  - Különbség Diszkriminancia analízis, Klaszter analízis és Faktor analízis között:
    - FA változókat csoportosít, KA eseteket
    - DA és KA is eseteket csoportosít, de míg a DA esetén az új eset csoportba való besorolásához szükség van előzetes tudásról a csoportba tartozással kapcsolatban (már létező csoportokba sorolunk be új eseteket), addig KA esetében nincs priori tudás (meglévő esetekből alakítjuk ki a csoportokat)

	Kata	Peti	Eszti	Józsi	Timi	Laci	Zoli	Robi	Tomi
Kor	11	12	14	14	15	16	18	19	20
	A klaszter		B klaszter				C klaszter		

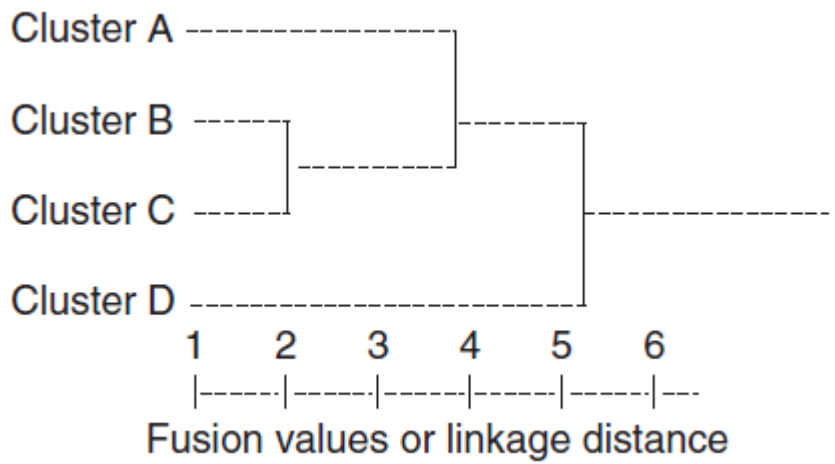

Complete Method
Lat
Lon
40.2
40.1
40.0
39.9
39.8
39.7
39.6
39.5
39.4
39.3
-105.4
-105.2
-105.0
-104.8
-104.6
-104.4
-104.2
Complete
1
2
3
4
5

KA módjai

- Hierarchikus módszer
  - Nem tudjuk előre, hány csoport lesz, majd kiadódik
  - Abból indulunk ki, hogy minden eset egy külön klaszter, majd elkezdjük a klasztereket összevonni
  - Kis dimenziószám esetén használható
  - Adatok lehetnek skálatípusúak, binárisak, gyakoriság értékek
- K-közép módszer
  - k számú feltételezett klasztert keres
  - A klaszterközéppontok helyzete egy iterációban folyamatosan változik, míg el nem ér egy stabil állapotot
  - Az esetek a legközelebbi klaszterközépponthoz lesznek rendelve
  - Sokszor először futtatunk egy feltáró hierarchikus klaszternalízist, meghatározzuk, hány klaszterünk van, mert újrafuttatjuk a klaszteranalízist k-közép módszerrel a meghatározott klaszterszámmal
  - Feltételei
    - Skála típusú adatokon végezhető
    - Csak nagy elemszámnál használható
    - Csak euklideszi távolságokkal számolható

Hierarchikus klaszteranalízis

- Kezdetben minden eset külön klasztert alkot – éppen annyi klaszter van, ahány eset
- Majd „lazít” a kritériumokon (mennyire kell a klaszteren belül az elemeknek hasonlítani egymásra), és megnézi össze lehet-e vonni két klasztert
- Majd a klasztereket addig csoportosítja, amíg egyetlen klaszter marad
- A csoportosításhoz a klaszterek különbözőségét, a közöttük lévő távolságot használjuk
- Dendrogram: fa diagram – a klaszterek csoportosulását ábrázolja

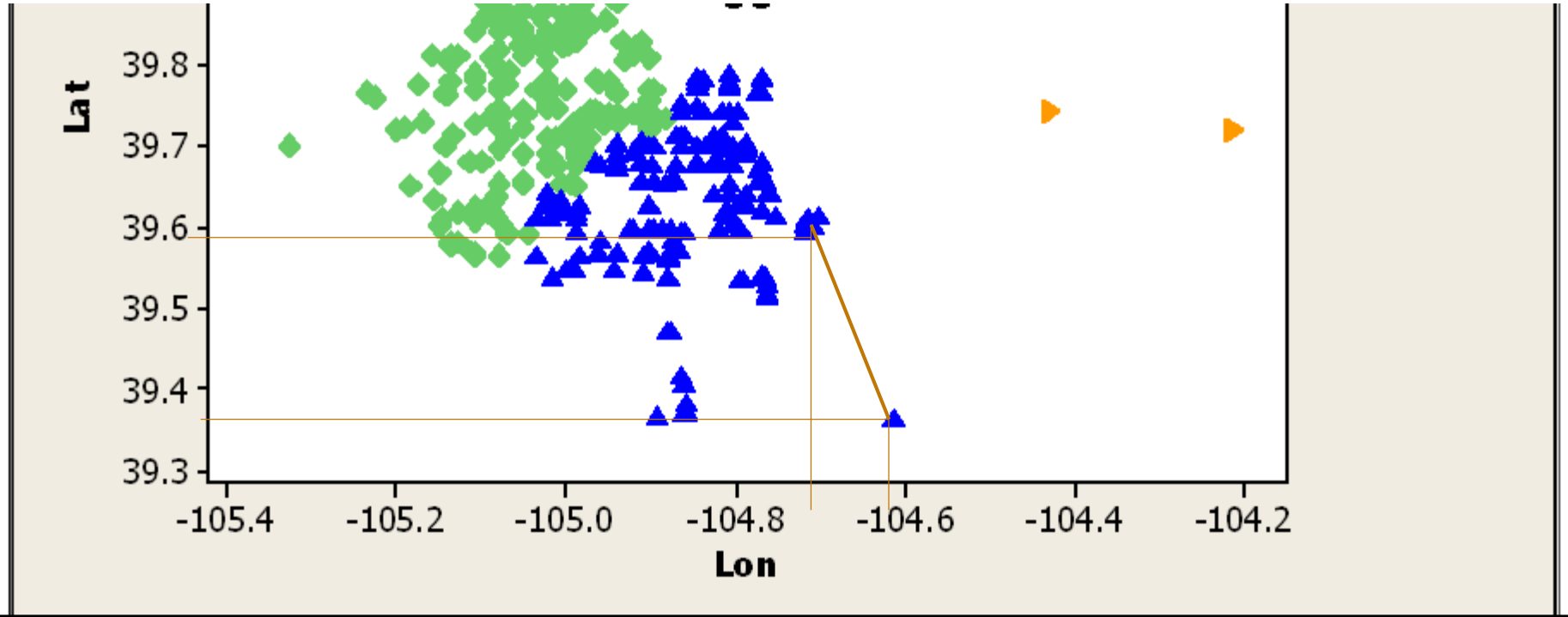


Távolság mérése

- Négyzetes Euklideszi távolság
  - Képzeljük el az adatokat egy annyi dimenziós térben, ahány változónk van, majd mérjük meg az adatpontok közötti távolságot
  - Euklideszi távolság a kibővített Pithagoras-tétel szerint számolódik
  - Például a kiválasztott 2 pont között ebben a 2D térben a távolság

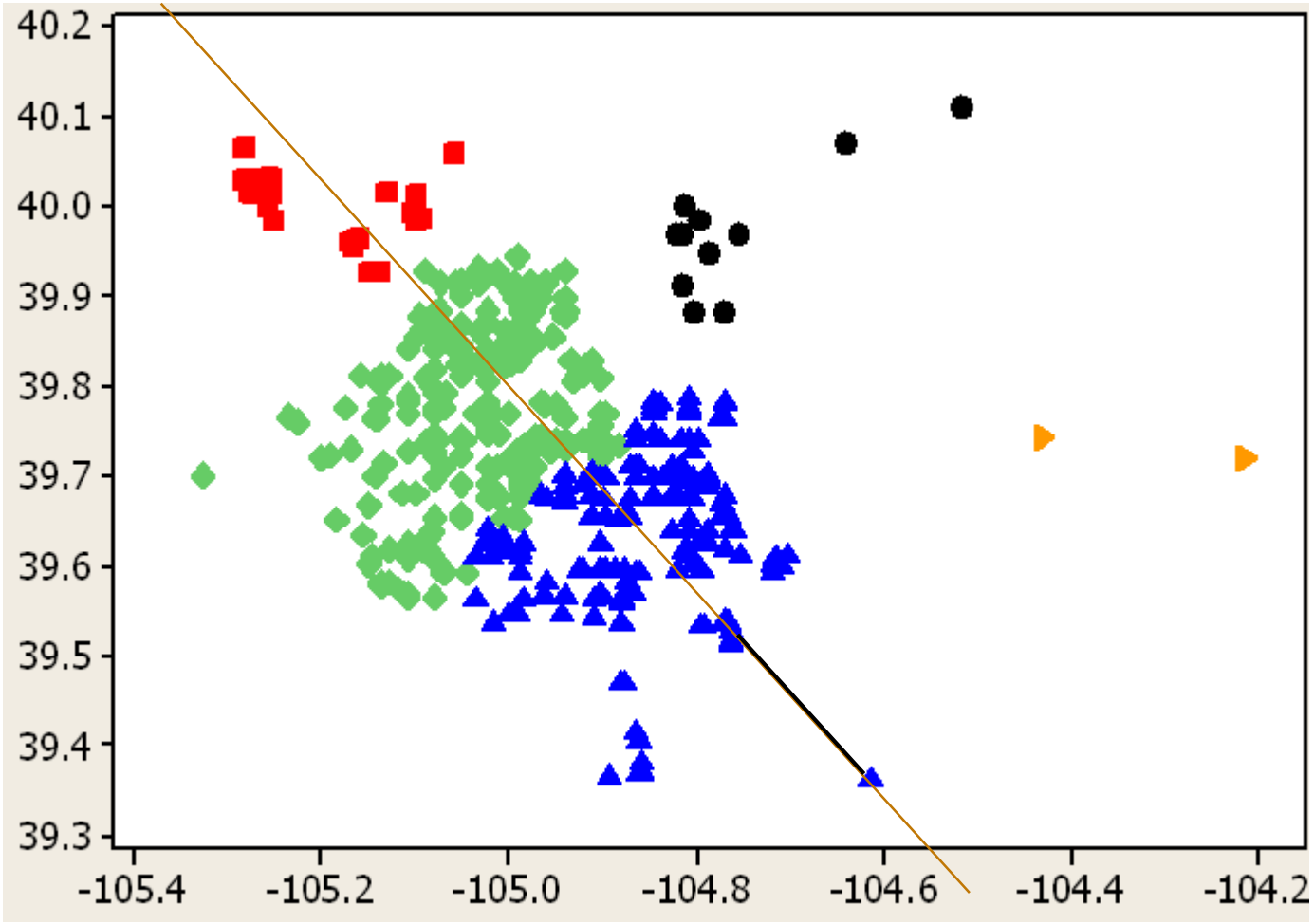
$$d = \sqrt{a^2 + b^2} = \sqrt{(39{,}59 - 39{,}37)^2 + ((-104{,}63) - (-104{,}72))^2}$$

  - Négyzetes Euklideszi távolság: gyakrabban használt, mert a nagy távolságokra nagyobb hangsúlyt helyez



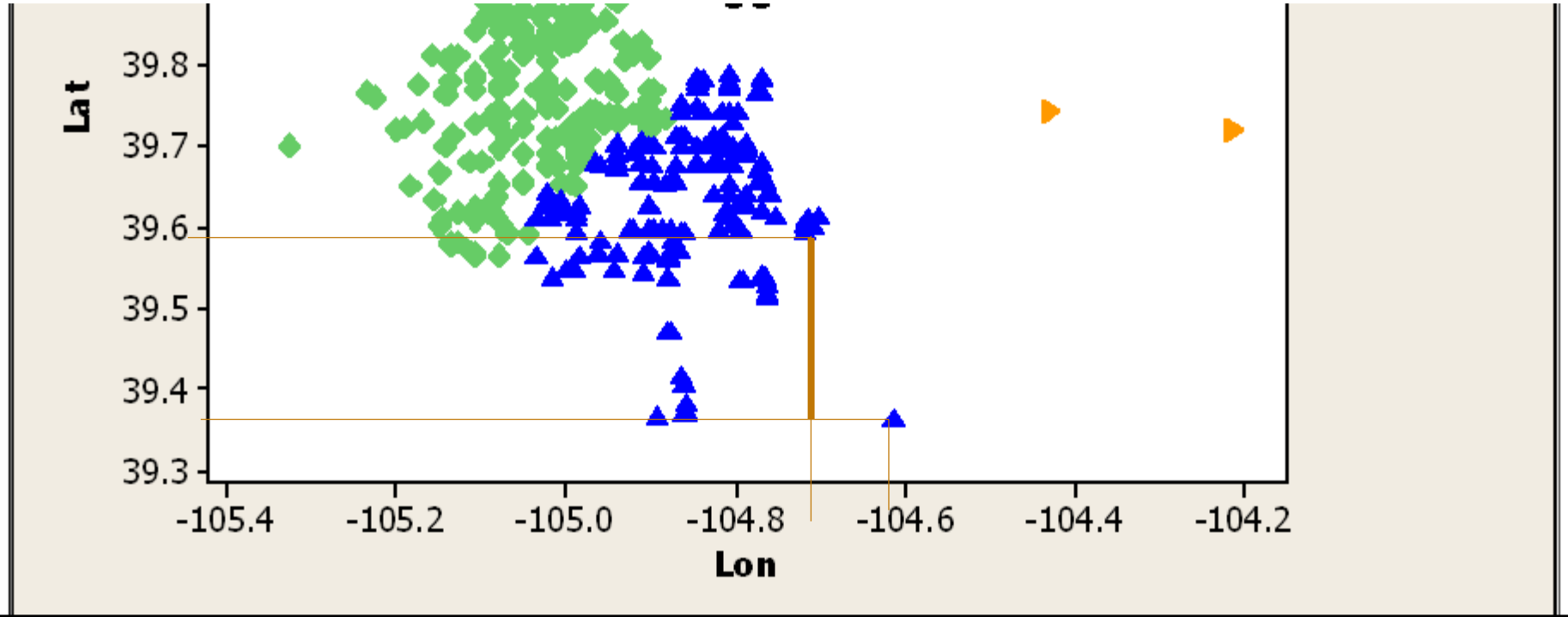
Távolság mérése

- Mahalanobis távolság / korrelációs módszer
  - Egy n dimenziós térben megnézzük a két változó közé húzható egyenes mentén, mekkora a szórás az adatokban
  - Ha egy irányba a szórás nagyobb, abban az irányban a Mahalanobis távolság kisebb lesz



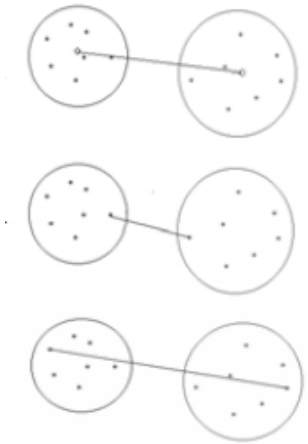
Távolság mérése

- Csebisev távolság
  - Képzeljük el az adatokat egy annyi dimenziós térben, ahány változónk van, majd mérjük meg az adatpontok közötti távolságot
  - A Csebisev távolság a legnagyobb távolság minden dimenzióban mérhető távolságok közül
  - Például a kiválasztott 2 pont között ebben a 2D térben a távolság
    X dimenzióban: (-104,63)-(-104,72)= 0,09 Y dimenzióban= 39,59-39,37=0,22
    Csebisev távolság= 0,22



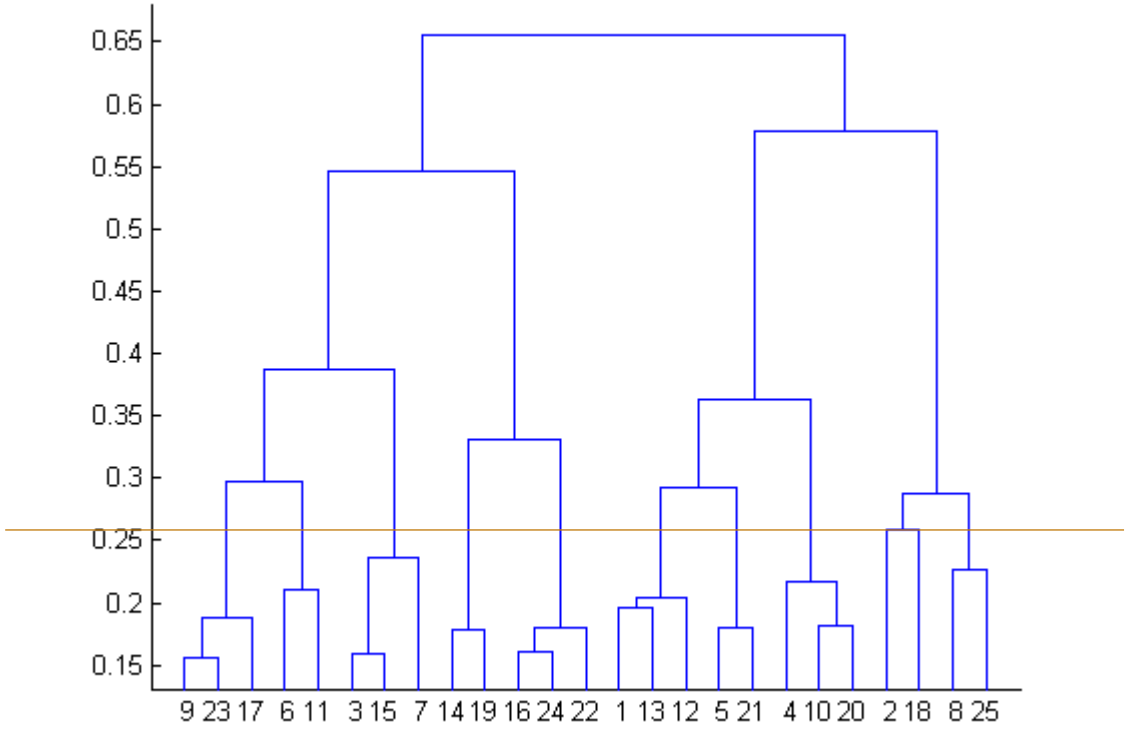
Távolságok értelmezése

- Elemek között tudunk távolságot nézni, de klaszterek között?
  - Centrum metrika (centroid clustering)
    - Klaszterek középső eleme közötti távolság
  - Legközelebbi elem metrika (nearest neighbor)
    - A két klaszter legközelebbi eleme közötti távolság
  - Legtávolabbi elem metrika (furthest neighbor)
    - A két klaszter legtávolabbi eleme közötti távolság
  - Median clustering
    - A két klaszter elemeinek mediánja közötti távolság
  - Between-groups linkage
    - A két klaszter elemeinek összes párosítása közötti távolság átlaga
  - Within groups linkage
    - Azt a két klasztert vonja össze, ahol az összevonás után a legkisebb az elemek közötti távolság
  - **Ward módszer**
    - **A létrejövő klaszter varianciájának elemzésén alapul**
    - **Az a két klaszter lesz összevonva, ahol az összevonást követően a legkisebb az átlagtól való távolság négyzetének összege (SS_R)**



Hol húzzuk meg a határt?

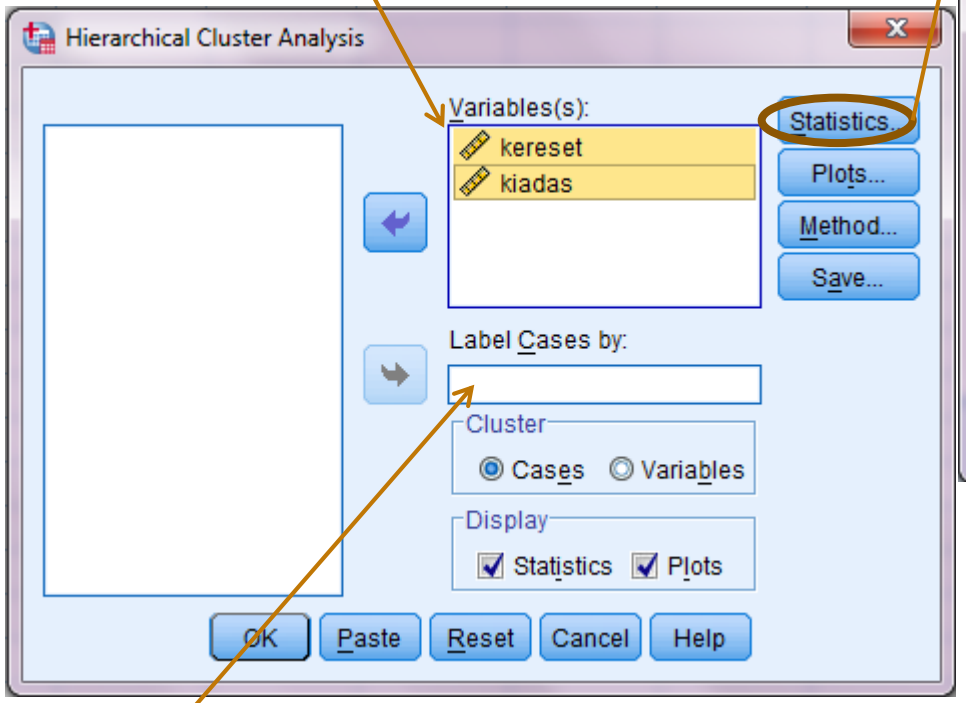
- A klaszterek számának eldöntése nagymértékben szubjektív
  - Dendrogram alapján döntünk
  - Értelmezhető mennyiségű klaszterünk legyen
  - Legalább 4 elem tartozzon egy klaszterbe



SPSS-ben

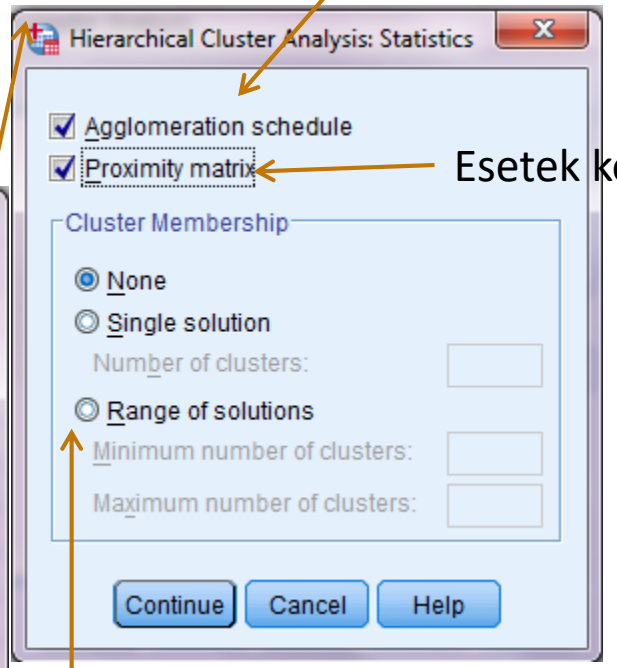
- Adatfájl: tobbval07_klaszter_kereset_kiadas .sav
- Analyze / Classify / Hierarchical Cluster

Változók, melyek mentén a klasztereket ki akarjuk alakítani



Ha van valamilyen azonosító, az kerül ide
Ha nincs, az azonosító az eset sora lesz

Csoportosítás menete (a dendrogram „szöveges változata”



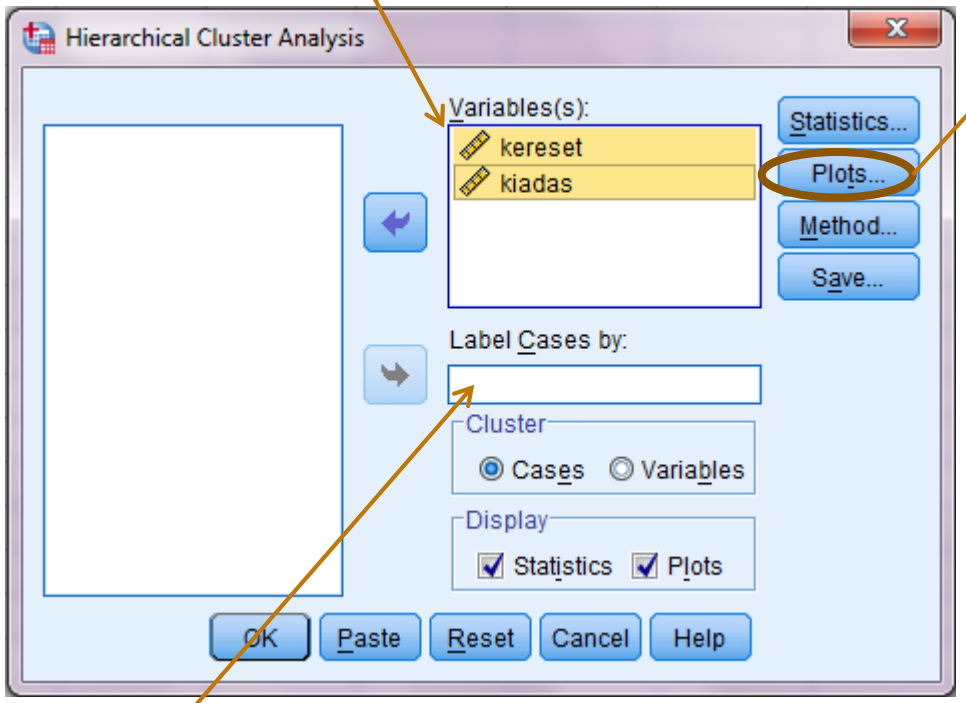
Esetek közötti távolság

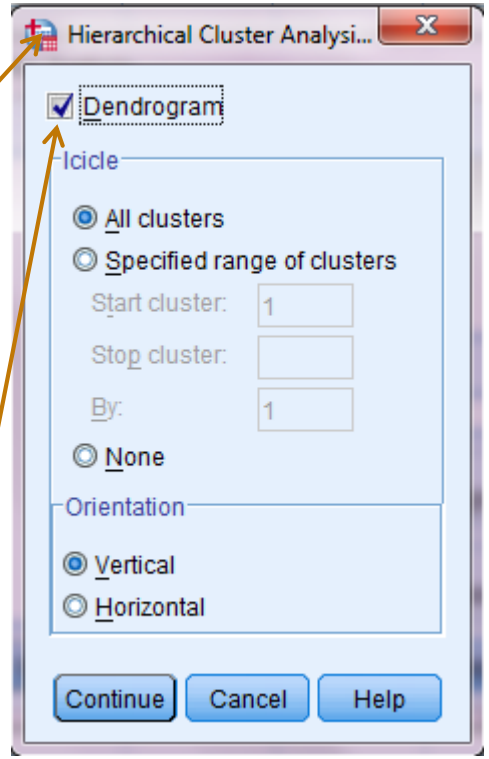
Mit jelenítsen meg?
None-Minden megoldást
Single solution – csak bizonyos klaszterszámú megoldást (például 3 klaszteres megoldást)
Range of solutions – több megoldást is (például 3, 4 és 5-klaszteres megoldásokat

SPSS-ben

- Kereset-kiadas.sav
- Analyze / Classify / Hierarchical Cluster

Változók, melyek mentén a klasztereket ki akarjuk alakítani





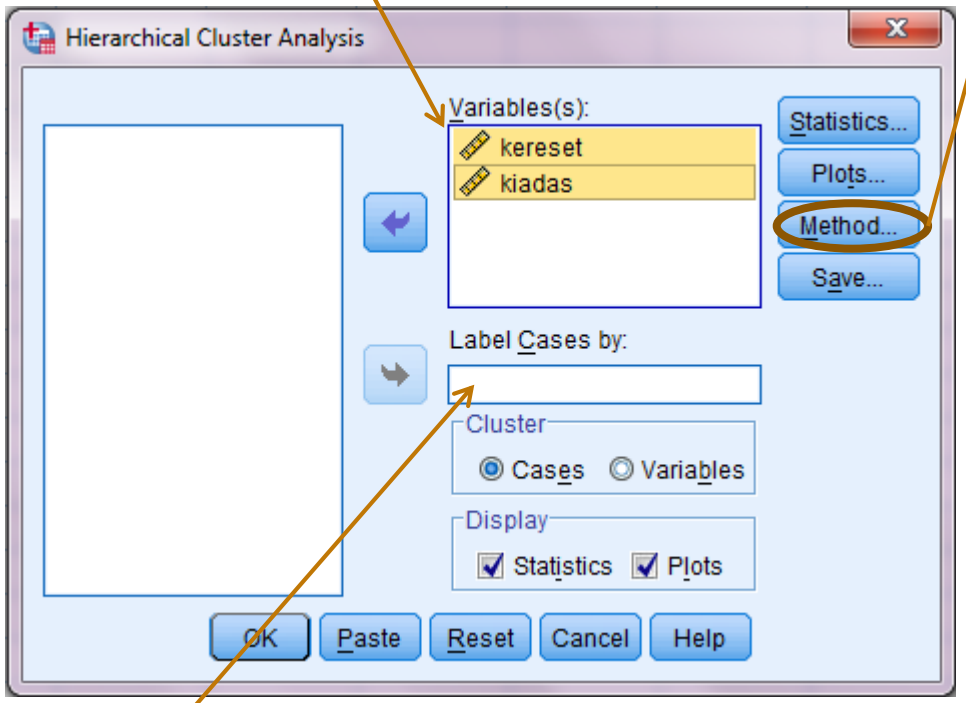
Fadiagram a klaszterek képződéséről

Ha van valamilyen azonosító, az kerül ide
Ha nincs, az azonosító az eset sora lesz

SPSS-ben

- Kereset-kiadas.sav
- Analyze / Classify / Hierarchical Cluster

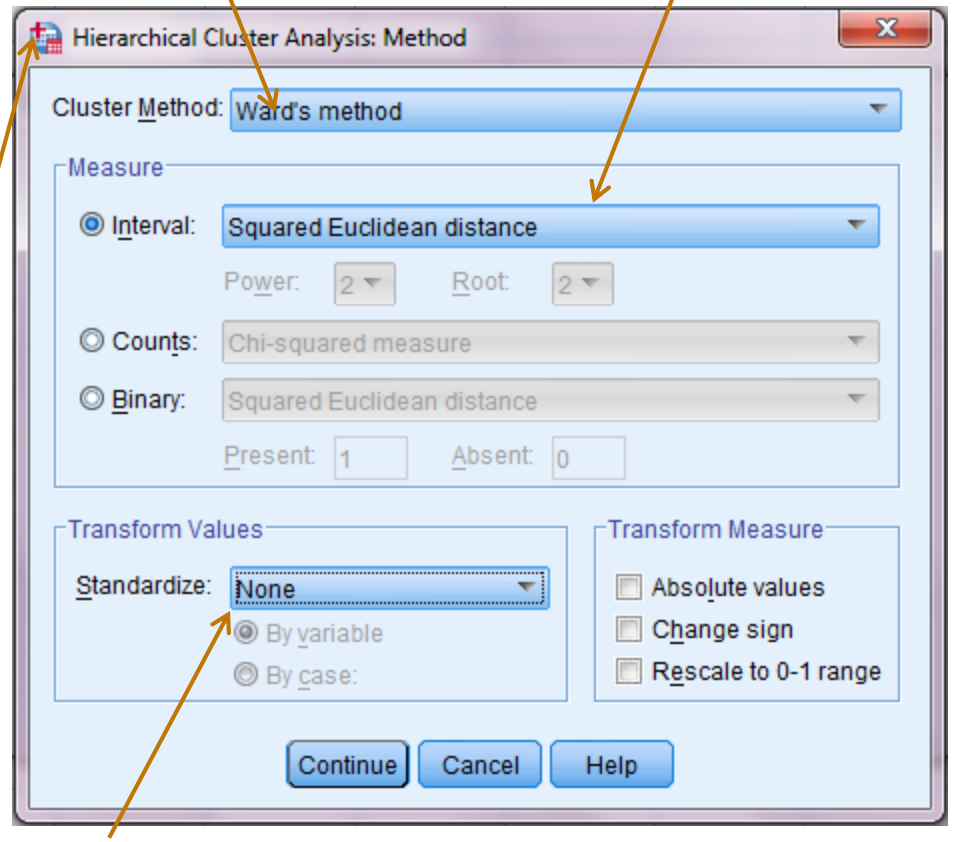
Változók, melyek mentén a klasztereket ki akarjuk alakítani



Ha van valamilyen azonosító, az kerül ide
Ha nincs, az azonosító az eset sora lesz

Klaszterezés módja

Távolságszámítás módja

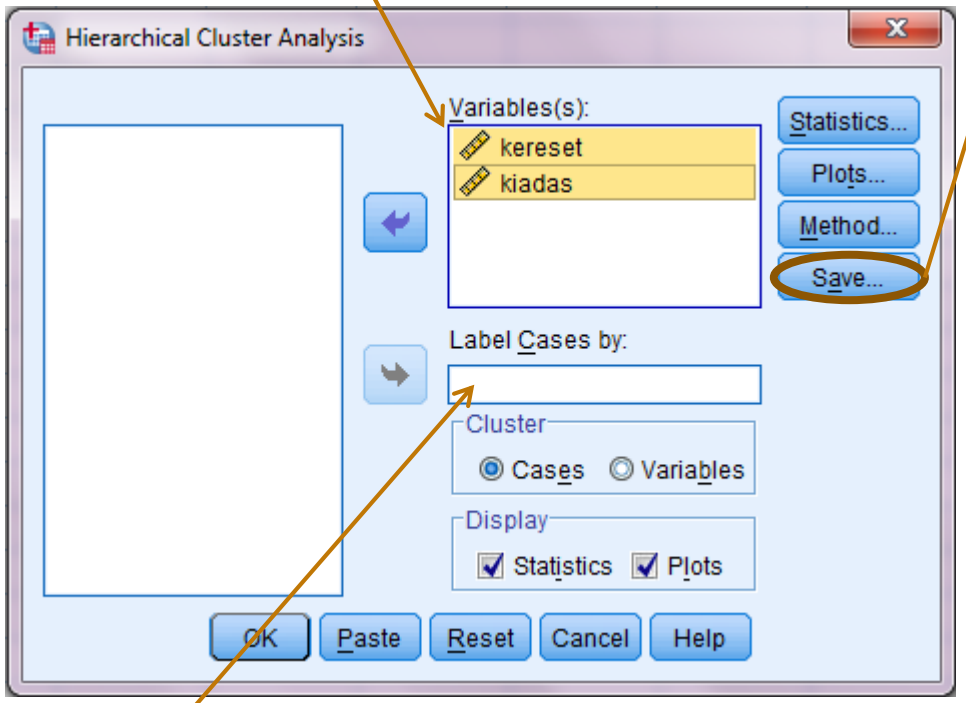


Standardizálás – ha a változók különböző skálán mozognak

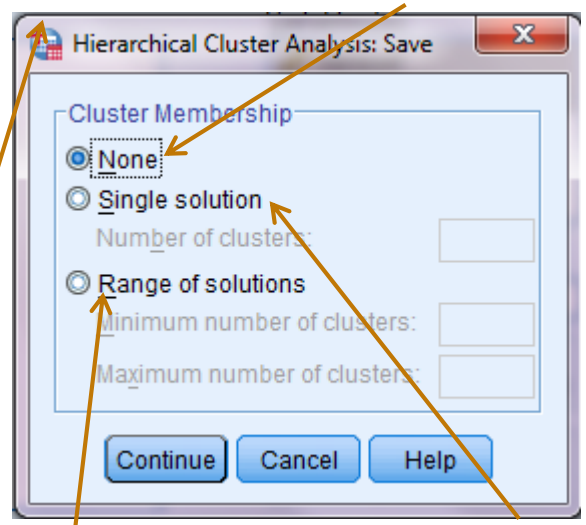
SPSS-ben

- Kereset-kiadas.sav
- Analyze / Classify / Hierarchical Cluster

Változók, melyek mentén a klasztereket ki akarjuk alakítani



Most még nem használjuk a Save-t, mert nem tudjuk, hány értelmezhető klaszterünk van



Ha egyszer lefuttattuk az elemzést, és megállapítottuk, hány értelmes klaszterünk van, újra futtatjuk az elemzést, és itt kérhetjük meg, hogy sorolja be az eseteinket az általunk meghatározott számú klaszterbe

Ha van valamilyen azonosító, az kerül ide
Ha nincs, az azonosító az eset sora lesz

Több megoldás is kérhető

Eredmények

Case Processing Summary[a]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
35	100,0	0	,0	35	100,0

a. Ward Linkage

Leíró adat – hány eset került az elemzésbe, hány maradt ki valamiért

Case	1	2	3	4	5	6	7
1	,000	881,000	565,000	1332,000	1040,000	173,000	986,000
2	881,000	,000	58,000	605,000	697,000	340,000	157,000
3	565,000	58,000	,000	421,000	853,000	130,000	97,000
4	1332,000	605,000	421,000	,000	2468,000	593,000	146,000
5	1040,000	697,000	853,000	2468,000	,000	981,000	1458,000
6	173,000	340,000	130,000	593,000	981,000	,000	333,000
7	986,000	157,000	97,000	146,000	1458,000	333,000	,000
8	458,000	905,000	509,000	410,000	2194,000	241,000	520,000
9	2290,000	349,000	585,000	778,000	1642,000	1261,000	400,000
10	1224,000	245,000	457,000	1620,000	200,000	809,000	794,000

Esetek egymástól való távolsága

Ward Linkage

Agglomeration Schedule

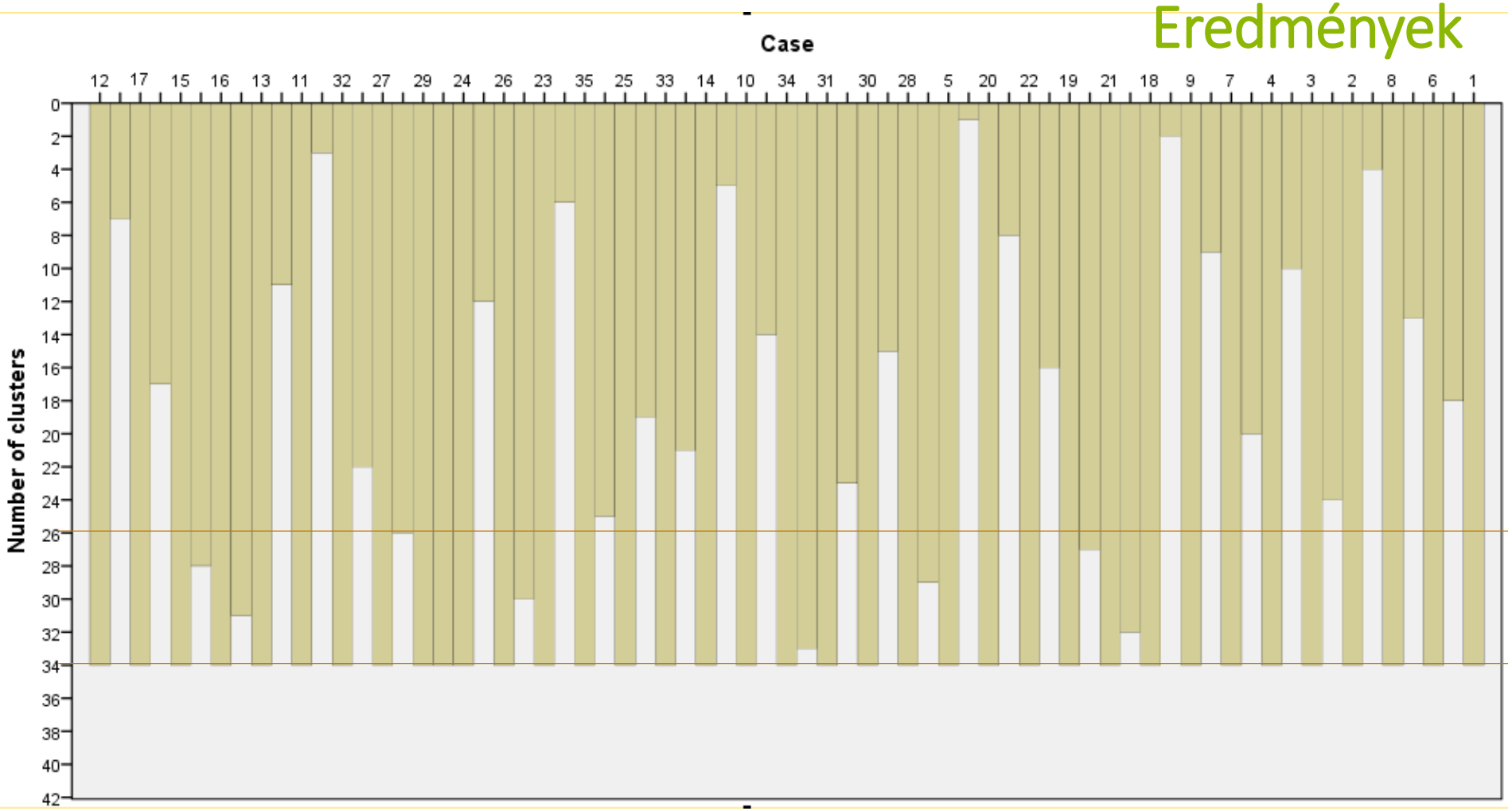
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	24	29	1,000	0	0	9
2	31	34	3,500	0	0	12
3	18	21	6,000	0	0	8
4	13	16	8,500	0	0	7
5	23	26	13,500	0	0	23
6	5	28	22,000	0	0	20
7	13	15	32,833	4	0	18
8	18	19	45,000	3	0	19
9	24	27	58,667	1	0	13
10	25	35	77,167	0	0	16
11	2	3	106,167	0	0	25
12	30	31	137,000	0	2	20
13	24	32	190,083	9	0	23
14	14	33	244,583	0	0	16
15	4	7	317,583	0	0	25
16	14	25	396,083	14	10	29
17	1	6	482,583	0	0	22
18	13	17	600,750	7	0	24
19	18	22	748,833	8	0	27
20	5	30	899,000	6	12	21
21	5	10	1073,000	20	0	30
22	1	8	1277,167	17	0	31
23	23	24	1512,583	5	13	29
24	11	13	1779,483	0	18	28

Eredmények

Klaszterek kialakulásának folyamata

Az esetek hogyan kerültek egy csoportba

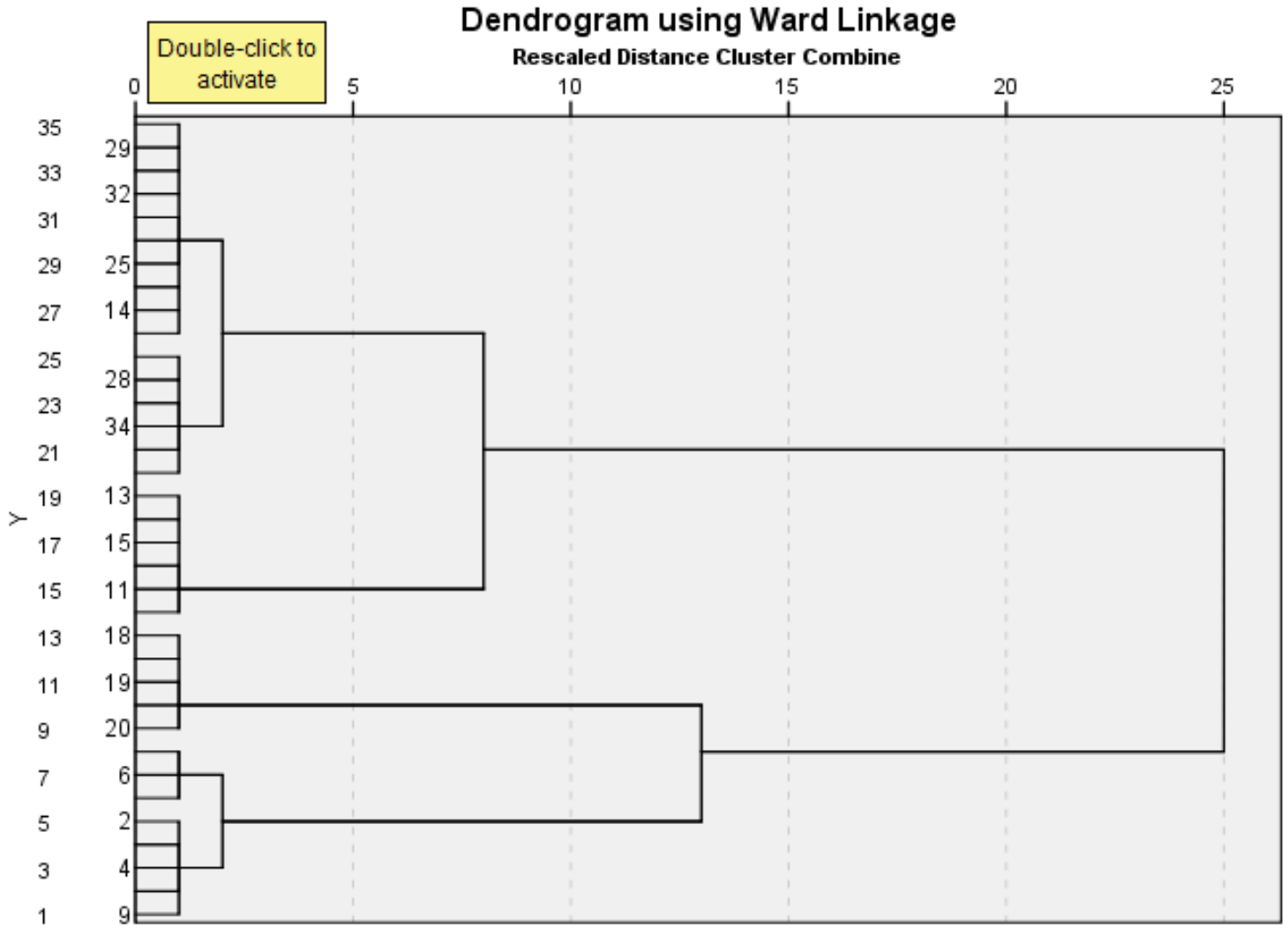
Eredmények



Klaszterek kialakulásának folyamata- Alulról felfele olvassuk. Megmutatja, hogy amikor x klaszter volt, akkor melyik változók alkottak egy klasztert.

Például, amikor 34 klaszter volt még, a 29-es és 24-es eset egy klaszterbe került, a többi eset még külön klaszter. Amikor 26 klaszter volt, akkor a 15, 16, 13 már egy klasztert alkotott, a 27, 29, 24 is, a 26, 23 is, a 34, 31 is és a 19, 21, 18 is. A többi eset még ekkor is külön klaszterként jelent meg.

Eredmények

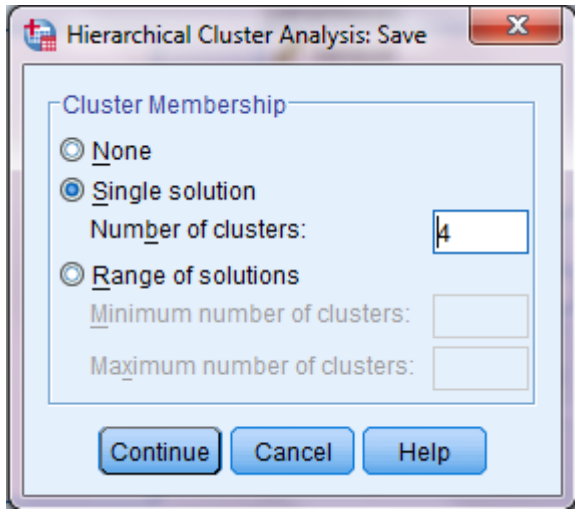


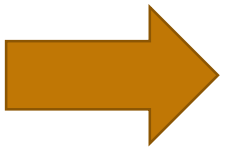
Dendrogram – a klaszterződés folyamatát mutatja meg.

A klaszterződés folyamatának három megjelenítési módja alapján eldöntjük, mennyi klaszter megfelelő számunkra.

Ki melyik klaszterbe tartozik

- Ha eldöntöttük, hány klaszter érdekel minket, újrafuttatjuk az elemzést, és a Save-ben kérünk egy új változót

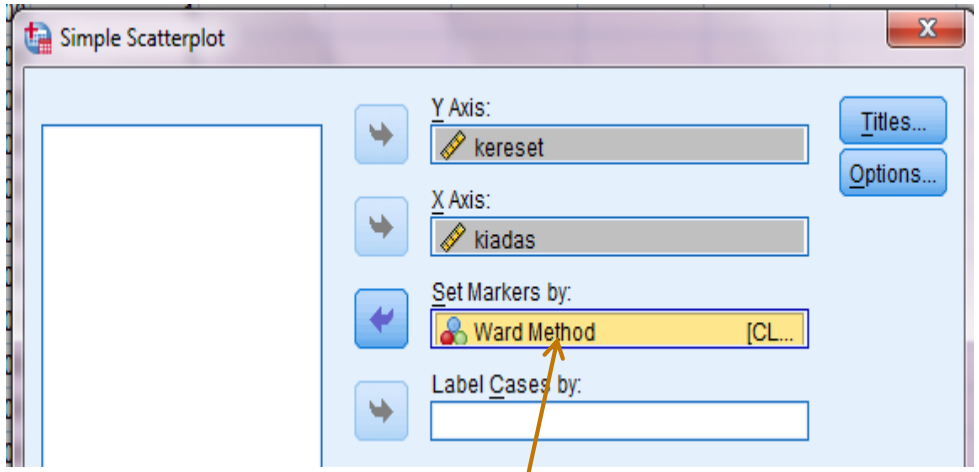




	kereset	kiadas	CLU4_1
1	89,00	49,00	1
2	73,00	24,00	1
3	80,00	27,00	1
4	95,00	13,00	1
5	57,00	45,00	2
6	87,00	36,00	1
7	84,00	18,00	1
8	102,00	32,00	1
9	68,00	6,00	1
10	59,00	31,00	2
11	2,00	3,00	3
12	29,00	6,00	3
13	10,00	18,00	3
14	36,00	32,00	2
15	11,00	13,00	3
16	11,00	16,00	3
17	13,00	28,00	3

Mit jelentenek a csoportjaink?

- Nézzünk rá, hogyan helyezkednek egy klaszterek
- 2D (amikor csak 2 változóm volt) esetén ez könnyen szemléltethető egy scatterplottal. Több dimenzió esetén (sok változó esetén) már nehezebb vizualizálni
- Graphs / Legacy dialogs / Scatter\Dot



Csoportosító változónak a klasztert tesszük meg