Cluster Analysis

- การวิเคราะห์กลุ่ม -

By
Suwanna Sayruamyat

Email: suwanna.s@ku.ac.th
Facebook: Suwanna Sayruamyat
Page: EatEcon
Website: www.eatecon.com

Objectives of cluster analysis

Cluster analysis is used for:

1. Taxonomy description: Identifying natural groups within the data.
2. Data simplification: The ability to analyse groups of similar observations instead of all individual observations.
3. Relationship identification: The simplified structure from cluster analysis portrays relationships not revealed otherwise.

Theoretical, conceptual, and practical considerations must be observed when selecting clustering variables for cluster analysis:

1. Only variables that relate specifically to objectives of the cluster analysis are included.
2. Variables selected characterise the individuals (objects) being clustered.

To conclude:

- “Cluster analysis classifies **objects** (e.g. respondents, products, or entities), on a set of user selected characteristics.”
- Very useful to create profile of responses.

Methods

Three methods for the cluster analysis

เป็นขั้นตอน

Hierarchical procedure

Hierarchical cluster

- is the most common method. It generates a series of models with cluster solutions from 1 (all cases in one cluster) to n (each case is an individual cluster).
- Hierarchical cluster also works with variables as opposed to cases; it can cluster variables together in a manner somewhat similar to factor analysis. In addition, hierarchical cluster analysis can **handle nominal, ordinal, and scale data**; however it is **not recommended to mix different levels of measurement.**

ไม่เป็นขั้นตอน

Non - hierarchical procedure

K-means cluster

- is a method to quickly cluster large data sets. **The researcher define the number of clusters in advance**. This is useful to test different models with a different assumed number of clusters.

Two-step cluster

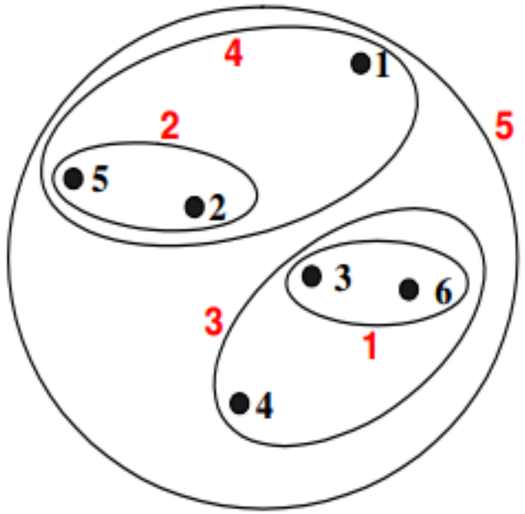
- analysis identifies groupings by running pre-clustering first and then by running hierarchical methods. Because it uses a quick cluster algorithm upfront, **it can handle large data sets** that would take a long time to compute with hierarchical cluster methods. In this respect, it is a combination of the previous two approaches. Two-step clustering can handle scale and ordinal data in the same model, and it automatically selects the number of clusters.

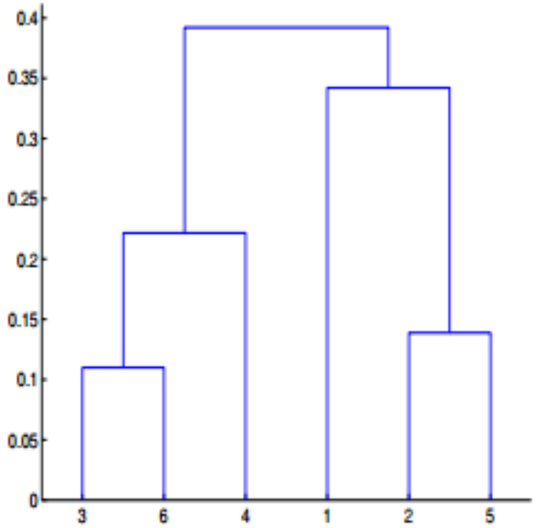
ขั้นตอนในการวิเคราะห์การจัดกลุ่ม

1. เลือกวิธีวัดระยะห่าง (Distence measure)
2. เลือกวิธีการจัดกลุ่ม (Clustering algorithm)
3. หาระยะห่างระหว่างกลุ่มสองกลุ่ม (Define the distance between two clusters)
4. ระบุจำนวนกลุ่ม (Determine the number of clusters)
5. ตรวจสอบความเหมาะสมของผลการวิเคราะห์ (Validate the analysis)

Hierarchical cluster analysis : HCA

- การวิเคราะห์จัดกลุ่มตามลำดับชั้น -





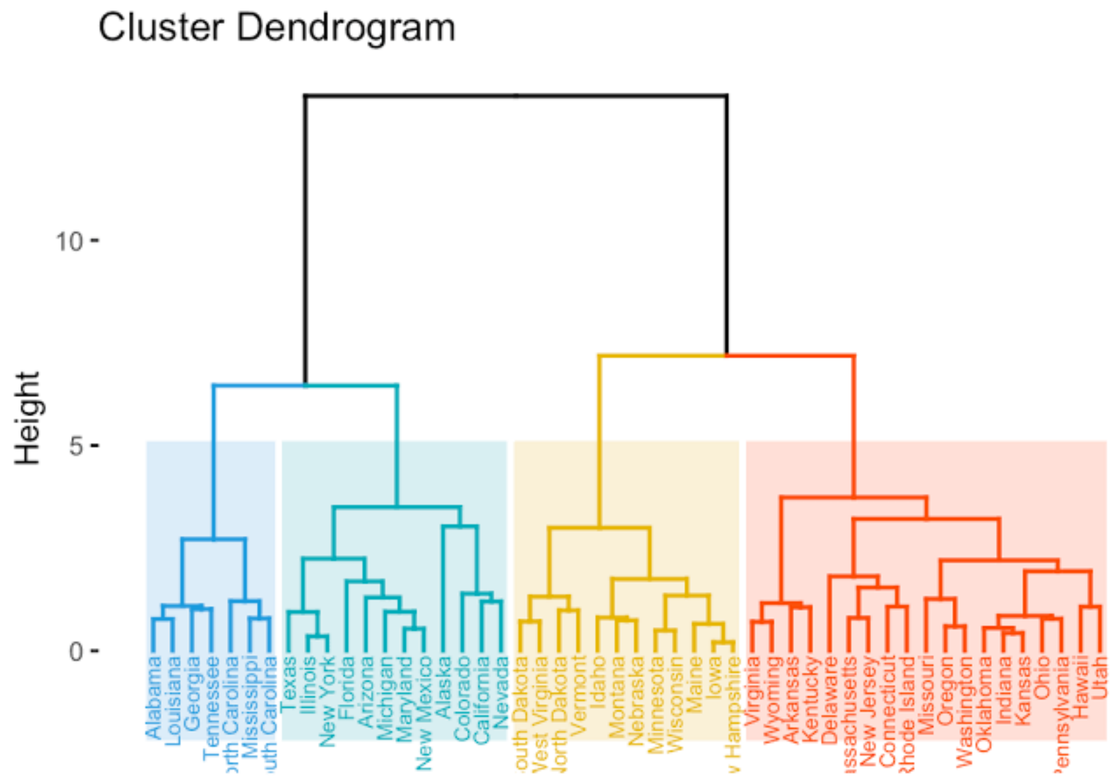
Source: https://www.analyticsvidhya.com/blog/2017/02/test-data-scientist-clustering/

เป็นการจัดกลุ่มที่นิยมใช้แบ่งกลุ่ม Case หรือกลุ่มตัวแปร โดยมีเงื่อนไข ดังนี้

1. **เหมาะกับข้อมูลขนาดเล็ก** (จำนวนเคส < 200 เคส/ตัวแปร)
   - ตัวอย่าง => แบ่งกลุ่ม (Classify cases)
   - ตัวแปร => ทดสอบความสัมพันธ์ระหว่างตัวแปร
2. ไม่จำเป็นต้องทราบจำนวนกลุ่มมาก่อน
3. ไม่จำเป็นต้องทราบว่าตัวแปรใดหรือเคสใดอยู่กลุ่มใดก่อน
4. ชนิดตัวแปรที่เหมาะสมคือ **nominal, ordinal, and scale data** และไม่ควรผสมชนิดของตัวแปร

HCA - เกณฑ์ในการจัดกลุ่ม

แยกแต่ละตัวอย่างไปยังกลุ่ม ด้วยระยะทาง (Distance) เริ่มต้นโดยแบ่งแยก 2 กลุ่มที่เหมือนกันมากที่สุดไปเรื่อย ๆ จนครบทุกตัวอย่าง



Source: https://stackoverflow.com/questions/55824002/d3-dendrogram-straight-edges

ขั้นตอนของเทคนิค Hierarchical cluster analysis สำหรับการแบ่งกลุ่มเคส

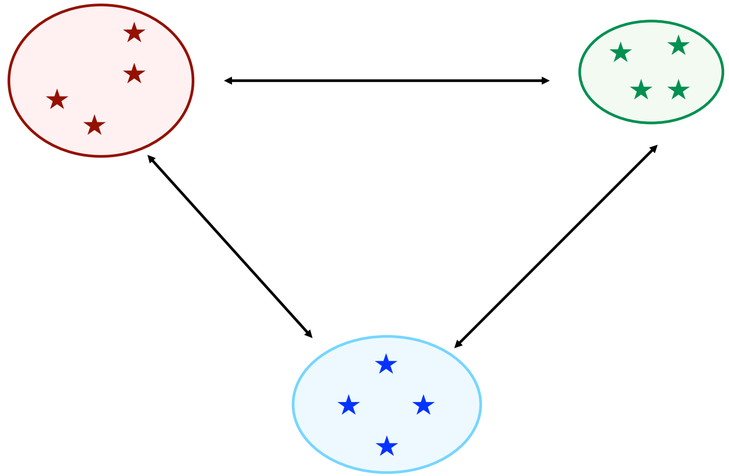
1. เลือกตัวแปรหรือปัจจัยที่คาดว่ามีอิทธิพลที่ทำให้เคสต่างกัน ตัวแปรจะทำให้สามารถแบ่งกลุ่มเคสได้ชัดเจน
2. เลือกวิธีการวัดระยะห่างระหว่างเคสแต่ละคู่ หรือเลือกวิธีการคำนวณเพื่อวัดค่า ความคล้ายของเคสแต่ละคู่
3. เลือกเกณฑ์ในการรวมกลุ่มหรือรวม Cluster

Note: หากมีตัวแปรที่มีการแจกแจงไม่ปกติ ให้ทำการแปลงเป็นค่ามาตรฐานก่อน (Z-score)

เกณฑ์ในการรวมกลุ่ม

1. **Between - group linkage** (or average linkage between group)
2. **Within-group linkage** (or average linkage within groups method)
   - วิธีนี้จะรวม cluster เข้าด้วยกันถ้าระยะห่างเฉลี่ยระหว่างทุกเคสใน cluster นั้นๆ มีค่าน้อยที่สุด
3. **Centroid clustering** - รวม 2 cluster เข้าด้วยกัน โดยพิจารณาจากระยะห่างของจุดกลาวของ cluster 2 cluster
4. **Ward's method** - พิจารณาค่า sum of the squared within-cluster distance โดยจะรวม cluster ที่ทำให้ค่า sum of the squared within-cluster distance เพิ่มขึ้นน้อยที่สุด โดยค่า square within-cluster distance คือค่า square Euclidean distance ของแต่ละเคสกับ cluster mean

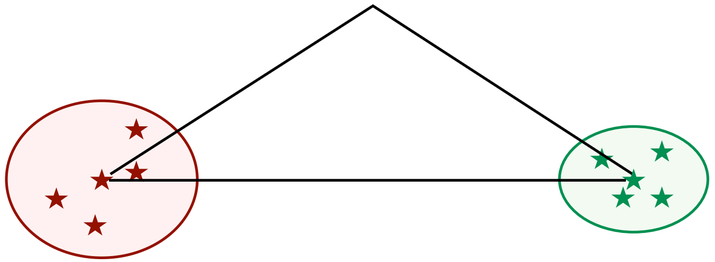
Between - group linkage



Cluster ที่ i

Cluster ที่ j

Centroid



Centroid clustering

Practice

Google classroom

Class code: 7dl3bgj

Practice

Learning objectives: To conduct cluster analysis in order to identify groups or segments of people having similar attitudes towards shopping, and to interpret and discuss the statistical output.

Exercise

To explore consumer attitudes towards shopping, a set of questions was asked to a random sample of respondents. Attitudes were measured on a seven point scale ranging from ‘extremely disagree’ to ‘extremely agree’ in the following way:

For me shopping is fun

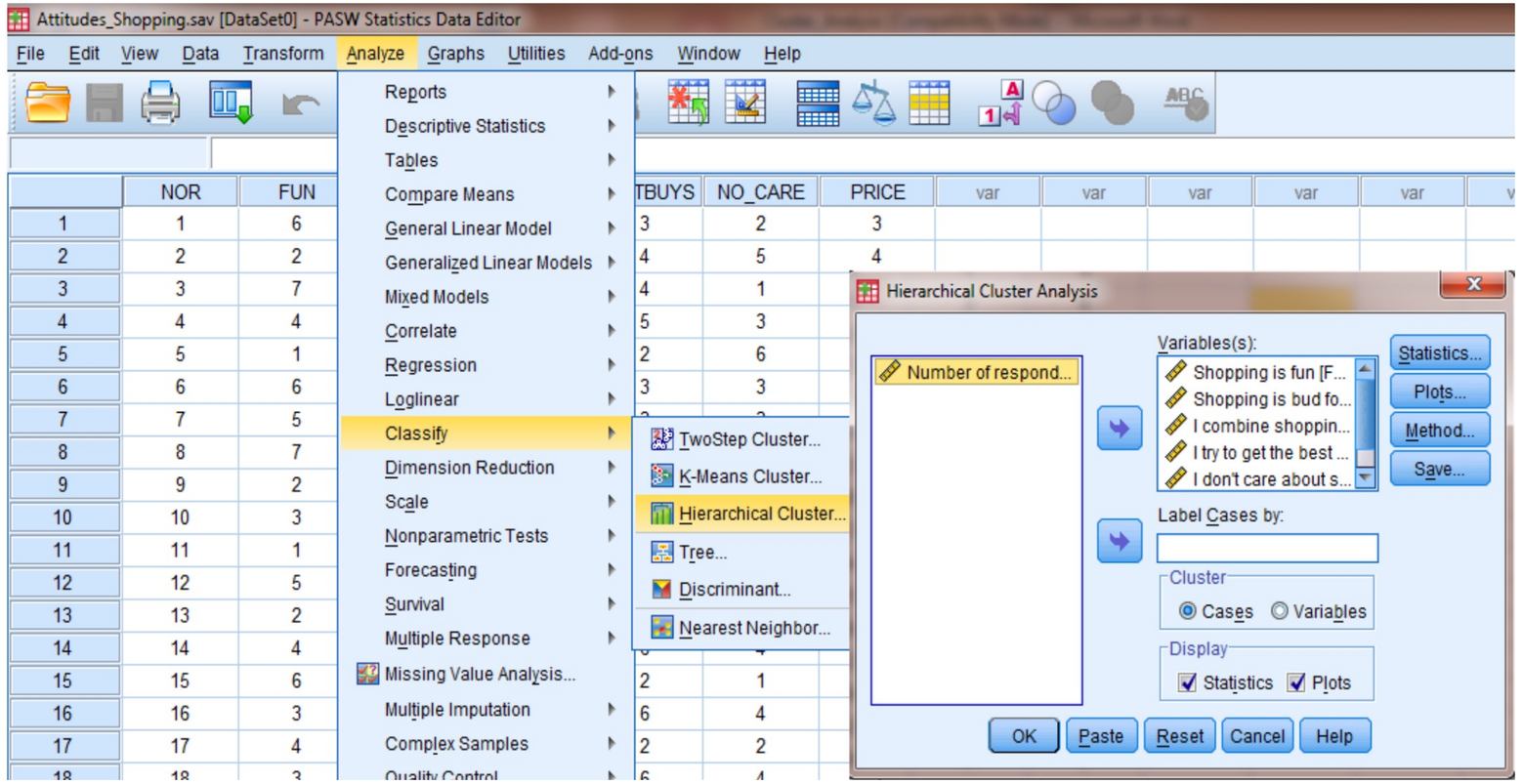
Extremely Disagree	Completely Disagree	Disagree	Neither disagree nor agree	Agree	Completely Agree	Extremely Agree
						

Table 1: Items used to measure consumers attitudes towards shopping.

Variables name	Attitude items	Type of data
1. NOR	Number of respondent	Scale
2. FUN	Shopping is fun	Scale
3. BUDGET	Shopping is bud for your budget	Scale
4. EATINGOUT	I combine shopping with eating out	Scale
5. BESTBUYS	I try to get the best buys when shopping	Scale
6. NO_CARE	I don’t care about shopping	Scale
7. PRICE	You can save a lot of money by comparing prices	Scale
8. GENDER		Nominal
9. EDUCATION		Ordinal
10. INCOME		Scale

10 Steps for cluster analysis in SPSS

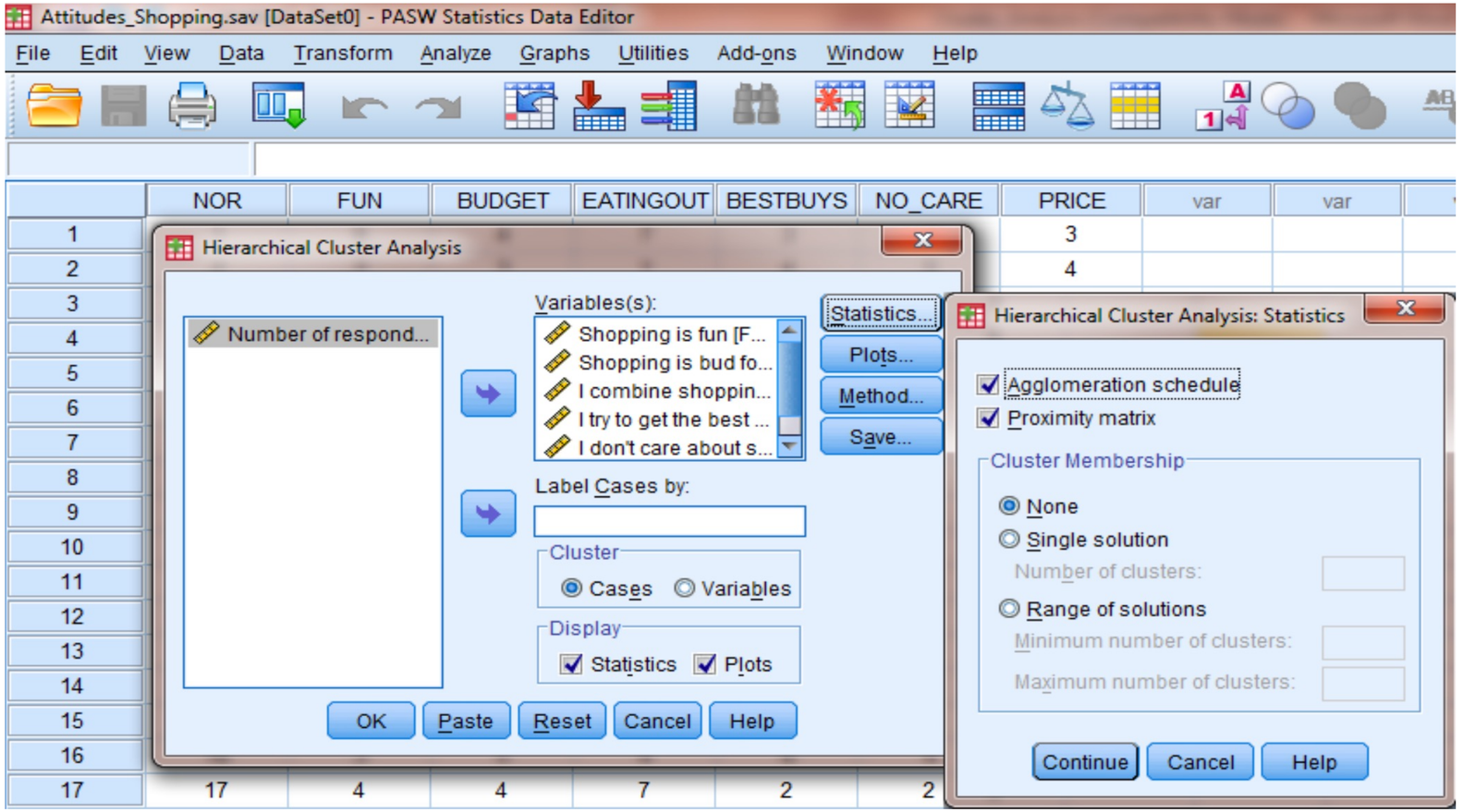
1. Select the **Analyze** menu;
2. Click on **Classify** and then **Hierarchical cluster...** to open the **Hierarchical** cluster Analysis dialogue box;
3. From the left hand side dialogue box select the 6 attitudinal variables and click on the blue arrow to move these variables into the **Variables** box;



10 Steps for cluster analysis in SPSS

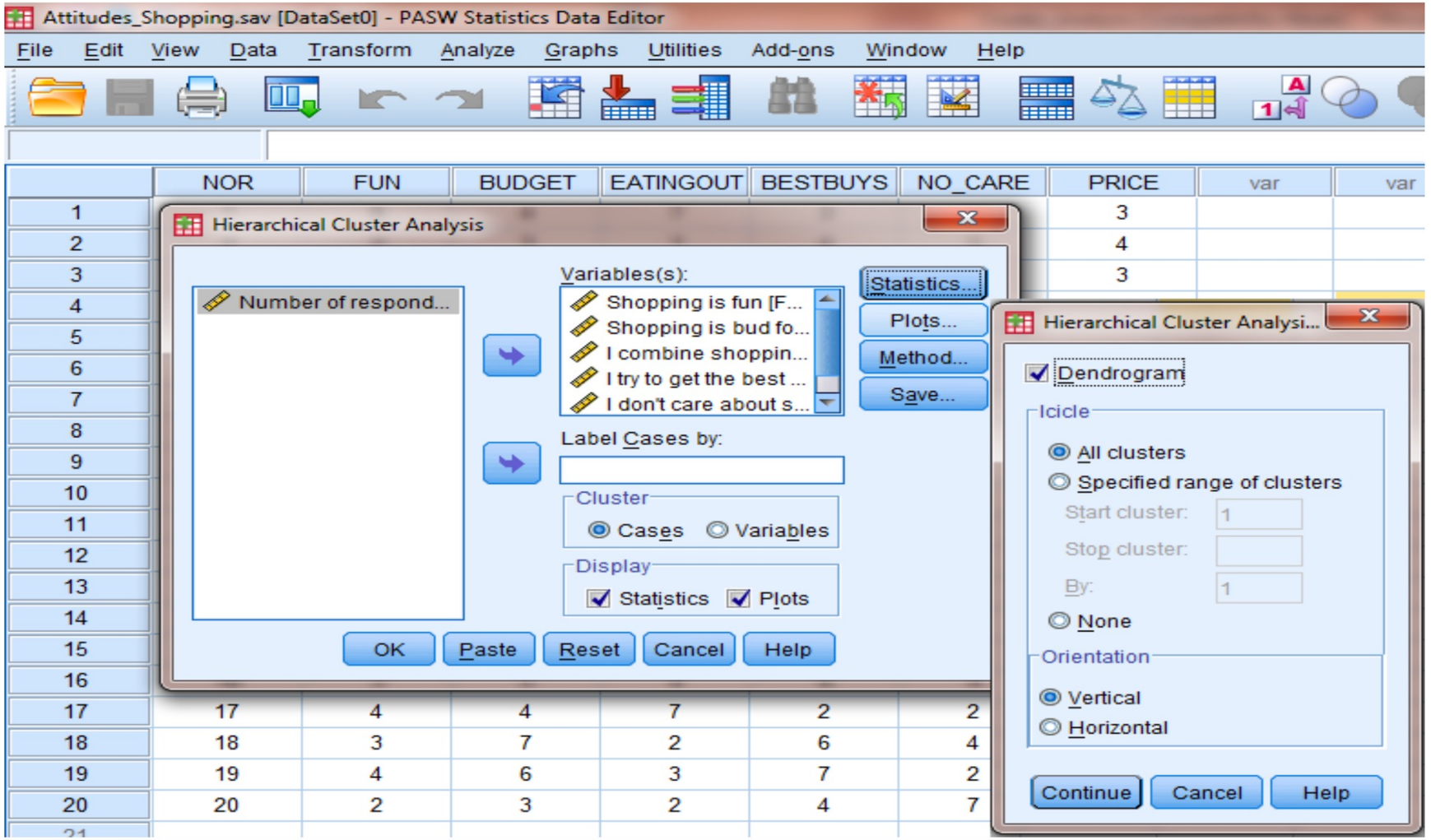
4. In the **Cluster** box, ensure that **Cases** radio button has been selected. As you can see you have also the option to cluster **Variables** i.e. the columns of your dataset;
5. In the **Display box**, ensure that **Statistics** and **Plot** check boxes have been selected;

6. Click on **Statistics** ... command push button to open the **Hierarchical Cluster Analysis: Statistics** sub-dialogue box, and ensure that **Agglomeration schedule** and **Proximity matrix** check boxes have been selected. At this stage do not select a cluster membership solution.



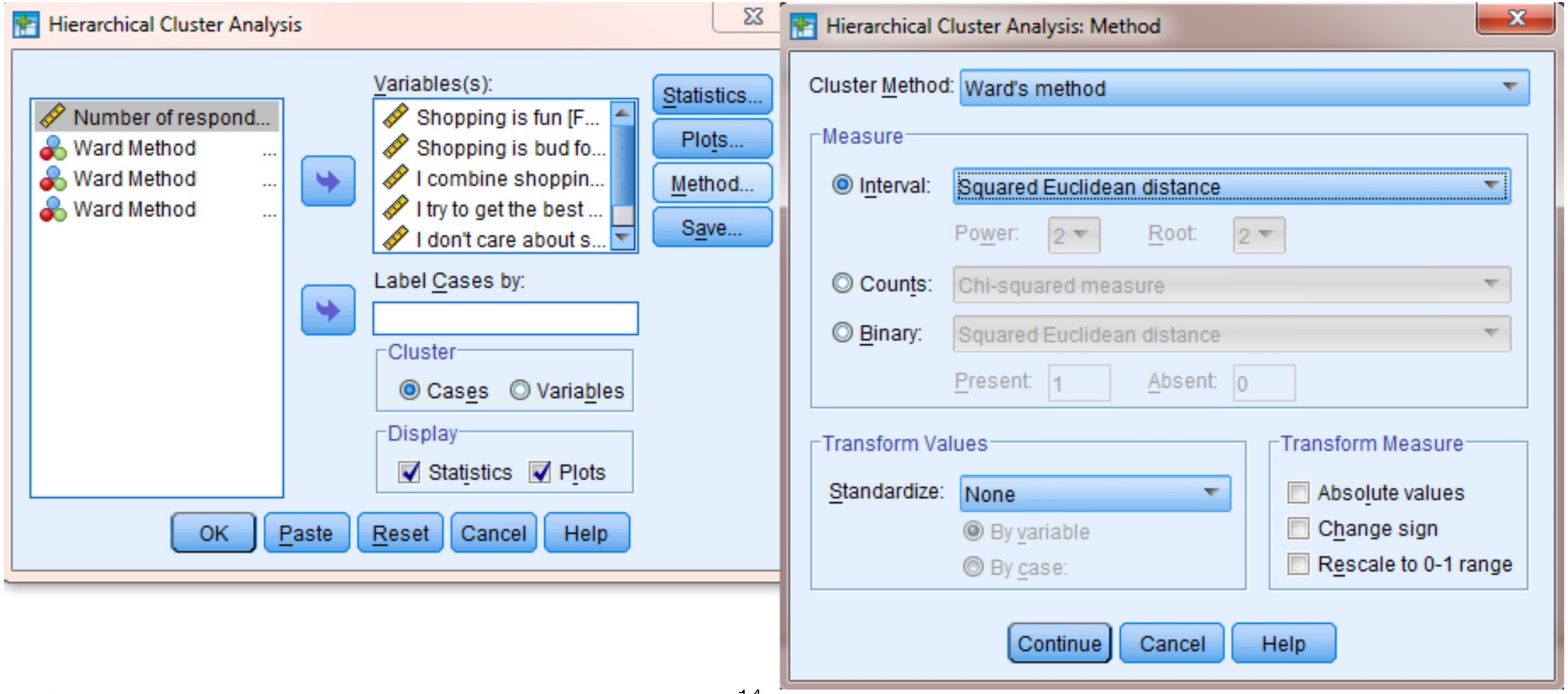
10 Steps for cluster analysis in SPSS

7. Click on the **Plots** command pushbutton to open the **Hierarchical Cluster Analysis: Plot** dialogue box and ensure that the **Dendogram** check box is selected ‘and then click on **Continue**;



10 Steps for cluster analysis in SPSS

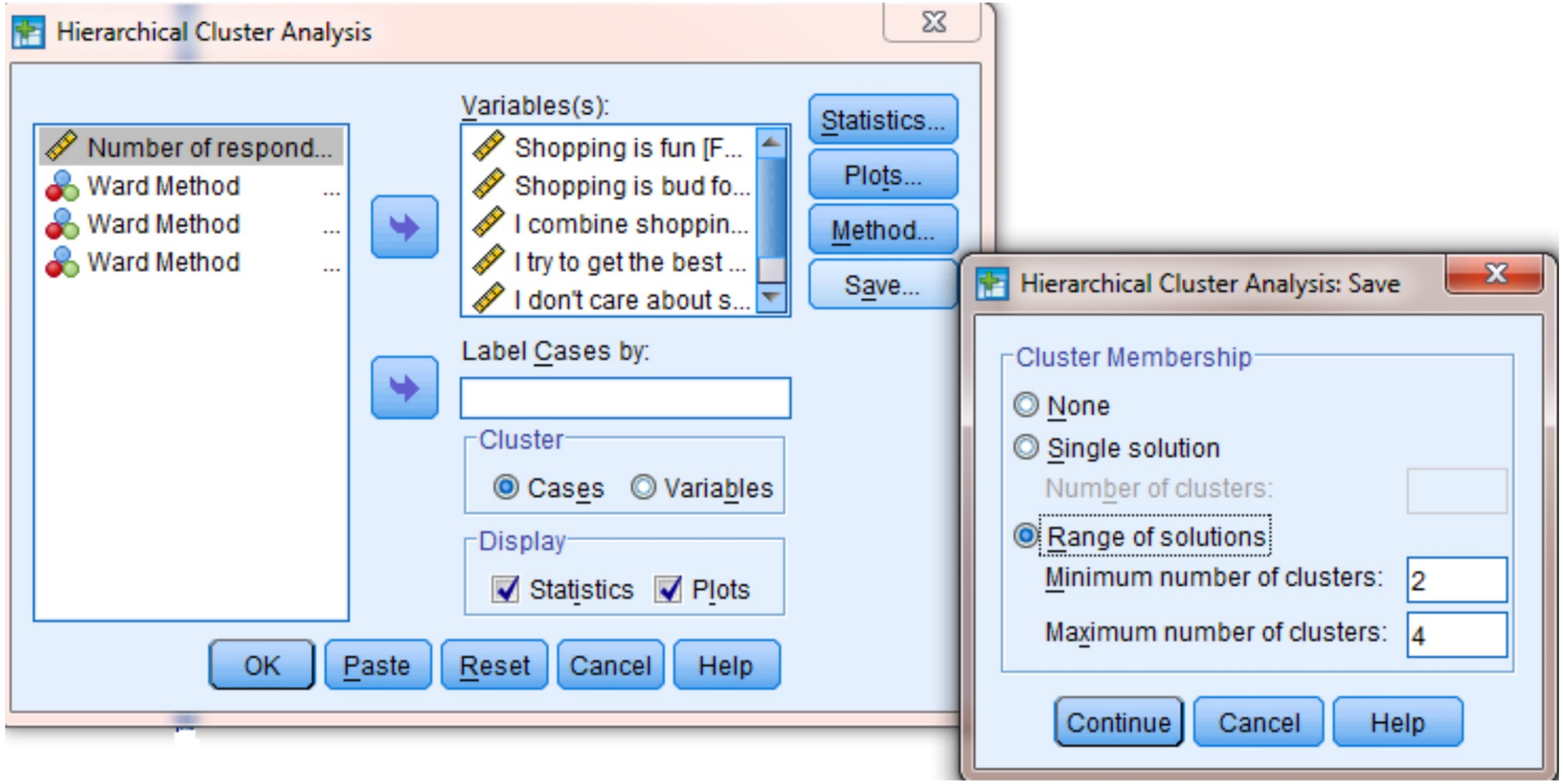
8. Click on the **Method...** command pushbutton to open the **Hierarchical Cluster: Method** sub-dialogue box, and in the **Cluster method:** dropdown list, and ensure both that the **Wards' method** and the **Squared Euclidean Distance are selected**;



10 Steps for cluster analysis in SPSS

9. Click on the **Save...** command pushbutton to open the **Hierarchical Cluster: Save** and check **Range of solutions** radio button (**Minimum number of clusters** = 2; **Maximum number of clusters** = 4), and then click on **Continue**;

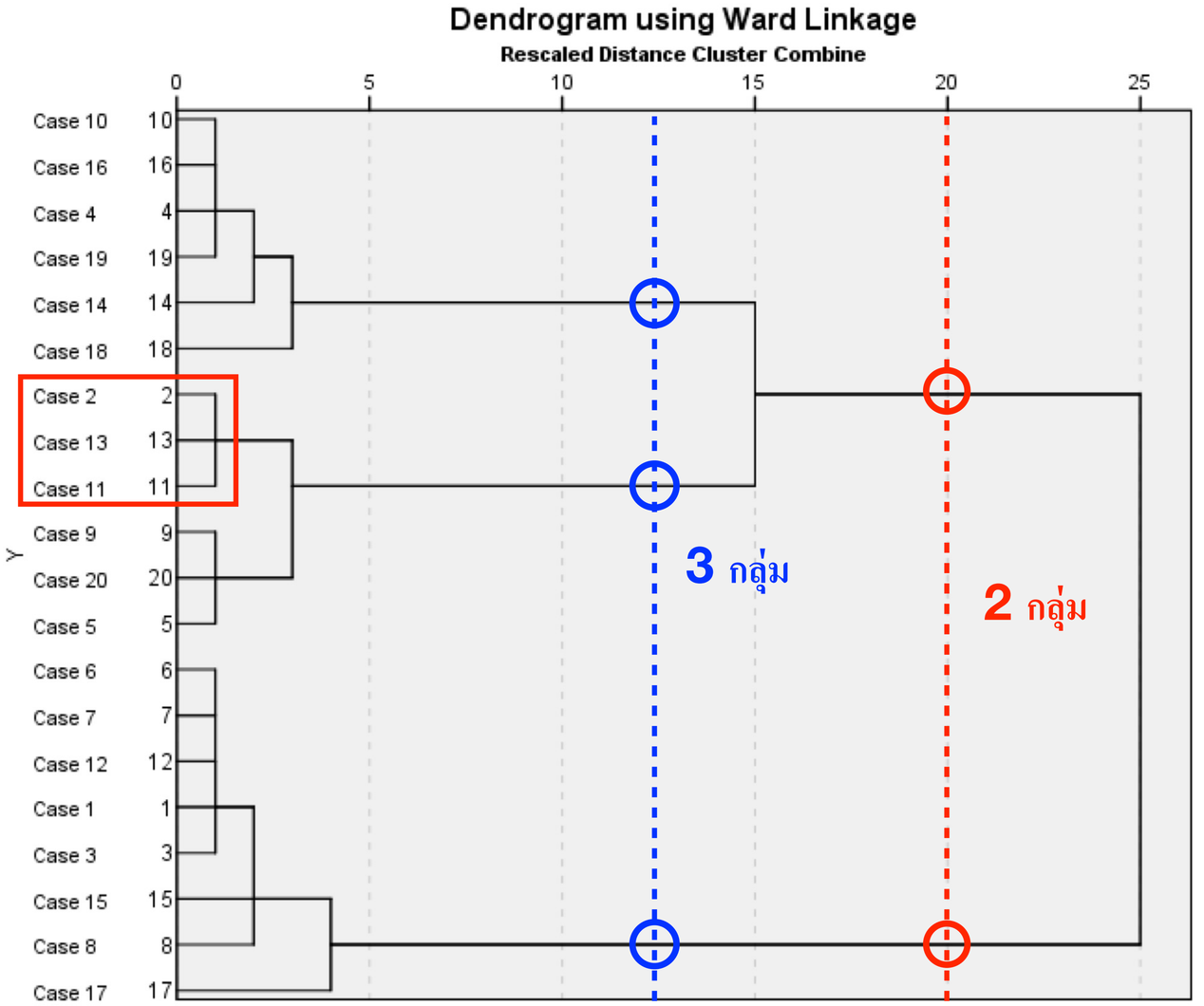
10.Click on **OK**.



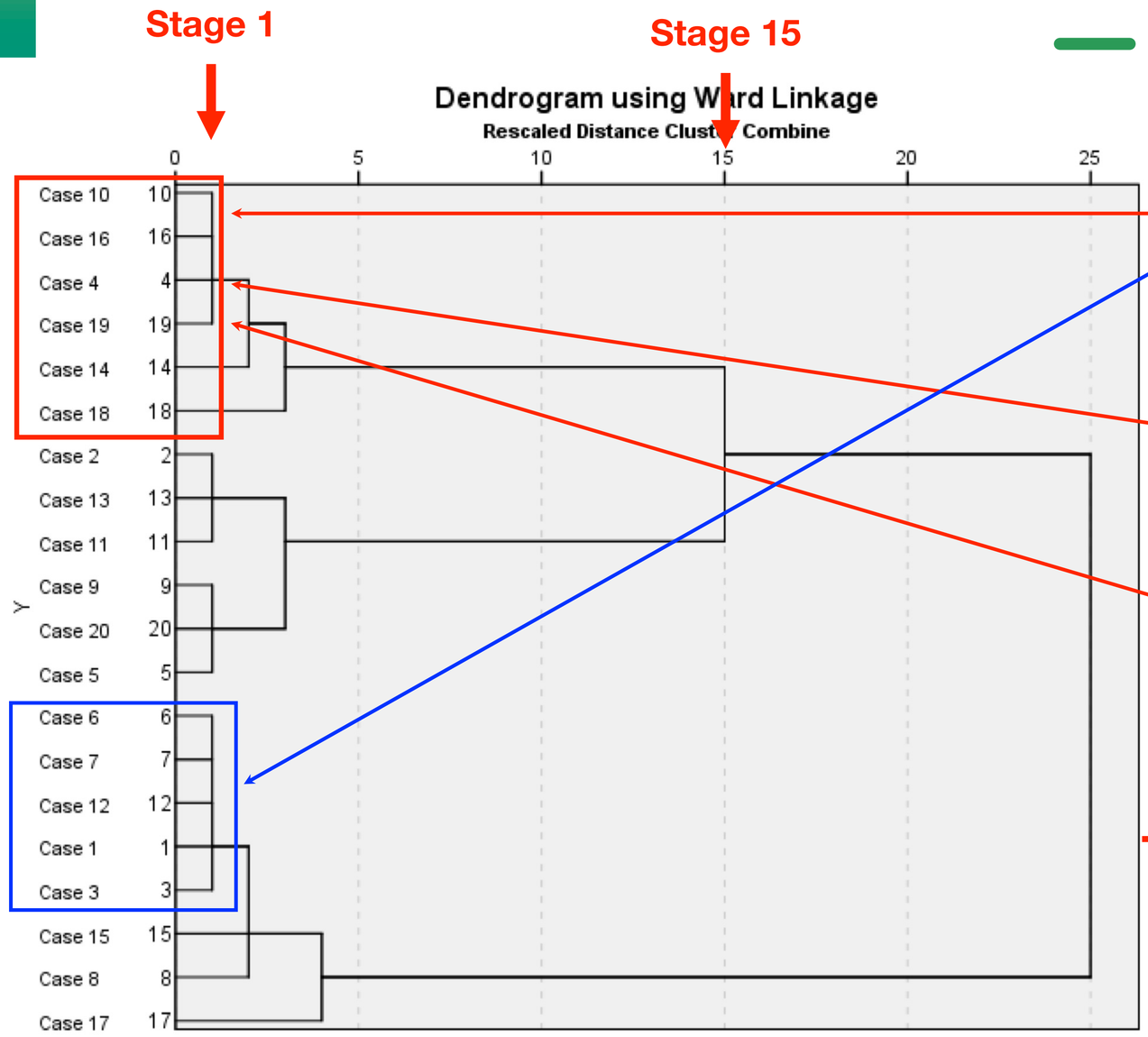
Dendrogram

แผนผังแสดงการจัดกลุ่ม เคสจะอยู่ด้านซ้าย โดยเคสที่มีความคล้ายกันจะอยู่ใกล้กัน

การพิจารณาจำนวนกลุ่มให้พิจารณาทางขวามือในแนวตั้งว่าต้องการกี่กลุ่ม



ผลการวิเคราะห์จัดกลุ่ม



Agglomeration Schedule

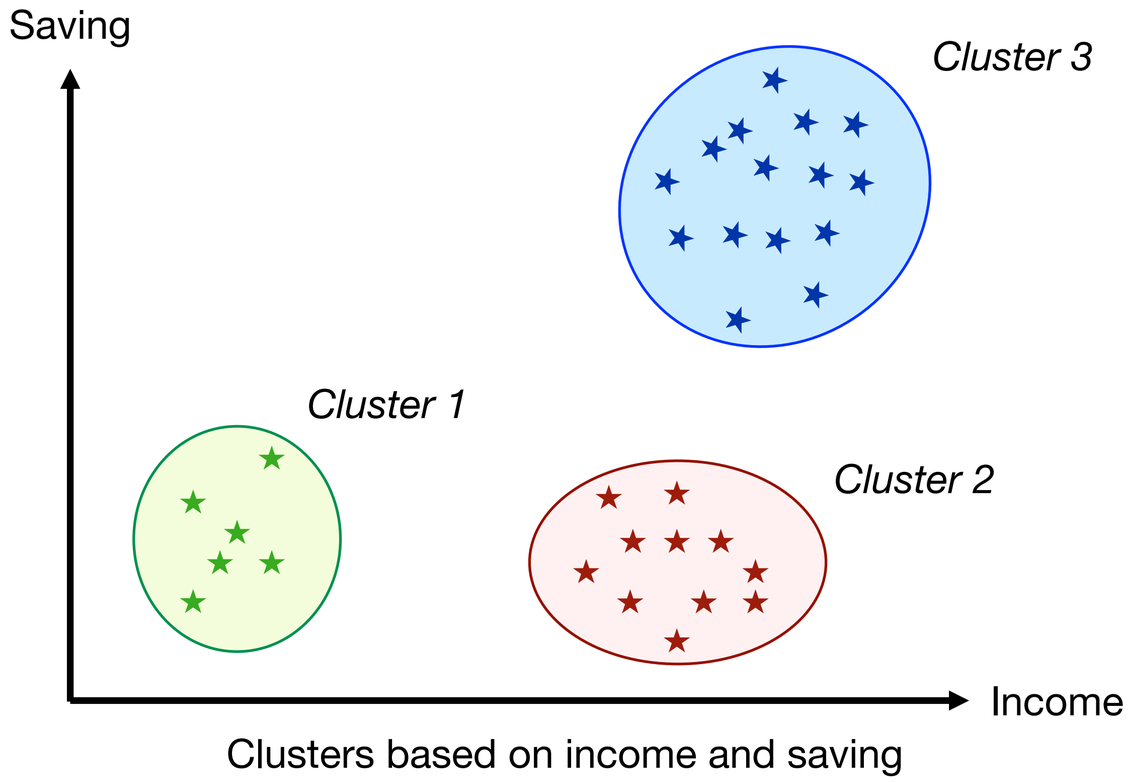
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	10	16	.348	0	0	6
2	6	7	.755	0	0	4
3	2	13	1.454	0	0	8
4	6	12	2.305	2	0	7
5	9	20	3.286	0	0	9
6	4	10	4.434	0	1	10
7	1	6	5.795	0	4	11
8	2	11	7.230	3	0	16
9	5	9	8.950	0	5	16
10	4	19	10.692	6	0	12
11	1	3	13.405	7	0	13
12	4	14	17.217	10	0	15
13	1	15	21.742	11	0	14
14	1	8	27.138	13	0	17
15	4	18	33.666	12	0	18
16	2	5	42.456	8	9	18
17	1	17	53.589	14	0	19
18	2	4	96.320	16	15	19
19	1	2	171.000	17	18	0

K-means cluster analysis

- การจัดกลุ่ม กรณีกลุ่มตัวอย่างขนาดใหญ่ -

- KCA is a method to quickly cluster large data sets. **The researcher define the number of clusters in advance**. This is useful to test different models with a different assumed number of clusters.

  - กำหนดจำนวนกลุ่ม (K) ที่ต้องการก่อน แล้วจัดตัวอย่างเข้ากลุ่ม
  - ใช้กับข้อมูลขนาดใหญ่ (> 200)
  - เหมาะกับตัวแปรค่าต่อเนื่อง (Scale)
  - ใช้ค่าเฉลี่ยเพื่อจัด Case เข้ากลุ่ม K
  - กลุ่มต่างกันน้อย กลุ่มต่างกันมาก
  - หาระยะห่างด้วยวิธี Euclidean distance



Clusters based on income and saving

Note: หากมีตัวแปรที่มีการแจกแจงไม่ปกติ ให้ทำการแปลงเป็นค่ามาตรฐานก่อน (Z-score)

K-means cluster analysis

10 Steps for cluster analysis in SPSS

1. Select the **Analyze** menu;
2. Click on **Classify** and then **K-means Cluster…**;
3. From the left hand side dialogue box select the 6 attitudinal variables and click on the blue arrow to move these variables into the **Variables** box;
4. Specify number of clusters in **Number of Clusters** dialogue.
5. Click **Iteration**: to specify Maximum iterations…, then **Continue**.
6. Click Options:
   - Select : **Initial cluster centres**
   - Select : **ANOVA table**
   - Select : **Cluster information for each case**
   - Select : **Exclude cases pairwise**
   - Then **Continue**.
7. **OK**

Initial Cluster Centers

	Cluster		
	1	2	3
Information about health risks caused by obesity, anorexia nervosa, bulimia and other illnesses linked to food	1	4	5
Information about foods related to health and beauty	1	4	5
Information about foods containing anti ageing properties	1	2	5
Information about life style, food tourism and eating out	1	1	5
Information about trends, consumptions evolution, food fads, and underscoring ethnicity, cultural, social diversity of Italian population	1	1	5
Information about food safety issues caused by bacteria and other substances	1	5	5
Information about food regulations affecting consumer choices and the food industry	1	2	5
Information about tradition, regional typical products and quality foods that are disappearing from the Italian market	1	4	5
Information about production techniques used in the primary sector	1	5	5
Information about the food processing industry and innovations in terms of products and processes	1	4	5
Information about Italian and international cuisine, food culture and good living	1	1	5

การวิเคราะห์ผล

Initial Cluster Centers

เป็นระยะห่างของตัวแปรในแต่ละกลุ่ม

Iteration History

แสดงการจัดกลุ่มในแต่ละรอบของการหมุน ซึ่งจะแตกต่างกันมากในช่วงแรก และลดลงเรื่อยๆ จนเป็น 0

หยุดหมุนรอบที่ 28
โดยค่า Max Iteration จะผันตามจำนวนเคส

Iteration History[a]

Iteration	Change in Cluster Centers		
	1	2	3
1	4.362	4.457	3.369
2	.804	.318	.210
3	.464	.159	.088
4	.325	.118	.055
5	.391	.130	.042
6	.332	.130	.038
7	.371	.160	.040
8	.248	.137	.035
9	.219	.137	.034
10	.156	.136	.061
11	.112	.125	.060
12	.094	.107	.051
13	.128	.140	.052
14	.159	.147	.043
15	.058	.071	.039
16	.063	.058	.032
17	.041	.066	.038
18	.025	.034	.018
19	.028	.039	.022
20	.013	.016	.008
21	.030	.026	.007
22	.035	.042	.022
23	.010	.026	.022
24	.014	.018	.014
25	.000	.017	.016
26	.020	.021	.011
27	.015	.013	.010
28	.000	.000	.000

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Information about health risks caused by obesity, anorexia nervosa, bulimia and other illnesses linked to food	13.171	2	.416	754	31.674	.000
Information about foods related to health and beauty	138.492	2	.700	752	197.937	.000
Information about foods containing anti ageing properties	47.559	2	.643	751	74.018	.000
Information about life style, food tourism and eating out	75.393	2	.658	754	114.566	.000
Information about trends, consumptions evolution, food fads, and underscoring ethnicity, cultural, social diversity of Italian population	99.363	2	.651	750	152.653	.000
Information about food safety issues caused by bacteria and other substances	27.729	2	.364	754	76.144	.000
Information about food regulations affecting consumer choices and the food industry	111.066	2	.654	746	169.738	.000
Information about tradition, regional typical products and quality foods that are disappearing from the Italian market	78.180	2	.536	752	145.961	.000
Information about production techniques used in the primary sector	121.989	2	.512	751	238.167	.000
Information about the food processing industry and innovations in terms of products and processes	138.918	2	.567	745	244.982	.000
Information about Italian and international cuisine, food culture and good living	71.212	2	.661	751	107.777	.000

การวิเคราะห์ผล

ANOVA

ใช้ชี้วัดว่าตัวแปรใดมีผลมากที่สุดต่อการกำหนดกลุ่มที่กำหนดไว้

ซึ่งพิจารณาจากตัวแปรที่มีค่า F-stat สูงสุด แสดงว่าเป็นตัวแปรที่ใช้แบ่งกลุ่มได้ดีที่สุด

Final Cluster Centers

	Cluster		
	1	2	3
Information about health risks caused by obesity, anorexia nervosa, bulimia and other illnesses linked to food	4	5	5
Information about foods related to health and beauty	3	3	4
Information about foods containing anti ageing properties	4	4	4
Information about life style, food tourism and eating out	3	3	4
Information about trends, consumptions evolution, food fads, and underscoring ethnicity, cultural, social diversity of Italian population	3	3	4
Information about food safety issues caused by bacteria and other substances	4	5	5
Information about food regulations affecting consumer choices and the food industry	3	4	4
Information about tradition, regional typical products and quality foods that are disappearing from the Italian market	3	4	4
Information about production techniques used in the primary sector	3	4	4
Information about the food processing industry and innovations in terms of products and processes	2	4	4
Information about Italian and international cuisine, food culture and good living	3	3	4

การวิเคราะห์ผล

Final Cluster Centers

แสดงค่าเฉลี่ยของตัวแปรแต่ละตัวแปรในการแบ่งกลุ่มครั้งสุดท้าย ซึ่งสามารถสะท้อนคุณลักษณะของตัวอย่างในแต่ละกลุ่มได้

Distances between Final Cluster Centers

Cluster	1	2	3
1		2.562	3.673
2	2.562		2.127
3	3.673	2.127	

กลุ่ม 1 กับกลุ่ม 3 ต่างกันมากที่สุด
กลุ่ม 2 คล้ายกับกลุ่ม 1 และกลุ่ม 3

Number of Cases in each Cluster

Cluster	1	173.000
	2	283.000
	3	301.000
Valid		757.000
Missing		.000

แสดงจำนวนตัวอย่างของแต่ละกลุ่ม โดยกลุ่มที่ 3 มีกลุ่มตัวอย่างมากที่สุด
ซึ่งข้อมูลชุดนี้อาจแบ่งออกเป็น 4 กลุ่มได้ ซึ่งสามารถวิเคราะห์อีกครั้งโดยกำหนดจำนวนกลุ่มที่ 4 กลุ่ม

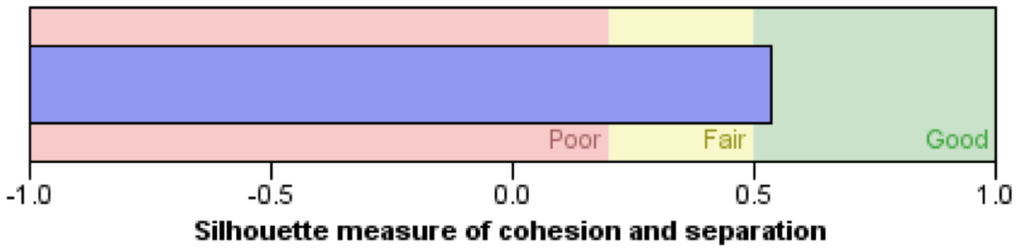
Two-step cluster analysis

- analysis identifies groupings by running pre-clustering first and then by running hierarchical methods. Because it uses a quick cluster algorithm upfront, **it can handle large data sets** that would take a long time to compute with hierarchical cluster methods. In this respect, it is a combination of the previous two approaches. Two-step clustering **can handle scale and ordinal data in the same model**, and it automatically selects the number of clusters.

Model Summary

Algorithm	TwoStep
Inputs	4
Clusters	4

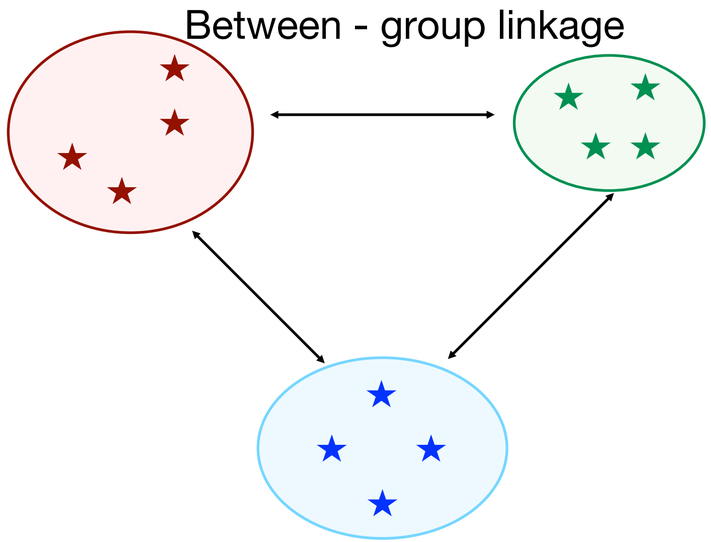
Cluster Quality

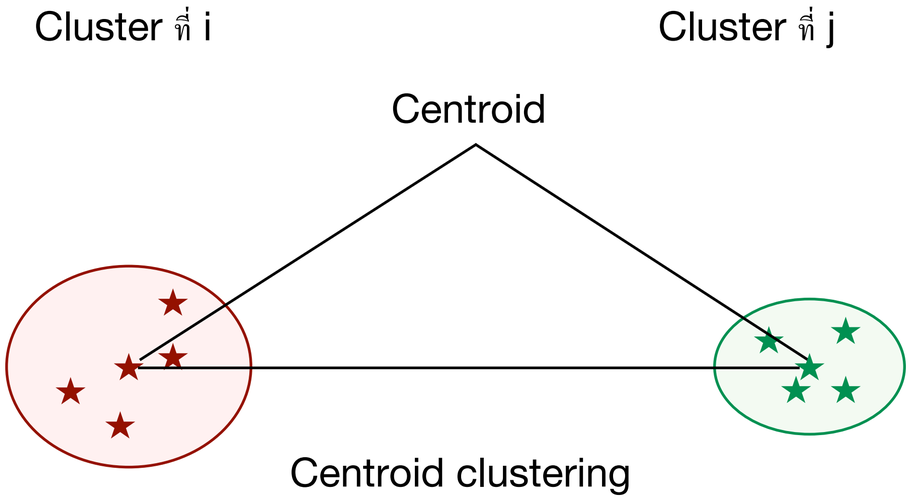


เกณฑ์ในการรวมกลุ่ม

1. **Between - group linkage** (or average linkage between group)
2. **Within-group linkage** (or average linkage within groups method) - วิธีนี้จะรวม cluster เข้าด้วยกันถ้าระยะห่างเฉลี่ยระหว่างทุกเคสใน cluster นั้นๆ มีค่าน้อยที่สุด
3. **Nearest neighbour (Single linkage)** รวม Cluster 2 Cluster เข้าด้วยกันโดยพิจารณาจากระยะห่างที่สั้นที่สุด
4. **Furthest neighbour** (Complete linkage) รวม Cluster 2 Cluster เข้าด้วยกันโดยพิจารณาจากระยะห่างที่ยาวที่สุด
5. **Centroid clustering** - รวม 2 cluster เข้าด้วยกัน โดยพิจารณาจากระยะห่างของจุดกลาวของ cluster 2 cluster
6. **Median clustering** รวม Cluster 2 Cluster เข้าด้วยกัน โดยให้แต่ละ Cluster สาคัญเท่ากัน

7. **Ward's method** - พิจารณาค่า sum of the squared within-cluster distance โดยจะรวม cluster ที่ทำให้ค่า sum of the squared within-cluster distance เพิ่มขึ้นน้อยที่สุด โดยค่า square within-cluster distance คือค่า **square Euclidean distance** ของแต่ละเคสกับ cluster mean





ข้อแตกต่างระหว่างการจำแนกกลุ่มด้วย เทคนิค Cluster analysis และเทคนิค Discriminant Analysis

Cluster Analysis (การวิเคราะห์กลุ่ม)	Discriminant Analysis (การวิเคราะห์จำแนกกลุ่ม)
1. ไม่จำเป็นต้องทราบก่อนว่ามีกี่กลุ่ม	1. ทราบจำนวนกลุ่มก่อน (สามารถกำหนดเองได้)
2. ไม่ทราบมาก่อนว่า Case ใดอยู่กลุ่มไหน	2. ทราบมาก่อนว่า Case ใดอยู่กลุ่มไหน
3. ไม่มีการแสดงความสัมพันธ์	3. มีสมการแสดงความสัมพันธ์

References

1. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). Multivariate data analysis (Vol. 7). Upper Saddle River, NJ: Pearson Prentice Hall

2. Wongsaichue, T. 2018. Cluster analysis - กรณีกลุ่มตัวอย่างขนาดใหญ่ (K-means cluster analysis) online: https://www.youtube.com/watch?v=5OaZM6x29U0.

3. Gaskin, J. 2012. Two-step Cluster Analysis in SPSS. Online: https://www.youtube.com/watch?v=DpucueFsigA.