Cluster analysis

Based on H.C. Romesburg: *Cluster analysis for researchers,*
Lifetime Learning Publications, Belmont, CA, 1984
P.H.A. Sneath and R.R. Sokal: *Numericxal Taxonomy*, Freeman,
San Fransisco, CA, 1973

Jens C. Frisvad
BioCentrum-DTU

Biological data analysis and chemometrics

Two primary methods

- Cluster analysis (no projection)
  - Hierarchical clustering
  - Divisive clustering
  - Fuzzy clustering
- Ordination (projection)
  - Principal component analysis
  - Correspondence analysis
  - Multidimensional scaling

Advantages of cluster analysis

- Good for a quick overview of data
- Good if there are many groups in data
- Good if unusual similarity measures are needed
- Can be added on ordination plots (often as a minimum spanning tree, however)
- Good for the nearest neighbours, ordination better for the deeper relationships

Different clustering methods

- NCLAS: Agglomerative clustering by distance optimization
- HMCL: Agglomerative clustering by homogeneity optimization
- INFCL: Agglomerative clustering by information theory criteria
- MINGFC: Agglomerative clustering by global optimization
- ASSIN: Divisive monothetic clustering
- PARREL: Partitioning by global optimization
- FCM: Fuzzy c-means clustering
- MINSPAN: Minimum spanning tree
- REBLOCK: Block clustering (k-means clustering)

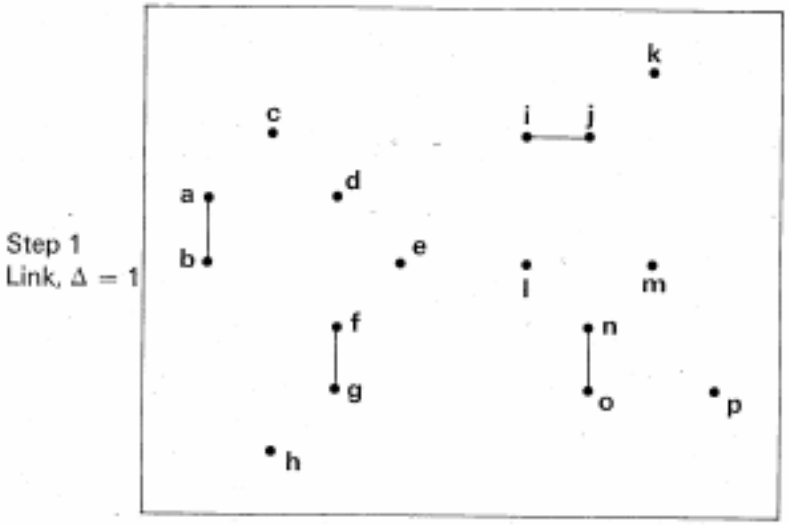
SAHN clustering

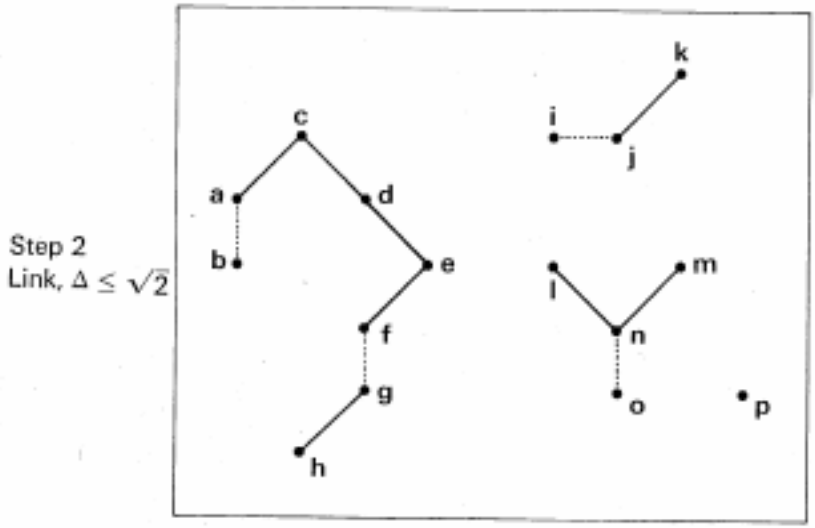
- Sequential agglomerative hierarchic nonoverlapping clustering

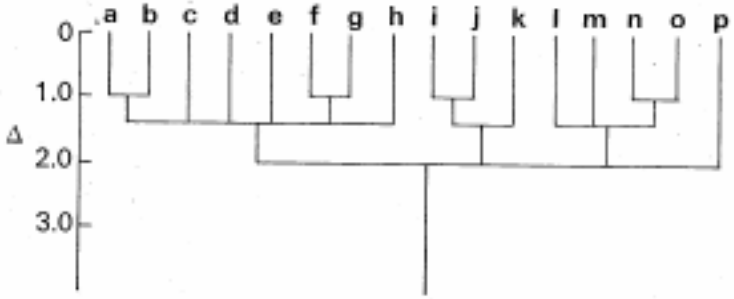
Single linkage

- Nearest neighbor, minimum method
- Close to minimum spanning tree
- Contracting space
- Chaining possible
- $\alpha_J = 0.5$, $\alpha_K = 0.5$, $\beta = 0$, $\gamma = -0.5$
- $U_{J,K} = \min U_{jk}$

$$U_{(J,K)L} = \alpha_J U_{J,L} + \alpha_K U_{K,L} + \beta U_{J,K} + \gamma \left| U_{J,L} - U_{K,L} \right|$$







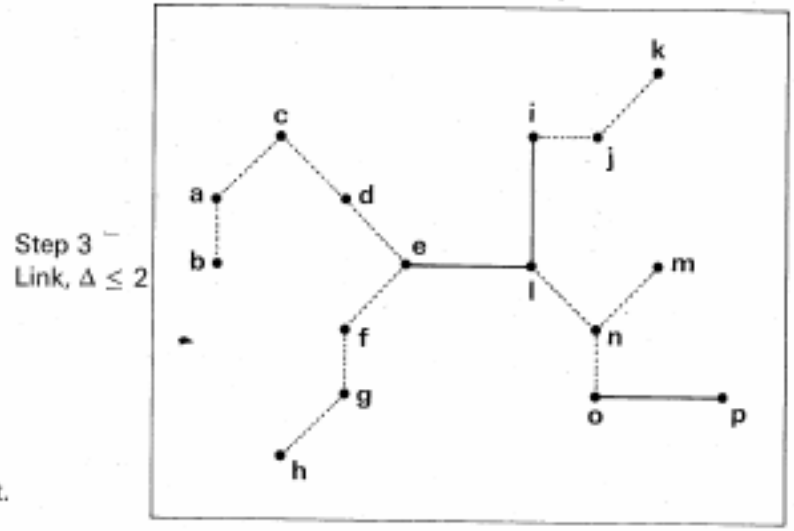
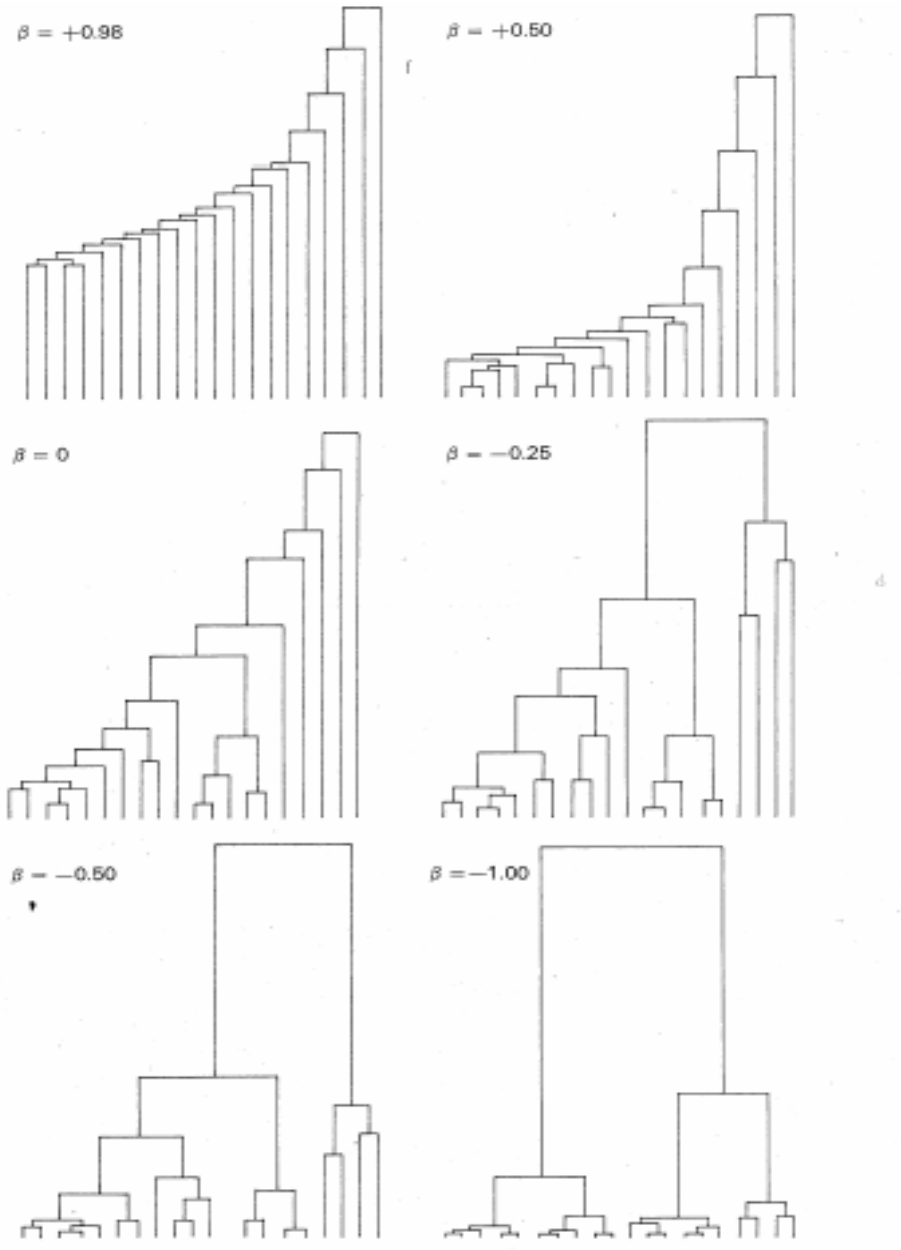


FIGURE 5-3
Single linkage clustering of the data in Table 5-1. For explanation, see text.


β = +0.98
β = +0.50
β = 0
β = −0.25
β = −0.50
β = −1.00

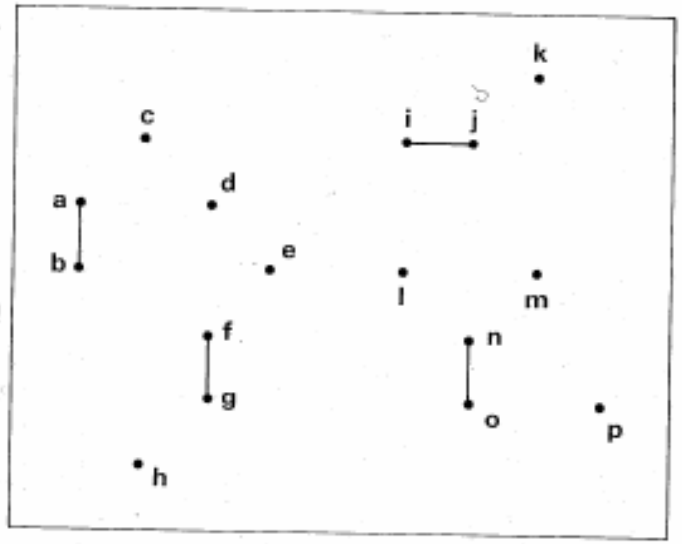
Complete linkage

- Furthest neighbor, maximum method
- Dilating space
- $\alpha_J = 0.5$, $\alpha_K = 0.5$, $\beta = 0$, $\gamma = 0.5$
- $U_{J,K} = \max U_{jk}$

Step 1

Level of cohesion

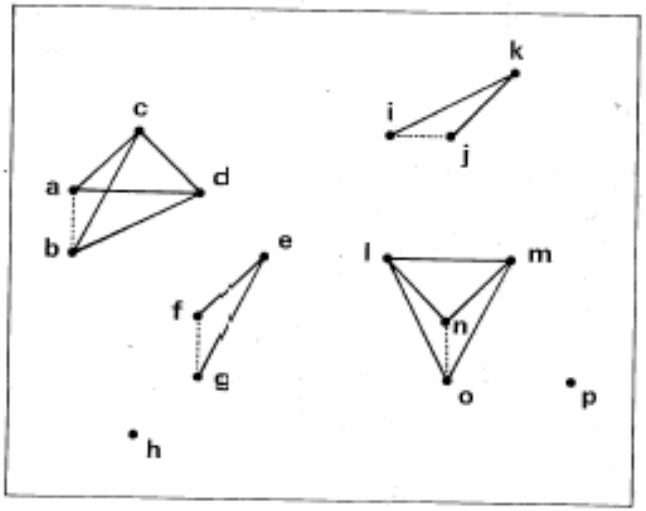
(maximum link), $\Delta = 1$



Step 2

Level of cohesion

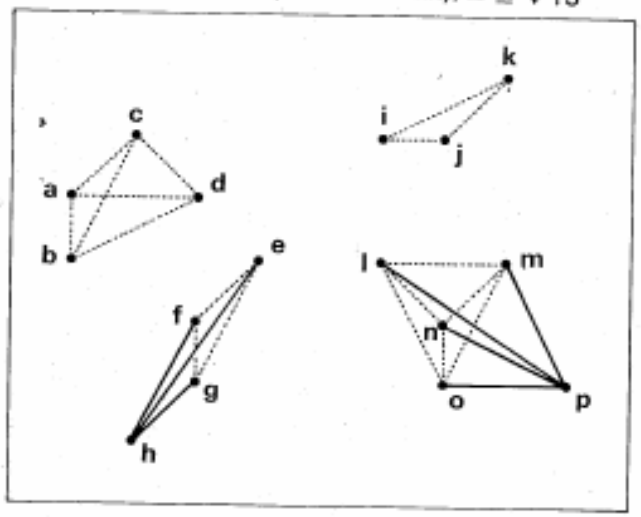
(maximum link), $\Delta \leq \sqrt{5}$



Step 3

Level of cohesion

(maximum link), $\Delta \leq \sqrt{13}$

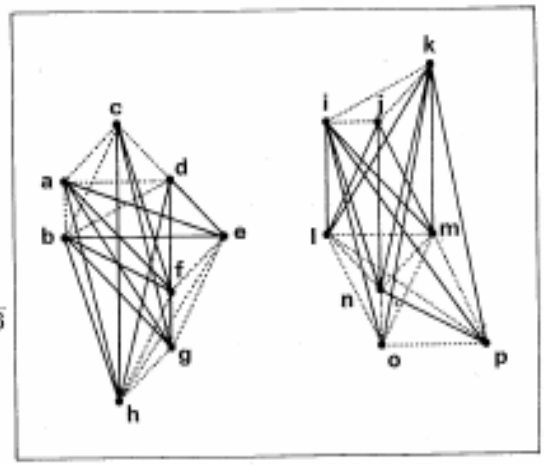


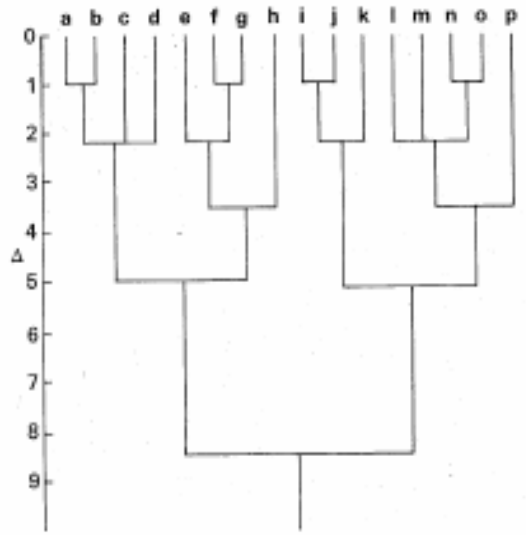
Step 4

Levels of cohesion (maximum links)

Left cluster, $\Delta \leq 5$

Right cluster, $\Delta \leq \sqrt{26}$

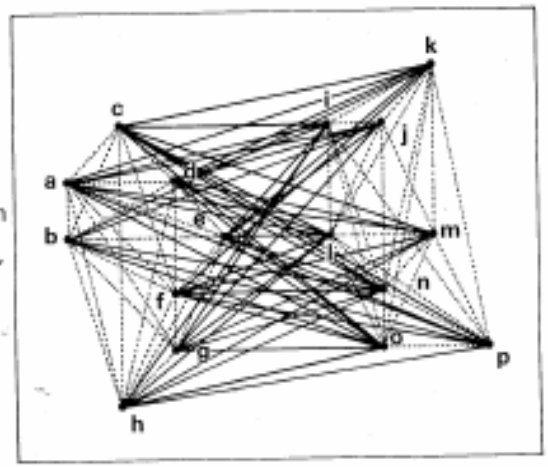


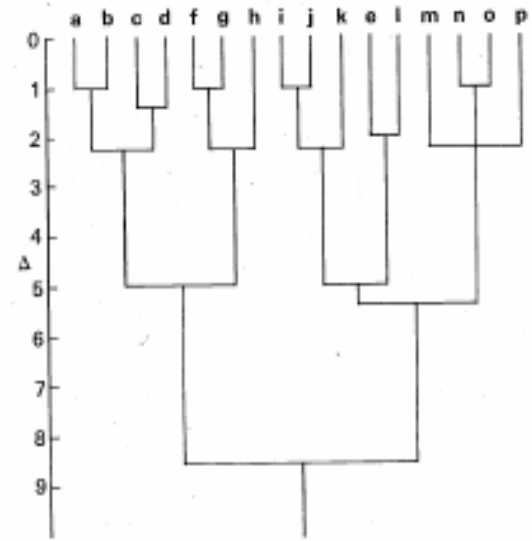


Phenogram A

Step 5

Level of cohesion (maximum link), $\Delta \leq \sqrt{73}$





Phenogram B

Average linkage

- Aritmetic average
  - Unweighted: UPGMA (group average)
  - Weighted: WPGMA
- Centroid
  - Unweighted centroid (Centroid)
  - Weighted centroid (Median)

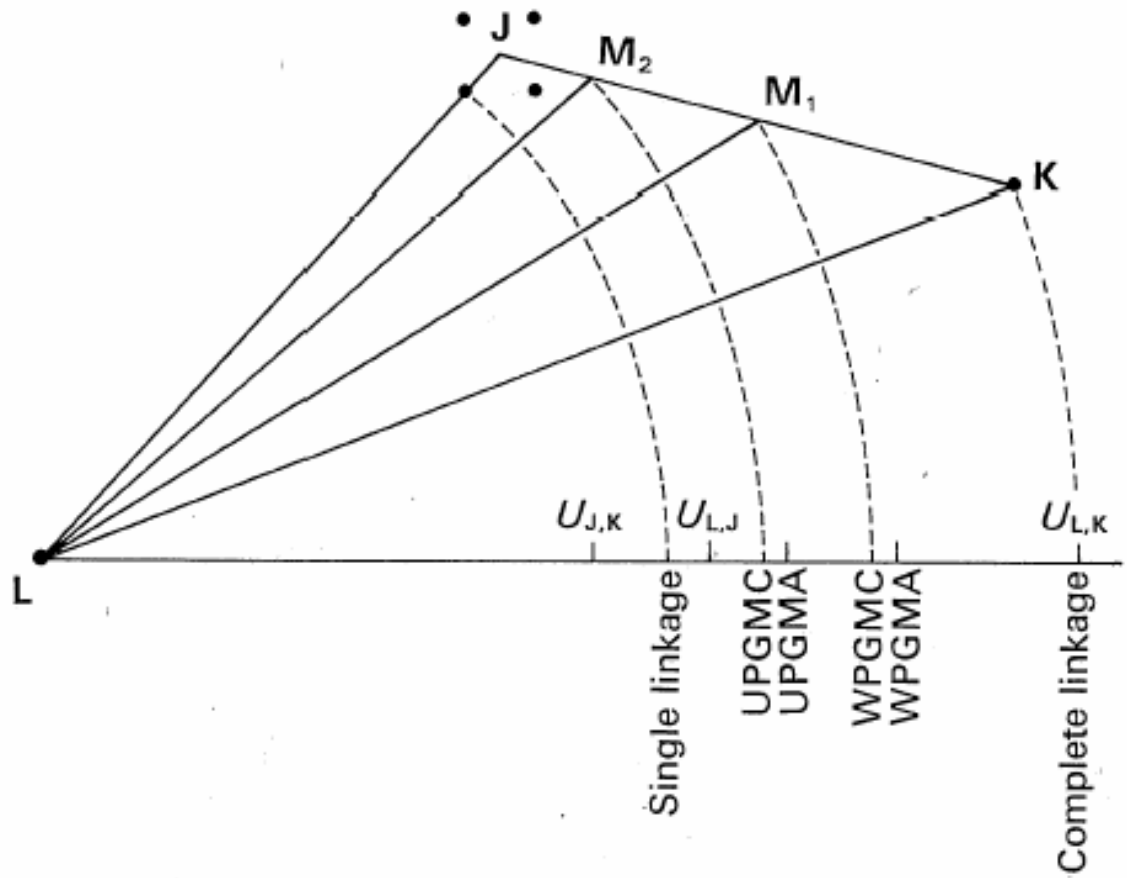


FIGURE 5-11
The effects of several clustering methods (for explanation see text) on the criterion for admitting **L** (containing one OTU) to the cluster formed of four OTU's in **J** plus one in **K**. OTU's are indicated by solid circles. Abscissa is distance Δ.

From Sneath and Sokal, 1973, Numerical taxonomy

Ordinary clustering

- Obtain the data matrix
- Transform or standardize the data matrix
- Select the best resemblance or distance measure
- Compute the resemblance matrix
- Execute the clustering method (often UPGMA = average linkage)
- Rearrange the data and resemblance matrices
- Compute the cophenetic correlation coefficient

Binary similarity coefficients

(between two objects i and j)

i \ j	1	0
1	a	b
0	c	d

Matches and mismatches

- m = a + b (number of matches)
- u = c + d (number of mismatches)
- n = m + u = a + b + c + d (total sample size)

- Similarity (often 0 to 1)
- Dissimilarity (distance) (often 0 to 1)
- Correlation (-1 to 1)

Simple matching coefficient

- $SM = (a + d) / (a + b + c + d) = m / n$

- Euclidean distance for binary data:
- $D = 1\text{-}SM = (b + c) / (a + b + c + d) = u / n$

Avoiding zero zero comparisons

- Jaccard = J = a / (a +b +c)
- Sørensen or Dice: DICE = 2a / (2a + b + c)

Correlation coefficients

Yule: (ad – bc) / (ad + bc)

$$PHI = (ad - bc) / \sqrt{(a+b)(c+d)(a+c)(b+d)}$$

Other binary coefficients

- Hamann = H = (a + d – b –c) / (a + b + c + d)
- Rogers and Tanimoto = RT = (a + d) / (a + 2b + 2c + d)
- Russel and Rao = RR = a / (a + b + c + d)
- Kulzynski 1 = K1 = a / (b + c)
- UN1 = (2a + 2d) / (2a + b + c + 2d)
- UN2 = a / (a + 2b + 2c)
- UN3 = (a + d) / (b + c)

Distances for quantitative (interval) data
Euclidean and taxonomic distance

$$EUCLID = E_{ij} = \sqrt{\sum_k (x_{ki} + x_{kj})^2}$$

$$DIST = d_{ij} = \sqrt{\frac{1}{n}\sum_k (x_{ki} + x_{kj})^2}$$

Bray-Curtis and Canberra distance

$$BRAYCURT = d_{ij} = \sum_k \left| x_{ki} - x_{kj} \right| / \sum_k (x_{ki} + x_{kj})$$

$$CANBERRA = \frac{1}{n} \sum_k \left| x_{ki} - x_{kj} \right| / \sum_k (x_{ki} + x_{kj})$$

Average Manhattan distance (city block)

$$MANHAT = M_{ij} = \frac{1}{n}\sum_{k}\left|x_{ki} - x_{kj}\right|$$

Chi-squared distance

$$CHISQ = d_{ij} = \sqrt{\sum_k \frac{\left(\frac{x_{ki}}{x_{\cdot i}} - \frac{x_{kj}}{x_{\cdot j}} \right)^2}{x_k}}$$

Cosine coefficient

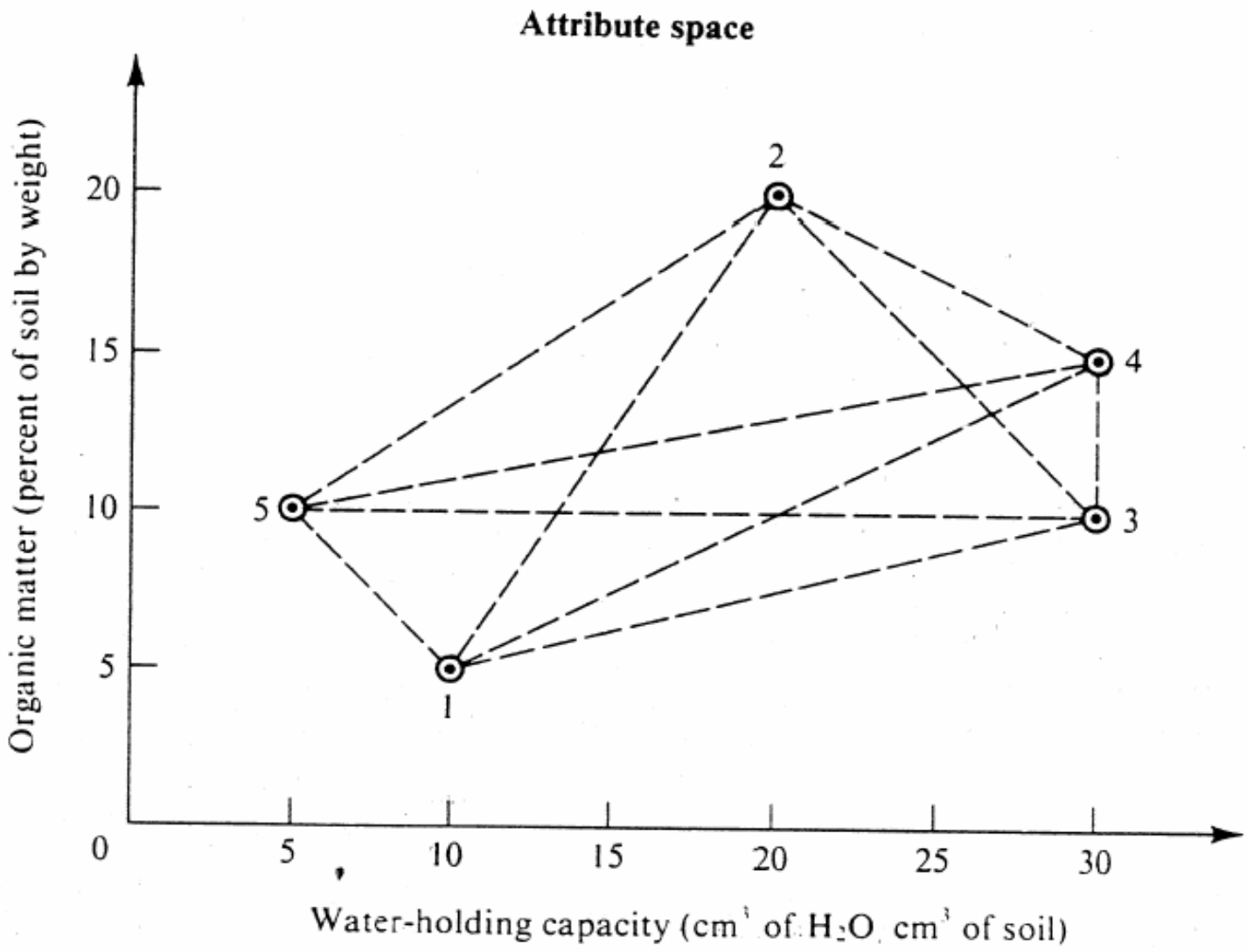
$$COSINE = c_{ij} = \sum_k x_{ki} x_{kj} \Big/ \sqrt{\sum_k x_{ki}^{\,2} \sum_k x_{kj}^{\,2}}$$

Step 1. Obtain the data matrix

Feature	Object 1	2	3	4	5
1	10	20	30	30	5
2	5	20	10	15	10

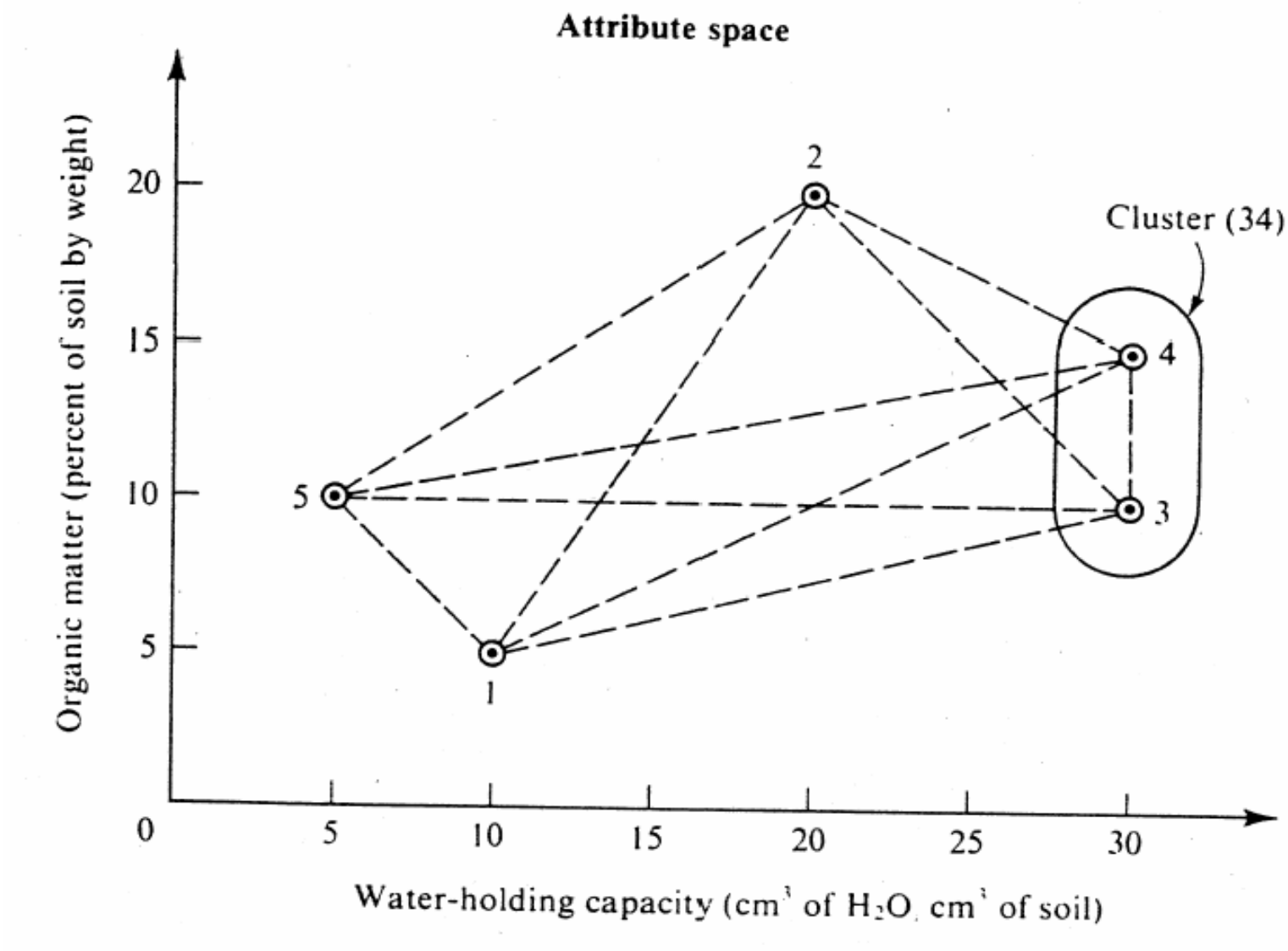
Objects and features

- The five objects are plots of farm land
- The features are
  - 1. Water-holding capacity (%)
  - 2. Weight % soil organic matter
- Objective: find the two most similar plots


Attribute space
Organic matter (percent of soil by weight)
0
5
10
15
20
1
2
3
4
5
5
10
15
20
25
30
Water-holding capacity (cm³ of H₂O, cm³ of soil)

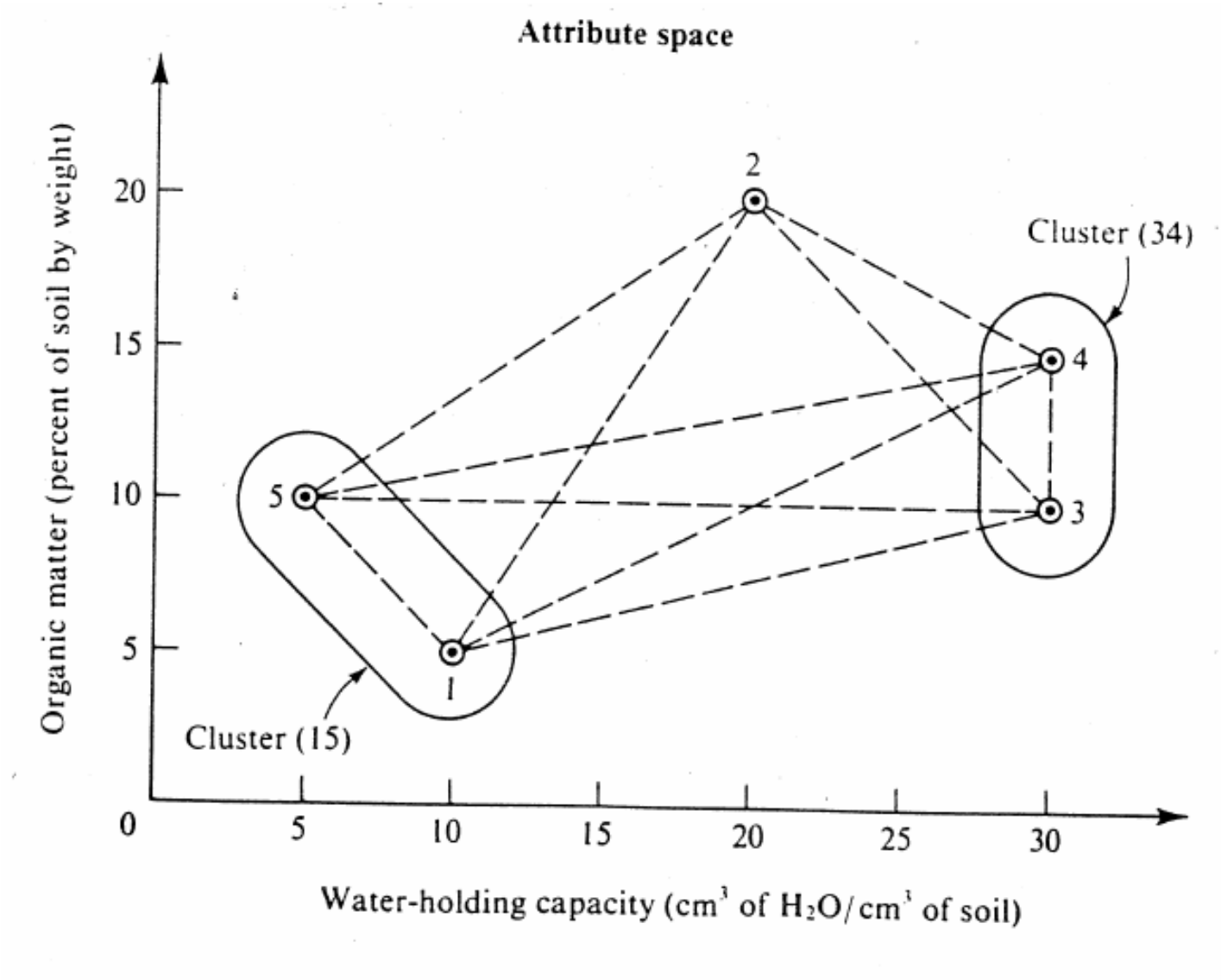
Resemblance matrix

	1	2	3	4	5
1	-	-	-	-	-
2	18.0	-	-	-	-
3	20.6	14.1	-	-	-
4	22.4	11.2	5.00	-	-
5	7.07	18.0	25.0	25.5	-


Attribute space
Organic matter (percent of soil by weight)
20
15
10
5
0
5
10
15
20
25
30
Water-holding capacity (cm³ of H₂O, cm³ of soil)
1
2
3
4
5
Cluster (34)

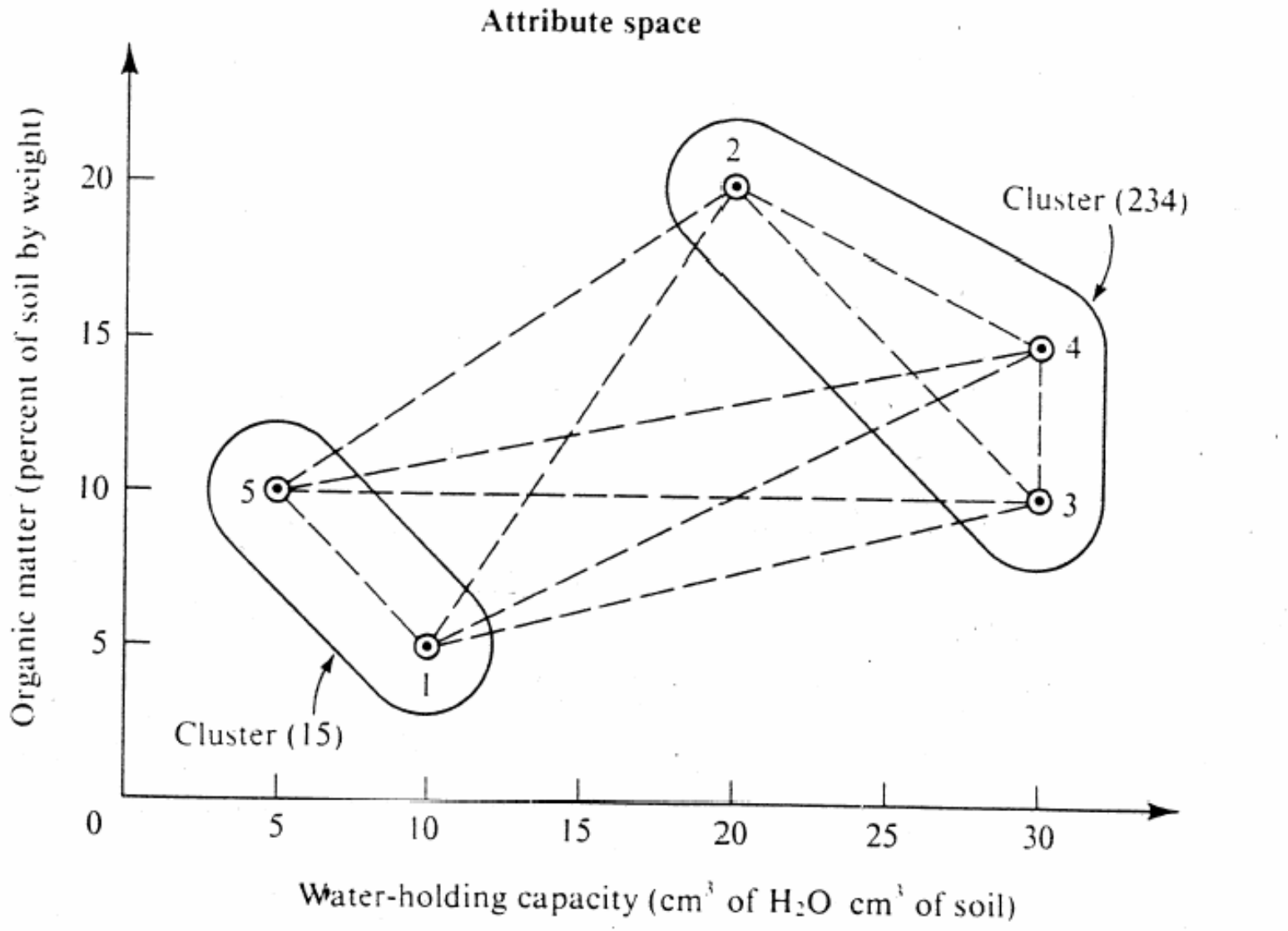
Revised resemblance matrix

	1	2	5	(34)
1	-	-	-	-
2	18.0	-	-	-
5	7.07	18.0	-	-
(34)	21.5	12.7	25.3	-


Attribute space
Organic matter (percent of soil by weight)
20
15
10
5
0
5
10
15
20
25
30
Water-holding capacity (cm³ of H₂O/cm³ of soil)
1
2
3
4
5
Cluster (15)
Cluster (34)

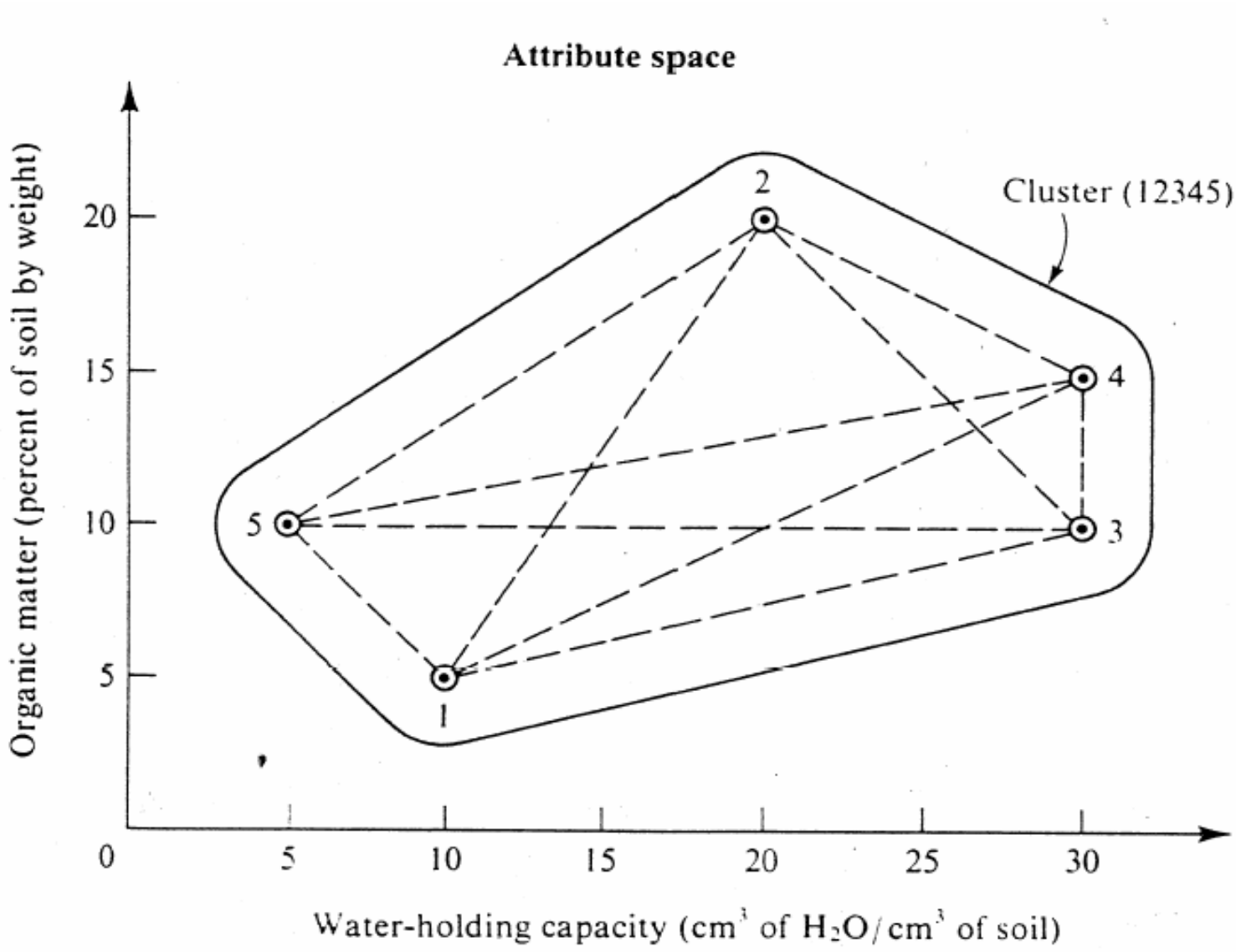
Revised resemblance matrix

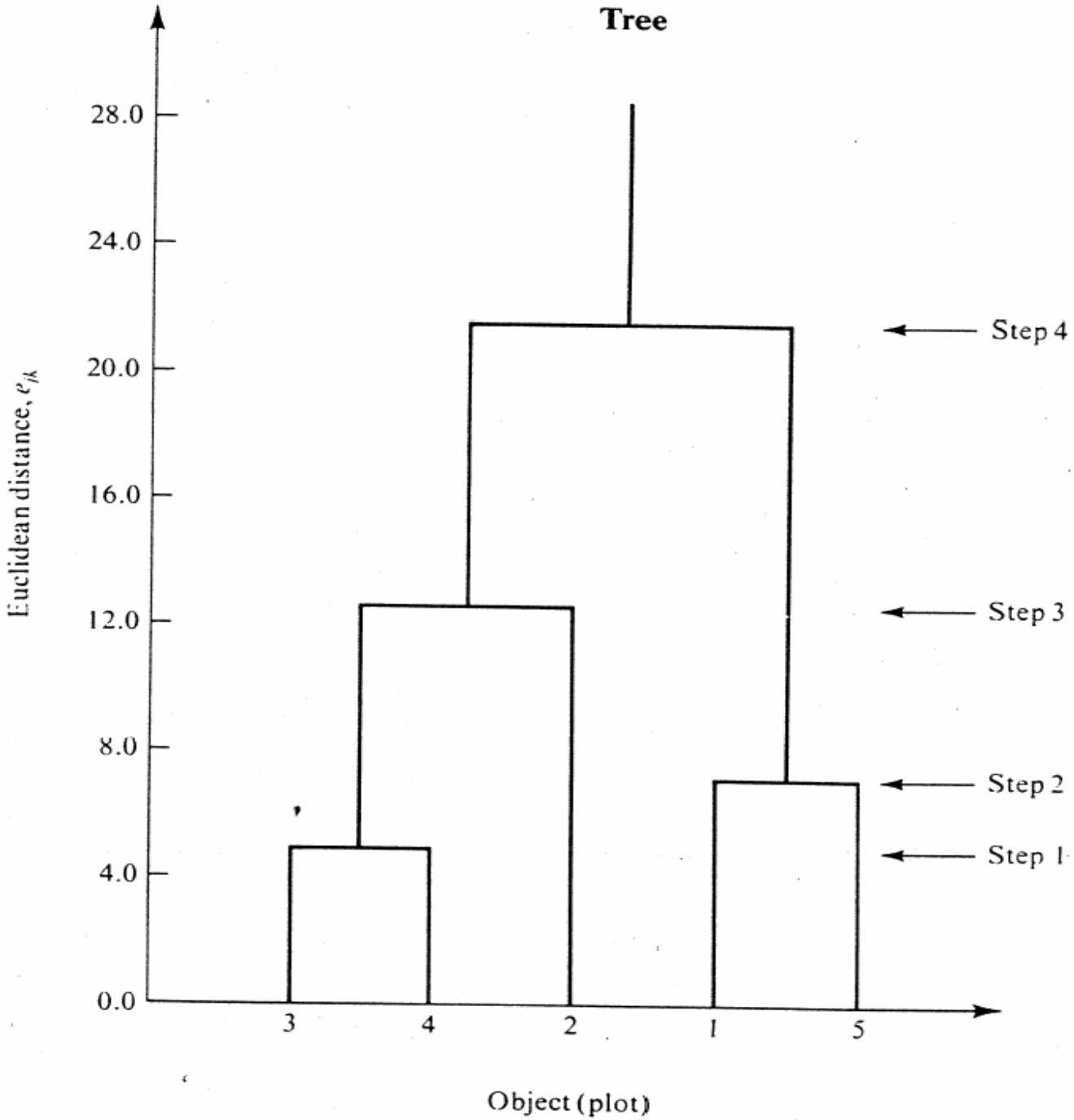
	2	(34)	(15)
2	-	-	-
(34)	12.7	-	-
(15)	18.0	23.4	-


Attribute space
Organic matter (percent of soil by weight)
20
15
10
5
0
5
10
15
20
25
30
Water-holding capacity (cm³ of H₂O cm³ of soil)
Cluster (234)
Cluster (15)
1
2
3
4
5

Rvised resemblance matrix

	(15)	(234)
(15)	-	-
(234)	21.6	-


Attribute space
Cluster (12345)
2
4
5
3
1
20
15
10
5
0
5
10
15
20
25
30
Organic matter (percent of soil by weight)
Water-holding capacity (cm³ of H₂O/cm³ of soil)


Tree
28.0
24.0
20.0
16.0
12.0
8.0
4.0
0.0
Euclidean distance, e_{jk}
Step 4
Step 3
Step 2
Step 1
3
4
2
1
5
Object (plot)

Rearranged data matrix

Attribute	Object 3	Object 4	Object 2	Object 1	Object 5
1	30	30	20	10	5
2	10	15	20	5	10

Rearranged resemblance matrix

Object	Object 3	Object 4	Object 2	Object 1	Object 5
3	—	—	—	—	—
4	5.00	—	—	—	—
2	14.1	11.2	—	—	—
1	20.6	22.4	18.0	—	—
5	25.0	25.5	18.0	7.07	—

Cophenetic correlation coefficient (Pearson product-moment correlation coefficient)

- A comparison of the similarities according to the similarity matrix and the similarities according to the dendrogram

$$r_{X,Y} = \frac{\sum xy - (1/n)(\sum x)(\sum y)}{\sqrt{(\sum x^2 - (1/n)(\sum x)^2)(\sum y^2 - (1/n)(\sum y)^2)}}$$

NTSYS

- Import matrix
- Transpose matrix if objects are rows (they are supposed to be columns in NTSYS) (transp in transformation / **general**)
- Consider log1 or autoscaling (standardization)
- Select similarity or distance measure (**similarity**)
- Produce similarity matrix

NTSYS (continued)

- Select clustering procedure (often UPGMA) (**clustering**)
- Calculate cophenetic matrix (**clustering**)
- Compare similarity matrix with cophenetic matix (made from the dendrogram) and write down the cophenetic correlation (**graphics, matrix comparison**)
- Write dendrogram (**graphics, treeplot**)