Cluster Analysis

Multivariate Analysis

Mathematical Background, Worked Example, Tool Usage

English

Wolfgang Boehme

wolfgang.boehme@economic-modeling.de

www.economic-modeling.de

8 November 2026

The documentation on this website was developed with the support of the AI Claude (Anthropic).

Table of Contents

1 Introduction and Application Areas **2**

2 Mathematical Background **2**
2.1 Distance Measures . . . 2
2.2 Hierarchical Methods . . . 2
2.2.1 Ward's Method (Variance Minimisation) . . . 2
2.2.2 Complete Linkage (Maximum Distance) . . . 3
2.2.3 Average Linkage (Mean Distance) . . . 3
2.3 k-Means Algorithm . . . 3
2.4 Choosing the Number of Clusters . . . 3
2.4.1 Elbow Method . . . 3
2.4.2 Silhouette Coefficient . . . 3

3 Worked Example: Balance-Sheet Ratios **4**
3.1 Data . . . 4
3.2 Step 1: Distance Matrix (Euclidean, Standardised) . . . 4
3.3 Step 2: Hierarchical Clustering (Ward) . . . 4
3.4 Step 3: Cluster Profiles ($k = 3$) . . . 5
3.5 Step 4: Silhouette Coefficient and Choice of k . . . 5

4 Using the Interactive Cluster Analysis Tool **6**
4.1 Data Input . . . 6
4.2 Parameter Selection . . . 6
4.3 Results Tabs . . . 7
4.4 Typical Workflow . . . 7

5 Formula Reference **8**

6 Further Reading **8**

Introduction and Application Areas

Cluster analysis is an unsupervised classification method that partitions observations into homogeneous groups (*clusters*) based on their similarity – without prior knowledge of the group structure.

Typical application areas:

- Market segmentation (customer typologies)
- Biological taxonomy (species classification)
- Document clustering (text mining)
- Anomaly detection (outlier identification)
- Image processing (colour quantisation)

Mathematical Background

Distance Measures

Euclidean distance (standard):

$$d_E(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{1}$$

Manhattan distance (more robust to outliers):

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}| \tag{2}$$

Cosine distance (for proportional patterns):

$$d_C(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\| \, \|\mathbf{x}_j\|} \tag{3}$$

Hierarchical Methods

2.2.1 Ward's Method (Variance Minimisation)

Ward's method minimises the increase in **total within-cluster scatter**:

$$d_W(C_i, C_j) = \sqrt{\frac{n_i n_j}{n_i + n_j}} \, d_E(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j) \tag{4}$$

with cluster means $\bar{\mathbf{x}}_i$, $\bar{\mathbf{x}}_j$ and cluster sizes n_i, n_j.

2.2.2 Complete Linkage (Maximum Distance)

$$d_{CL}(C_i, C_j) = \max_{a \in C_i,\, b \in C_j} d(\mathbf{x}_a, \mathbf{x}_b) \tag{5}$$

2.2.3 Average Linkage (Mean Distance)

$$d_{AL}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{a \in C_i} \sum_{b \in C_j} d(\mathbf{x}_a, \mathbf{x}_b) \tag{6}$$

Comparison of methods:

Table 1: Properties of the hierarchical methods

Method	Cluster shape	Outliers	Recommendation
Ward	Compact, similar size	Sensitive	General use
Complete	Round clusters	Moderate	With outliers
Average	Compromise	Robust	General use

k-Means Algorithm

1. **Initialisation** (k-Means++): Choose the first centroid at random; further centroids with probability proportional to $d^2(\mathbf{x}_i, \text{nearest centroid})$.
2. **Assignment**: Assign each observation to the nearest centroid: $c_i = \arg\min_k d(\mathbf{x}_i, \boldsymbol{\mu}_k)$.
3. **Update**: Recompute centroids: $\boldsymbol{\mu}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \mathbf{x}_i$.
4. Repeat steps 2–3 until convergence.

Objective function (WCSS – Within-Cluster Sum of Squares):

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \tag{7}$$

Choosing the Number of Clusters

2.4.1 Elbow Method

Plot WCSS against k; the kink (elbow) in the graph marks the optimal k.

2.4.2 Silhouette Coefficient

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \quad s_i \in [-1, 1] \tag{8}$$

where a_i = mean distance to observations in the same cluster, b_i = minimum mean distance to the nearest foreign cluster.

Table 2: Interpretation of the mean silhouette coefficient

$\bar{s}$	Interpretation
≥ 0.70	Strong cluster structure
$0.50 - 0.69$	Reasonable structure
$0.25 - 0.49$	Weak structure
< 0.25	No structure

Worked Example: Balance-Sheet Ratios

Data

The same 24 synthetic companies U01–U24 with six balance-sheet ratios as in Handbook 1 are used (full data table there): X_1 equity ratio, X_2 quick ratio, X_3 asset coverage II, X_4 return on sales, X_5 return on assets, X_6 sales growth. Tool settings: Ward, Euclidean distance, z-transformation, 3 clusters (defaults).

Goal of the cluster analysis: group the companies into homogeneous types based on the ratios alone – *without* knowing the rating. Only afterwards is it checked how well the clusters found agree with the ratings A/B/C.

Step 1: Distance Matrix (Euclidean, Standardised)

The ratios are z-standardised (means and standard deviations see Handbook 1); the Euclidean distance matrix $\mathbf{D}$ has dimension 24×24. Some values:

$$d(U11, U14) = 0.726, \quad d(U01, U02) = 0.979, \quad d(U01, U12) = 3.007,$$
$$d(U01, U19) = 3.848, \quad d(U12, U19) = 5.514$$

Companies of the same type lie close together; the highly profitable B company U12 and the C company at risk U19 are furthest apart.

Step 2: Hierarchical Clustering (Ward)

Table 3: First and last fusion steps (Ward)

Step	Cluster A	Cluster B	d_W
1	U11	U14	0.513
2	U01	U02	0.692
3	U18	U19	0.717
...	...	...	...
20	{U17, U20}	{U03, U18, U19, U21–U24}	2.468
21	{U13, U16}	{U09–U12, U14, U15}	2.523
22	{U01, U02, U04–U08}	{U09–U16}	5.972
23	{U03, U17–U24}	{U01, U02, U04–U16}	7.652

Example, step 1: $d_W(U11, U14) = \sqrt{\frac{1 \cdot 1}{2}}\, d_E(U11, U14) = 0.7071 \cdot 0.726 = 0.513$.

Up to step 21 the fusion distances rise only slowly (at most 2.52). Then there is a large jump to 5.97 and 7.65: the data split **naturally into 3 clusters**. The cut at $k = 3$ (default) lies between steps 21 and 22 and separates exactly these three groups.

Step 3: Cluster Profiles ($k = 3$)

Table 4: Cluster means (original data, in %)

Cluster	X_1 Equity	X_2 Liqu.	X_3 ACII	X_4 ROS	X_5 ROA	X_6 Growth
1 ($n = 7$)	47.01	130.29	128.71	5.33	7.59	5.30
2 ($n = 9$)	28.44	87.11	90.11	2.20	4.64	0.18
3 ($n = 8$)	34.12	99.38	96.75	8.55	11.16	11.71
Total ($n = 24$)	35.75	103.79	103.58	5.23	7.68	5.52

Interpretation:

- Cluster 1 (U01, U02, U04–U08) – **solid asset-rich companies**: clearly above-average stability, average profitability.
- Cluster 2 (U03, U17–U24) – **companies at risk**: below average on all six ratios, stagnating sales.
- Cluster 3 (U09–U16) – **highly profitable growth companies**: highest returns and growth rates with slightly below-average stability.

Table 5: Comparison of clusters and ratings

	Rating A	Rating B	Rating C
Cluster 1	**7**	0	0
Cluster 2	1	0	**8**
Cluster 3	0	**8**	0

23 of 24 companies end up in the cluster of "their" rating, although the method does not know the rating. The exception is **U03** (rating A): with an equity ratio of 32.0 %, asset coverage of 100 % and a return on assets of only 5.0 %, it is below average on almost all ratios. Its distance to the centre of cluster 2 (1.76 without U03) is smaller than to the centre of cluster 1 (2.45). In practice this means: either the A rating is based on information not contained in the six ratios, or it is a warning signal of a possible downgrade.

Step 4: Silhouette Coefficient and Choice of k

$\bar{s} = 0.441$ for $k = 3$ – according to Table 2 a **weak structure** (rated "acceptable" in the tool, thresholds 0.5/0.25). The best individual values belong to U02 ($s_i = 0.593$) and U22 (0.588); the lowest to U07 (0.195) and U03 (0.212) – the same borderline cases that stand out in the PCA (Handbook 1) and in discriminant analysis (Handbook 4).

Table 6: Silhouettes and WCSS for different k

	$k = 2$	$k = 3$	$k = 4$
$\bar{s}$ (Ward)	0.366	**0.441**	0.381
WCSS (k-Means)	79.45	43.79	37.31

$k = 3$ yields the highest silhouette coefficient, and the k-Means elbow plot bends clearly at $k = 3$ (WCSS falls from 79.45 to 43.79, then only to 37.31). k-Means with $k = 3$ yields the same partition as Ward. All criteria thus consistently point to three company types.

Using the Interactive Cluster Analysis Tool

Data Input

In addition to the standard tabs (Manual, CSV, Examples), the cluster analysis tool provides a **label column** ("Label") on the left of the manual input table. Meaningful names for observations can be entered here; they appear in the dendrogram and scatter plot.

Pre-loaded: Balance-sheet ratios (24 companies U01–U24, 6 variables). Further examples: Iris (15 obs., 4 variables), City profile (12 cities, 4 variables) and Customer segmentation (20 customers, 4 variables). For each example, an explanatory box above the input table describes the data, the goal of the analysis and the expected result; it disappears when "Clear" is clicked or a CSV file is loaded.

Parameter Selection

Method
Hierarchical: Ward (default), Complete Linkage, Average Linkage. Non-hierarchical: k-Means. When k-Means is selected, the Dendrogram tab is hidden.

Distance measure
Euclidean (default), Manhattan, Cosine similarity. Ward uses the selected distance between the cluster centres; it is theoretically justified only for the Euclidean distance. k-Means always works with Euclidean distance (selector locked). The silhouette coefficient uses the selected distance measure.

Standardisation
Z-transformation strongly recommended when variables are on different scales.

Number of clusters (hierarchical)
Determines the cut in the dendrogram. The golden dashed line lies between the two fusion levels at which the tree is cut; branches below the cut are green, above it red.

k and $k_{\max}$ (k-Means)
k: desired number of clusters. $k_{\max}$: upper bound for the elbow plot and silhouette analysis (up to 12).

Results Tabs

Overview
Key figures (n, p, clusters, silhouette) + complete assignment table.

Dendrogram
Dendrogram with cut line; observations are arranged in tree order and coloured by assigned cluster. Available for hierarchical methods only.

Charts
k-Means: elbow plot (WCSS) and silhouette bars per k. Both methods: scatter plot of the first two variables (standardised), observations coloured by cluster.

Cluster Profile
Mean table per cluster (original data) + radar chart with standardised cluster profiles.

Interpretation
Automatic assessment: silhouette coefficient, dendrogram notes, k-Means elbow interpretation.

Export
CSV: cluster assignment, cluster profile. PNG: scatter plot, radar chart, dendrogram.

Typical Workflow

1. Load data; check standardisation setting.
2. Select Ward method and Euclidean distance.
3. Set number of clusters to 3 (or adjust after inspecting the dendrogram).
4. Click **"Calculate Cluster Analysis"**.
5. Dendrogram tab: large fusion distances indicate natural boundaries.
6. Check the silhouette value ($\bar{s} \geq 0.5$ = good).
7. If needed, confirm with k-Means and the elbow plot.
8. Cluster Profile tab: interpret the groups substantively.

Important notes:
- At least $n = 4$ complete observations are required.
- k-Means: 10 restarts (k-Means++) prevent local minima.
- Ward assumes spherical clusters of similar size.
- For $p > 2$, the scatter plot displays only the first two variables.
- Dendrogram: download as PNG for publications.

Formula Reference

Table 7: Cluster analysis formulae

Name	**Formula**
Euclidean distance	$d_E = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$
Manhattan distance	$d_M = \sum_k \lvert x_{ik} - x_{jk} \rvert$
Ward fusion distance	$d_W = \sqrt{\frac{n_i n_j}{n_i + n_j}}\, d_E(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j)$
Cosine distance	$d_C = 1 - \mathbf{x}_i^\top \mathbf{x}_j / (\lVert \mathbf{x}_i \rVert \, \lVert \mathbf{x}_j \rVert)$
WCSS (k-Means)	$\sum_k \sum_{i \in C_k} \lVert \mathbf{x}_i - \boldsymbol{\mu}_k \rVert^2$
Silhouette	$s_i = (b_i - a_i) / \max(a_i, b_i)$

Further Reading

- Kaufman, L. & Rousseeuw, P. J. (2005). *Finding Groups in Data*. Wiley.
- Backhaus, K. et al. (2021). *Multivariate Analysemethoden*, 16th ed. Springer Gabler.
- Arthur, D. & Vassilvitskii, S. (2007). k-Means++: The Advantages of Careful Seeding. *SODA 2007*, 1027–1035.