HANDBOOK OF COMPUTATIONAL SOCIAL SCIENCE, VOLUME 2

Data Science, Statistical Modelling, and Machine Learning Methods

Edited by Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu and Lars Lyberg

First published 2022

ISBN: 978-0-367-45780-8 (hbk)
ISBN: 978-1-032-07770-3 (pbk)
ISBN: 978-1-003-02524-5 (ebk)

19

UNSUPERVISED METHODS

Clustering methods

Johann Bacher, Andreas Pöge and Knut Wenzig

DOI: 10.4324/9781003025245-23

The funder for this chapter is Johannes Kepler University Linz, Institute of Sociology

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

19
UNSUPERVISED METHODS
Clustering methods

Johann Bacher, Andreas Pöge and Knut Wenzig

1 Introduction

The basic aim of all clustering methods is to assign objects to groups (clusters) according to similarities in their specific characteristics. Two objects assigned to the same cluster should share similar specified characteristics (variables, patterns, symbols, etc.), whereas two objects allocated to different clusters should be less similar. Objects might be cases of either a data matrix or variables. For example, countries (cases) might be classified in clusters according to their values in selected variables. Alternatively, variables might be clustered into groups, so that cluster 1 contains variable *X1*, *X2*, and *X4*, cluster 2 variables *X3*, *X5*, etc. In most applications, cases are clustered. Therefore, these two terms (objects, cases) will be used here synonymously.

The development of clustering methods has varied in intensity and innovation since the 1960s when they first became popular. For example, Ward proposed his well-known minimum variance method in 1963. The 1970s saw a flurry of textbooks on the subject (e.g., Everitt, 1974; Hartigan, 1974; Jardine & Sibson, 1971), which tended to focus on algorithms to generate the clusters and proposed some formal criteria to decide the number of clusters. However, several problems remained unsolved at the end of this first period of intensive development (Everitt, 1979). They included the selection of appropriate variables and appropriate clustering methods, the determination of the number of clusters, and the evaluation of the clustering results. A further practical problem was the limited computer capacity at the time. In the 1980s, with the increase in computer capacity, cluster analysis techniques were included in standard statistical packages.

In the 1990s, inroads were made into addressing these early problems. This period was marked by the development of so-called model-based and probabilistic clustering techniques (e.g., Fraley & Raftery, 1999; Vermunt & Magidson, 2000) on the one hand and density clustering methods (Ester et al., 1996) on the other.

Today, in the early 21st century, with huge advancements in computer capacity and capability, elaborate and computationally intensive methods have become the norm (Wierzchoń & Kłopotek, 2018; Zgurovsky & Zaychenko, 2020) and the literature has exploded (Murtagh, 2016). Clustering methods are available in most statistical software packages, as well as in machine-learning software and data-mining packages such as RapidMiner. The most comprehensive collection of clustering methods is available in the software package R (Leisch & Gruen, 2020).

DOI: 10.4324/9781003025245-23

This article provides an overview of clustering methods and covers the following topics:

- Steps toward an appropriate cluster solution
- Clustering methods
- Criteria to determine the number of clusters
- Methods to validate cluster solutions
- Computer programs
- Application
- Summary and recommendations

An in-depth insight into the discussed topics is provided by the excellent handbook by Hennig et al. (2016) and the reader by Wierzchoń and Kłopotek (2018).

2 Steps toward an appropriate cluster solution

In order to arrive at an appropriate cluster solution, the following steps are necessary:

1. *Selection of appropriate variables, cases, and clustering method*. The selection of appropriate variables and cases is a substantive decision that depends on the research question. Sometimes researchers can collect variables and cases by themselves, but in many applications, the data already exist and the researchers merely have to select the variables. From a formal perspective, the variables should be able to differentiate between the clusters. However, whether this is the case can only be judged a posteriori after completing the next steps. The selection of an appropriate clustering method depends on the selected data (size, measurement level of variables) and the researcher's assumption of what the cluster should look like.
2. *Running the cluster analysis*. Sometimes the selected method is not available in the researcher's statistical package, making it necessary for him/her to familiarize him-/herself with a new computer program.
3. *Selection of one or more appropriate cluster solutions*. Sometimes, only one cluster solution comes into consideration for subsequent steps, but very often, there is more than one appropriate cluster solution for further consideration.
4. *Validation of cluster solution(s)*. The selected cluster solutions are validated using external and internal techniques. If more than one cluster solution remains after step 3, this step should help to make a final decision.
5. *Final decision for a specific cluster solution*. If one cluster solution meets formal and substantive criteria, this cluster solution is selected and the resulting classification can be used. If this is not the case, the researcher can return to step 1 and opt to select additional variables or to exclude variables, and/or to choose a different clustering method.

The application of these steps will be demonstrated in section 8.

3 Clustering methods

There are different ways to classify clustering methods (e.g., Saxena et al., 2017). One prominent distinction (Saxena et al., 2017) is between *hierarchical* and *partitioning techniques* (see Figure 19.1). *Hierarchical methods* can be further divided into divisive and agglomerative methods. *Divisive hierarchical methods* start with the assumption that all objects belong to one large cluster and divide the clusters stepwise until each object builds a distinct cluster. In contrast,

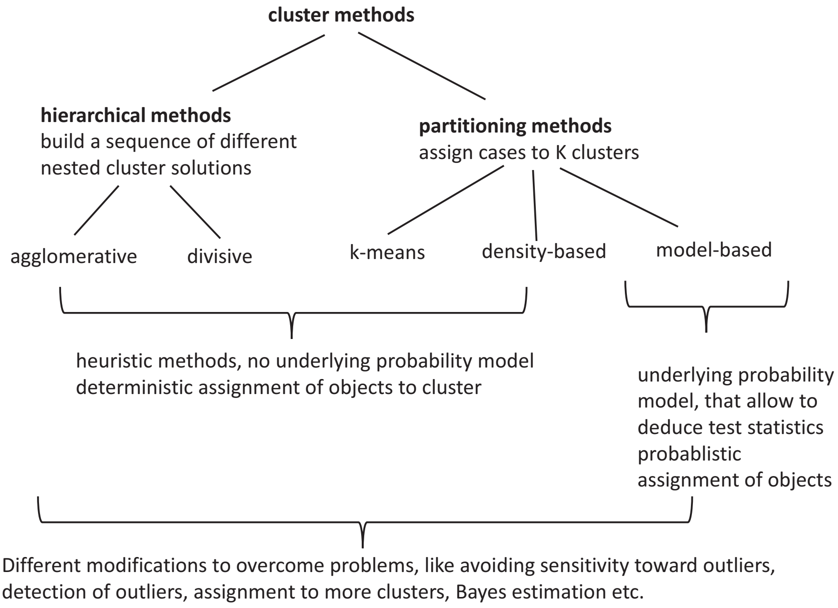
cluster methods
hierarchical methods
build a sequence of different nested cluster solutions
partitioning methods
assign cases to K clusters
agglomerative
divisive
k-means
density-based
model-based
heuristic methods, no underlying probability model deterministic assignment of objects to cluster
underlying probability model, that allow to deduce test statistics probablistic assignment of objects
Different modifications to overcome problems, like avoiding sensitivity toward outliers, detection of outliers, assignment to more clusters, Bayes estimation etc.

Figure 19.1 Overview of clustering methods

agglomerative hierarchical methods assign each case to a distinct cluster at the beginning and then combine the clusters stepwise until all cases belong to one large cluster. *Partitioning techniques* start with a given number of clusters *K* and assign the cases iteratively to these *K* clusters by minimizing or maximizing a certain criterion. The most popular partitioning technique is *k-means clustering*.

In order to demonstrate the logic of *hierarchical methods*, it is important to compute a similarity or dissimilarity matrix for the cases. A large number of similarity and dissimilarity measures exists. They are well documented in most textbooks (e.g., Bacher et al., 2010; Everitt et al., 2001).

For *quantitative variables*, product and distance measures can be distinguished. The most prominent product measure is *Pearson's correlation* coefficient:

$$s_{ij} = \frac{\sum_{l=1}^{m}\left(x_{il} - \bar{x}_i\right)\left(x_{jl} - \bar{x}_j\right)}{\left(\sum_{l=1}^{m}\left(x_{il} - \bar{x}_i\right)^2 \sum_{l=1}^{m}\left(x_{jl} - \bar{x}_j\right)^2\right)^{1/2}}$$

Pearson's correlation coefficient is a similarity measure. A higher value indicates a greater similarity between two objects.

Prominent examples of distance measures are the Euclidean distance and the city-block distance, which can be derived from the general *Minkowski metric* for two cases *i* and *j*:

$$d_{ij} = \left(\sum_{l=1}^{m}\left|x_{il} - x_{jl}\right|^p\right)^{1/p}$$

For $p=1$ the Minkowski metric results in the city-block distance, sometimes denoted as L1 metric, for $p=2$ the Euclidean distance, also denoted as L2. Distance measures are dissimilarity measures.

If the variables have another nonquantitative measurement level, distance and product measures are available, too. Especially for dichotomous variables, numerous measures have been developed that differ as to how the presence and absence of an attribute is evaluated. Already in the 1970s, Gower (1971) proposed a similarity measure for variables with mixed measurement levels. If cases are clustered, the measures are computed for each pair of cases in contrast to the usual analysis whereby the correlation coefficient is computed for pairs of variables.

An example of a dissimilarity matrix is given in Table 19.1. A higher value indicates a larger dissimilarity. In the example, objects 5 and 6 with a value of $d(5,6)=1$ have the smallest dissimilarity; they are the most similar of the six objects. The largest dissimilarity occurs for objects 1 and 6 ($d(1,6)=44$). These objects are the least similar of the six objects.

Agglomerative methods start with the assumption that each object/case builds a cluster. For *n* cases, there are *n* clusters. The algorithms search the pair of clusters with the smallest dissimilarity and agglomerate them into one cluster. The number of clusters reduces to *n*−1 and a new dissimilarity matrix is computed; therefore, the dissimilarity between two clusters has to be defined (see later). Afterward, the aforementioned steps are repeated until all cases build one large cluster. The process is usually reported in an agglomeration schedule (see Table 19.2) and graphically visualized in a dendrogram (Figure 19.2). *Divisive methods* follow the opposite principle and start with one large cluster that contains all cases. This large cluster is split stepwise into subclusters until each case builds a cluster.

Table 19.2 reports the *agglomeration schedule of single linkage* for the dissimilarity matrix of Table 19.1. At the beginning, each object builds one cluster. *C1*={1}, *C2*={2}, . . ., *C6*={6}. In

Table 19.1 Dissimilarity matrix for six objects

	1	*2*	*3*	*4*	*5*	*6*
1	0					
2	10	0				
3	30	20	0			
4	38	28	8	0		
5	43	33	13	5	0	
6	44	34	14	6	1	0

Note: Table was generated by the authors.

Table 19.2 Agglomeration schedule for dissimilarity matrix in Table 19.1

Stage	*Cluster combined*		*Coefficients (agglomeration level v_k)*
	Cluster 1	*Cluster 2*	
1	5	6	1.000
2	4	5	5.000
3	3	4	8.000
4	1	2	10.000
5	1	3	20.000

Note: The cluster analysis was performed with IBM SPSS (version 26), module CLUSTER.

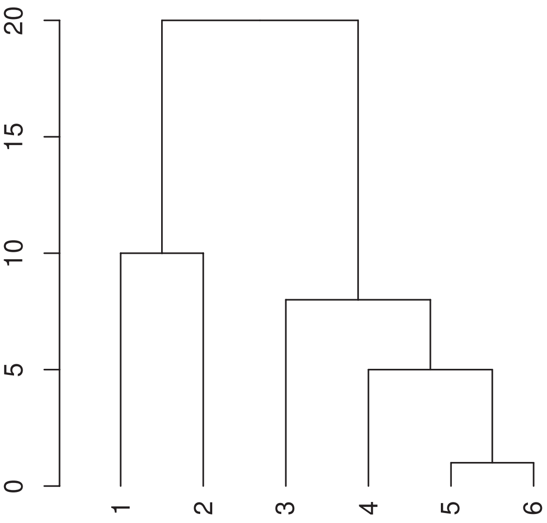

Figure 19.2 Dendrogram for the agglomeration schedule of Table 19.2 (result for single linkage)

the first step, the nearest cases 5 and 6 are combined at a level of 1.0. Now we have five clusters: *C1*={1}, *C2*={2}, . . ., *C5*={5, 6}. In the next step, the cluster with case 4 and the cluster with case 5 (and 6, not reported) are combined at a level of 5.0, and so on. The *dendrogram* (Figure 19.2) reports the process graphically. Objects 5 and 6 are combined at a low level; the next object, 4, is assigned to this cluster, afterward object 3. In the next step a new cluster is formed by objects 1 and 2, and finally the two clusters *C1*={3,4,5,6} and *C2*={1,2} are combined into one cluster. The length of the line where two clusters merge represents the dissimilarity that occurs after the two clusters are combined. It corresponds to the agglomeration level in the agglomeration schedule.

The agglomeration methods differ in how they compute the new dissimilarities after two clusters are combined. Three main approaches exist (Everitt et al., 2001, pp. 55–67):

- *Single linkage*. In a specific step, the new dissimilarities between a cluster *k* and the new cluster (*i, j*), which agglomerates clusters *i* and *j*, is computed as: $d_{k(ij)} = \min(d_{ki}, d_{kj})$. This procedure results in clustering whereby each object of a cluster has at least one nearest neighbor within the cluster with a dissimilarity less than/equal to the reported agglomeration level v_k at a certain step. Due to this property, single linkage is also referred to as the nearest neighbor method. It is able to produce chains and to find outliers.
- *Complete linkage*. The new dissimilarities between a cluster *k* and the new cluster (*i, j*) are computed as: $d_{k(ij)} = \max(d_{ki}, d_{kj})$. This procedure results in a clustering where the dissimilarities between all objects of a cluster are less than/equal to the reported agglomeration level v_k at a certain step. Due to this property, complete linkage is known as the furthest neighbor method, because the furthest object in a cluster is a neighbor. Complete linkage results in very homogenous clusters. The structure of the cluster is unimportant.
- *(Weighted) average linkage*. The new dissimilarities between a cluster *k* and the new cluster (*i, j*) are computed as a weighted average. Different formulas are used.

The aforementioned Ward's method can be seen as a special agglomerative method. It requires quantitative variables and uses squared Euclidean distance. At a certain step, those two clusters are combined that minimize the sum of squares within clusters.

K-means methods do not require a similarity or dissimilarity matrix to be computed. Rather, the number of clusters and a starting configuration must be specified at the outset. The starting configuration can be generated randomly or empirically with another cluster or statistical method. It is also possible to use results or theoretical considerations. The results may depend on the starting values and the ordering of the cases.

K-means clustering assigns the cases to K clusters so that the within-cluster variation $SSE\left(K\right)$ ("sum of squares of error") is minimized:

$$SSE(K) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} \sum_{l=1}^{p}\left(x_{il} - \overline{x}_{kl}\right)^2 \rightarrow \min$$

where

w_{ik} = membership function ($w_{ik} = 1$ if case *i* belongs to cluster *k*, else 0)

x_{ij} = value of case *i* in variable *l*

$\overline{x}_{kj}$ = mean of cluster *k* in variable *l*

Table 19.3 reports the result of k-means clustering for a data set with 25 cases. In the example, a cluster solution with three clusters was computed. The means of cluster 1 and cluster 2 in variable *x1* are similar (1.97 and 2.01). Hence, the two clusters do not differ with respect to *x1*. However, they do differ in *x2*. The mean of cluster 1 is 2.62, whereas cluster 2 has a mean of 1.27. In contrast, cluster 3 differs from cluster 1 and 2 in *x1*, and from cluster 1 in *x2*, too.

Density-based clustering (Ester et al., 1996; Schubert et al., 2017). K-means clustering tends to build spherical clusters (Steinley, 2016) and is sensitive to outliers (Kaufinan & Rousseeuw, 1990, p. 117) like every procedure that works with the sum of squares. Density-based clustering overcomes this problem. It can detect clusters of different shapes. It assumes that areas of higher and lower density exist in the data space and requires the definition of two parameters: the radius ε and the number of points $NPts$ that should occur within the radius of objects that build a region with high density. Figure 19.3 reports the results of density-based clustering.

Hierarchical methods, k-means, and density-based clustering methods are all *heuristic methods*. They use no underlying statistical model, like a normal distribution, and hence they are unable to deduce model-based measures to select a specific cluster solution and to evaluate this solution. They usually require decisions by the user that are ambiguous. For example, DBSCAN requires the definition of the minimal number of cases that should belong to a cluster and is very sensitive to this specification. If we increase the number of points *NPts* in the previous example, more objects are labeled as outliers even if they seem to be close to a cluster.

Table 19.3 Results of k-means clustering

Cluster centers				*Test statistics*				
				Cluster K	*SSE*	*ETA²*	*PRE*	*FMAX*
	Cluster			1	18.19	0.0%		
	1	2	3	2	8.86	51.3%	51.3%	24.24
N	6	9	10	3	4.10	77.5%	53.7%	37.81
x1	1.97	2.01	1.00	4	2.67	85.3%	34.8%	40.65
x2	2.62	1.27	1.29	5	1.69	90.7%	36.8%	48.85
				6	1.39	92.4%	17.7%	45,93

Note: k-means clustering was performed with IBM SPSS (version 26), module QUICK CLUSTER, option UPDATE. The test statistics are computed via additional syntax. Data are generated by the authors.

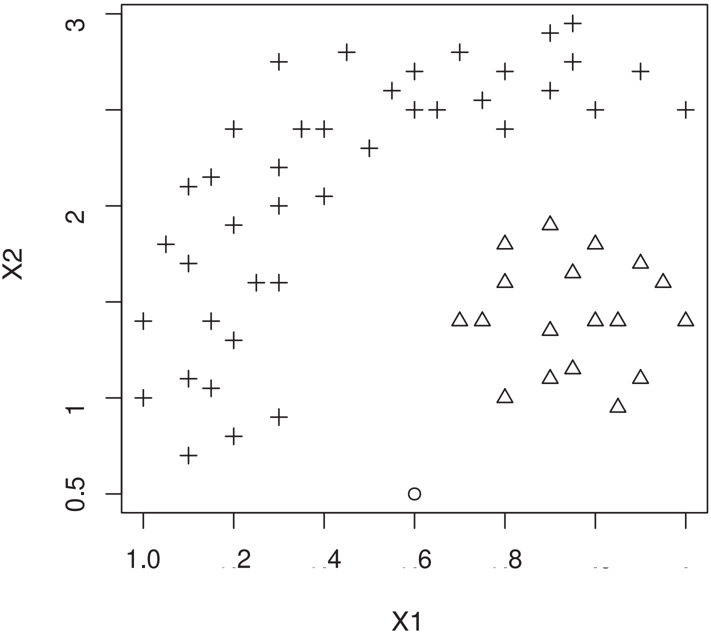

Figure 19.3 Graphical results of DBSCAN

Note. The solution was generated with R-function HDBSCAN from the package DBSCAN. The number of points *NPts* was set to 5. The cases of cluster 1 are drawn with crosses. The cases of cluster 2 are drawn with triangles. One case (circle) is detected as an outlier. Even without the outlier, the K-means solution with 2 clusters would not differentiate between the two geographic shapes.

Model-based clustering methods use an underlying statistical model. They assume that a mixture of probability distributions underlies the analyzed data. In the case of quantitative data, a mixture of normal distributions is usually assumed. In model-based clustering methods the general model is:

$$f(X/\theta_K = \{\theta_1, \theta_2, ..., \theta_K,\}, K) = \prod_{k=1}^{K} f(X/\theta_k)\, p_k$$

where

$f\left(X/\theta_k\right)$ = joint distributions of the variables X in cluster k. The distribution depends on the parameters θ_K

p_k = proportion of cluster k

If all variables are dichotomous, the *latent class model* evolves. If all variables are quantitative, we arrive at the *latent profile model*. Both models were proposed by Lazarsfeld and Henry (1968) in the 1960s. The approach includes *finite mixture models* (McLachlan & Peel, 2005), which assumes in its classical approach that the observed distribution is a mixture of K p-dimensional normal distributions with a mean vector $\boldsymbol{\mu_k}$ and variance-covariance matrix $\boldsymbol{\Sigma_k}$ (Everitt et al., 2001, pp. 120–122). Nowadays, adequate software is available, like LatentGold (see section 7), and it is possible to analyze variables of a mixed measurement type.

The primary objectives in the model-based clustering method are to estimate the parameters θ_K of the conditional distributions and the number K of the clusters. Due to the assumption of probability distributions, statistically based measures are available. Information measures are most frequently used for this purpose (see section 5).

4 Modifications and recent developments

The methods described previously can be regarded as prototypes that have been modified in several ways, especially in recent years. Some of these modifications are now described:

- *Enhanced hierarchical methods for large data sets*. Hierarchical methods require the computation and storage of a dissimilarity matrix and therefore need considerable computer memory. Consequently, enhanced hierarchical methods have been developed to handle large data sets, like BIRCH or CURE (Saxena et al., 2017).
- *Using other measures of central tendency instead of the mean in k-means*. Clustering methods exist that use medians (k-median clustering, Bradley et al., 1996), modes (K-modes clustering, Huang, 1998), or representative data points (k-medoids clustering, Kaufman & Rousseeuw, 1990, pp. 68–123) instead of means.
- *Using other distance measures in k-means*. K-means clustering uses squared Euclidean distances and consequently the centers are sensitive to outliers. Therefore, the use of the city-block distance instead of the squared Euclidean distance was proposed. For instance, k-medoids clustering minimizes: $F\left(K\right)=\sum_{i=1}^{n}\sum_{k=1}^{K}w_{ik}\sum_{l=1}^{p}\left|x_{il}-m_{kl}\right|$, where m_{kl} is a representative data point.
- *Detection and elimination of outliers*. Another way to avoid dependence on outliers is to detect and eliminate them (Hautamäki et al., 2005). The iterative k-means procedure (Holmgren et al., 2020) is one example.
- *Assignment of cases to more than one cluster*. Some cases are difficult to assign to a single cluster. They may be located between two clusters. Hence, it may be reasonable to assign them to two or even more clusters. Fuzzy clustering methods (Jain & Dubes, 1988, pp. 130–133) are one approach that allows multiple assignment.
- *Bayes estimation methods*. In accordance with the general trend toward Bayesian statistics, Bayes estimation methods have been introduced for clustering techniques. One cluster program that has fully implemented a Bayes approach is AutoClass (Bacher et al., 2010, pp. 439–446).

5 Criteria to determine the number of clusters

The *determination of the number of clusters* is critical in clustering methods. Generally, the user must decide the number of clusters. Most implementations of cluster methods in software packages provide formal criteria. Some implementations have automatized the decision and propose a certain cluster solution. However, these suggestions depend on the specified parameters and if these parameters are changed, the number of cluster changes, too. Therefore, even in these cases, the proposed cluster solution should be validated (see the next section). It might be the case that a different solution is more appropriate than that proposed. Further criteria, which are not used in the automatic proposal, are relevant. These further criteria might be interpretability, comprehensibility (small number of clusters is preferred), and minimal cluster size (each cluster should have at least a certain number of cases).

The available formal criteria depend on the clustering methods. For *hierarchical clustering methods* the number of clusters is commonly fixed graphically. The user sets the number of clusters equal to the number of "hills" in the dendrogram. In Figure 19.2, two hills can be seen. Another graphical method is the inverse scree test (Lathrop & Williams, 1989), sometimes called the elbow test. Similar to explorative factor analysis, a scatter plot is generated. The number of clusters defines the x-axis, the value of the agglomeration schedule the y-axis (see Figure 19.4).

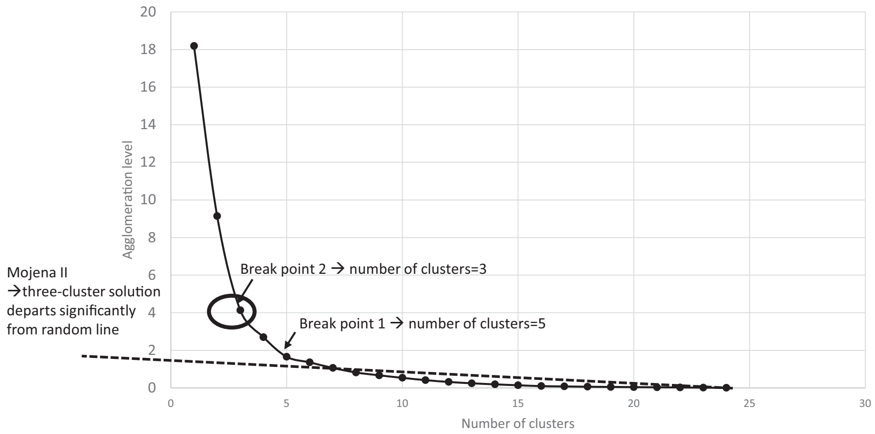

Figure 19.4 Scree diagram generated for agglomerative hierarchical clustering method

The diagram is read from right to left, starting with the highest number of clusters. Moving left, one stops when the first break point ("elbow") is observed. The number of clusters is set equal to the number of clusters where the first break point occurs. In Figure 19.4, the first break point occurs for five clusters. For a large data set, this procedure can result in a high number of clusters. In this case, one can continue and look for the next break point. In our example, five clusters might be acceptable and the number is not too high. The example indicates a further break point at three clusters. Hence, we can conclude that two cluster solutions are appropriate. Mojena (1977) formalized the decision based on the scree diagram. One of his criteria is to estimate a regression line until K clusters, to predict the value for $K-1$ clusters based on the results of this regression line and to test whether the empirical value significantly departs form the predicted value. Mojena proposes a threshold of 2.75 for the standardized residuals for K clusters from a regression line from 1 to $K-1$ clusters. In Figure 19.4, the significant departure occurs for three clusters.

For *k-means clustering*, it is obvious to use the sum of squares of error and some derived measures to determine the number of clusters. For this purpose it is necessary to generate a series of k-means solutions. We recommend starting with one cluster as the lowest value and set the highest value as one that would not be expected from a substantive perspective, e.g., 10 to 20 clusters. The decision for a cluster solution might now be based on:

- Explained variance ($ETA_K^2 = 1 - \frac{SSE_K}{SSE_1}$) : the user defines a threshold in advance, starts with K=1 cluster and selects the first solution with K clusters that meets this threshold.
- PRE statistic ($PRE_K = 1 - \frac{SSE_K}{SSE_{K-1}}$) : the user selects the solution with K clusters if PRE_K is large and the following PRE for K+1, K+2 are small.
- FMAX statistic ($FMAX_K = \frac{\left(SSE_1 - SSE_K\right) / \left(K-1\right)}{SSE_K / \left(n-K\right)}$) : the user selects the cluster solution with the highest F value.

In Table 19.3, FMAX suggests a five-cluster solution, whereas we can observe a clear drop after three clusters for PRE (from 53.7% to 34.8%). ETA^2 already reaches a high level of 77.5% for three clusters. Hence, it might be useful to further analyze a three- and a five-cluster solution. It is also possible to draw a scree diagram for the different criteria mentioned earlier. If FMAX statistics are used as the y-axis, for example, the solutions with the highest peak are selected. If a modified method is used, alternative measures can be used. We will demonstrate the use of this graphical method later for information criteria.

Density-based clustering methods propose a certain number of clusters. Therefore, the user does not need to decide the number of clusters at first. However, s/he must test validity. One or more alternative solutions may be more appropriate. In addition, as already mentioned, the solution depends on the specified parameter.

Model-based clustering methods have the advantage that the underlying statistical probability model enables the deduction of formal criteria. On the one hand they make it possible to run chi-square-based tests, like likelihood-ratio (LR) test; on the other information criteria are available. In order to select a certain cluster solution, it is necessary to generate a series of possible solutions. Again, we recommend starting with K=1.

The LR test is defined as

$$LR\left(K, K-1\right) = LL_{K-1} - LL_K,$$

where
LL_x = log-likelihood function for a solution with x clusters.

The LR statistic makes it possible to test whether a solution with K clusters significantly improves a solution with K−1 (or K−x) clusters. Wolfe (1970) proposed a modification for the LR statistic. From a theoretical point of view, the LR statistic or its modification by Wolfe has a chi-square distribution. Results by McLachlan and Peel (2005) and McLachlan and Basford (2000) suggest that this is unfortunately mostly not the case. Therefore, the bootstrap method is recommended nowadays.

Information measures are most frequently used to decide the number of clusters. Popular information measures are

$$\begin{aligned} AIC_K &= -2LL_K + 2m_K \\ BIC_K &= -2LL_K + m_K \log(n) \\ CAIC_K &= -2LL_K + m_K\left(\log(n)+1\right) \\ AIC3_K &= -2LL_K + 3m_K \end{aligned}$$

where
LL_K = value of the log-likelihood function that maximizes $LL = \sum_{i=1}^{n} w_i \ln\left(f\left(X_i / \theta_k\right)\right)$
m_k = number of parameters that must be estimated
n = number of cases

The underlying idea behind these measures is to correct for the fact that more clusters will automatically provide a better fit. Again, a scree diagram can be drawn. The most appropriate solution is the solution with the lowest value (inverse peak, see Figure 19.5).

An evaluation study (Fonseca & Cardoso, 2007) suggests that AIC3 performs best for categorical data, whereas BIC performs best for quantitative (metric) variables. AIC has a tendency to select too many clusters (McLachlan & Peel, 2005, p. 220). For mixed scaled variables, the integrated completed likelihood criterion $ICL-BIC$ (Biernacki et al., 2000) outperforms. It is defined as

$$ICL-BIC = BIC + 2\,EN\left(S\right)$$

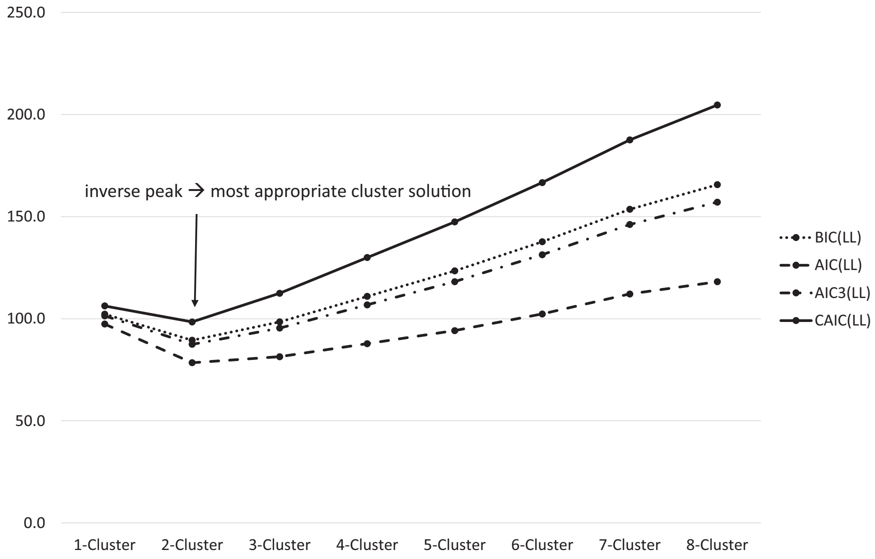

Figure 19.5 Scree diagram generated for different information measures (results from model-based clustering)

and additionally integrates the entropy of the probability of class membership $EN(S_K) = -\sum_{i=1}^{n}\sum_{k=1}^{K}\pi(k/i)\ \log(\pi(k/i))$.

According to McLachlan and Peel (2005, pp. 217–220) $ICL-BIC$ outperforms in the case of quantitative variables, too. Akogul and Erisoglu (2016) report that the Kullback information criterion (KIC) performs best for quantitative variables. KIC is defined as

$$KIC_K = -2LL_K + 3\left(m_K + 1\right)$$

and differs from AIC3 only by adding 3. In a further paper, the authors (Akogul & Erisoglu, 2017) propose using an analytic hierarchy process (AHP) that combines different information measures. This procedure is similar to the consensus method proposed in Bacher et al. (2010).

6 Criteria to validate a cluster solution

After the decision for one or more possible cluster solutions, the selected solutions must be evaluated or rather validated. Validation involves (Everitt et al., 2001, pp. 180–196; e.g., Jain & Dubes, 1988; Wierzchoń & Kłopotek, 2018):

1. *Formal internal validation of the selected solutions*. An index is computed that measures the homogeneity of the clusters of the different solutions. If only one solution is evaluated, thresholds must be defined in order to be able to judge whether the solution can be accepted. If more than one solution is validated, one speaks of a relative validation. In this case, the solutions with the highest formal validity can be selected.
2. *Stability test*. Cluster analysis requires decisions where the user is uncertain which decision is correct. These uncertain decisions should have no or only a small influence on the results. Therefore, this criterion is labeled also as robustness.

3. *Interpretability*. This is the most important criterion. Ideally, formal validation and the stability test result in a decision for a certain cluster solution. If this solution should be used for further analysis, the clusters must be substantively interpretable. It must be possible to give the clusters substantive meaningful names. Sometimes it may occur that this is not possible for all clusters.
4. *Validation by external criteria*. Interpretation very often results in the specification of hypotheses about the association of one or more clusters with other variables, for example "Cluster *C1* is associated with variable *Z*." It may also be possible that these hypotheses exist in advance. The researcher expects certain clusters and associations. In very rare cases, it might be possible to use another classification for validation. The task in this step is to test whether the hypotheses hold.

In the last two decades, many coefficients for formal validation have been proposed (Liu et al., 2010; Satre-Meloy et al., 2020). One frequently used coefficient for formal validation is the silhouette coefficient SC. SC reports how much the objects of one cluster differ from the objects of the cluster that is closest to them. The SC for one object *i*, cluster *k*, and finally the cluster solution *K* is defined as

$$SC(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))}, \quad SC(k) = \frac{1}{n_k}\sum_{i=1}^{n_k} SC(i) \text{ and } SC(K) = \frac{1}{n}\sum_{k=1}^{K} n_k SC(k),$$

where

$b(i)$ = average distance of object *i* to all objects in its nearest cluster.

$a(i)$ = average distance of object *i* to all other objects in the cluster to which object *i* belongs.

Different distance measures can be used and will result in different scores. If the clusters are well separated, *SC(K)* should be large. Kaufman and Rousseeuw (1990, p. 88) propose the following threshold values:

```
0.71 ≤ SC(K) ≤ 1.00 strong structure
0.51 ≤ SC(K) ≤ 0.70 reasonable structure
0.26 ≤ SC(K) ≤ 0.50 weak structure (could be artificial,try ad-
                    ditional methods)
  SC(K) ≤ 0.25       no substantial structure
```

A further frequently cited index is the Dunn index (Kaufman & Rousseeuw, 1990, p. 171). Similar to *SC(K)*, higher values of the Dunn index indicate a better separation. Recent literature recommends to use at least one of these indices to determine the number of clusters, for example by drawing a scree diagram for the silhouette coefficient for different cluster solutions and selecting the solution with the highest silhouette coefficient. Several further indices are available for this task.

In order to compare different cluster solutions, the Rand index is available (Everitt et al., 2001, pp. 181–183). The Rand index depends on the marginal distributions of the classification. Therefore, the adjusted (or corrected) Rand index is recommended. It corrects for purely random agreement and is able to discriminate good solutions. Thresholds are *RAND* > 0.7 (Frabioni & Saltstone, 1992). For the Hubert-Arabie Adjusted Rand index, Steinley (2004) gives the following thresholds:

```
0.90 < adj.Rand          excellent recovery
0.80 < adj.Rand ≤ 0.90   good recovery
0.65 < adj.Rand ≤ 0.80   moderate recovery
       adj.Rand ≤ 0.65   poor recovery.
```

7 Software

Modules in standard statistical software and special, stand-alone software programs are available for cluster analysis. A short and narrative overview will be provided here because the implementation and availability of a program can change during a program upgrade.

IBM SPSS (version 24.0 and above, www.ibm.com/analytics/spss-statistics-software) offers three procedures for clustering: agglomerative hierarchical methods, k-means clustering, and model-based clustering (TSC, two-step cluster). TSC is a hierarchical (divisive) model-based program. It starts with one cluster and splits the clusters as long as the increase in the BIC or AIC change falls below a certain threshold (Bacher et al., 2004). It enables users to handle outliers and to analyze variables with mixed measurement levels. However, ordinal variables have to be treated as nominal scaled variables.

STATA (version 15 and above, www.stata.com/) offers agglomerative hierarchical methods and k-means and model-based clustering similar to IBM SPSS. Model-based clustering is available via generalized structural equation models and corresponds to the described approach. In addition, *k*-median is available, as are special modules for computing the silhouette coefficients and adjusted R.

R offers the most powerful implementation for clustering methods. Leisch and Gruen (2020) provide an overview.

LatentGold (www.statisticalinnovations.com/latent-gold-5-1/) is a stand-alone software that enables model-based clustering. Variables with different measurement levels can be used. Correct standard errors are computed for complex sample designs (like multistage sampling). Bayes elements are integrated in order to avoid local minima and degeneration of solutions. The same models are available in *MPLUS* (www.statmodel.com/).

A comparison by Kent et al. (2014) between TSC and LatentGold favors LatentGold. This result corresponds to Bacher et al. (2004). Rodriguez et al. (2019) compare nine clustering methods that are implemented in R. The studied methods cover all discussed types of clusters. They found a small difference if the dimensionality of the data is small (Rodriguez et al., 2019).

8 Application

We reanalyze data from the Austrian Social Survey (SSÖ) from 2018 (Hadler et al., 2019), which are described in more detail in Eder et al. (2020).[1] The authors use this data set of 1,200 respondents aged 18 and above to analyze their positional, moral, and emotional subjective recognition with the aim of identifying social groups that feel unrecognized. With the help of eight dichotomized indicator items, the authors perform model-based clustering (latent class analysis) and select a solution with four classes. The indicators are shown in Figure 19.6. Their analyses were performed with the statistical software R and the package poLCA.

The first class (cluster) is described as "almost entirely recognized," the second class as "positionally recognized but emotionally unrecognized," the third class as "emotionally recognized but positionally unrecognized," and the fourth class as "poorly recognized." The profile of the four clusters is visualized in a diagram by the authors.

The authors validated their interpretation with a multinomial logistic regression. The four clusters were used as dependent variables. Variables deduced from theory were used as independent variables. The multivariate analysis confirms the interpretation.

Hence, the following criteria are fulfilled:

- Formal internal validity (the four-cluster solution has the lowest BIC and entropy-R-squared is sufficient high).

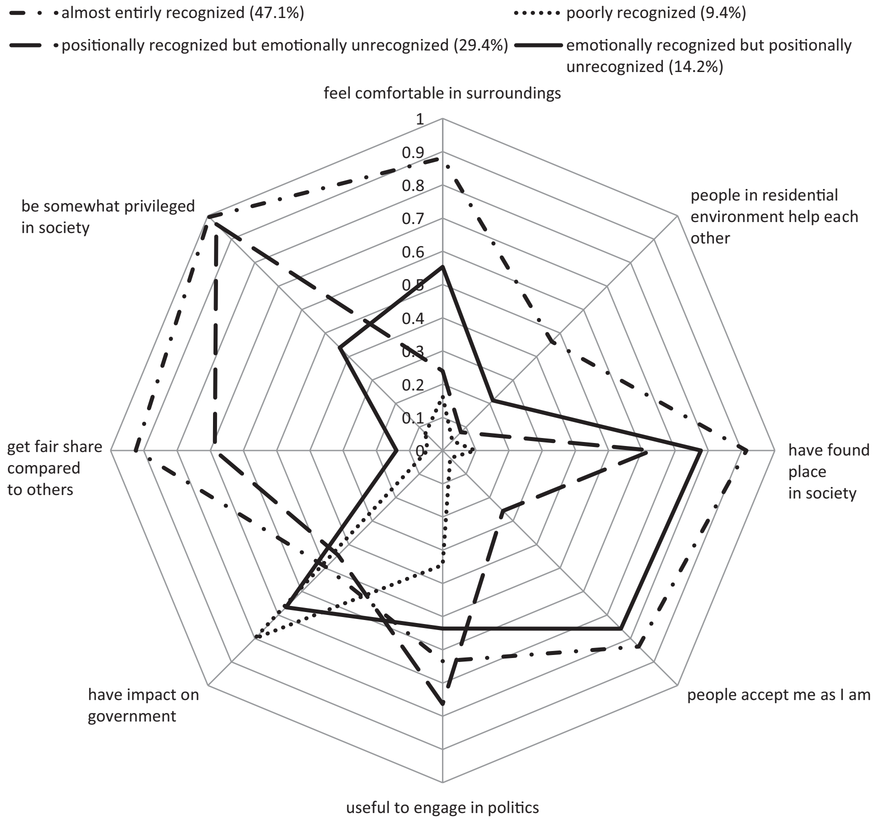

Figure 19.6 Profiles of positional, moral, and subjective recognition for four clusters

- Interpretability is given.
- Validity by external criteria is given.

Relative validity and stability were not tested. This is not necessary, especially if all or nearly all criteria suggest a four-cluster solution. Nonetheless, we will perform a relative internal validity test and test stability. For this purpose, we applied another computer program, namely LatentGold 5.0, and computed up to 10 clusters. The information measures (see Table 19.4) suggest a four-cluster solution (BIC, CAIC) and a six-cluster solution (AIC3). Some simulation results (see earlier) suggest AIC3 for categorical variables. Therefore, we further analyze the four- and six-cluster solution to come to a final decision.

Table 19.5 reports the results of internal validity tests. The four-cluster solution results in a slightly higher silhouette coefficient and Dunn index than the six-cluster solution. The average weighted evidence prefers the four-cluster solution, too. According to the thresholds for the silhouette coefficient, this is a weak cluster solution and further tests should be conducted.

Stability was tested by varying the measurement level of the variables. The use of the non-dichotomized indicator items defined as continuous variables leads to computational problems in LatentGold and cannot be further analyzed. Defining these indicator items as ordinal, the BIC suggests a seven- and the CAIC a five-cluster solution, whereas the AIC3 does not indicate a solution within one to 10 clusters. To remain comparable to the aforementioned

Table 19.4 LatentGold output (eight indicators, nominal scale type, missing values excluded)

Classes	*LL*	*BIC(LL)*	*AIC3(LL)*	*CAIC(LL)*	*Entropy-R$^{2\ a)}$*
1	-5379.1005	10814.0638	10782.2009	10822.0638	---
2	-5028.8155	10176.3396	10108.6309	10193.3396	0.6589
3	-4933.6455	10048.8454	9945.2910	10074.8454	0.6763
4	-4891.8868	10028.1737	9888.7735	10063.1737	0.6586
5	-4864.1043	10035.4545	9860.2085	10079.4545	0.6391
6	-4848.3874	10066.8666	9855.7749	10119.8666	0.6536
7	-4839.4093	10111.7561	9864.8186	10173.7561	0.6822
8	-4832.7305	10161.2443	9878.4611	10232.2443	0.6333
9	-4825.9397	10210.5083	9891.8793	10290.5083	0.6538
10	-4820.3521	10262.1789	9907.7041	10351.1789	0.7069

Note: a) Entropy-R^2 (Vermunt & Magidson, 2013, p. 70) measures the explained variance for a cluster solution with K clusters. The results were generated with LatentGold version 5.0. Calculation was done by the authors. Data are taken from Social Survey Austria 2018 (Hadler et al., 2019), which is freely available at AUSSDA (https://aussda.at/).

Table 19.5 Results of further internal validity tests

	Internal validity		
Cluster	*SC (K)*	*DUNN*	*AWE*
4r	0.2937	0.6919	– a)
4l	0.2956	0.6891	11274.24
6r	0.2766	0.6489	– a)
6l	0.2766	0.6397	11788.93

Note: r = poLCA, l = LatentGold, SC was computed using city-block distance.

a) AWE (approximate weight of evidence, Vermunt & Magidson, 2013, p. 71) not available in poLCA. Calculation was done by the authors. Data are taken from Social Survey Austria 2018 (Hadler et al., 2019), which is freely available at AUSSDA (https://aussda.at/).

solutions, we test the stability of four- and six-cluster solutions based on the ordinal scale type (see Table 19.6). All solutions seem to be somewhat stable, whereby the solutions based on the dichotomized indicator are very similar between poLCA and LatentGold. However, the Rand index and especially the adjusted Rand index indicates bigger differences between the solutions based on the original ordinal indicators and the dichotomized solutions. All values for the four-cluster solution are above the quoted thresholds and we can regard the solutions as stable. For the six-cluster solution, the adjusted Rand index falls under the threshold of 0.65 and the recovery is poor.

Finally, we checked the interpretability of a six-cluster solution with dichotomized indicators. It is possible to describe the class profiles as follows: Class 1 (40.3%): “positionally recognized but morally and emotionally unrecognized”; class 2 (23.4%): “almost entirely recognized”; class 3 (14.6%): “positionally recognized but emotionally unrecognized”; class 4 (8.8%): “emotionally and morally recognized but positionally unrecognized”; class 5 (8.5%): “somewhat recognized, somewhat unrecognized”; and class 6 (4.6%): “poorly recognized.” The interpretability is not as clear as that for the four-class solution, as the additional classes do not add substantial improvements. In summary, our additional analysis supports the authors’ decision.

Table 19.6 Rand indices and adjusted Rand indices

	4l	*4r*		*6l*	*6r*
4r	0.9984 (0.9960)		6r	0.9987 (0.9967)	
4l ord	0.8710 (0.7142)	0.8708 (0.7139)	6l ord	0.7700 (0.3809)	0.7695 (0.3801)

Note: r: poLCA, l: LatentGold, l ord: Latent Gold, ordinal indicators. First line Rand index, second line adjusted Rand index. Calculation was done by the authors. Data are taken from Social Survey Austria 2018 (Hadler et al., 2019), which is freely available at AUSSDA (https://aussda.at/).

9 Summary and recommendations

In the last two decades, a large variety of cluster algorithms and criteria for selecting and validating cluster solutions have been developed. This development focuses on computer science and informatics. The innovations are partially used in applied science. It is difficult to predict which innovations will become established.

Therefore, we can recommend using only those methods that are well documented. It is important to read the definition and computation criteria because they may have the same name but in fact differ in computation method.

Among the discussed methods, we prefer – where appropriate – to use model-based methods because they have a statistical base. Exceptions are a small data set or geographical data. For the first case, hierarchical methods are recommended; for the second, density-based clustering methods can be considered.

The aforementioned innovations may help to solve some of the problems that were mentioned by Everitt (1979). Sometimes, the picture may become unclear when more measures are used. In our experience, the main reason for this unsatisfactory situation is missing substantive theory that enables valid variables and expected clusters to be deduced. As long as this theory is lacking, some cluster analysis problems will remain unsolved.

10 Appendix: data sets and syntax for examples

The used data sets and the syntax are available at doi:10.5281/zenodo.5031638 [1.8.2021].

Note

[1] We would like to thank Robert Moosbrugger as one of the authors for his help in preparing the data for analysis.

References

Akogul, S., & Erisoglu, M. (2016). A comparison of information criteria in clustering based on mixture of multivariate normal distributions. *Mathematical and Computational Applications*, *21*(3), 34. https://doi.org/10.3390/mca21030034

Akogul, S., & Erisoglu, M. (2017). An approach for determining the number of clusters in a model-based cluster analysis. *Entropy*, *19*(9), 452. https://doi.org/10.3390/e19090452

Bacher, J., Pöge, A., & Wenzig, K. (2010). *Clusteranalyse*. Oldenbourg Wissenschaftsverlag GmbH. https://doi.org/10.1524/9783486710236

Bacher, J., Wenzig, K., & Vogler, M. (2004). *SPSS TwoStep Cluster – a first evaluation*. Lehrstuhl für Soziologie, WISO-Fakultät, Universität Erlangen-Nürnnberg. Retrieved from www.ssoar.info/ssoar/handle/document/32715

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725. https://doi.org/10.1109/34.865189

Bradley, P. S., Mangasarian, O. L., & Street, W. N. (1996). Clustering via concave minimization. In *Proceedings of the 9th international conference on neural information processing systems* (pp. 368–374). MIT Press.

Eder, A., Moosbrugger, R., & Hadler, M. (2020). An enquiry in the importance of positional, moral and emotional recognition for social integration in Austria. *Österreichische Zeitschrift Für Soziologie*, *45*(2), 213–233. https://link.springer.com/article/10.1007/s11614-020-00415-y

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd ACM international conference on knowledge discovery and data mining (KDD)* (pp. 226–231). AAAI Press.

Everitt, B. (1974). *Cluster analysis. Reviews of current research* (Vol. 11). John Wiley & Sons.

Everitt, B. (1979). Unresolved problems in cluster analysis. *Biometrics*, *35*(1), 169–181.

Everitt, B., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). Arnold Publishers.

Fonseca, J. R. S., & Cardoso, M. G. M. S. (2007). Mixture-model cluster analysis using information theoretical criteria. *Intelligent Data Analysis*, *11*(2), 155–173. https://doi.org/10.3233/IDA-2007-11204

Frabioni, M., & Saltstone, R. (1992). The WAIS-R number-of-factors quandary: A cluster analxtic approach to construct validation. *Educational and Psychological Measurement*, *52*(3), 603–613.

Fraley, C., & Raftery, A. E. (1999). MCLUST: Software of model-based cluster analysis. *Journal of Classification*, *16*(2), 297–306.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871.

Hadler, M., Höllinger, F., & Muckenhuber, J. (2019). *Social Survey Austria 2018 (SUF edition)*. https://doi.org/10.11587/ERDG3O

Hartigan, J. (1974). *Clustering algorithms*. John Wiley & Sons.

Hautamäki, V., Cherednichenko, S., Kärkkäinen, I., Kinnunen, T., & Fränti, P. (2005). Improving K-means by outlier removal. In A. Kaarna, H. Kalviainen, & J. Parkkinen (Eds.), *Lecture notes in computer science: Vol. 3540. Image analysis* (Vol. 3540, pp. 978–987). Springer-Verlag. https://doi.org/10.1007/11499145_99

Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (Eds.). (2016). *Chapman & Hall/CRC handbooks of modern statistical methods. Handbook of cluster analysis*. CRC Press a Chapman & Hall book.

Holmgren, J., Knapen, L., Olsson, V., & Masud, A. P. (2020). On the use of clustering analysis for identification of unsafe places in an urban traffic network. *Procedia Computer Science*, *170*, 187–194. https://doi.org/10.1016/j.procs.2020.03.024

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, *2*, 283–304.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data. Prentice-Hall advanced reference series*. Prentice Hall.

Jardine, N., & Sibson, R. (1971). *Mathematical taxonomy*. John Wiley & Sons.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data*. John Wiley & Sons. https://doi.org/10.1002/9780470316801

Kent, P., Jensen, R. K., & Kongsted, A. (2014). A comparison of three clustering methods for finding subgroups in MRI, SMS or clinical data: Spss TwoStep cluster analysis, latent gold and SNOB. *BMC Medical Research Methodology*, *14*, 113. https://doi.org/10.1186/1471-2288-14–113

Lathrop, R. G., & Williams, J. E. (1989). The Sahpe of the inverse scree test for cluster analysis. *Educational and Psychological Measurement*, *50*(2), 325–330.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Mifflin.

Leisch, F., & Gruen, B. (2020). *CRAN task view: Cluster analysis & finite mixture models*. JKU. https://cran.r-project.org/web/views/Cluster.html

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of internal clustering validation measures. In *IEEE international conference 13.12.2010–17.12.2010* (pp. 911–916). https://doi.org/10.1109/ICDM.2010.35

McLachlan, G. J., & Basford, K. E. (2000). *Mixture models: Inferences and application to clustering*. Marcel Dekker.

McLachlan, G. J., & Peel, D. (2005). *Finite mixture models. Wiley series in probability and statistics Applied probability and statistics section*. John Wiley & Sons. Retrieved from http://gbv.eblib.com/patron/FullRecord.aspx?p=219004 https://doi.org/10.1002/0471721182

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal*, *20*(4), 359–363. https://doi.org/10.1093/comjnl/20.4.359

Murtagh, F. (2016). A brief history of cluster analysis. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Chapman & hall/CRC handbooks of modern statistical methods. Handbook of cluster analysis* (pp. 21–30). CRC Press a Chapman & Hall book.

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS One*, *14*(1), e0210236. https://doi.org/10.1371/journal.pone.0210236

Satre-Meloy, A., Diakonova, M., & Grünewald, P. (2020). Cluster analysis and prediction of residential peak demand profiles using occupant activity data. *Applied Energy*, *260*, 114246. https://doi.org/10.1016/j.apenergy.2019.114246

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., . . . Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664–681. https://doi.org/10.1016/j.neucom.2017.06.053

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited. *ACM Transactions on Database Systems*, *42*(3), 1–21. https://doi.org/10.1145/3068335

Steinley, D. (2004). Properties of the Hubert-Arabie adjusted rand index. *Psychological Methods*, *9*(3), 386–396. https://doi.org/10.1037/1082-989X.9.3.386

Steinley, D. (2016). K-mediods and other criteria for crisp clusters. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Chapman & hall/CRC handbooks of modern statistical methods. Handbook of cluster analysis* (pp. 55–66). CRC Press a Chapman & Hall book.

Vermunt, J. K., & Magidson, J. (2000). *Latent GOLD user's guide*. Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2013). *Technical guide for latent GOLD 5.0: Basic, advanced, and syntax*. Statistical Innovations Inc.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Wierzchoń, S. T., & Kłopotek, M. (2018). *Modern algorithms of cluster analysis. Studies in big data* (Vol. 34). Springer-Verlag. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&AN=1670339

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, *5*(3), 329–350. https://doi.org/10.1207/s15327906mbr0503_6

Zgurovsky, M. Z., & Zaychenko, Y. P. (2020). *Big data: Conceptual analysis and applications* (Vol. 58). Springer International Publishing. https://doi.org/10.1007/978-3-030-14298-8