PRAGUE UNIVERSITY OF ECONOMICS AND BUSINESS
FACULTY OF INFORMATICS AND STATISTICS

HABILITATION THESIS

2023 **ZDENĚK ŠULC**

PRAGUE UNIVERSITY OF ECONOMICS AND BUSINESS
FACULTY OF INFORMATICS AND STATISTICS

HIERARCHICAL CLUSTER ANALYSIS OF CATEGORICAL DATA

Habilitation Thesis

Author: Ing. Zdeněk Šulc, Ph.D.
Field of habilitation: Statistics

Prague, September 2023

Declaration

I declare that I carried out this habilitation thesis independently and cited all used sources and literature.

Prague, September 7, 2023

Ing. Zdeněk Šulc, Ph.D.

Abstract

This habilitation thesis deals with two important areas of hierarchical clustering of categorical data, namely similarity measures for categorical data represented by nominal variables with more than two categories and evaluation criteria for the cluster quality assessment. The conducted literature review shows more research needs to be done in this area. Thus, the thesis explores these topics deeply using two experiments based on generated datasets with controlled properties, such as the number of variables or clusters. The first experiment performed on 2,700 datasets analyzes 16 similarity measures concerning their ability to produce good-quality clusters in different dataset properties and linkage methods. Some of the analyzed similarity measures are analyzed for the very first time in the domain of cluster analysis. The second experiment performed on 8,100 datasets compares 11 evaluation criteria for categorical data proposed in various papers. Two of them are newly proposed in this thesis. The criteria are examined from different perspectives, such as their mutual similarity or dependence on the clustered dataset's properties. In the conclusions of both experiments, the most appropriate similarity measures for a specific dataset's properties and evaluation criteria for several intended tasks are recommended. Since the thesis focuses on a practical application of the research outcomes, it presents and further improves a convenient software application that enables researchers to easily replicate the results in the thesis and, more importantly, to perform advanced approaches to categorical data clustering on their own.

Key words: categorical data, hierarchical cluster analysis, comparison, similarity measures, evaluation criteria, R package

Abstrakt

Tato habilitační práce se věnuje dvěma důležitým oblastem hierarchického shlukování kategoriálních dat, a to mírám podobnosti pro kategoriální data obsahující nominální proměnné s více než dvěma kategoriemi a hodnoticím kritériím pro posouzení kvality shluků. Provedená rešerše literatury ukazuje, že v této oblasti je třeba provést další výzkum. Práce se tedy těmto tématům věnuje hlouběji prostřednictvím dvou experimentů založených na generovaných datových souborech s předem stanovenými parametry, jako je počet proměnných nebo shluků. V prvním experimentu, provedeném na 2 700 datových souborů, je analyzováno 16 měr podobnosti ohledně jejich schopnosti vytvářet kvalitní shluky u datových souborů s různými parametry a u různých metod shlukové analýzy. Některé z analyzovaných měr podobnosti jsou v oblasti shlukové analýzy zkoumány vůbec poprvé. Ve druhém experimentu, založeném na analýze 8 100 datových souborů, se porovnává 11 hodnoticích kritérií určených pro kategoriální data, která byla představena v různých článcích. Dvě z nich jsou navržena v této práci. Kritéria jsou zkoumána z různých hledisek, například na základě vzájemné podobnosti nebo závislosti na parametrech shlukovaného souboru dat. V závěrech obou experimentů jsou doporučeny nejvhodnější míry podobnosti pro typické situace a hodnoticí kritéria na základě zamýšlené úlohy. Vzhledem k tomu, že se práce zaměřuje na praktické využití výsledků výzkumu, je v ní představena a dále vylepšena softwarová aplikace, která umožňuje výzkumným pracovníkům snadno zopakovat výsledky uvedené v práci, a především samostatně provádět pokročilé metody shlukování kategoriálních dat.

Klíčová slova: kategoriální data, hierarchická shluková analýza, porovnání, míry podobnosti, hodnoticí kritéria, R balíček

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

AIC	Akaike information criterion for categorical data
ANOVA	analysis of variance
ALM	average linkage method in HCA
AN	Anderberg similarity measure
ARI	adjusted Rand index
BIC	Bayesian information criterion for categorical data
BK	BK (best *k*) index
BU	Burnaby similarity measure
CAT	number of categories
CI	category information
CLM	complete linkage method in HCA
CLU	number of clusters
COOLCAT	entropy-based algorithm for categorical clustering
CU	category utility
DI	Dunn index
DIST	minimal between-cluster distance
ES	Eskin similarity measure
GA	Gambaryan similarity measure
G1	Goodall 1 similarity measure
G2	Goodall 2 similarity measure
G3	Goodall 3 similarity measure
G4	Goodall 4 similarity measure
HCA	agglomerative hierarchical cluster analysis
HE	Hartigan entropy
HM	Hartigan mutability
IOF	inverse occurrence frequency similarity measure
IQR	inter-quartile range
LCA	latent class analysis
LIN	Lin similarity measure
LIN1	Lin 1 similarity measure
LINK	linkage method

MI	mutual information
MRS	mean ranked scores
MZ	Morlini and Zani similarity measure
OF	occurrence frequency similarity measure
PSFE	pseudo F index based on entropy
PSFM	pseudo F index based on mutability
Q1	first (lower) quartile
Q3	third (upper) quartile
RI	Rand index
ROCK	robust clustering using links
SI	silhouette index
SLM	single linkage method in HCA
SM	simple matching similarity measure
SV	Smirnov similarity measure
TwoStep	two-step cluster analysis
VAR	number of variables
VE	variable entropy similarity measure
VM	variable mutability similarity measure
WCE	within-cluster entropy
WCE_s	standardized within-cluster entropy
WCM	within-cluster mutability
WCM_s	standardized within-cluster mutability

Symbols

$a(i)$	average dissimilarity of the ith object to the other objects in the same cluster
$b(i)$	minimum average dissimilarity of the ith object to other objects in any cluster not containing the ith object
c	index of a variable, $c = 1, 2, \ldots, m$
C_g	gth cluster
C_h	hth cluster
$D\left(\mathbf{x}_i, \mathbf{x}_j\right)$	dissimilarity between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$
$f(x_{ic})$	absolute frequency of the value x_{ic} by the cth variable
f_{cu}	absolute frequency of the uth category
FN	number of false-negative decisions
FP	number of false-positive decisions
G_{gc}	Gini coefficient of the cth variable in the gth cluster
g	index of a cluster, $g = 1, 2, \ldots, k$
H_{gc}	entropy of the c-th variable in the gth cluster
$H_E(k)$	expected entropy in a dataset with k clusters
h	index of a cluster, $h = 1, 2, \ldots, k$
i	index of an object, $i = 1, 2, \ldots, n$
$I(k)$	expected incremental entropy between the datasets with k and $k+1$ clusters
j	index of an object, $j = 1, 2, \ldots, n$
K_c	number of categories by the cth variable
k	number of clusters
m	number of variables
n	number of objects
n_g	number of objects in the gth cluster
n_h	number of objects in the hth cluster
n_{cu}	number of objects by the cth variable with the uth category
n_{gcu}	number of objects in the gth cluster by the cth variable with the uth category
$p(x_{ic})$	relative frequency of the value x_{ic} by the cth variable
p_{cu}	relative frequency of the uth category
$\hat{p}^2$	adjusted relative frequency
q	subset of categories
Q	subset of the data matrix $\mathbf{X}$
$S_c\left(x_{ic}, x_{jc}\right)$	similarity between the categories x_{ic} and x_{jc}
$SW_c\left(x_{ic}, x_{jc}\right)$	weighted similarity between the categories x_{ic} and x_{jc}
$S\left(\mathbf{x}_i, \mathbf{x}_j\right)$	similarity between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$
SS_B	between-group sum of squares
SS_T	total sum of squares

SS_W	within-group sum of squares
t	index of a variable, $t = 1, 2, \ldots, c, \ldots, m$
TN	number of true-negative decisions
TP	number of true-positive decisions
u	index of category, $u = 1, 2, \ldots, K_c$
v	index of a cluster, $v = 1, 2, \ldots, \ldots, k$
$\mathbf{w}$	weight vector containing weights for m variables
w_c	weight for the cth variable
$\mathbf{X}$	data matrix with n rows and m columns
$\mathbf{x}_i$	ith object in the dataset
$\mathbf{x}_j$	jth object in the dataset

Preface

My primary motivation for writing this habilitation thesis was to thoroughly cover the not-well-explored topic of hierarchical cluster analysis (HCA) of categorical data in one scientific text. I wanted to focus on parts of the research that differed substantially from quantitative data clustering, namely dissimilarity matrix calculation and assessment of final clusters. It meant exploring the areas of similarity measures for categorical data and evaluation criteria for categorical data clustering. I aimed to ensure that my conclusions would impact how the HCA of categorical data was used in practice. Therefore, an essential part of the thesis is the recommendation to other researchers of the best approaches. In addition, the thesis presents a software application that enables users to perform all the procedures in the thesis, so researchers have easy access to the examined methods without a need to program them.

The thesis deals with three main topics: similarity measures for categorical data, evaluation criteria for categorical data, and an R package for categorical data clustering. Since I already dealt with the issues of similarity measures and the R package in my dissertation thesis (Šulc, 2016), I want to clarify the main contributions of the habilitation thesis compared to the dissertation.

The first topic of the thesis, similarity measures for categorical data, is primarily based on the paper written by Šulc and Řezanková (2019). Compared to my dissertation thesis (Šulc, 2016), it analyzes similarity measures and linkage methods mutually and provides recommendations on which similarity measures are most appropriate for given properties of a dataset. The habilitation thesis further extends this research by four similarity measures that have not been examined before in the domain o hierarchical clustering. It also provides an updated methodology of similarity measures quality assessment in HCA based on boxplots.

The second topic, evaluation criteria for categorical data, is the new research that examines the properties of evaluation criteria and their mutual relationships. A minor part of this research focuses on the criteria's ability to determine the optimal number of clusters, which builds on the paper of Šulc et al. (2018).

The third topic, the `nomclust` R package for categorical data clustering, is based on the paper prepared by Šulc et al. (2022), which introduced the second generation of the `nomclust` package to the scientific community. The habilitation thesis extends this research mostly about new similarity measures, evaluation criteria, and variable weighting, presented in the latest package version. Compared to the first generation of the package proposed by Šulc (2016), the second generation is completely reworked and contains many features unavailable in the first generation. For instance, new evaluation criteria, the ability to produce graphical outputs or the support for S3 generic functions can be emphasized.

Although I am the primary author of the papers (Šulc and Řezanková, 2019) and (Šulc et al., 2022), I did not write them alone. Therefore, I would like to thank prof. Hana Řezanková for her priceless pieces of advice regarding the theoretical background of the research and the precise check of both papers. I also want to thank Mgr. Jana Cibulková for her assistance with the graphical functions in the `nomclust` package, and Ing. Jaroslav Horníček for the help with C++ programming.

Introduction

Cluster analysis is a multivariate statistical method that reveals an underlying data structure by identifying homogeneous groups (clusters) of objects. The choice of a suitable clustering method depends on the type of clustered data. The quantitative data enables a researcher to choose from many well-examined clustering methods based on various principles, such as distance, model, density, or grid. The list of suitable approaches for categorical data is considerably shorter, and the available methods are less examined than the methods for quantitative data. Therefore, this thesis focuses on categorical data clustering.

The clustering of objects characterized by categorical variables has become an important issue in recent years. There are many economic areas where a growing demand for categorical data clustering occurs. In marketing research and sociology, data from questionnaire surveys are often processed, e.g., to perform market segmentation, study customer behavior, or recognize societal opinions. There is a large application area in official statistics where cluster analysis can be used, for instance, for grouping similar regions based on the answers obtained via the Business Tendency Survey conducted by OECD (Organization for Economic Co-operation and Development) or to indicate household types in a society based on government surveys, such as EU-SILC (European Union Statistics on Income and Living Conditions). However, the use of cluster analysis for datasets with categorical data goes far beyond the field of economy. For instance, in artificial intelligence research, the clustering of categorical data is used for text-mining tasks to understand a written text or to analyze unstructured data.

This thesis deals with the agglomerative hierarchical cluster analysis (HCA) of categorical data. Under the term *categorical data*, only the data containing *nominal* variables are considered. This type of data has its specifics, among others, the inability to determine the order of categories. In fact, categories of nominal variables can be differentiated only by their equality or inequality. This often results in a very simplistic determination of dissimilarities between pairs of objects that does not reflect any additional information about the clustered dataset, such as the number of categories or frequency distributions of variables. The categorical nature of data also complicates the use of many evaluation criteria for cluster quality assessment proposed for quantitative data, which are usually based on the sum of squares or other variability concepts that are not applicable to categorical data. Although there have been developed categorical alternatives for some of these criteria, there is still no comparative study that assesses their quality and properties. The abovementioned problems show that HCA of data with categorical variables is one of the least investigated major clustering algorithms.

In the HCA of categorical data, the selection of linkage methods, similarity measures, and evaluation criteria is vastly limited compared to the approaches designed for quantitative data. Some linkage methods for quantitative data work with concepts that do not apply to data with categorical variables, such as cluster centroids by the centroid method or the within-cluster variability by Ward's method.

When selecting similarity measures to get dissimilarities between pairs of objects in HCA, there arises a problem of similarity definition between categories in categorical data. Currently, the most common way is to transform the variables into binary ones and use similarity measures for binary or quantitative data. However, these measures cannot utilize the information in sets of binary variables in the same way as the untransformed variables, see (Goodall, 1966). Thus, it is necessary to consider the use of this approach carefully. The other option is to use similarity measures directly determined for categorical data. Currently, the approaches introduced more than 50 years ago, such as the simple matching measure (Sokal and Michener, 1958) and the categorical (nominal) part of the Gower distance (Gower, 1971) for mixed-type data, are commonly used for this task. They only recognize if two categories match or not. Since then, many more sophisticated approaches have been introduced, e.g., by Eskin et al. (2002) or Boriah et al. (2008), that take into account various dataset properties for better similarity definition, such as frequency distributions of variables or their numbers of categories. Some of these approaches were examined by Šulc (2016). Still, there is room for deeper analysis and extending the research to more similarity measures for categorical data.

The important part of cluster analysis is the evaluation of the produced clusters, where the internal evaluation criteria (using intrinsic properties of a dataset) are the most appropriate since cluster analysis is an unsupervised method. Again, the number of evaluation criteria for categorical data is vastly limited compared to the quantitative data. Currently, only a few established evaluation criteria are determined for this data type. For purely categorical data, a few variability-based criteria, such as pseudo-F index based on mutability (PSFM) and pseudo-F index based on entropy (PSFE) (Řezanková et al., 2011), can be used. For categorical or mixed-type data, the Akaike information criterion (AIC) and Bayesian information criterion (BIC), proposed in SPSS, Inc. (2001), which are based on likelihood approximation, are available. The other option is to use evaluation criteria based on a dissimilarity (proximity, distance) matrix, which is a typical output of hierarchical clustering that does not require the original dataset, such as the Dunn index (Dunn, 1973) or silhouette coefficient (Rousseeuw, 1987). Internal evaluation criteria for categorical data clustering were examined only by Šulc et al. (2018) regarding the optimal number of clusters determination. However, a paper comparing criteria for cluster quality evaluation is still missing.

Another limitation of the hierarchical clustering of categorical data is its lack of implementation in commercial and non-commercial software. Currently, two convenient ways (apart from programming) to cluster such data exist. The first one lies in binary data transformation and using one of many similarity measures for binary data, e.g., in SPSS. The drawback of this approach is that similarity measures for binary-coded data often provide the same cluster partitions as the simple matching measure, as it was discovered by Šulc (2016). The second way is to use the `nomclust` package created by Šulc and Řezanková (2015) for the R environment (R Core Team, 2021). The package covers the whole clustering process, i.e., from proximity matrix calculation, over the clustering method selection, to an assessment of the created clusters, and it is freely available on CRAN (The Comprehensive R Archive Network). Although the package works relatively well (and some researchers use it), it suffers from low computational speed, and there is still room for adding many improvements, such as new evaluation criteria and graphical outputs.

It is evident that the HCA of categorical data has many unsolved issues that limit its use compared to the well-established methods for quantitative data clustering. Thus, the habilitation thesis aims to extend the research of Šulc (2016) through three main research goals.

Goal 1. Similarity measures for categorical data. The first objective is to inspect additional similarity measures for categorical data compared to those examined by Šulc (2016), such as the measures proposed by Burnaby (1970) or Gambaryan (1964), and to compare them using internal evaluation criteria for categorical data. The aim is to recommend several suitable similarity measures for a given dataset's properties. Since the performance of similarity measures is strongly influenced by the linkage method used (Šulc, 2016), the combinations of similarity measures and three different linkage methods will also be examined. The outcomes of the experiment performed on generated datasets will help researchers reduce the number of the considered similarity measures when conducting cluster analysis.

Goal 2. Internal evaluation criteria for the clusters created from categorical data. The second objective, which is partly based on the research by Šulc et al. (2018), is to compare the commonly used internal evaluation criteria for categorical data, analyze their mutual relationships from different perspectives, and examine the relationship between the investigated internal criteria and the adjusted Rand index, a typical representative of the external criteria. A partial goal is to propose new internal criteria based on the variability of the clustered variables. The outcomes based on generated datasets should help a researcher decide which evaluation criterion is suitable for a particular situation or inform which criteria assess the cluster quality almost identically.

Goal 3. nomclust 2.0. The third objective is to present the second generation of the `nomclust` R package (Šulc et al., 2022) to the scientific community, further improve it, and illustrate its use. The second generation of the package deals with the drawbacks outlined in the previous text. The aim is that researchers can replicate the research in this thesis and use the package for categorical data clustering on their own.

Apart from Introduction and Conclusion, the thesis consists of six thematic chapters. The first one describes the current knowledge of the selected categorical data clustering areas. The second one outlines theoretical approaches to the examined similarity measures for categorical data clustering. The third one deals with evaluation criteria for assessing the created clusters. The fourth one presents the second generation of the `nomclust` R package. The fifth and sixth chapters carry out the experiments comparing the examined similarity measures and evaluation criteria for the categorical data. A more detailed description of the thesis's chapters is presented in the paragraphs below.

Chapter 1 presents the current state of knowledge in categorical data clustering divided into three sections that correspond to three goals set in the thesis. The text is not restricted to hierarchical clustering, but it also describes alternative methods to categorical data clustering, including the software solutions where they can be found.

Chapter 2 focuses on the similarity measures that can be used in HCA of categorical data. It is divided into four sections. The first one describes the calculation steps of the examined similarity measures from a mathematical perspective. In the second one, the similarity measures are categorized according to the principles they are based on, and their history and properties are described. The third section proposes a new variable weighting concept based on adjusting similarity measure values, later demonstrated in Chapter 4. The fourth section presents the methods of hierarchical cluster analysis that determine how the dissimilarity (based on the similarity measures) between two clusters is calculated.

Chapter 3 deals with evaluation criteria used in categorical data clustering, where they assess the created clusters' quality. It is divided into two sections. The first one describes external evaluation criteria; one of them is used for internal evaluation criteria assessment in Chapter 6. The second one provides an extensive overview of the internal evaluation criteria determined for categorical data, which are divided according to the principle they are built on. Moreover, two new internal criteria are proposed in this section.

Chapter 4 presents the `nomclust` R package. It is divided into two sections. The first section mainly describes a theoretical background of the package's functionalities. The second one demonstrates the typical use of the package on several practical examples that illustrate different scenarios of the possible package use.

Chapter 5 contains the first experiment dealing with comparing the examined similarity measures for categorical data presented in Chapter 2. It is divided into three sections. The first describes the dataset generation for the experiment, the second defines the research methodology, and the third one contains the experiment itself. The analysis determines the generally well-performing similarity measures and recommends which combinations of similarity measures and linkage methods are the most suitable for the specific dataset properties.

Chapter 6 carries out the experiment comparing and assessing the internal evaluation criteria for categorical data presented in Chapter 3. It is divided into three sections. The first one defines the data generation process and similarity measures selection for the second experiment. The second one describes the statistical methods used for the evaluation criteria comparison. The third one comprises the experiment, which examines the mutual similarity of the evaluation criteria, their ability to determine the optimal number of clusters and their dependence on clustered dataset properties. Eventually, specific evaluation criteria for the typical tasks are recommended.

1 State of Knowledge

This chapter presents a state of knowledge for the three main research areas. The first part describes advances in categorical data clustering and similarity measures for categorical data. The second part is devoted to evaluation criteria applicable to outputs of the categorical data clustering. The third one covers different software solutions that can be used for categorical data clustering.

1.1 Categorical Data Clustering

Many approaches can be used for categorical data clustering. One can use HCA with the similarity measures for binary or nominal variables, some flat clustering algorithms, such as the k-modes clustering, or even model-based approaches, i.e., latent class analysis. In this subsection, the options will be briefly described.

When clustering categorical data, currently, the most common way is to transform the variables into binary ones and use HCA with similarity measures for binary variables, such as the Jaccard coefficient (Jaccard, 1912). In the 1950s and 1960s, researchers from different fields introduced many similarity measures for binary data. Therefore, some measures were known under several names. These measures were summarized, e.g., by Warrens (2008) or Cibulková et al. (2020). Unfortunately, many of these measures are identical or strongly linearly dependent (Todeschini et al., 2012). Cibulková et al. (2020) divided 66 similarity measures for binary-coded data into four distinct groups, the largest of which provided the same cluster partitions as the simple matching approach. Other interesting approaches for similarity measures for categorical data use binary data transformation. For instance, Morlini and Zani (2012) developed two similarity measures for nominal variables which use a binary transformation. They are information-based similarity measures that were evaluated in clustering on real-world and simulated datasets.

Goodall (1966) introduced a new similarity measure directly determined for categorical data, where the rare matching pairs of values contributed more to the total similarity. In his study, the new measure was compared with commonly used measures for binary data, such as the Jaccard coefficient, the first similarity measure for binary data, on a numerical taxonomy dataset. In the following years, many similarity measures for categorical data that considered the dataset properties were introduced. For instance, Spärck Jones (1972) proposed a similarity measure based on an inverse document frequency principle, which assigned higher weights to less frequent matches compared to more frequent ones, or Lin (1998) attempted to create a universal probabilistic similarity measure. Another interesting example is the measure proposed by Eskin et al. (2002), which uses the number of categories of a variable for similarity determination. The measure was originally introduced for unsupervised anomaly detection in the density-based clustering algorithm. It was evaluated on two intrusion datasets, and it showed promising results. Boriah et al. (2008) proposed several modifications of original similarity measures introduced by Goodall (1966) and Lin (1998). New similarity measures and algorithms based on this principle are constantly proposed, e.g., (Desai et al., 2011) or (Yi et al., 2016). Usually, they are primarily determined for the text-mining tasks. There are only a few papers where the measures for categorical data were independently compared and evaluated. For instance, the research of Boriah et al. (2008) and Chandola et al. (2009), where the selected similarity measures for categorical data were evaluated in a domain of outlier detection. The performance of these measures in HCA was evaluated only by Šulc (2016), who found out which similarity measures perform well and under which circumstances. However, there are still some similarity measures whose performance in hierarchical clustering needs to be examined.

From the distance-based algorithms, one can also use partitioning methods, also known as *flat* clustering, that iteratively assign objects to the closest cluster center. In categorical data, this approach is typically represented by the k-modes method (Chaturvedi et al., 2001), where the cluster centers are defined as the vector of modes for variable values belonging to the clusters. The other option is to use the k-prototypes clustering that deals with quantitative and categorical variables. The main advantage of the partitioning method is its computation speed, enabling a researcher to cluster objects in large datasets. Among the drawbacks, a tendency to find a local optimum and the necessity to set the number of clusters in advance can be mentioned. The research in this area is still active. Currently, it focuses mainly on mixed-type data clustering, (e.g., Ahmad and Khan, 2019).

There are some alternative approaches to categorical data clustering that can be either model- or distance-based. The model-based clustering assumes that the clusters are defined by parametric distributions, and the whole dataset is a mixture of such distributions (Anderlucci and Hennig, 2014). An object is assigned to a cluster with the highest conditional probability. The most commonly used representative is latent class analysis (LCA) (Hagenaars and McCutcheon, 2002), which usually provides good-quality clusters. LCA is a complex method determined not only for clustering but also for other statistical tasks, such as causal analysis (Hagenaars and McCutcheon, 2002); see also (Šulc, 2016).

Regarding the distance-based methods, one can use relatively known TwoStep cluster analysis

(Bacher et al., 2004), which is implemented in IBM SPSS (SPSS, Inc., 2001). It was designed to cluster large datasets comprising quantitative and categorical variables. Thus, the algorithm uses the log-likelihood distance, whose formula contains parts for quantitative and categorical variables. The algorithm is fast because it only goes through a whole dataset once. It consists of two steps. In the first one, preliminary clusters are created sequentially. In the second one, the hierarchical cluster algorithm is applied to the preliminary clusters. The method also contains a procedure that enables determining the optimal cluster solutions. The TwoStep method and the similarity measures for categorical data in HCA were compared by Šulc (2016).

There are some additional approaches for categorical data that are not very known. One of them is the ROCK algorithm (Guha et al., 1999) that transforms the categorical data into sets of binary variables, and consequently, the Jaccard coefficient is used. It is based on principles of the graph theory, in particular links, which is the number of common neighbors between a pair of objects. Two objects are considered neighbors if their distance is equal to or lower than a user-set cutoff value. The ROCK algorithm generally creates good clusters. However, some objects in the dataset do not have to be assigned to any cluster, which can be considered a severe drawback in certain situations. One can also use the COOLCAT algorithm (Barbará et al., 2002) that aims to minimize the entropy of the created clusters.

1.2 Cluster Assessment in Categorical Data

External or internal evaluation criteria, also known as *cluster validation indices*, can assess the obtained cluster assignments. The external criteria (see, e.g., de Souto et al., 2012; Draszawka and Szymański, 2011), compare a cluster assignment to a priori-known class variable. Since the class variable is usually unknown in clustering tasks, the external criteria are unsuitable for practical application. However, they are helpful in simulation studies, where the properties of the clustering algorithms are assessed. The internal criteria (see, e.g., Liu et al., 2010; Miligan and Cooper, 1985; Vendramin et al., 2010), use intrinsic properties of a dataset to determine the cluster quality or suggest the optimal number of clusters. It makes them more suitable for practical application than the external criteria because the correct cluster assignment is usually unknown. For their calculation, an original dataset or a distance matrix based on an original dataset is necessary. Some internal criteria, such as the BK (best *k*) index (Chen and Liu, 2009) or the Hartigan's rule (Hartigan, 1975), were primarily designed to suggest the optimal number of clusters in a dataset, while another (Arbelaitz et al., 2013; Dimitriadou et al., 2002) to assess the quality of the created clusters. A big topic is the analysis of relationships between external and internal evaluation criteria, which often leads to the question if internal criteria provide comparable results to the external criteria. In quantitative data, these relationships were studied, e.g., by Hennig (2022); Kargar et al. (2019); Tomasini. et al. (2017); Rendón et al. (2011); Halkidi et al. (2001). Unfortunately, there is no such analysis for categorical data.

When dealing with categorical data clustering, a researcher can use the PSFE and PSFM criteria (Řezanková et al., 2011), which are the modifications of the pseudo-F index (Caliński and

Harabasz, 1974) for datasets with categorical variables. One can also use Category Utility (CU) and Category Information (CI) criteria introduced by Corter and Gluck (1992) or the BK index (Chen and Liu, 2009). Next, there are the modifications of the AIC and BIC indices (SPSS, Inc., 2001) derived from the original information criteria proposed by Akaike (1973) and Schwarz (1978). Another approach is to use one of several criteria based on a dissimilarity matrix, e.g., the silhouette index (Rousseeuw, 1987) or the Dunn index (Dunn, 1973). Thus, overall, there are many internal evaluation criteria. The problem is that they are not sufficiently examined, and therefore, only a few papers use them when evaluating categorical clustering results. For instance, the applications presented by Xavier et al. (2013) and Bontemps and Toussile (2013) can be mentioned.

Although many papers compare internal evaluation criteria for quantitative data (e.g., Miligan and Cooper, 1985; Brun et al., 2007; Vendramin et al., 2010), there are only several papers where the internal evaluation criteria suitable for categorical data are compared or assessed. For instance, Šulc et al. (2018) examined 11 internal evaluation criteria suitable for categorical data from an aspect of the determination of the optimal number of clusters in the generated datasets. However, this research did not analyze the internal criteria' ability to judge the created clusters' quality. Another example is a paper prepared by Bai and Liang (2015), who inspected the performance of three cluster validity functions determined to optimize the k-modes algorithm on several real datasets. Since these functions were used within the k-modes algorithm, they cannot be used to compare different clustering algorithms, such as HCA, on a given dataset.

1.3 Software for Categorical Data Clustering

If a given algorithm is to be considered for the general public or common researchers' use, providing a suitable software implementation is essential. Only a few researchers will program the algorithm on their own. This subsection provides an overview of the software used for categorical data clustering.

In HCA of categorical data, one can use the `hclust()` function from the `stats` R package or the `agnes()` function from the `cluster` package (Maechler et al., 2022). The functions enable a researcher to use any pre-calculated dissimilarity matrix as an input based, for instance, on one of ten similarity measures for binary data in the `Mercator` R package (Coombes and Coombes, 2022), or one of more than twenty similarity measures in IBM SPSS. Unfortunately, there are not many dissimilarity matrix functions for categorical data. One of the few is the *Gower* distance (Gower, 1971) in the `daisy()` function in the `cluster` package, which calculates the simple matching approach for nominal variables in a dataset. The next option is to calculate the dissimilarity matrix using one of 13 similarity measures for categorical data in the first generation of the `nomclust` package (Šulc and Řezanková, 2015). The package also contains an option to use the `nomclust()` function that covers the whole clustering process of categorical data clustering, including, e.g., assessment of the cluster quality. Considering the partitioning methods, the k-modes clustering can be performed by the function `kmodes()` in the `klaR` R package (Weihs et al., 2005). The k-prototypes algorithm is complexly covered in the `clustMixType` R package (Szepannek, 2018).

The alternative approaches to categorical data clustering are primarily implemented in R, some in commercial software. Latent class analysis is available in various software, e.g., in LatentGold (Vermunt and Magidson, 2016). In R, LCA can be run using the function `poLCA()` in the `poLCA` package (Linzer and Lewis, 2011). The TwoStep cluster analysis procedure has been present in IBM SPSS since 2000 (version 11.5). There are not many alternatives to this method in other software. In R, a researcher can use the `prcr` package (Rosenberg et al., 2020), which is based on the research by Bergman and El-Khouri (1999), but this method is not entirely equivalent alternative, so the analysis setting and the provided clusters can differ. The ROCK algorithm can be applied using the function `rockCluster()` in the `cba` R package (Buchta, Hahsler, 2019), and the COOLCAT algorithm using the function `coolcat()` in the `coolcat` R package that is not available on CRAN (The Comprehensive R Archive Network), but it can be found in GitHub (Github, 2020).

2 Similarity Measures and Methods for Categorical Data Clustering

Similarity measures play a key role in many multivariate methods, such as multidimensional scaling, outlier detection, or cluster analysis. They serve to determine (dis)similarity between objects characterized by vectors of values. Commonly used similarity measures for datasets with quantitative variables, e.g., the Euclidean distance or the Manhattan distance (see, e.g., Deza and Deza, 2009; Warrens, 2016) are well examined. A more complicated situation occurs when nominal variables are used. Currently, a researcher can choose from many not-well-explored approaches that were analyzed mainly in papers where they were introduced, often with outstanding results. However, there are only a few papers where these measures were independently compared and evaluated. For instance, Boriah et al. (2008); Chandola et al. (2009) assessed the selected similarity measures for categorical data in a domain of outlier detection, or Šulc (2016) analyzed the selected similarity measures in the area of HCA of categorical data.

The chapter extends the research by Šulc and Řezanková (2019). It presents the 16 similarity measures for categorical data, which will be used in the first experiment in Chapter 5. It contains four sections. In the first one, the similarity measures are presented. Second one describes their properties in greater detail. The third proposes a method for adjusting the similarity measures for the task of variable weighting. The last section presents the linkage methods that can be used with the similarity measures for categorical data.

2.1 Similarity Measures for Categorical Data

In this thesis, similarity measures for categorical data are defined as those which are determined to deal with nominal variables with more than two categories and do not need a dummy transformation. Table 2.1 presents the examined similarity measures overview with their full names, abbreviations that will be used throughout the thesis, and the papers where they were introduced.

Table 2.1: Overview of the examined similarity measures

Name	Abbreviation	Introduced by
Anderberg	AN	Anderberg (1973)
Burnaby	BU	Burnaby (1970)
Eskin	ES	Eskin et al. (2002)
Goodall 1	G1	Boriah et al. (2008) based on (Goodall, 1966)
Goodall 2	G2	Boriah et al. (2008) based on (Goodall, 1966)
Goodall 3	G3	Boriah et al. (2008)
Goodall 4	G4	Boriah et al. (2008)
Gambaryan	GA	Gambaryan (1964)
Inverse Occurrence Frequency	IOF	Spärck Jones (1972)
Lin	LIN	Boriah et al. (2008) based on (Lin, 1998)
Lin 1	LIN1	Boriah et al. (2008) based on (Lin, 1998)
Occurrence Frequency	OF	Spärck Jones (1972)
Simple Matching Coefficient	SM	Sokal and Michener (1958)
Smirnov	SV	Smirnov (1968)
Variable Entropy	VE	Šulc (2016)
Variable Mutability	VM	Šulc (2016)

All the investigated similarity measures are applied directly to the data matrix $\mathbf{X} = [x_{ic}]$, where $i = 1,2,\dots,n$ (n is the total number of objects); $c = 1,2,\dots,m$ (m is the total number of variables). The number of categories of the cth variable is denoted as K_c, an absolute frequency of a category equal to the value x_{ic} as $f(x_{ic})$, and a relative frequency as $p(x_{ic})$. An absolute frequency of the uth category ($u = 1,2,...K_c$) in the cth variable is marked as f_{cu} and the relative frequency as p_{cu}. The G1, G2, G3, and G4 measures use the adjusted relative frequencies according to the formula

$$\hat{p}^2 = \frac{f(f-1)}{n(n-1)}. \tag{2.1}$$

Fifteen of the examined similarity measures are calculated in two steps. In the first one, similarities between values of the cth variable for the ith and jth objects $S_c(x_{ic}, x_{jc})$ are computed separately. The S_c computation differs based on the match $x_{ic} = x_{jc}$ or mismatch $x_{ic} \neq x_{jc}$ of categories as described in Table 2.2.

In the second step, the similarity $S(\mathbf{x}_i, \mathbf{x}_j)$ between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$ is determined. The calculation depends on the similarity measure, as described in Table 2.3. The majority of the examined similarity measures use the first way (column *Type*), which is defined as the arithmetic mean of the similarities $S_c(x_{ic}, x_{jc})$. The measures LIN and LIN1 use the approach based on relative frequencies, and the measures GA and SV utilize the numbers of categories in the calculation.

The thesis examines one similarity measure, AN, which does not follow the two-step approach

Table 2.2: Formulas for the similarity measures for categorical data

Measure	S_c for $x_{ic} = x_{jc}$	S_c for $x_{ic} \neq x_{jc}$
BU	1	$\dfrac{\sum_{u=1}^{K_c} 2\ln(1-p_{cu})}{\ln \frac{p(x_{ic})p(x_{jc})}{(1-p(x_{ic}))(1-p(x_{ic}))} \sum_{u=1}^{K_c} 2\ln(1-p_{cu})}$
ES	1	$\dfrac{K_c^2}{K_c^2+2}$
G1	$1-\sum\limits_{q\in Q} \hat{p}^2\left(q\right)$	0
G2	$1-\sum\limits_{q\in Q} \hat{p}^2\left(q\right)$	0
G3	$1-\hat{p}^2\left(x_{ic}\right)$	0
G4	$\hat{p}^2\left(x_{ic}\right)$	0
GA	$-\left[p(x_{ic})\log_2 p(x_{ic}) + (1-p(x_{ic}))\log_2(1-p(x_{ic}))\right]$	1
IOF	1	$\dfrac{1}{1+\ln f\left(x_{ic}\right)\cdot \ln f\left(x_{jc}\right)}$
LIN	$2\ln p\left(x_{ic}\right)$	$2\ln\left(p\left(x_{ic}\right)+p\left(x_{jc}\right)\right)$
LIN1	$\sum\limits_{q\in Q} \ln p\left(q\right)$	$2\ln \sum\limits_{q\in Q} p\left(q\right)$
OF	1	$\dfrac{1}{1+\ln \frac{n}{f(x_{ic})}\cdot \ln \frac{n}{f(x_{jc})}}$
SM	1	0
SV	$2+\dfrac{n-f\left(x_{ic}\right)}{f\left(x_{ic}\right)}+\sum\limits_{u=1:u\neq x_{ic}}^{K_c} \dfrac{f_{cu}}{n-f_{cu}}$	$\sum\limits_{u=1:u\neq x_{ic},x_{jc}}^{K_c} \dfrac{f_{cu}}{n-f_{cu}}$
VE	$-\dfrac{1}{\ln K_c}\sum\limits_{u=1}^{K_c} p_{cu}\ln p_{cu}$	0
VM	$\dfrac{K_c}{K_c-1}\left(1-\sum\limits_{u=1}^{K_c} p_{cu}^2\right)$	0

Q is a subset of X containing all q fulfilling a certain condition. For the G1 measure, it is defined as $Q \subseteq X_c : \forall q, \hat{p}\left(q\right) \leq \hat{p}\left(x_{ic}\right)$, for the G2 measure as $Q \subseteq X_c : \forall q, \hat{p}\left(q\right) \geq \hat{p}\left(x_{ic}\right)$, and for the LIN1 measure as $Q \subseteq X_c : \forall q, p\left(x_{ic}\right) \leq p\left(q\right) \leq p\left(x_{jc}\right)$. For the VE measure, if $p_{cu} = 0$, the corresponding addend equals zero.

of similarity calculation as the measures in Table 2.2. Instead, it enables a researcher to determine the similarity $S\left(\mathbf{x}_i, \mathbf{x}_j\right)$ in one step using the formula:

$$S_{AN}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \frac{\sum_{c=1:x_{ic}=x_{jc}}^{m} \left(\frac{1}{p(x_{ic})}\right)^2 \frac{2}{K_c(K_c+1)}}{\sum_{c=1:x_{ic}=x_{jc}}^{m} \left(\frac{1}{p(x_{ic})}\right)^2 \frac{2}{K_c(K_c+1)} + \sum_{c=1:x_{ic}\neq x_{jc}}^{m} \left(\frac{1}{2p(x_{ic})p(x_{jc})}\right) \frac{2}{K_c(K_c+1)}}. \tag{2.2}$$

Table 2.3: Similarity definition between two objects

Type	Measures	$S\left(\mathbf{x}_i,\mathbf{x}_j\right)$
I	BU, ES, G1, G2, G3, G4, IOF, OF, SM, VE, VM	$S\left(\mathbf{x}_i,\mathbf{x}_j\right)=\frac{\sum_{c=1}^{m}S_c\left(x_{ic},x_{jc}\right)}{m}$
II	LIN, LIN1	$S\left(\mathbf{x}_i,\mathbf{x}_j\right)=\frac{\sum_{c=1}^{m}S_c\left(x_{ic},x_{jc}\right)}{\sum_{c=1}^{m}\left(\ln p\left(x_{ic}\right)+\ln p\left(x_{jc}\right)\right)}$
III	GA, SV	$S\left(\mathbf{x}_i,\mathbf{x}_j\right)=\frac{\sum_{c=1}^{m}S_c\left(x_{ic},x_{jc}\right)}{\sum_{c=1}^{m}K_c}$

In order to create a dissimilarity matrix, which is required by the majority of software solutions to perform HCA, it is necessary to compute dissimilarities $D\left(\mathbf{x}_i,\mathbf{x}_j\right)$ between all pairs of objects, which can be obtained from similarities $S\left(\mathbf{x}_i,\mathbf{x}_j\right)$. The dissimilarity calculation depends on the range of values a given similarity measure can attain; see Table 2.4. For most similarity measures that take on values between zero and one, the dissimilarity is defined as a complement to one. The second way is suitable for the similarity measures whose possible maximum is lower than one. The SV measure can take on values exceeding one; thus, it uses the third way of dissimilarity determination.

Dissimilarities derived from specific similarity measures, namely G1, G2, G3, G4, LIN1, VE, and VM, do not reach zero dissimilarity for two identical objects as might be expected. The reason is that these similarity measures use weights in the case of a match of categories that cannot reach the value one in most situations. Then, the dissimilarities, calculated using the formulas in Table 2.4, are higher than one. Fortunately, the positive dissimilarities for identical objects are not an obstacle for performing HCA since the dissimilarity matrix requirements are met, i.e., $D\left(\mathbf{x}_i,\mathbf{x}_i\right)=0$, $D\left(\mathbf{x}_i,\mathbf{x}_j\right)\geq 0$ and $D\left(\mathbf{x}_i,\mathbf{x}_j\right)=D\left(\mathbf{x}_j,\mathbf{x}_i\right)$ (Everitt et al., 2009). However, if one wants to guarantee zero dissimilarities between identical objects, the rest of the similarity measures in Table 2.2 and the AN measure can be used for this purpose.

Table 2.4: Dissimilarity definition between two objects

Type	Measures	$D\left(\mathbf{x}_i,\mathbf{x}_j\right)$
I	AN, BU, GA, G1, G2, G3, G4, SM, VE, VM	$D\left(\mathbf{x}_i,\mathbf{x}_j\right)=1-S\left(\mathbf{x}_i,\mathbf{x}_j\right)$
II	ES, IOF, LIN, LIN1, OF	$D\left(\mathbf{x}_i,\mathbf{x}_j\right)=\frac{1}{S\left(\mathbf{x}_i,\mathbf{x}_j\right)}-1$
III	SV	$D\left(\mathbf{x}_i,\mathbf{x}_j\right)=\frac{1}{S\left(\mathbf{x}_i,\mathbf{x}_j\right)}+1$

For the LIN and LIN1 measures, if $S\left(\mathbf{x}_i,\mathbf{x}_j\right)=0$, the $D\left(\mathbf{x}_i,\mathbf{x}_j\right)$ is calculated as $\max_{1\leq i,j\leq n}D\left(\mathbf{x}_i,\mathbf{x}_j\right)+1$.

2.2 Properties of the Examined Similarity Measures

This section provides a detailed description of the examined similarity measures. There are different ways of possible classification. For instance, the similarity measures can be divided according to the datasets' characteristics used for similarity definition. Table 2.5 shows which datasets' properties are utilized by the similarity measures to improve the definition of similarity between objects. Most similarity measures are based on one principle, for example, ES on the number of categories or IOF on absolute frequencies. The measures AN and SV utilize two different characteristics. The SM measure representing the simple matching approach does not use any dataset property to improve the similarity definition. Another classification can be based on how the similarity measures express the similarity between objects, i.e., in matches of categories, mismatches of categories, or both ways. In this section, the measures are classified according to the latter way.

Table 2.5: Classification of the similarity measures based on their principle

Based on ...	$D(\mathbf{x}_i, \mathbf{x}_j)$
none	SM
number of categories	ES, AN
absolute frequencies	IOF, OF, SV
relative frequencies	LIN, LIN1, GA, AN
adjusted relative frequencies	G1, G2, G3, G4
number of objects	SV

2.2.1 Reference similarity measure

The SM (simple matching) measure, introduced by Sokal and Michener (1958), only recognizes whether two categories match or not. Thus, it neglects important dataset characteristics, such as the number of categories or the absolute frequencies of categories, which could be utilized for better similarity determination. Nevertheless, the SM measure is considered the standard because it is still the most used similarity measure for categorical data. It is used as a part of other similarity measures as well, for instance, in the Gower similarity coefficient (see Gower, 1971), which serves for similarity determination between objects characterized by the mixed-type variables. Therefore, SM will be used as a reference similarity measure in this thesis.

Šulc (2016) found out that the hierarchical clustering with the SM measure provides the same clusters as the clustering with most of similarity measures determined for the binary-coded data, namely the measures Dice (Dice, 1945), Jaccard (Jaccard, 1912), Russel and Rao (Russel and Rao, 1940), Rogers and Tanimoto (Rogers and Tanimoto, 1960), Sokal and Michener (Sokal and Michener, 1958), Sokal and Sneath 1, Sokal and Sneath 2, Sokal and Sneath 4, Sokal and

Sneath 5 proposed by Sokal and Sneath (1963), Hamman (Hamann, 1961), Yule Q (Yule, 1900) and Yule Y (Yule, 1912). This topic was further analyzed by Cibulková et al. (2020), who studied similarity measures for binary data and found that the clusters obtained with 66 similarity measures for binary data can be classified into four groups that provide the same cluster assignment. The largest one, named *Euclid-based measures*, provides the same clusters as the SM measure.

2.2.2 Similarity measures evaluating the matches of categories

The G1 measure was introduced by Boriah et al. (2008) as a derivative of the original Goodall's measure (see Goodall, 1966). The first step of the calculation is the same as in the original measure. The second step is calculated as the arithmetic mean (Type I in Table 2.3) instead of a more complicated approach based on dependencies of used variables. The G1 measure uses the adjusted relative frequency of the observed category according to Eq. 2.1, and all the adjusted relative frequencies that are lower than the observed one. It takes on values from zero to $1 - \frac{2}{n(n-1)}$.

The G2 measure (Boriah et al., 2008) is another variant of Goodall's measure. It gives higher weight to matches of infrequent categories if the variable contains even rarer values. The measure uses adjusted relative frequency of categories that are equal or higher to the observed one. It takes on values from zero to $1 - \frac{2}{n(n-1)}$.

The G3 measure (Boriah et al., 2008) assigns higher weights to matches of infrequent categories. It does not consider frequencies of non-matching categories, and it takes on values from zero to $1 - \frac{2}{n(n-1)}$.

The G4 measure (Boriah et al., 2008) puts higher weights if the matching values are frequent. Similarly, as G3, it does not consider frequencies of non-matching categories. It takes on values from $\frac{2}{n(n-1)}$ to one.

The GA measure was proposed by Gambaryan (1964), and it assigns higher weights to matches of categories that are not rare nor frequent (Boriah et al., 2008). Its formula is related to Shannon information theory (Shannon, 1948). The measure takes on values in the range from zero to one.

The VE and VM measures (Šulc, 2016) are two variability-based similarity measures for categorical data. The measures are based on a new concept where the similarity between categories x_{ic} and x_{jc} by the cth variable is based on the within-cluster variability of the cth variable. Let us assume two different variables, one with high variability, i.e., with approximately evenly distributed categories, and the other with small variability, where one category is dominant, and the rest are sparsely represented. Both similarity measures praise the match of two categories in the variable with the high variability because it is rarer than the match in the low-variability variable. For individual variables variability definition, the VE measure uses the entropy, and the VM measure uses the nominal variance, also known as the mutability or the

Gini coefficient. Both of these variability measures use relative frequencies of all categories.

The VE and VM measures were developed to be simple, not computationally demanding, and theory-based. They take on values from zero to one in the case of a match and zero otherwise. In the case of a match, the zero value can be attained if a given variable takes on only one category for all cases in a dataset. In this case, the variable becomes redundant for clustering, and the zero value for the appropriate weight is fully justified. The value one occurs in matches if all categories by a given variable are equally distributed. When determining the overall similarity between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$, both the similarity measures can take on values from the range zero to one as well. Zero similarity is obtained if there is no match over all variables (or they contain only one category, as mentioned before). The value one is achieved if there are matches over all variables and all of them have the maximum possible variability.

2.2.3 Similarity measures evaluating the mismatches of categories

The BU measure (Burnaby, 1970) is based on the information theory where the rarely observed values are considered more informative. Thus, the measure assigns low similarity to mismatches of frequent categories and high similarity to mismatches of rare categories. For mismatches, it takes on values from $\frac{n\ln\left(1-\frac{1}{n}\right)}{n\ln\left(1-\frac{1}{n}\right)-\ln(n-1)}$ to one.

The ES measure was originally proposed by Eskin et al. (2002) as a distance measure. As a similarity measure, it uses the number of categories K_c of the cth variable to determine the similarity between two categories. It assigns higher weights to mismatches for variables with more categories. It takes on values from $\frac{2}{3}$ to $1-\frac{n^2}{n^2+2}$.

The IOF (inverse occurrence frequency) measure, proposed by Spärck Jones (1972), uses the absolute frequencies of the observed categories to achieve a more precise similarity definition between two categories in the case of mismatches. Initially, the measure was introduced in the information retrieval field, where it determined a relative number of documents containing a specific word. The original measure was designed to deal only with binary variables; later, it was adjusted to deal with nominal variables. IOF assigns higher weights to less frequent mismatches, and it takes on values from $\frac{1}{1+\left(\ln\frac{n}{2}\right)^2}$ to one.

The OF (occurrence frequency) measure (Spärck Jones, 1972) is based on the same principle as IOF, but it differs in the weight system. It assigns higher weights to more frequent mismatches, and it takes on values from $\frac{1}{1+(\ln n)^2}$ to $\frac{1}{1+(\ln 2)^2}$.

2.2.4 Similarity measures evaluating the matches and mismatches of categories

The AN measure, proposed by Anderberg (1973), is based on the idea that the rare values are the key to determining the similarity. Therefore, it assigns a high weight to matches of rare values and a lower weight to mismatches of rare values. It takes on values from zero to one.

The LIN measure was introduced by Boriah et al. (2008) based on the framework outlined by Lin (1998). It represents an information-theoretic definition of similarity based on relative frequencies. It assigns higher weights to more frequent categories in the case of the match and lower weights to less frequent categories in the case of the mismatch. In the case of a match, it takes on values from $-2\ln n$ to zero; in the case of a mismatch, it can attain values from $-2\ln\frac{n}{2}$ to zero.

The LIN1 measure is another measure proposed by Boriah et al. (2008) based on Lin's framework. It has a complex weight system. Boriah et al. (2008) described it in the following way: "*It gives lower weight to mismatches if either of the mismatching values is very frequent, or if there are several values that have a frequency in between those of mismatching values. Higher weight is given when there are mismatches on infrequent values, and there are a few other infrequent values. For matches, lower weight is given for matches on frequent categories or matches on values that have many other values of the same frequency. Higher weight is given to matches on rare values.*" The range for matches and mismatches is the same by this measure, which takes on values from $-2\ln n$ to zero.

The SV measure, proposed by Smirnov (1968), represents a probabilistic approach to determining the similarity between two objects. It utilizes the absolute frequencies of the match or mismatch and the absolute frequencies of all categories of a given variable. It assigns higher weights to rare matches. For matches, it takes on values from two to $2n$, and for mismatches, from zero to $\frac{n}{2}-1$.

2.3 Variable Weighting in Hierarchical Clustering

Variable weighting is typically helpful if one of the variables is too influential that it substantially affects the created clusters. Setting a lower weight to such a variable can diminish its importance in clustering. In some situations, a researcher has external information about the examined problem. Variable weighting can help him set the variable importance accordingly to the external information provided. However, no variable weighting approach is developed for the hierarchical clustering of categorical data. Therefore, a simple method for variable weighting in HCA is proposed in this thesis.

In HCA of categorical data, the proposed variable weighting method is applied at the lowest level of similarity determination, i.e., at the level where the similarity $S_c\left(x_{ic}, x_{jc}\right)$ between two categories of the cth variable is calculated, see Table 2.2. Let us have a vector $\mathbf{w}$ of the length m containing weights for all variables in a dataset, which is restricted to contain values in a range from zero to one. Then, the weighted similarity SW_c between the categories x_{ic} and x_{jc} by the cth variable can be obtained using the formula

$$SW_c\left(x_{ic}, x_{jc}\right) = \frac{S_c\left(x_{ic}, x_{jc}\right) \cdot w_c}{\sum_t w_t}, \tag{2.3}$$

where $t = 1, 2, \ldots, c, \ldots, m$. The rest of the dissimilarity matrix calculation remains unchanged.

The presented approach to the variable weighting can be applied to the Type I and Type II similarity measures in Table 2.3. Thus, the variable weights cannot be currently applied to the measures AN, GA, and SV.

2.4 Methods of Hierarchical Cluster Analysis

Hierarchical cluster analysis (see, e.g., Everitt et al., 2009; Hennig et al., 2015), is based either on an agglomerative or divisive clustering process. Usually, the agglomerative process is used. This type of clustering considers each object a cluster at the start of the clustering process. Then, the two most similar clusters are joined into a new one, and the dissimilarity matrix is recalculated. This algorithm repeats itself until there is one cluster left. Then, the created hierarchy of clusters can be cut at any point to get the desired number of clusters. The linkage method plays an important role in cluster hierarchy creation because it defines the value of dissimilarity, which will be used after merging two clusters.

Compared to the quantitative data, where various linkage methods are available, the number of linkage methods for categorical data is limited. The reason is that some linkage methods for the quantitative data work with concepts that do not apply to categorical data, such as cluster centroids by the centroid method or the within-cluster variance by Ward's method. Although there are ways to deal with these limitations, (see, e.g., Strauss and von Maltitz, 2017; Chen and Guo, 2014), for a comparison of the examined similarity measures in this thesis, three commonly used linkage methods using between-cluster distances based on dissimilarities are used, namely, the single, complete and average linkage methods.

The *single linkage method* (SLM) defines the dissimilarity between clusters C_g and C_h as the distance between the two closest objects of two different clusters. This linkage is often associated with a problem called *chaining phenomenon*, in which clusters are merged based on their closest elements, even though some are very distant. However, this linkage method performs better than other linkages when the clusters are not spherical or elliptical in shape. The *complete linkage method* (CLM) considers a dissimilarity between two clusters $D(C_g, C_h)$ as the dissimilarity between two farthest objects from these clusters. This between-cluster distance usually provides compact clusters with approximately equal diameters. However, it is sensitive to outliers. The *average linkage method* (ALM) takes the average pairwise dissimilarity between objects in two clusters. It is a robust method, considered a compromise between the single and the complete linkages. An overview of the used linkage methods formulas occurs in Table 2.6, where n_g and n_h are the numbers of objects in clusters C_g and C_h.

To illustrate a relationship between dissimilarity matrix values and the presented linkage methods in a graphical way, an example based on the procedure by Anselin et al. (2006) is shown. Table 2.7 contains the original dissimilarity matrix, where each object represents a standalone cluster. Objects 3 and 5 are the most similar since their mutual dissimilarity

Table 2.6: An overview of the linkage methods

linkage method	$D(C_g, C_h)$
single	$\min_{\mathbf{x}_i \in C_g, \mathbf{x}_j \in C_h} D(\mathbf{x}_i, \mathbf{x}_j)$
complete	$\max_{\mathbf{x}_i \in C_g, \mathbf{x}_j \in C_h} D(\mathbf{x}_i, \mathbf{x}_j)$
average	$\frac{1}{n_g n_h} \sum_{\mathbf{x}_i \in C_g} \sum_{\mathbf{x}_j \in C_h} D(\mathbf{x}_i, \mathbf{x}_j)$

equal to 0.466 is the lowest from all pairs of objects (the field outlined in purple). Thus, these two objects will be merged in the next calculation step, and the dissimilarity matrix will be recalculated. The selected linkage method determines the way of dissimilarity matrix recalculation.

Table 2.7: Original dissimilarity matrix – step 1 (example)

	1	2	3	4	5	6
1	0	0.617	1.000	0.796	0.889	0.907
2		0	0.841	0.637	0.730	0.748
3			0	0.841	0.466	0.654
4				0	0.730	0.594
5					0	0.841
6						0

Table 2.8 shows the second step of dissimilarity matrix calculation when SLM is used. Due to merging, the table contains one less row and column. The highlighted cells indicate the newly calculated dissimilarities for the new row (and column) "3+5". The values are calculated according to the formula for SLM in Table 2.6. For instance, the dissimilarity 0.889 between the clusters "1" and "3+5" is defined as the minimum of the distances 1.000 (between clusters "1" and "3") and 0.889 (between clusters "1" and "5") in Table 2.7. After that, the procedure is repeated until only one cluster is left. All the calculations steps of SLM occur in Table I in Appendix A.

Table 2.8: Single linkage method – step 2 (example)

	1	2	3+5	4	6
1	0	0.617	0.889	0.796	0.907
2		0	0.730	0.637	0.748
3+5			0	0.730	0.654
4				0	0.594
6					0

The second step of CLM is presented in Table 2.9. According to the formula for CLM in Table

2.6, the dissimilarity 1.000 between the clusters “1” and “3+5” is defined as the maximum of the distances 1.000 (between clusters “1” and “3”) and 0.889 (between clusters “1” and “5”) in Table 2.7.

Table 2.9: Complete linkage method – step 2 (example)

	1	2	3+5	4	6
1	0	0.617	1.000	0.796	0.907
2		0	0.841	0.637	0.748
3+5			0	0.841	0.654
4				0	0.594
6					0

Table 2.10 contains the second step of the ALM calculation, which uses the weighted average of the dissimilarities for two merged clusters, as stated in Table 2.6. Since this is the second step of the calculation, where the original objects are merged, the dissimilarity 0.945 between the clusters “1” and “3+5” is defined as the simple arithmetic mean of the distances 1.000 (between clusters “1” and “3”) and 0.889 (between clusters “1” and “5”). However, in the following steps of calculations, the arithmetic mean needs to be weighted by the number of objects in the merged clusters.

Table 2.10: Average linkage method – step 2 (example)

	1	2	3+5	4	6
1	0	0.617	0.945	0.796	0.907
2		0	0.785	0.637	0.748
3+5			0	0.785	0.747
4				0	0.594
6					0

3 Cluster Evaluation Criteria

Cluster analysis comprises a set of often very distinct approaches which have in common that they produce a vector with cluster memberships for the objects in a dataset. The created clusters can considerably differ using different algorithms or methods. Therefore, comparing several cluster assignments using at least one evaluation criterion can help a researcher choose the most suitable cluster partition.

This chapter is partly based on the paper written by Šulc et al. (2018), and it is divided into two sections. The first section deals with external evaluation criteria suitable for the research based on simulation studies, such as those presented in the experiment. The second one offers one of the most extensive overviews of internal evaluation criteria for categorical data. Moreover, two new internal criteria are proposed there.

3.1 External Evaluation Criteria

This section presents two external indices commonly used by many researchers, namely the Rand index and the adjusted Rand index. Their use is demonstrated in the example illustrated in Figure 3.1, see (Manning et al., 2008).

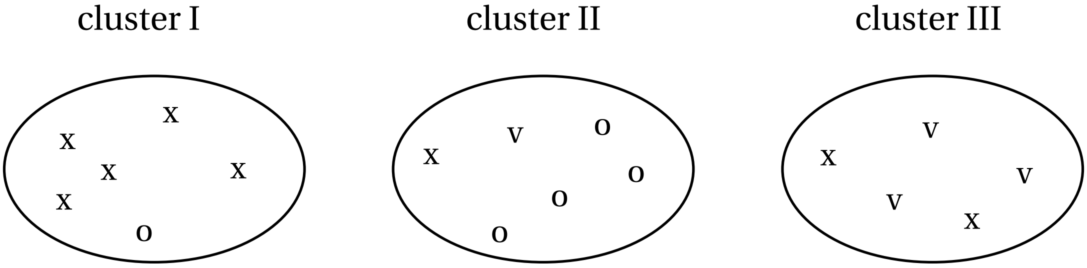


Figure 3.1: An example of object assignment into three clusters

The commonly used Rand and adjusted Rand indices are based on an approach, which can be interpreted as a series of decisions for each of $n \times (n-1)/2$ pairs of less or more similar objects in a dataset. Four possible types of decisions for pairs assignment can be made. A true positive (TP) decision expresses the number of all combinations of pairs correctly assigned to the same clusters. A true negative (TN) decision is the number of all combinations of pairs that are correctly assigned into different clusters. There can occur two types of error decisions. A false positive (FP) decision represents the number of combinations of pairs of dissimilar objects that are falsely assigned to the same cluster, and a false negative (FN) decision is the number of combinations for similar pairs that are falsely assigned into two different clusters. These four types of decisions can be summarized in the form of the 2 × 2 confusion table, see Table 3.1.

Table 3.1: Confusion table

	Actual	
Predicted	Positive	Negative
Positive	TP	FP
Negative	FN	TN

The *Rand* (RI) index (Rand, 1971), with the formula

$$RI = \frac{TP + TN}{TP + FP + FN + TN}, \tag{3.1}$$

represents a ratio of correctly assigned objects, both positively and negatively, out of all possible pairs. It takes on values from zero to one.

Although Eq. 3.1 seems to be simple, the calculation of this index is relatively complex. For instance, if one wants to calculate RI based on Figure 3.1, a cooccurrence matrix needs to be created first; see Table 3.2.

Table 3.2: Cooccurrence matrix (example)

	Actual		
Prediction	I	II	III
x	5	1	2
o	1	4	0
v	0	1	3

The quantity TP can be calculated as the sum of all combination numbers (choose two) in the matrix higher than one, i.e., $TP = \binom{5}{2} + \binom{4}{2} + \binom{3}{2} + \binom{2}{2} = 20$. Then, the quantity $TP + FP$ can be calculated as the sum of combination numbers of the row sums of the matrix, i.e., $TP + FP = \binom{6}{2} + \binom{6}{2} + \binom{5}{2} = 40$. Analogically, the quantity $TP + FN$ is determined as $TP + FN = \binom{8}{2} + \binom{5}{2} + \binom{4}{2} = 44$. Finally, the quantity $TP + FP + FN + TN$ is calculated as the total number of objects choose two, i.e., $TP + FP + FN + TN = \binom{17}{2} = 136$. Based on the calculated quantities, the frequencies in the confusion table can be obtained; see Table 3.3. Thus, the RI value, based on Eq. (3.1), is 0.676.

Table 3.3: Confusion table (example)

	Actual	
Predicted	Positive	Negative
Positive	20	20
Negative	24	72

The *adjusted Rand index* (ARI) (Hubert and Arabie, 1985) can also be used for a comparison of two membership partitions. Compared to the original Rand index, it is corrected for a chance. Theoretically, it takes on values between minus one and one, but only slightly negative values usually occur. The value one indicates identical partitions, and zero the randomly assigned partitions. The positive values express that the predicted values are better than the random chance. It can be expressed as

$$ARI = \frac{TP - \frac{(TP+FP)(TP+FN)}{TP+FP+FN+TN}}{\frac{(TP+FP)+(TP+FN)}{2} - \frac{(TP+FP)(TP+FN)}{TP+FP+FN+TN}}. \tag{3.2}$$

Based on the quantities in Table 3.3, the ARI value is 0.243.

3.2 Internal Evaluation Criteria

The internal evaluation criteria are usually constructed to satisfy the principles of *compactness* or *separation* of the created clusters (Liu et al., 2010; Zhao et al., 2005). The compactness declares the similarity of objects in clusters. It can be expressed by the closeness of objects in clusters, low within-cluster variability, or the high value of a likelihood function. The separation measures cluster distinctness. It is usually represented by the high between-cluster variability or the high distance between clusters. Some criteria are built on both principles, some just on one. This section classifies the criteria into variability-, likelihood-, and distance-based.

3.2.1 Evaluation criteria based on the variability

In categorical data, variability can be measured by mutability (the Gini coefficient) or entropy. Whereas mutability is a more straightforward measure whose forms can be interpreted as the probabilities of distinctions, entropy is a more complex approach that can be interpreted as the average information needed to distinguish all the information in the data (Ellerman, 2013).

The mutability of the cth variable ($c = 1, 2, \dots, m$) in the gth cluster ($g = 1, 2, \dots, k$) can be calculated as

$$G_{gc} = 1 - \sum_{u=1}^{K_c} \left(\frac{n_{gcu}}{n_g} \right)^2 \tag{3.3}$$

and the entropy as

$$H_{gc} = -\sum_{u=1}^{K_c} \left(\frac{n_{gcu}}{n_g} \ln \frac{n_{gcu}}{n_g} \right). \tag{3.4}$$

In the formulas, n_g is the number of objects in the gth cluster, n_{gcu} is the number of objects in the gth cluster by the cth variable with the uth category ($u = 1, 2, \dots, K_c$), and K_c is the number of categories by the cth variable. The corresponding addend equals zero if $n_{gcu} = 0$.

The *within-cluster mutability* (WCM) of the whole dataset with m variables broken down into k clusters is defined as

$$WCM(k) = \sum_{g=1}^{k} \frac{n_g}{n} \sum_{c=1}^{m} G_{gc} \tag{3.5}$$

and the *within-cluster entropy* (WCE) as

$$WCE(k) = \sum_{g=1}^{k} \frac{n_g}{n} \sum_{c=1}^{m} H_{gc}, \tag{3.6}$$

where m is the total number of variables and n is the number of objects in a dataset. The special cases $WCM(1)$ and $WCE(1)$ represent total variability in a dataset expressed by mutability respective entropy, and the differences $WCM(1)–WCM(k)$ and $WCE(1)–WCE(k)$ the between-cluster mutability respective entropy in the k-cluster solution.

The standardized forms of the WCM and WCE criteria take on values in a range from zero to one and have the formulas

$$WCM_s(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} \frac{K_c}{K_c - 1} G_{gc} \tag{3.7}$$

and

$$WCE_s(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} \frac{H_{gc}}{\ln K_c}. \tag{3.8}$$

The standardized forms are helpful when comparing the variability of datasets with different numbers of categories in categorical variables. When multiplied by 100, the criteria's values can be easily interpreted as percentages.

Several evaluation criteria are based on Eq. (3.5) and Eq. (3.6), namely, *pseudo-F coefficients* *based on mutability* (PSFM), *pseudo-F coefficients based on entropy* (PSFE), *category utility* (CU), and *category information* (CI).

The PSFM and PSFE criteria, proposed by Řezanková et al. (2011), are the modifications of the pseudo-F index (Caliński and Harabasz, 1974) for datasets with categorical variables. This index is defined as the ratio of the weighted between-cluster variability (separation principle) and the weighted within-cluster variability (compactness principle) in the k-cluster solution. The PSFM criterion is defined as

$$PSFM(k) = \frac{(n-k)\,[WCM(1) - WCM(k)]}{(k-1)\,WCM(k)} \tag{3.9}$$

and the PSFE criterion as

$$PSFE(k) = \frac{(n-k)\,[WCE(1) - WCE(k)]}{(k-1)\,WCE(k)}. \tag{3.10}$$

The maximal value across all the examined cluster solutions suggests the optimal number of clusters for both criteria.

Corter and Gluck (1992) proposed two evaluation criteria that measure category goodness, namely category utility (CU) and category information (CI). The CU criterion measures the overall quality of partitioning clustered objects into clusters. It is based on the principle that the ability to predict a given category is higher if the cluster memberships are known than unknown. Category utility summarizes the gain of conditional probabilities (calculated as relative frequencies) with the known cluster membership compared to the unconditional approach, as can be seen in the formula

$$CU(k) = \frac{1}{k}\sum_{g=1}^{k}\frac{n_g}{n}\left[\sum_{c=1}^{m}\sum_{u=1}^{K_c}\left(\frac{n_{gcu}}{n_g}\right)^2 - \sum_{c=1}^{m}\sum_{u=1}^{K_c}\left(\frac{n_{cu}}{n}\right)^2\right], \tag{3.11}$$

where $\frac{n_{gcu}}{n_g}$ represents the conditional relative frequency and $\frac{n_{cu}}{n}$ the unconditional one. The CI criterion is based on concepts of information theory (Shannon, 1948). It expresses the expected reduction of information needed to be provided if the cluster membership is known as follows

$$CI(k) = \frac{1}{k}\sum_{g=1}^{k}\frac{n_g}{n}\left[\sum_{c=1}^{m}\sum_{u=1}^{K_c}\left(\frac{n_{gcu}}{n_g}\ln\frac{n_{gcu}}{n_g}\right) - \sum_{c=1}^{m}\sum_{u=1}^{K_c}\left(\frac{n_{cu}}{n}\ln\frac{n_{cu}}{n}\right)\right]. \tag{3.12}$$

From Eq. (3.11) and Eq. (3.12), one can derive that the CU and CI criteria can also be expressed using the between-cluster mutability respective entropy (the separation principle), which can be defined as

$$CU(k) = \frac{1}{k}\,[WCM(1) - WCM(k)] \tag{3.13}$$

and

$$CI(k) = \frac{1}{k}\left[WCE(1) - WCE(k)\right]. \tag{3.14}$$

Thus, the CU and CI formulas can be interpreted in two ways satisfying two different principles to cluster quality definition. Moreover, the CI criterion is equivalent to *mutual information* (MI) introduced by Shannon (1948), as it is shown by the formula

$$CI(k) = \frac{1}{k}\sum_{c=1}^{m}\left[\sum_{g=1}^{k}\frac{n_g}{n}\left[\sum_{u=1}^{K_c}\left(\frac{n_{gcu}}{n_g}\ln\frac{n_{gcu}}{n_g}\right) - \sum_{u=1}^{K_c}\left(\frac{n_{cu}}{n}\ln\frac{n_{cu}}{n}\right)\right]\right] = \frac{1}{k}\sum_{c=1}^{m} MI_c(k) = MI(k). \tag{3.15}$$

The CU and CI criteria take on non-negative values, and their maximal values indicate the optimal number of clusters.

The BK (best k) index (Chen and Liu, 2009) is based on the incremental expected entropy, which represents the information gain between the expected entropy in the k-cluster and $(k+1)$-cluster solutions. The expected incremental entropy can be expressed as

$$I(k) = H_E(k) - H_E(k+1), \tag{3.16}$$

where H_E is the expected entropy in a dataset with the formula

$$H_E(k) = \frac{1}{k}\sum_{g=1}^{k}\frac{n_g}{n}\sum_{c=1}^{m}\frac{H_{gc}}{\ln K_c}. \tag{3.17}$$

The BK index is defined as the second-order difference of the incremental expected entropy of the dataset with k clusters.

$$BK(k) = \Delta^2 I(k) = \left[I(k-1) - I(k)\right] - \left[I(k) - I(k+1)\right]. \tag{3.18}$$

The highest value of the index indicates the optimal number of clusters.

3.2.2 Evaluation criteria based on likelihood

The Bayesian information criterion (BIC) and the Akaike information criterion (AIC) for the categorical data were presented in the SPSS technical report (SPSS, Inc., 2001) and further described by Bacher et al. (2004). Both criteria maximize the likelihood function (compactness principle) while inflicting a penalty on complex models (Biem, 2003). In terms of clustering, they penalize solutions with more clusters. Both criteria indicate the optimal number of clusters by their minimal value.

The modification of the BIC index (Schwarz, 1978) for categorical data can be calculated using the formula

$$BIC(k) = 2\sum_{g=1}^{k} n_g \sum_{c=1}^{m} H_{gc} + k\sum_{c=1}^{m}(K_c - 1)\ln n \tag{3.19}$$

and the modification of the AIC index (Akaike, 1973) is defined as

$$AIC(k) = 2\sum_{g=1}^{k} n_g \sum_{c=1}^{m} H_{gc} + 2k\sum_{c=1}^{m}(K_c - 1). \tag{3.20}$$

3.2.3 Evaluation criteria based on distances

The *silhouette* index (SI) (Rousseeuw, 1987), also known as the *average silhouette width*, is defined as the average relative difference between between-cluster (separation principle) and within-cluster distances (compactness principle). It takes on values from –1 to 1. The high positive values indicate well-separated clusters with low within-cluster and high between-cluster distances. The values close to zero or the negative ones suggest badly separated clusters. The maximal value of the criterion across all the examined cluster solutions indicates the optimal number of clusters. It can be expressed as

$$SI(k) = \frac{1}{n}\sum_{i=1}^{n} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{3.21}$$

where $a(i)$ is the average dissimilarity of the ith object to the other objects in the same cluster, and $b(i)$ is the minimum average dissimilarity of the ith object to other objects in any cluster not containing the ith object.

The Dunn index (DI) (Dunn, 1973) is calculated as a ratio of the smallest within-cluster distance (compactness principle) to the largest between-cluster distance (separation principle). It takes on values from zero to infinity. The highest value indicates the optimal cluster solution. For the cluster solution with k clusters, it can be expressed by the formula

$$DI(k) = \min_{1\leq g<h\leq k}\left(\frac{D(C_g, C_h)}{\max_{1\leq v\leq k} diam(C_v)}\right), \tag{3.22}$$

where $D(C_g, C_h)$ is the distance between the gth and hth clusters (expressed by a given linkage method), and $diam(C_v)$ is the maximal distance expressed by a given similarity measure between two objects in the vth cluster.

The values of distance-based criteria, SI and DI, depend on the similarity measure used for calculating the distance matrix. Thus, if two different similarity measures do not have a monotonous relationship, the resulting clusters are not directly comparable using the distance-based criterion. This issue is more severe in the categorical data, where the distances are not defined as straightforwardly as in quantitative data. The influence of the used similarity measures on the evaluation criteria values will be examined in Chapter 6.

3.2.4 New variability-based evaluation criteria

This thesis proposes two modifications of Hartigan's rule (Hartigan, 1975) for categorical data, with the names *Hartigan mutability* (HM) and *Hartigan entropy* (HE). The original index for quantitative data is one of the variability-based methods that a researcher can use for the optimal number of clusters determination in the k-means algorithm. According to Chiang and Mirkin (2010), it is one of the best methods developed for this purpose. It is determined as a ratio of within-cluster variabilities, expressed as sums of squares, in the k and $k+1$ solutions. The resulting ratio approximately follows the F distribution with n and n–k–1 degrees of freedom. The algorithm runs on a series of cluster solutions (starting with $k = 2$). It stops when it finds k, whose incremental decrease of the criterion's value is sufficiently large (usually higher than 10).

The newly proposed criteria HM and HE are based on the same principle as the original Hartigan's rule. They differ in using mutability respective entropy instead of the sum of squares. However, they still evaluate the marginal gain in cluster compactness when the number of clusters increases. The criteria are constructed as a ratio of the within-cluster variability in the k-cluster and $(k+1)$-cluster solutions. In the case of HM, the within-cluster variability is expressed by WCM (Eq. (3.5)) using the formula

$$HM(k) = \left(\frac{WCM(k)}{WCM(k+1)} - 1\right)(n-k-1). \tag{3.23}$$

HE utilizes WCE (Eq. (3.6)), which can be written as

$$HE(k) = \left(\frac{WCE(k)}{WCE(k+1)} - 1\right)(n-k-1). \tag{3.24}$$

Both criteria take on values from zero to infinity, and their minimal values express the optimal number of clusters over the examined cluster solutions. Similarly, as the original Hartigan's rule, the proposed criteria are expected to perform well in the optimal number of clusters determination. However, they can also be used when judging cluster quality. The performance of the HM and HE criteria will be examined in the conducted experiment.

4 The nomclust Package

This chapter presents the second generation of the `nomclust` R package, proposed by Šulc et al. (2022). The package was developed for HCA of categorical data. It completely covers the hierarchical clustering process, from dissimilarity matrix calculation, over the choice of a clustering method, to the evaluation of the final clusters. The whole clustering process utilized by the package uses similarity measures, clustering methods, and evaluation criteria developed solely for categorical data, which makes this package unique.

Compared to the first generation of the package (Šulc and Řezanková, 2015), the second generation represents a considerable step forward. Among the most noteworthy changes can be named the wholly redesigned evaluation criteria based on concepts of *variability*, *likelihood* (adjusted for categorical data), and *distance*. Some of these criteria were examined by Šulc et al. (2018). Next, the issue with the low calculation speed of hierarchical clustering was addressed by rewriting the critical parts of the code into C++ to substantially increase the clustering speed. Finally, the support for S3 generic functions and the ability to draw dendrograms and values of evaluation criteria was added to the package. The package is available on the *Comprehensive R Archive Network* (CRAN) web site[1].

4.1 Methods Used in nomclust 2.0

The process of hierarchical clustering consists of three main areas: calculating a dissimilarity matrix using a selected (dis)similarity measure, creating a hierarchy of clusters by a chosen linkage method, and optionally evaluating created clusters using one or more evaluation criteria. In this section, the theoretical background for all these areas in the `nomclust` package is described.

[1] https://cran.r-project.org/package=nomclust

4.1.1 Dissimilarity matrix

The `nomclust` package contains 16 similarity measures for categorical data; 14 of them were summarized by Boriah et al. (2008), and two of them by Šulc and Řezanková (2019). Their formulas can be found in Section 2.1. In the package, dissimilarity matrices for all the available similarity measures can be calculated separately from other analysis steps, e.g., as an input for other R packages. Table 4.1 presents the function calls for these measures and their most important properties.

4.1.2 Linkage methods

After a dissimilarity matrix is calculated, a hierarchy of clusters needs to be created. The `nomclust` package uses *agglomerative* clustering from the `cluster` package. Since it is determined to cluster the categorical data only, only three linkage methods using between-cluster distances based on dissimilarities suitable for categorical data are available, namely *average*, *complete*, and *singe* linkage methods; see Section 2.4.

4.1.3 Evaluation criteria

The resulting clusters can be evaluated up to 13 evaluation criteria presented in Section 3.2. Variability-based coefficients based on mutability and entropy usually do not differ very much. They are included in the package to provide two independent ways of variability computation. Substantial differences between mutability- and entropy-based coefficients should attract the researcher's attention.

The overview of the used evaluation criteria with their important properties occurs in Table 4.2. The column *Optimum* indicates if the maximal or minimal value of the criterion indicates the optimal number of clusters or if one should rely on the elbow of the curve with the evaluation criterion values. The *elbow method* (Thorndike, 1953) allows a researcher to choose the optimal number of clusters subjectively. It requires a researcher to find a point representing a certain number of clusters where the curve of criterion values visibly bends from a high to a low slope. At this point, increasing the number of clusters by one does not bring a sufficient decrease in the total within-cluster variability. However, the elbow does not have to be visible in some situations. Then, it is recommended to use a different evaluation criterion.

The recommended number of clusters may differ for different evaluation criteria in the package. There are no strict guidelines on how to proceed in such a case. One can use the number of clusters recommended by most of the evaluation criteria. It is also good to inspect one lower and one higher number of clusters than the recommended one. A researcher should pay attention to the situation when almost every criterion suggests a different number of clusters. It may indicate that the clusters in the data are badly separated or that there are no clusters at all.

Table 4.1: Function calls and properties for the similarity measures in the `nomclust` package

Measure	Function call	It uses ...
AN	`anderberg()`	the number of categories and relative frequencies of the observed categories; assigns the high weight to matches of rare values and the lower weight to mismatches of rare values.
BU	`burnaby()`	relative frequencies of all categories; assigns low similarity to mismatches of frequent categories and high similarity to mismatches of rare categories.
ES	`eskin()`	the number of categories; assigns higher weights to mismatches by variables with higher number of categories.
G1	`goodall1()`	relative frequencies of selected categories that are lower than the observed one; assigns higher weights to infrequent categories in the case of match.
G2	`goodall2()`	relative frequencies of selected categories that are higher than the observed one; assigns higher weights to infrequent categories in the case of match.
G3	`goodall3()`	relative frequencies of the observed categories; assigns higher weights to infrequent categories in the case of match.
G4	`goodall4()`	relative frequencies of the observed categories; assigns higher weights to frequent categories in the case of match.
GA	`gambaryan()`	relative frequencies of the observed categories; assigns higher weights to matches of categories that are not rare nor frequent.
IOF	`iof()`	absolute frequencies of the observed categories; assigns higher weights to infrequent mismatches of categories.
LIN	`lin()`	relative frequencies of the observed categories; assigns higher weights to more frequent categories in the case of the match, and lower weights to less frequent categories in the case of the mismatch.
LIN1	`lin1()`	relative frequencies of the selected categories; a complex weight system, see (Boriah et al., 2008).
OF	`of()`	absolute frequencies of the observed categories; assigns higher weights to frequent mismatches of categories.
SM	`sm()`	the simple matching approach; no weight system; a reference measure.
SV	`smirnov()`	absolute frequencies of the observed categories and the total frequency; assigns higher weights to rare matches.
VE	`ve()`	relative frequencies of all categories; assigns higher weights to matches in variables with high variability expressed by the entropy.
VM	`vm()`	relative frequencies of all categories; assigns higher weights to matches in variables with high variability expressed by the mutability.

Table 4.2: Overview of the internal evaluation criteria in the `nomclust` package

Criterion	Optimum	Properties
WCM	elbow	Standardized form. Suitable for measurement of the cluster quality in different cluster solutions; mutability-based.
WCE	elbow	Standardized form. Suitable for measurement of the cluster quality in different cluster solutions; entropy-based.
PSFM	max	Categorical alternative to the pseudo F-index; mutability-based.
PSFE	max	Categorical alternative to the pseudo F-index; entropy-based.
CU	max	The weighted between-cluster variability; mutability-based.
CI	max	The weighted between-cluster variability; entropy-based.
BK	max	Defined as the second-order difference of the incremental entropy of the dataset with k clusters; entropy-based.
BIC	min	Categorical alternative to BIC; entropy-based.
AIC	min	Categorical alternative to AIC; entropy-based.
SI	max	Based on a comparison of the within-cluster and between-cluster distances.
DI	max	Based on a comparison of the within-cluster and between-cluster distances.
HM	min	Categorical alternative to the Hartigan's rule; mutability-based.
HE	min	Categorical alternative to the Hartigan's rule; entropy-based.

4.1.4 Optimization

Dissimilarity matrix calculation is the most time-demanding part of the hierarchical clustering process because the number of values in this matrix, which need to be calculated, increases with the square of the number of observations. The calculation can take minutes, even with relatively small datasets (several thousand objects). Therefore, hierarchical cluster analysis is usually recommended for up to 10,000 rows in clustered datasets. Apart from the number of objects, the dissimilarity matrix calculation time depends on a used similarity measure, dataset properties (number of variables or categories), and computer speed.

The high complexity of the dissimilarity matrices calculation is caused by many loops in their code that are not processed efficiently in the R language. Therefore, critical parts of the code of all the used similarity measures were rewritten to the C++ language, which handles the loops more effectively. The implementation of the C++ code was performed using the `Rcpp` package (Eddelbuettel and Francois, 2011).

To assess the effect of the C++ language implementation, an experiment on 60 generated datasets[1] was conducted. For the experiment, the same datasets were used as those used by Šulc and Řezanková (2019). The average calculation times of clustering with a certain

[1] The datasets contained four numbers of variables (four, six, eight, ten), three ranges of categories (2–4, 2–6, 6–10), and the number of cases varied from 300 to 700. Each of the datasets contained four clusters with a middle between-cluster distance. All the combinations were replicated five times.

Table 4.3: Performance comparison of the first and second generations of the package

Measure	nomclust 1.0	nomclust 2.0	Speed-up
AN	–	0.38 s	–
BU	–	0.42 s	–
ES	42.54 s	0.37 s	116×
G1	65.17 s	0.37 s	175×
G2	65.64 s	0.37 s	176×
G3	59.52 s	0.37 s	160×
G4	59.67 s	0.37 s	160×
GA	–	0.38 s	–
IOF	79.21 s	0.39 s	205×
LIN	80.44 s	0.40 s	200×
LIN1	45.84 s	0.43 s	108×
MZ	430.0 s	–	–
OF	80.07 s	0.38 s	209×
SM	40.07 s	0.36 s	111×
SV	–	0.39 s	–
VE	41.42 s	0.37 s	112×
VM	41.27 s	0.37 s	112×
Total	58.41 s	0.38 s	154×

The calculations were performed using R (version 4.2.2) on a machine with the processor 3.7 GHz 6-Core Intel Core i5 and 24 GB of RAM.

similarity measure for the old and new versions of the package are placed in the columns *nomclust 1.0* and *nomclust 2.0* in Table 4.3. Considering much lower computation times by the new version, it is clear that the C++ implementation was successful. Whereas the average calculation time was 58.41 s by the first release, the average calculation time was just 0.38 s by the second version of the package. For the similarity measures presented in both generations of the package, the *Speed-up* column shows how many times the second release is faster when using a certain similarity measure compared to the previous version. On average, the new package version performed 154 times faster in the experiment.

4.2 Illustrations of Use

The nomclust package can be used either to perform the complete clustering process from dissimilarity matrix calculation to the evaluation of the obtained clusters using the nomclust() function or to perform only a part of the clustering using one of the subsidiary functions. Typical use will be demonstrated using the CA.methods dataset, which is included in the package.

4.2.1 The whole clustering process

The most crucial function in the nomclust package is the nomclust() function, which completely covers the hierarchical clustering of categorical data with the function call below.

```
R> nomclust(data, measure = "lin", method = "average",
 clu.high = 6, eval = TRUE, prox = 100, var.weights = NULL)
```

The only mandatory input argument is data, representing a categorical dataset in a class of a *data.frame* or a *matrix* entering the cluster analysis. The measure argument stands for a similarity measure used for dissimilarity matrix calculation. Any measure in Table 4.1 can be used here. The method argument enables a researcher to choose one from one of the three linkage methods presented in Section 4.1.2. Regarding these two arguments, the *average* linkage with the *lin* similarity measure was set as default because this combination usually provides the most coherent clusters (Šulc and Řezanková, 2019). The clu.high argument defines the upper limit for the number of cluster solutions provided in the final output. A logical argument eval indicates if the evaluation criteria presented in Section 4.1.3 will be calculated in the *nomclust* object. The prox argument indicates if a dissimilarity matrix will be saved in the *nomclust* object. The argument can be set as a logical value or an integer specifying the maximum number of objects in a dataset for which a dissimilarity matrix will be kept in the output. Since a dissimilarity matrix is needed for dendrogram construction, 100 was chosen as the default value of this argument. Thus, one can display dendrograms in small datasets, where they are most helpful. On the other hand, a large dissimilarity matrix based on a sizable dataset will not be saved in the *nomclust* object. The var.weights argument enables a researcher to set a vector of weights for the clustered variables. Its default value is set to NULL, indicating that all the variables entering the analysis have the same weight. The var.weights argument must be assigned to a vector with variable weights for each clustered variable. The weights take on values from zero to one.

The use of the nomclust() function is demonstrated on the CA.methods dataset, which contains five different characteristics of 24 clustering algorithms; six of them are displayed below.

```
R> library(nomclust)
R> data("CA.methods")
R> head(CA.methods)

                Type OptClu Large  TypicalType MoreTypes
AGNES   hierarchical     no    no quantitative       yes
BIRCH   hierarchical     no   yes quantitative        no
CACTUS          grid    yes   yes  categorical        no
CLARA   partitioning     no   yes quantitative        no
CLIQUE          grid    yes   yes quantitative        no
COOLCAT partitioning     no   yes  categorical        no
```

Hierarchical clustering with the G1 similarity measure and the average linkage method is then performed using the following syntax.

```
R> hca.G1 <- nomclust(CA.methods, measure = "goodall1")
```

The resulting output comprises six components in a *list*. The `mem` component contains cluster membership partitions for two to six clusters. For instance, the four-cluster solution can be obtained using the syntax below.

```
R> hca.G1$mem$clu_4

[1] 1 1 2 3 2 3 1 2 2 1 2 3 3 3 3 4 4 1 2 3 3 1 2 4
```

The `eval` component contains 13 evaluation criteria as vectors in a *list*. To see them all at once, the form of a *data.frame* is more appropriate.

```
R> as.data.frame(hca.G1$eval)

  names  WCM  WCE PSFM PSFE    BIC    AIC    BK   SI   DI   CU   CI   HE   HM
1 clu_1 0.83 0.85   NA   NA 226.39 215.79    NA   NA   NA   NA   NA 9.34 7.90
2 clu_2 0.59 0.60 7.90 9.34 196.04 174.84  0.47 0.21 0.63 0.33 0.61 7.10 5.59
3 clu_3 0.46 0.43 7.57 9.52 189.56 157.76  0.48 0.21 0.60 0.35 0.65 5.42 5.69
4 clu_4 0.37 0.37 8.07 9.49 196.04 153.63  0.08 0.24 0.71 0.34 0.61 3.22 2.43
5 clu_5 0.33 0.32 7.09 8.71 212.82 159.81 -0.10 0.19 0.75 0.30 0.53 3.87 3.63
6 clu_6 0.27 0.25 7.18 8.80 229.06 165.44    NA 0.18 0.78 0.27 0.49   NA   NA
```

The `opt` component is always present in the output together with the `eval` component. It displays the optimal number of clusters for the evaluation criteria from the `eval` component, except for WCM and WCE, where the optimal number of clusters can be determined only by the elbow method.

```
R> as.data.frame(hca.G1$opt)

  PSFM PSFE BIC AIC BK SI DI CU CI HE HM
1    4    3   3   4  3  4  6  3  3  4  4
```

Since the number of objects in a dataset is lower than 100, the prox component containing the dissimilarity matrix is present in the output, where the first six rows and columns in a *matrix* form are displayed.

```
R> as.matrix(hca.G1$prox)[1:6, 1:6]

         AGNES BIRCH  CACTUS  CLARA CLIQUE COOLCAT
AGNES   0.0000 0.6167 1.0000 0.7964 0.8891  0.9072
BIRCH   0.6167 0.0000 0.8406 0.6370 0.7297  0.7478
CACTUS  1.0000 0.8406 0.0000 0.8406 0.4659  0.6536
CLARA   0.7964 0.6370 0.8406 0.0000 0.7297  0.5942
CLIQUE  0.8891 0.7297 0.4659 0.7297 0.0000  0.8406
COOLCAT 0.9072 0.7478 0.6536 0.5942 0.8406  0.0000
```

The dend component contains all the necessary information for dendrogram creation, and the call component includes the function call.

The following syntax demonstrates the change of the dissimilarity matrix when the first variable *Type* in the dataset has four times higher weight than the remaining variables.

```
R> hca.G1w <- nomclust(CA.methods, measure = "goodall1",
                       var.weights = c(1, 0.25, 0.25, 0.25, 0.25))
R> as.matrix(hca.G1w$prox)[1:6, 1:6]

         AGNES BIRCH  CACTUS  CLARA CLIQUE COOLCAT
AGNES   0.0000 0.4235 1.0000 0.8727 0.9307  0.9420
BIRCH   0.4235 0.0000 0.9004 0.7731 0.8311  0.8424
CACTUS  1.0000 0.9004 0.0000 0.9004 0.3007  0.7835
CLARA   0.8727 0.7731 0.9004 0.0000 0.8311  0.4583
CLIQUE  0.9307 0.8311 0.3007 0.8311 0.0000  0.9004
COOLCAT 0.9420 0.8424 0.7835 0.4583 0.9004  0.0000
```

A graphical comparison of the non-weighted and weighted approach is presented in Subsection 4.2.3.

4.2.2 Subsidiary functions

In case one needs to produce only a part of the clustering process, e.g., to calculate a dissimilarity matrix for a different clustering algorithm or to evaluate a cluster solution that was not produced by the `nomclust` package, one of the available subsidiary functions can be used.

Dissimilarity matrices based on the available similarity measures in the package can be calculated using the function calls from Table 4.1. The resulting matrices are objects of the *dist* class. They can be used as an input for hierarchical clustering functions in other R packages as well, e.g., the `agnes()` function in the `cluster` package or the `hclust()` function in the `stat` package.

A dissimilarity matrix based on the simple matching measure, the reference measure in this thesis, is obtained using the `sm()` function.

```
R> prox.SM <- sm(CA.methods)
R> as.matrix(prox.SM)[1:6, 1:6]

        AGNES  BIRCH CACTUS CLARA CLIQUE COOLCAT
AGNES     0.0    0.4    1.0   0.6    0.8     0.8
BIRCH     0.4    0.0    0.6   0.2    0.4     0.4
CACTUS    1.0    0.6    0.0   0.6    0.2     0.4
CLARA     0.6    0.2    0.6   0.0    0.4     0.2
CLIQUE    0.8    0.4    0.2   0.4    0.0     0.6
COOLCAT   0.8    0.4    0.4   0.2    0.6     0.0
```

From the output, it is clear that the SM measure, which does not use any additional information about the categorical variables, provided the dissimilarity matrix with many identical values. This makes the objects challenging to cluster unambiguously with different clustering algorithms, e.g., `hclust()` vs. `agnes()`.

Sometimes, the use of an own-calculated dissimilarity matrix is necessary. The `nomprox()` function can be used in such a situation. It enables a user to run hierarchical clustering based on a provided dissimilarity matrix and, if the original dataset is available, to calculate a unique set of evaluation criteria available in the `nomclust` package. It has the following syntax.

```
R> nomprox(diss, data = NULL, method = "average",
  clu.high = 6, eval = TRUE, prox = 100)
```

The `diss` argument stands for a dissimilarity matrix either of a class *matrix* or *dist*. The argument `data` represents the data from which the dissimilarity matrix was calculated. It is not mandatory for the cluster partitions calculation, but it is necessary for evaluation criteria calculation. The method, clu.high, eval, and prox arguments work in the same way as in the nomclust() function.

Hierarchical clustering based on the already calculated dissimilarity matrix of the SM measure and the original dataset can be performed using the following syntax.

```
R> hca.SM <- nomprox(diss = prox.SM, data = CA.methods)
```

The resulting object contains mem, eval, opt, dend, prox and call components. For instance, the cluster membership variables in the form of a *data.frame* can be obtained using the following syntax.

```
R> clu.SM <- as.data.frame(hca.SM$mem)
R> head(clu.SM)

  clu_2 clu_3 clu_4 clu_5 clu_6
1     1     1     1     1     1
2     2     2     2     2     2
3     2     2     3     3     3
4     2     2     2     2     2
5     2     2     3     3     3
6     2     2     2     2     2
```

In certain situations, the need may arise to apply a set of evaluation criteria from the nomclust package to cluster membership partitions obtained by different clustering algorithms, e.g., LCA or *k*-modes. The evalclust() function can be used for such cases. It has the syntax expressed below.

```
R> evalclust(data, clusters, diss = NULL)
```

The data argument represents the dataset used for clustering, the clusters argument stands for a *data.frame* or a *list* with cluster membership partitions, and the optional diss argument denotes the dissimilarity matrix for the objects in the dataset.

The output of the evalclust() function is the eval component with a set of evaluation criteria, the opt component with the optimal number of clusters based on these criteria, and the call component with the function call. Suppose the dissimilarity matrix is not defined in the diss argument. In that case, the function does not provide outcomes for the distance-based criteria SI and DI, as demonstrated in the following example.

```
R> eval.SM <- evalclust(CA.methods, clu.SM)
R> as.data.frame(eval.SM$eval)

  names WCM  WCE PSFM PSFE    BIC    AIC    BK   CU   CI    HE   HM
1 clu_1 0.83 0.85   NA   NA 226.39 215.79    NA   NA   NA  6.01 5.12
2 clu_2 0.66 0.65 5.12 6.01 212.53 191.32  0.74 0.23 0.44  1.66 1.73
3 clu_3 0.60 0.58 3.51 3.93 229.73 197.92 -0.90 0.21 0.37 12.51 9.87
4 clu_4 0.38 0.34 6.62 8.23 202.95 160.54  0.95 0.31 0.57  2.93 2.16
5 clu_5 0.34 0.29 5.79 7.49 219.74 166.73  0.13 0.27 0.50  1.99 1.72
6 clu_6 0.31 0.27 5.15 6.71 240.70 177.09    NA 0.24 0.45    NA   NA

R> as.data.frame(eval.SM$opt)

  PSFM PSFE BIC AIC BK CU CI HE HM
1    4    4   4   4  4  4  4  2  5
```

When comparing the outputs for the G1 and SM similarity measures on this dataset, it seems that WCM and WCE criteria variability decreases faster by G1, suggesting that the clusters of the G1 measure are more homogeneous. On the other hand, the recommended numbers of clusters in the opt component are more consistent by the SM measure, where most of the evaluation criteria prefer the four-cluster solution.

4.2.3 Graphical functions

Graphical outputs can help a researcher choose the optimal number of clusters or evaluate the quality (and interpretability) of the created clusters. The nomclust package offers two graphical functions, one for evaluation criteria visualization and the second one for dendrogram creation.

To visualize the evaluation criteria from the eval component, the eval.plot() function with the following syntax can be used.

```
R> eval.plot(x, criteria = "all", style = "greys",
 opt.col = "red", main = "Cluster Evaluation", ...)
```

The x argument represents an output of the functions nomclust(), nomprox() or evalclust() containing the eval and opt components. The argument criteria specifies the evaluation criteria which are to be visualized. It can be selected by one particular criterion, a vector of criteria, or all the available criteria. The argument style defines a graphical style of the produced plots. There are two predefined styles in the nomclust package, namely *greys* and *dark*, but a custom color scheme can be set by a user as a vector with colors of a length four.

The `opt.col` argument specifies a color used for the optimal number of clusters identification and the `main` argument determines the title of a chart. The symbol `...` indicates that it is possible to use specific graphical arguments from a generic `plot()` function.

The `eval.plot()` can be used to obtain graphical representations of the AIC evaluation criteria from the *hca.G1* and *hca.SM* objects using the syntax below.

```
R> par(mfrow = c(1,2))
R> eval.plot(hca.G1, criteria = "AIC", main = "G1 measure")
R> eval.plot(hca.SM, criteria = "AIC", main = "SM measure")
```

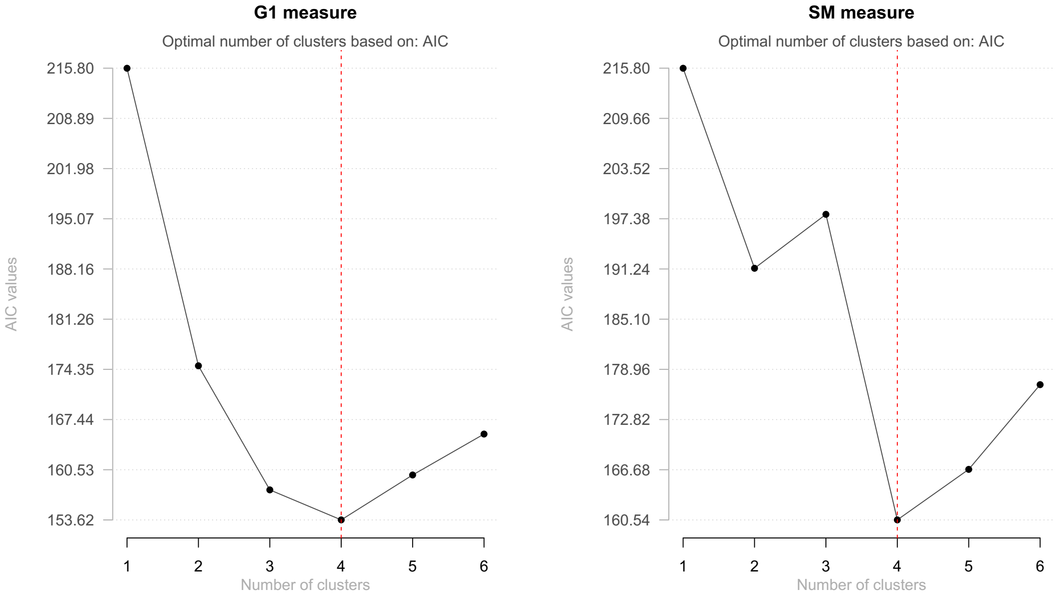


Figure 4.1: The optimal number of clusters for HCA with G1 and SM measures based on AIC

Figure 4.1 shows that the lowest value for both similarity measures is four, so the solution with four clusters should be preferred. One can also examine the cluster solutions with slightly higher evaluation criteria values, e.g., the three- or five-cluster solutions. A graphical representation can help a researcher notice minor differences between the values, and thus, it makes determining the optimal number of clusters easier.

A dendrogram visualizes a hierarchy of clusters, and it can help a researcher decide on the number of clusters, especially with datasets of smaller sizes. To produce a dendrogram in the `nomclust` package, the `dend.plot()` function is used.

```
R> dend.plot(x, clusters = "BIC", style = "greys", colorful = TRUE,
 clu.col = NA, main = "Dendrogram", ac = TRUE, ...)
```

The x argument represents an output of the functions nomclust() or nomprox() containing the dend component. The clusters argument determines the number of clusters displayed in a dendrogram. It can be either set as a number or as a name of the evaluation criterion if the eval and opt components are present in the output. The style and main arguments work in the same way as by the eval.plot() function. A logical argument colorful specifies if the output will be colorful or black and white. An optional argument clu.col allows a researcher to apply user-defined colors to distinguish clusters in a dendrogram. The ac argument indicates if the value of an *agglomerative coefficient* from the cluster package is displayed below the dendrogram.

The dend.plot() function is demonstrated on a comparison of the four-cluster solutions provided by the G1 and SM measures.

```
R> dend.plot(hca.G1, clusters = 4, main = "G1 measure")
R> dend.plot(hca.SM, clusters = 4, main = "SM measure")
```

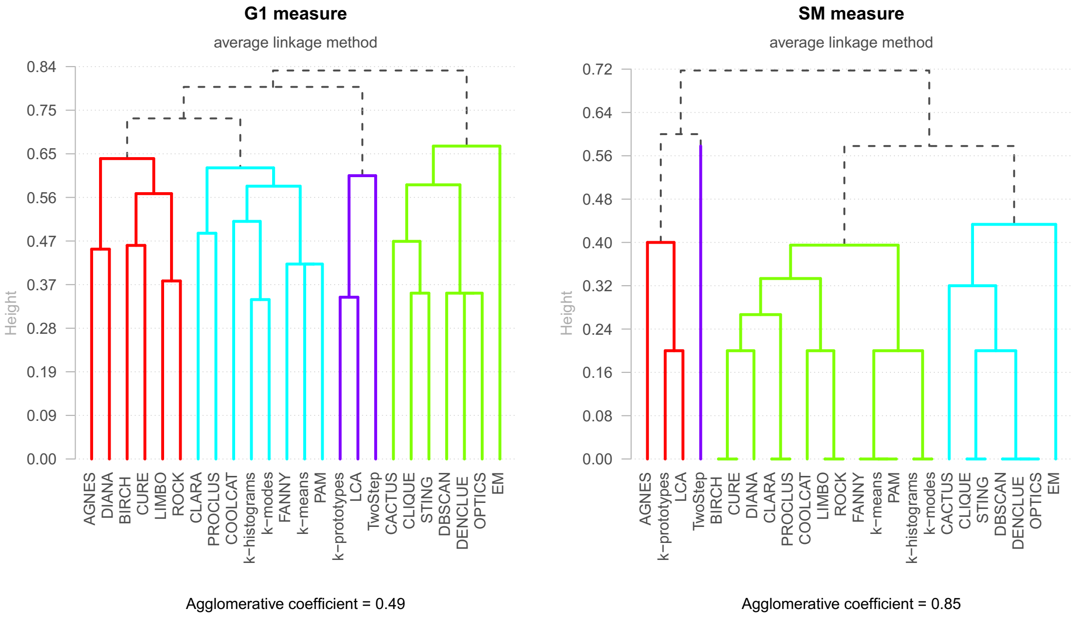


Figure 4.2: Comparison of dendrograms for the G1 and SM measures

Figure 4.2 shows that clusters of SM are better separated, but by G1, they seem more meaningful. Regarding the G1 clusters, the first one contains the hierarchical clustering methods. The second one comprises the majority of partitioning methods. The third one includes the clustering algorithms primarily determined for the mixed-type data. The fourth cluster consists of clustering methods based on grid and density. Most of the clusters provided by the SM measure do not have a logical structure, despite the higher value of the agglomerative coefficient.

Considering Figure 4.2, it is apparent that different similarity measures may lead to substantially different results. Therefore, examining more combinations of similarity measures and linkage methods is always good. A common practice is finding well-separated clusters that are meaningful to a researcher. In a small dataset, such as the demonstrated one, displaying a dendrogram is sufficient. In a large dataset, the created clusters can be characterized by a series of contingency tables containing the variable categories broken down by the cluster membership variable.

Another example may be a graphical comparison of the non-weighted and weighted approaches presented in Subsection 4.2.1. Figure 4.3 visualizes the calculated dissimilarity matrices using dendrograms. Since the variable *Type* has four times higher weight in clustering with the weighted approach, the produced clusters mostly correspond to the categories of this variable. The first cluster contains only hierarchical methods, the second one the partitioning methods, the third the model-based methods, and the last the density-based and grid methods.

```
R> dend.plot(hca.G1, clusters = 4, main = "G1 (non-weighted)")
R> dend.plot(hca.G1w, clusters = 4, main = "G1 (weighted)")
```

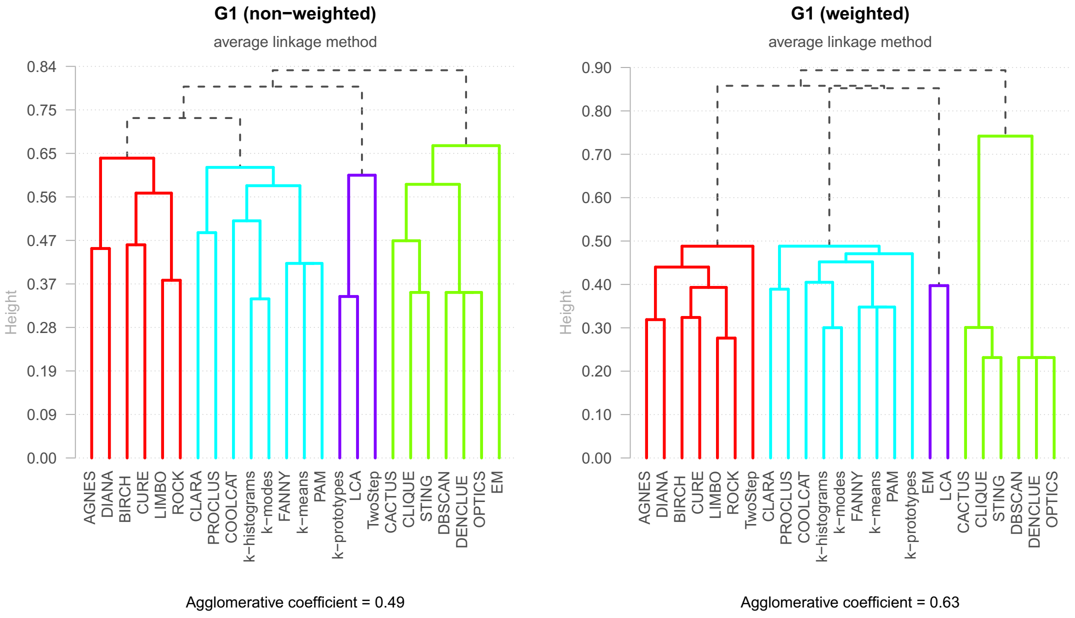


Figure 4.3: Comparison of dendrograms for the non-weighted and weighted approach

4.2.4 Generic functions

The second generation of the nomclust package added support for standard generic functions, such as summary() or print(). They can be applied to the outputs of the functions nomclust(), nomprox(), or evalclust(), which are of the class *nomclust*. The range of the generic function outcomes differs by the number of components in the output. Thus, the most complex outputs are provided by the nomclust() function, whereas the evalclust() function usually provides the most limited outcomes.

The summary() function can be used if one wants quick information about the clustering results. The outcome contains frequency distribution tables for the created numbers of clusters, so a researcher can see if the cluster sizes are balanced or if there are one-object clusters. The function also provides the optimal numbers of clusters according to all calculated evaluation criteria and the value of the agglomerative coefficient. The outcome of the summary() function is shown in the following example.

```
R > hca.IOF <- nomclust(CA.methods, measure = "iof",
 method = "complete", clu.high = 3)
R> summary(hca.IOF)

Sizes of the created clusters:

2 clusters:
 1  2
 9 15

3 clusters:
 1  2  3
 9 10  5

Optimal number of clusters based on the evaluation criteria:
  PSFM PSFE BIC AIC BK SI
1    2    2   2   3  2  2

Agglomerative coefficient: 0.9685352
```

The print() function offers all the necessary information for cluster quality evaluation, namely values of the calculated evaluation criteria, the corresponding optimal numbers of clusters, and the agglomerative coefficient.

Sometimes, a researcher may need to apply additional functionalities to the obtained clustering outcomes, often requiring the object of the *hclust* class as an input. Therefore, the class of the clustering object can be easily changed by the as.hclust() function. Moreover, we introduce the `as.agnes()` function which transforms a *nomclust* object into an *agnes, twins* object, which is occasionally also required. The introduced function is not generic, but it works as expected. The use of both functions is shown below.

```
R> hca.IOF.hclust <- as.hclust(hca.IOF)
R> hca.IOF.agnes <- as.agnes(hca.IOF)
```

The function `plot()` applied to the output of the `nomclust()` or `nomprox()` functions creates a basic dendrogram that is suitable for a quick look at the hierarchy of clusters. Compared to the more complex function `dend.plot()`, it allows a researcher only to change the chart's title.

```
R> plot(hca.IOF, main = "plot() in nomclust")
R> plot(hca.IOF.hclust, main = "plot() in hclust")
```

Figure 4.4 depicts the outcomes of the `plot()` function applied to the objects of the *nomclust* and *hclust* classes.

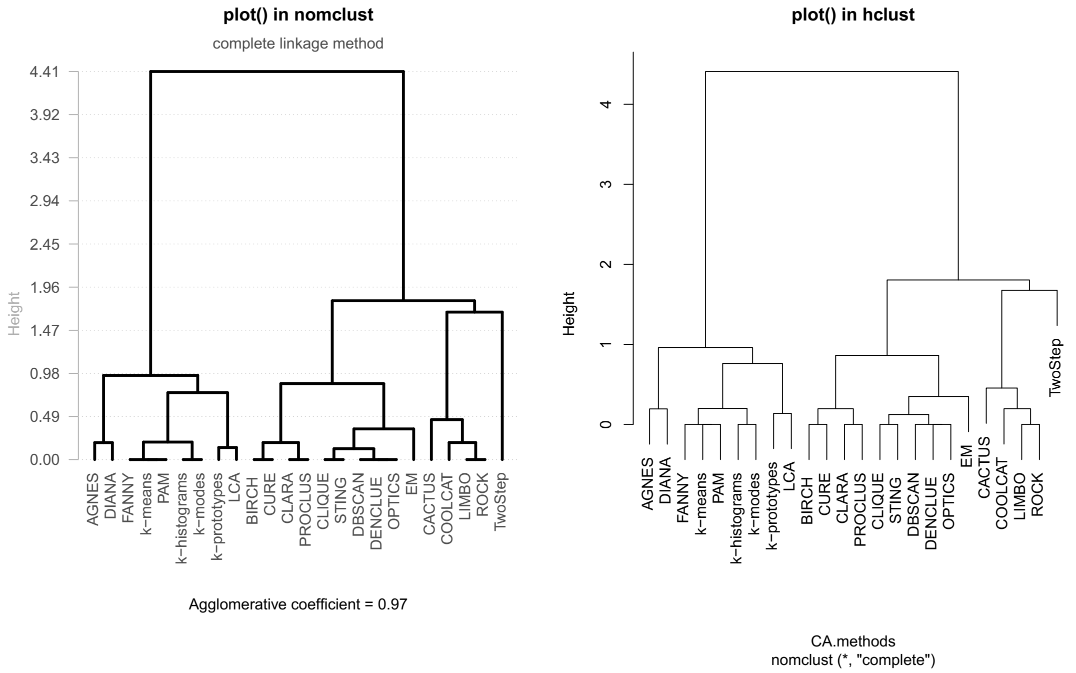


Figure 4.4: Outputs of the `plot()` function applied on the *nomclust* and *hclust* objects

5 Comparison of Similarity Measures for Categorical Data

This chapter deals with the first goal of this thesis, i.e., the comparison of the similarity measures for categorical data presented in Introduction. The similarity measures are evaluated according to their ability to create good-quality clusters in HCA. The analysis aims to determine in which situations a certain similarity measure forms good-quality clusters and when not. Therefore, the experiment is performed on 2,700 generated datasets with controlled properties, such as the number of variables or natural clusters in a dataset.

The chapter is divided into three sections. The first one describes the data generation process and the generated datasets' properties. The second one defines the methodology for the similarity measures evaluation, and the third one contains the conducted experiment.

5.1 Data Generation Process

The generated datasets for the experiment are obtained using the updated `gen_object()` function introduced by Šulc (2016), which is based on the `genRandomClust()` function from the `clusterGeneration` R package (Qiu and Joe, 2006) and the `discretize()` function from the `arules` (Hahsler et al., 2015) R package. The function has the following function call:

```
R> gen_object(n_per_clu = 150, nclu = 4, nvar = 4, ncat = c(3, 3, 3, 3),
    dist = 0.34, discretize = "interval", mem = 1),
```

and it contains seven parameters. The argument *n_per_clu* sets the cluster size equal for all generated clusters. The argument *nclu* specifies the number of original clusters in a dataset,*nvar* defines the number of variables in a dataset, *ncat* specifies the number of categories for every variable in a dataset. It has a form of a vector, and its length must be equal to the number of variables. The argument *dist* defines the distance between two neighboring clusters from the interval (–1,1). The closer the value is to one, the more separated the clusters are. The *discretize* parameter allows a researcher to choose a way of dataset categorization. The default value "interval" creates equal intervals from the original quantitative values of a given variable differing in the number of objects. The value “frequency” creates a given number of intervals with unequal width, each containing approximately equal number of objects. The last parameter *mem* is a logical operator. If this parameter is *TRUE*, a cluster membership variable will be added to the generated dataset.

The dataset generation based on the `gen_object()` function uses a two-step approach. In the first step, quantitative data with a multidimensional correlation structure reflecting the given properties (between-cluster distances, the number of clusters, variables, and categories) is created. In the second step, the variables in datasets are categorized using the equal-width intervals approach, which constructs more naturally-looking datasets. This generation approach was already used by Šulc (2016) and Šulc and Řezanková (2019). The datasets generated this way contain, in fact, ordinal variables. Fortunately, this is not a problem for the planned experiment since the similarity measures for categorical data do not consider the order of categories.

For the experiment, 27 different dataset settings were used; see Figure 5.1. All the datasets were generated with four original clusters. The reason is that most internal evaluation criteria depend on the number of clusters; thus, their values would be incomparable by different numbers of clusters. Three minimal between-cluster distances (0.21, 0.34, 0.50)[1] were used, representing intersecting, partly intersecting, and almost non-intersecting clusters. In this thesis, these distances are also referred to as *small*, *medium*, and *large* minimal between-cluster distances. Next, the datasets were generated with three different numbers of variables (4, 7, 10) covering the typical range of clustering of categorical datasets. Finally, there are three different numbers of categories (3, 5, 7) in the generated datasets illustrating *simple*, *medium*, and *complex* dataset structure. The number of objects in generated datasets was firmly set to 600 cases. Each dataset setting combination was replicated one hundred times to ensure the robustness of the obtained results. In total, this makes 2,700 generated datasets used for the analysis.

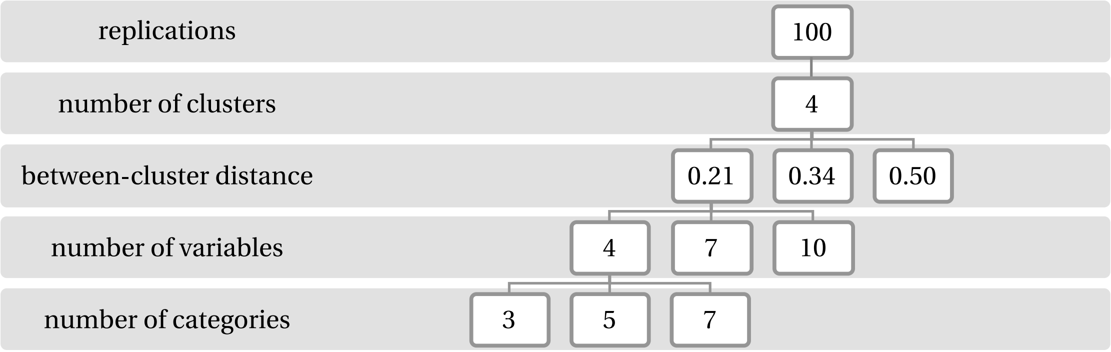


Figure 5.1: Dataset generation scheme for the first experiment

[1] Due to some information loss by categorization of the dataset, the values of the *dist* argument are higher than the recommended ones for quantitative data.

5.2 Research Methodology

The examined similarity measures are compared and evaluated based on their clustering performance (ability to produce good-quality clusters) in HCA. It implies that the measures are compared indirectly using the obtained clusters' quality measured by internal evaluation criteria. When dealing with a large number of HCA outputs[2], a need to process and interpret such an amount of data in a synoptic way arises. This thesis assesses the examined similarity measures in two different ways. The first represents a relative comparison based on the mean ranked scores methodology proposed by Šulc and Řezanková (2019), and the second focuses on absolute evaluation criteria differences expressed by boxplots.

Mean ranked scores methodology

The mean ranked scores methodology assesses the created clusters based on the mean ranked scores (MRS) of the internal evaluation criteria that are either variability- or likelihood-based[3]. In a given dataset, values of these criteria can be compared not only with their values in different cluster solutions of a certain similarity measure but also with their values in a particular cluster solution for different similarity measures. However, the values of variability- and likelihood-based criteria are non-standardized, so they are incomparable if the datasets' properties, such as the dataset's size, differ. That is why the constant number of objects was used in the datasets used in the experiment. Then, using MRS ensures comparable and easily interpretable internal evaluation criteria outcomes.

The MRS methodology consists of two steps. In the first one, a series of cluster partitions based on the first dataset is produced by HCAs with all the examined similarity measures, and the resulting clusters are then evaluated using a given internal criterion. The evaluation criterion's outcome scores are then ranked from the best to the worst (the direction depends on the evaluation criterion used) so that the best criterion value is ranked as one. In the same manner, MRS for the other datasets are obtained. In the second step, MRS are averaged over the number of replications and other properties that are not of interest in the given analysis. For example, to get MRS broken down by the number of variables, it is necessary to average MRS over all other properties that are not of interest (the number of replications, minimal between-cluster distances, and the number of categories). The resulting MRS are considered the main output that can be displayed as an easily interpretable table. The lower the MRS of a similarity measure, the better its clustering performance is.

The MRS methodology enables a researcher to order the examined similarity measures regarding their clustering performance from the best one to the worst one. However, it does not provide information about the extent to which the similarity measures differ. Thus, the differences among the similarity measures are evaluated relatively. Another issue is a limited

[2] In the experiment, 43,200 cluster membership partitions with four clusters is used (27 dataset types × 100 replications × 16 similarity measures).

[3] Distance-based criteria depend on the used similarity measure, see Chapter 6.

way of expressing the variability of the ranked scores. One can use a standard deviation, as it was performed by Šulc and Řezanková (2019), but its value is difficult to interpret. These issues are solved by the second way of the similarity measures evaluation by which the differences among the similarity measures are evaluated absolutely using the boxplots.

Boxplot assessment

Displaying the values of the evaluation criterion broken down by the examined similarity measures is a simple yet efficient way to show the absolute differences among the similarity measures represented by median values of a given criterion and also the variability of the criterion values expressed by the inter-quartile range (IQR). Moreover, the criterion values can be further broken down by a specific dataset's property, such as the number of variables. This way, the resulting boxplots can reveal the differences between different levels of such a property. If the differences are substantial, dependence on a specific dataset's property can be assumed

The *mean ranked scores methodology* and *boxplot assessment* analyze the similarity measures from both relative and absolute comparison perspectives. Thus, they complement each other well, which will be utilized in the experiment.

Evaluation criteria used

In the experiment performed in this thesis, the PSFE and CU criteria are used. The entropy-based PSFE criterion was chosen since it enables a direct comparison with the previous studies by Šulc (2016) and Šulc and Řezanková (2019), where this criterion was also used. The mutability-based CU criterion was chosen because it is one of the few widely known internal evaluation criteria that are correctly used for categorical clustering assessment. It is used in the COBWEB algorithm (Fisher, 1987) for the conceptual clustering, and also in many papers, e.g., (Murakoshi and Fujikawa, 2016), and books, e.g., (Witten et al., 2016), dealing with the categorical data clustering. Thus, using this criterion will allow many researchers to put its results in the context of their experience with this criterion.

Both the used criteria are variability-based, and they express the cluster quality in a non-standardized way, i.e., their maximal values are not the same. Based on Eq. (3.10) and Eq. (3.11), values of PSFE depend on the number of clusters k and the number of observations n, and the values of CU on the number of clusters k. Thus, in order to be able to compare their values across different datasets, the clustered datasets must have the same number of objects, and the evaluation criteria values must be for the same number of clusters (four in this experiment). These conditions are satisfied in the generated datasets presented in Section 5.1.

Experiment design

For each of the 2,700 generated datasets (see Section 5.1), a set of dissimilarity matrices was computed according to the 16 examined similarity measures for categorical data presented in Chapter 2. Next, HCA with ALM, CLM, and SLM for the four-cluster solution was performed on each dissimilarity matrix. Each HCA solution was evaluated by the evaluation criteria PSFE and CU presented in Section 3.2. The calculations were performed using the `nomclust` package for R, presented in Chapter 4.

The similarity measures are compared using the evaluation criteria values in the four-cluster solution, which corresponds to the original number of clusters in the generated datasets. For the comparison, mean ranked scores methodology and boxplot assessment presented in Section 5.2 are used. The described procedure offers an efficient approach to comparing a large number of similarity measures, both in relative and absolute ways. All the scripts used in the experiment are available as an electronic appendix, which is described in Table II in Appendix B.

5.3 Experiment

The experiment aims to determine in which situations a certain similarity measure creates good-quality clusters and when not. Many dataset properties can influence the clustering performance of a similarity measure, e.g., the number of variables, categories, or the minimal distance of clusters in a dataset. Some researchers' decisions can also influence the quality of clusters, mainly the choice of the linkage method.

Compared to the research performed by Šulc (2016), the current experiment applies a new way of similarity measures evaluation, the boxplot assessment. Additionally, mutual interactions between the linkage methods and similarity measures are investigated, which results in recommendations for researchers which combination of similarity measures and linkage methods works well for a dataset with specific properties. Moreover, the influence of different minimal between-cluster distances is explored in this thesis. Next, more similarity measures, a new evaluation criterion, and a substantially larger number of datasets are used in the experiment. Finally, the current experiment provides full results for all three linkage methods recommended for categorical data.

The experiment is spread across five subsections. The first one examines the influence of the linkage method on cluster quality. The following three subsections analyze the similarity measures regarding the quality of the created clusters separately for each linkage method. The last subsection recommends which combinations of similarity measures and linkage methods suit a given dataset with specific properties.

5.3.1 Linkage methods comparison

Table 5.1 and Table 5.2 present the mean ranked scores (MRS) for the examined similarity measures in three linkage methods (LINK), namely ALM, CLM, and SLM, based on the PSFE and CU evaluation criteria. The criteria measure the relative clustering performance of the examined similarity measures. Thus, a given mean ranked score is not comparable across different linkage methods in absolute terms. Since 16 similarity measures are assessed in the experiment, the average mean ranked score is 8.5. Thus, the measures evaluated by the lower score (represented by green shades) provide better clusters than the average similarity measure in a given linkage. On the other hand, the measures with the high MRS (represented by purple shades) create poor-quality clusters.

MRS based on the PSFE and CU criteria do not differ substantially, suggesting that the quality of the created clusters is measured well by these two criteria. Moreover, PSFE provides the same MRS outputs as the criteria BIC, AIC, and CI that are not used in this experiment, which further increases the validity of the results.

Table 5.1: MRS for three linkage methods based on PSFE

LINK	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
ALM	11.7	10.8	8.0	3.4	8.5	4.4	15.7	12.5	7.8	5.5	12.4	9.7	8.1	4.8	6.2	6.3
CLM	12.0	6.4	11.6	10.2	8.3	10.2	9.9	9.8	6.3	6.4	7.7	5.4	11.8	6.9	6.6	6.6
SLM	6.5	6.6	6.6	10.1	9.7	10.5	9.5	8.9	9.0	9.5	11.2	7.1	6.1	11.8	6.3	6.4

Table 5.2: MRS for three linkage methods based on CU

LINK	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
ALM	12.6	11.5	7.7	3.8	8.0	4.6	15.6	11.9	6.7	5.6	12.1	10.2	7.8	5.7	6.1	6.2
CLM	12.7	7.9	11.0	10.0	7.8	10.0	8.8	8.7	5.2	6.5	8.8	6.7	11.3	8.0	6.3	6.4
SLM	7.1	7.5	6.4	10.5	9.3	10.9	8.4	8.7	8.2	10.4	10.9	7.8	5.9	11.6	6.2	6.3

Both the tables show vast differences in MRS between the linkage methods by most of the similarity measures indicating that many similarity measures perform relatively well by a specific linkage method while they create poor-quality clusters in other linkages. The exceptions are the VM and VE measures that perform well (MRS are low) in all the linkage methods and the G2 measure that creates the average-quality clusters (MRS are around 8.5). Thus, it is clear that the performance of the similarity measures is closely related to the linkage method used. Since the linkage method is not a dataset property but a user-defined setting of the analysis, the results for different linkages will be analyzed separately in the following subsections.

In ALM, G1 produces the best clusters (with the lowest MRS) according to both evaluation criteria. It is closely followed by G3 and further by LIN and SV measures whose order differs according to the used criterion. The measures VE and VM, proposed by Šulc (2016), also perform very well. The measures IOF, ES, and SM also create better clusters than the average. The rest of the measures form clusters of below-average quality. The G4 measure produces the worst clusters with the mean ranked score equal to 15.7 out of the maximum of 16.

Figure 5.2: Boxplots of the PSFE values for three linkage methods

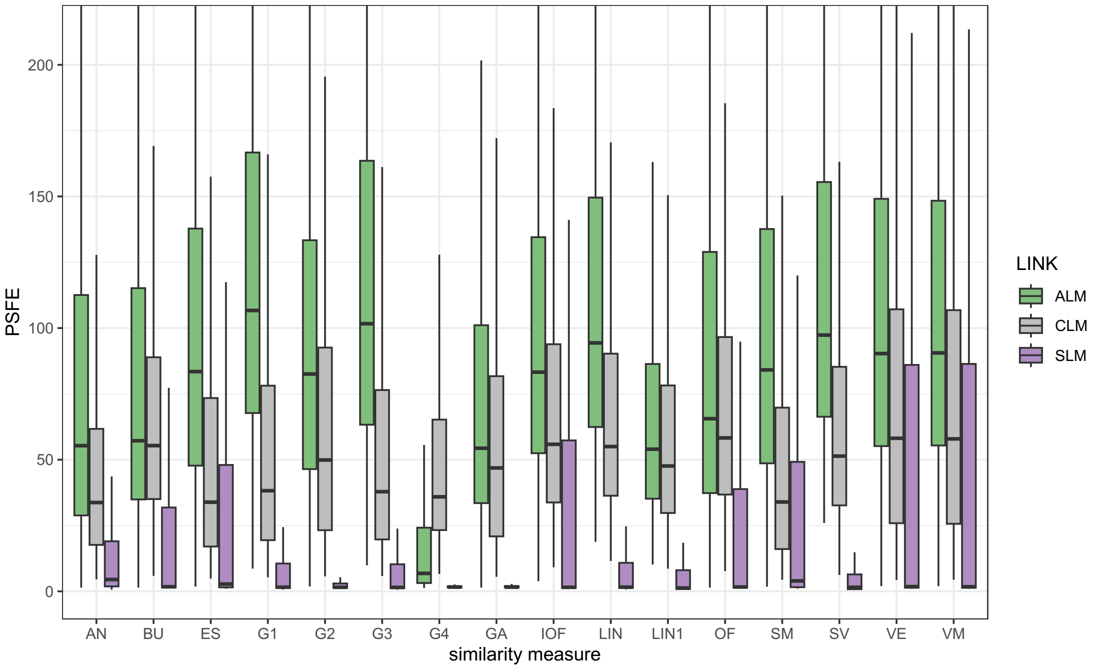
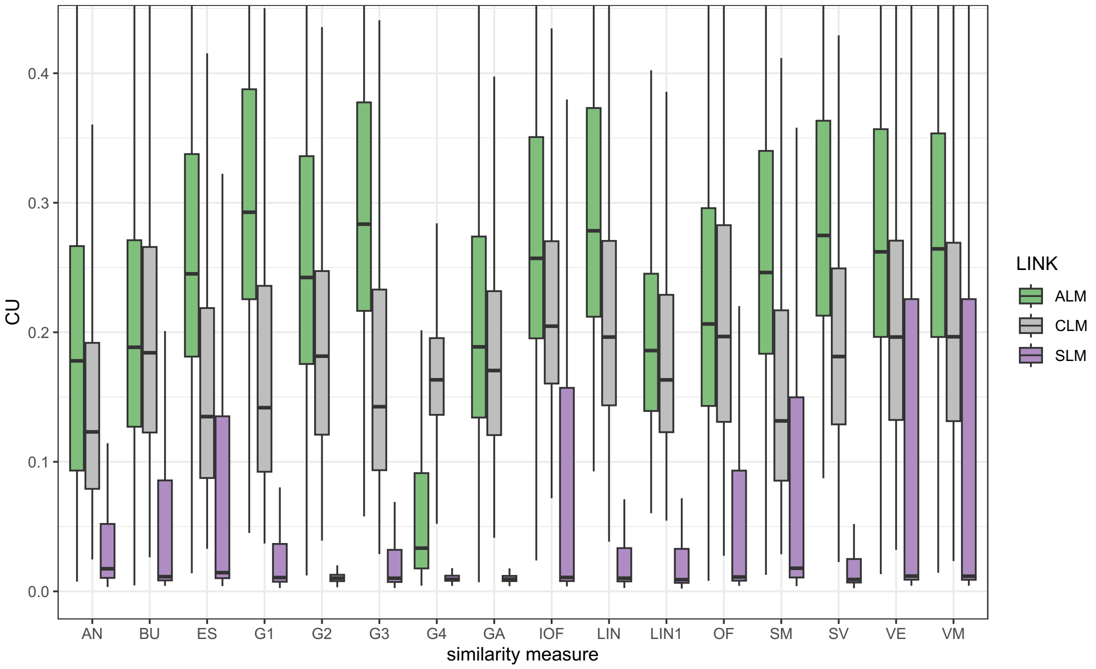


Figure 5.3: Boxplots of the CU values for three linkage methods



In CLM, IOF and OF create the best clusters, followed by the LIN measure. VM, VE, and BU are also well-performing measures. The rest of the measures produce clusters of moderate or poor quality. When comparing MRS of CLM with ALM, it is evident that some measures improved their relative clustering performance, e.g., BU or G4. On the other hand, some measures perform much worse. The most apparent is the drop by the G1 and G3 measures that form the best clusters in ALM.

In SLM, the reference measure SM creates the best clusters. It is closely followed by VM, VE, and ES. The outputs of SLM are the most distinct from the other two linkages. The largest differences are by the measures AN and BU, which produced good clusters in SLM but mediocre ones in the other linkages.

The relative comparison of similarity measures conducted in the previous paragraphs enables a researcher to order the examined similarity measures according to their suitability in a given dataset property, such as the used linkage. However, it does not show absolute differences between the dataset properties. Therefore, the boxplot assessment is used. Figure 5.2 and Figure 5.3 show the boxplots based on the original PSFE and CU criteria values. Three boxplots for each similarity measure are present according to three linkage methods. This way, each boxplot is based on 2,700 evaluation criteria values. The colorful boxes in the charts represent IQRs, i.e., the middle 50% of evaluation criteria values, and the lines in the boxes are the median criteria values. The main characteristics of the boxplots are presented in Table III and Table IV in Appendix C.

Both the charts show huge differences between all three linkage methods. According to both PSFE and CU criteria, most of the examined similarity measures produce the best clusters by ALM. The only exception is the G4 similarity measure which performs poorly overall. The variability expressed by IQR is usually the largest in ALM compared to the other two linkage methods, and it is very similar across the similarity measures (with the exceptions of G4 and LIN1). On the other hand, SLM generally creates poor-quality clusters with the median values of the criteria close to zero. The boxplots of SLM often contain the lowest values, and they do not overlap with the other linkages, which indicates that SLM rarely outperforms them. The outputs of CLM are in between ALM and SLM; however, they are more similar to ALM than to SLM.

5.3.2 Average linkage method

This subsection deals with the influence of different minimal between-cluster distances, numbers of variables, and numbers of categories on the clustering performance quality of the examined similarity measures when ALM is used.

First, the minimal between-cluster distances (DIST) with values 0.21, 0.34, and 0.5 representing the small, medium, and large DIST in generated datasets are examined. The aim is to determine if the performance of the examined similarity measures to provide good-quality clusters depends on how much the clusters in the datasets overlap. Table 5.3 and Table 5.4 provide MRS for the three examined DIST values based on the PSFE and CU. The DIST values are ordered descendingly since the datasets with the largest ones are the easiest to cluster.

The outputs show the relative differences in MRS between the three DISTs. This result indicates that most of the examined similarity measures perform consistently across datasets with different extents of clusters overlapping, e.g., the well-performing G1 and G3 measures, but also the poorly-performing G4 measure. Thus, by these measures, there is no necessity to analyze the overlaps of clusters before applying a particular similarity measure in ALM.

However, some similarity measures improve their relative clustering performance with the decreasing DIST. For instance, SV performs exceptionally well according to the PSFE criterion. Smaller improvements with the decreasing DIST can also be observed in the LIN and LIN1 measures. Some measures, such as SM, ES, or BU, show slightly worsening clustering performance with decreased DIST. However, these deteriorations are almost negligible.

Table 5.3: MRS for three minimal between-cluster distances based on PSFE (ALM)

DIST	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
0.50	12.3	10.3	7.4	3.7	7.9	4.6	15.8	11.7	7.4	5.8	13.9	9.2	7.5	6.3	5.8	6.2
0.34	11.6	10.5	8.2	3.3	8.6	4.3	15.8	12.8	7.9	5.4	12.8	9.5	8.2	4.7	6.2	6.1
0.21	11.3	11.7	8.5	3.3	9.0	4.3	15.6	13.1	8.0	5.2	10.5	10.3	8.5	3.5	6.5	6.7

Table 5.4: MRS for three minimal between-cluster distances based on CU (ALM)

DIST	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
0.50	12.7	10.8	7.2	4.0	7.5	4.7	15.7	11.3	6.8	6.0	13.5	9.6	7.3	6.6	6.0	6.3
0.34	12.6	11.2	7.9	3.7	8.2	4.5	15.7	12.1	6.8	5.4	12.3	10.1	7.9	5.5	6.1	6.1
0.21	12.6	12.5	7.9	3.6	8.3	4.5	15.4	12.2	6.5	5.3	10.5	11.0	8.1	5.0	6.3	6.4

Analysis of the absolute differences of the evaluation criteria using the boxplots occurs in Figure 5.4 and Figure 5.5, where each boxplot is based on 900 evaluation criteria values. Both charts show that all the examined similarity measures provide the best clusters by the largest DIST of 0.5 and the worst by the smallest DIST of 0.21. This conclusion is supported by almost non-overlapping bars in the boxplots (representing IQRs) by most measures, especially by the CU criterion. The variability of the outputs is the highest by the largest DIST and the lowest by the smallest DIST. Concerning the similarity measures, G4 creates the worst clusters from all examined measures. Even by the largest DIST, it cannot compete with the small and medium DISTs of other explored similarity measures. The numeric characteristics of both the boxplots are presented in Table V and Table VI.

Figure 5.4: Boxplots of the PSFE values for three minimal between-cluster distances (ALM)

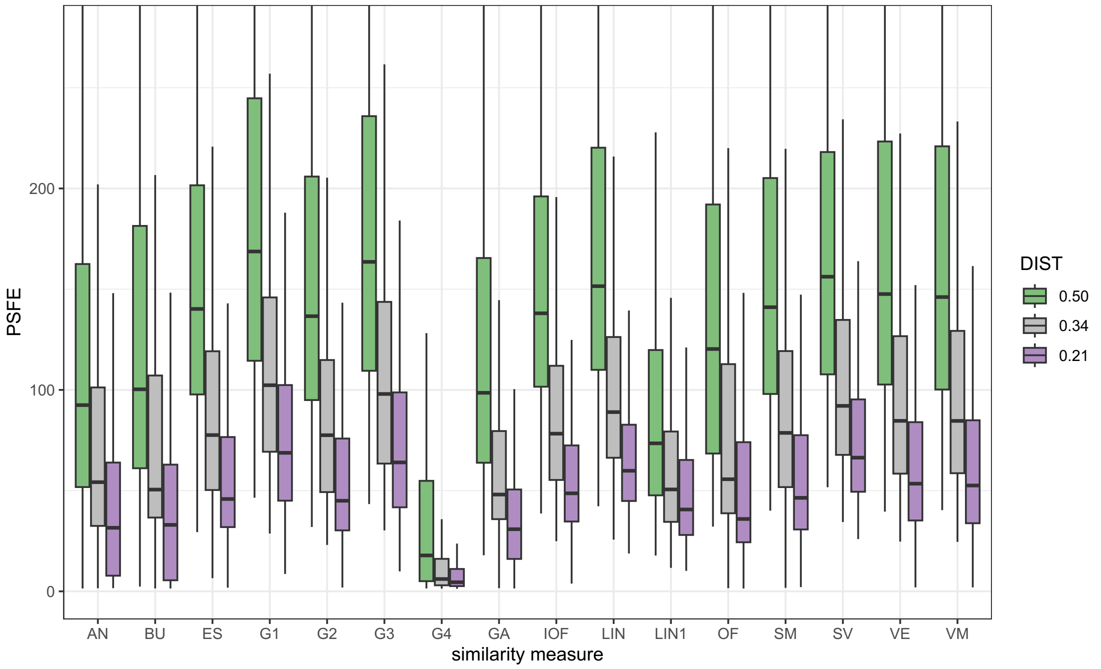
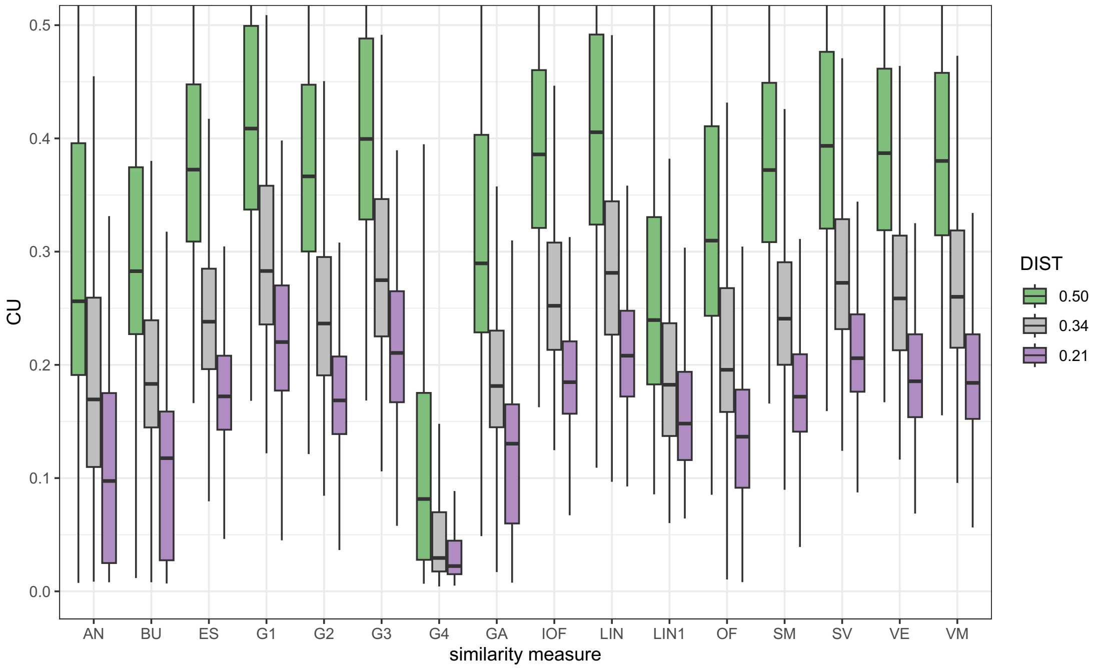


Figure 5.5: Boxplots of the CU values for three minimal between-cluster distances (ALM)



Second, the influence of three different numbers of variables (VAR) on the cluster quality by 16 similarity measures is explored. These numbers of variables, namely (4, 7, 10), were chosen to represent small, medium, and large numbers of the clustered variables in a typical HCA task.

Table 5.5 and Table 5.6 show MRS broken down by the numbers of variables for the PSFE and CU evaluation criteria, and they both describe the cluster quality in the same way. The outputs indicate that most similarity measures perform consistently across the examined variable numbers in ALM. The exceptions are the LIN and SV measures that substantially improve their relative clustering performance with an increasing number of variables. To a smaller extent, this type of behavior can be observed by the IOF measure. On the other hand, deterioration of the cluster quality with the increasing VAR can be expected by OF.

Table 5.5: MRS for three numbers of variables based on PSFE (ALM)

VAR	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
4	11.1	9.0	7.9	3.7	8.4	4.5	15.8	12.4	9.1	7.6	12.9	7.9	8.0	5.9	6.0	5.8
7	11.9	11.6	8.0	3.3	8.4	4.4	15.7	12.5	7.2	5.0	12.2	10.3	8.0	4.8	6.1	6.6
10	12.1	12.0	8.2	3.2	8.7	4.3	15.7	12.6	7.1	3.8	12.1	10.8	8.3	3.9	6.5	6.7

Table 5.6: MRS for three numbers of variables based on CU (ALM)

VAR	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
4	12.1	10.0	7.2	4.3	7.5	4.9	15.7	11.4	7.4	7.7	13.1	8.7	7.4	7.2	5.7	5.5
7	12.8	12.0	7.8	3.6	8.1	4.5	15.6	12.1	6.2	5.1	11.7	10.7	7.7	5.5	6.2	6.5
10	13.0	12.4	8.1	3.4	8.4	4.4	15.5	12.1	6.3	3.8	11.5	11.3	8.2	4.4	6.5	6.7

Figure 5.6 and Figure 5.7 represent the absolute differences of the used evaluation criteria broken down by the number of variables. In this analysis, the outputs of the PSFE and CU criteria differ extensively, which reduces their credibility in describing the absolute differences between the groups. Since there is no way to determine which of the criteria assesses the clusters in a better way, both outputs will be described separately.

According to PSFE, the best clusters are obtained when four (or a smaller) number of variables is used. Then, with the increasing VAR, the absolute clustering performance of all similarity measures decreases. The highest variability of outputs is when four variables are clustered, and the lowest is when ten variables are clustered.

On the contrary, the CU output shows an entirely different picture, where the created clusters are of similar quality by a majority of similarity measures by all three numbers of variables. The exception is G4 which performs poorly overall in ALM. Smaller differences can also be observed by specific similarity measures, e.g., G1, G3, or LIN. The result suggests that VAR does not influence the quality of the obtained clusters in a large scope. The numeric characteristics of both the boxplots are presented in Table VII and Table VIII in Appendix C.

Figure 5.6: Boxplots of the PSFE values for three variable numbers (ALM)

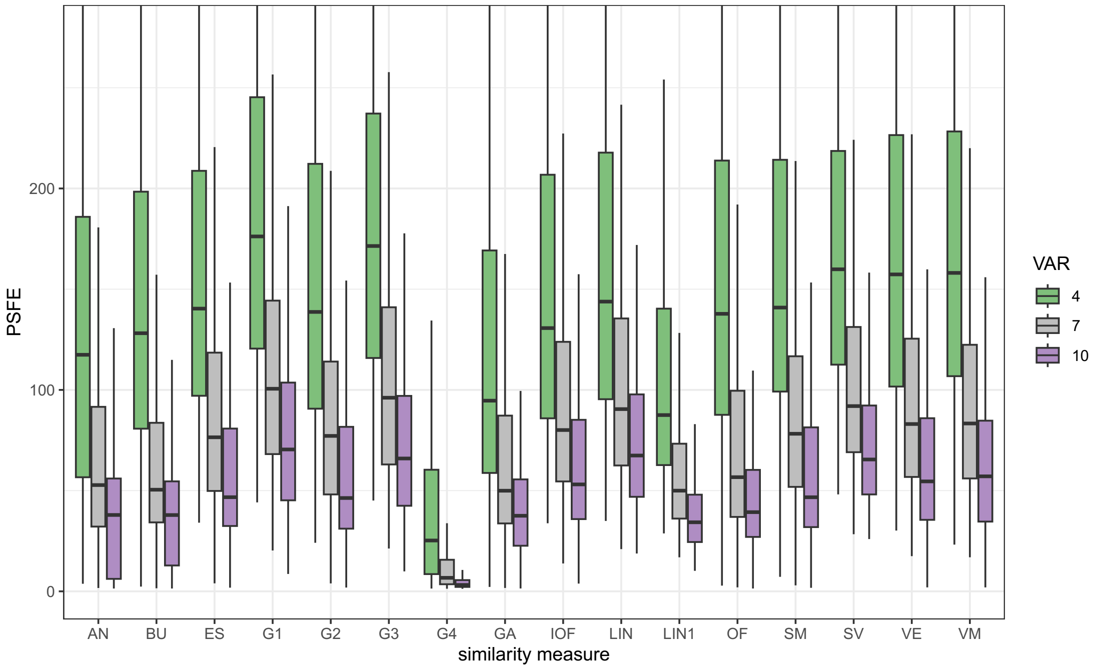
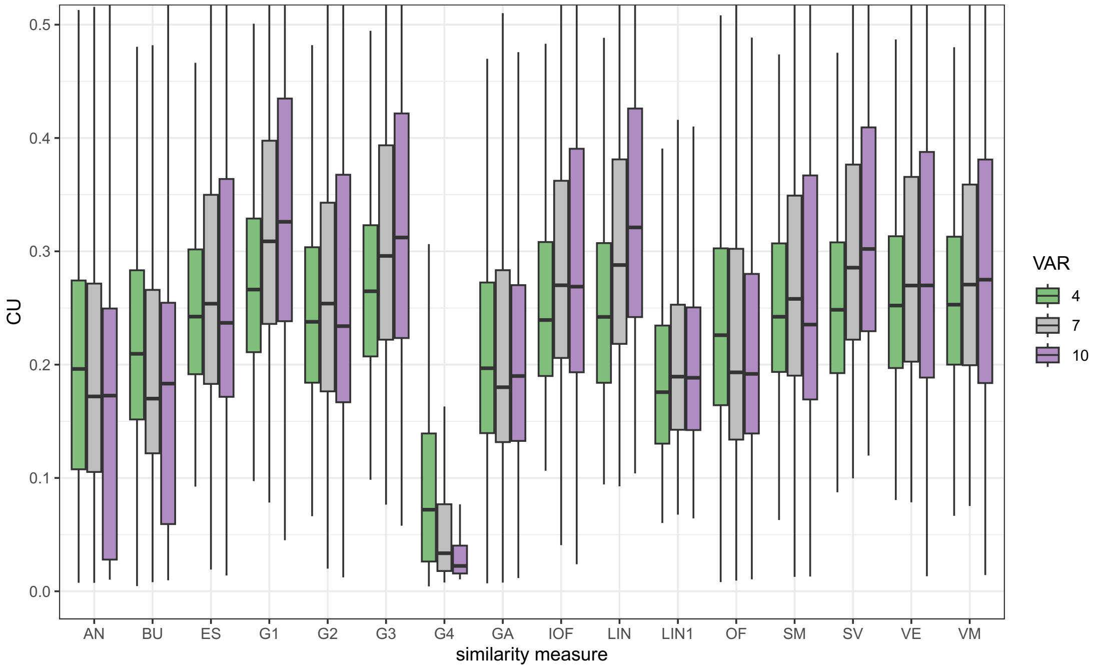


Figure 5.7: Boxplots of the CU values for three variable numbers (ALM)



Third, the influence of the number of categories (CAT), namely 3, 5, 7, of the clustered variables on the cluster quality was examined. The selected CAT levels cover the typical ranges of categories used in HCA tasks. Table 5.7 and Table 5.8 show MRS broken down by three CAT levels representing simple, medium, and complex dataset structures.

The outputs of both tables are in accordance, and they show that most of the similarity measures perform constantly across three numbers of variables in ALM. For instance, LIN performs well, no matter the number of categories. The most remarkable improvement of clustering performance with the increasing CAT is achieved by the measures ES and SM. These two measures perform almost identically, especially if the number of categories is constant. Then, their values have a monotonous relationship in the first step of their calculation; see their equations in Section 2.1. To a smaller extent, the measures IOF and SV also improve their relative performance when the number of categories is large.

On the contrary, the highest worsening of MRS is by the AN and OF measures, whose created clusters quality is overall poor in ALM. It is worth mentioning that G1 and G3 slightly deteriorate their clustering performance with the increasing CAT, but they still work outstandingly in ALM.

Table 5.7: MRS for three numbers of categories based on PSFE (ALM)

CAT	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
3	8.7	10.3	9.6	3.2	9.0	4.0	15.9	11.9	8.6	5.7	11.8	8.5	9.7	5.6	6.8	6.6
5	12.3	11.4	7.7	3.2	8.6	4.5	15.7	12.9	7.6	5.1	12.4	10.1	7.6	4.5	6.0	6.4
7	14.1	10.9	6.8	3.9	7.9	4.8	15.6	12.7	7.2	5.6	13.1	10.3	7.0	4.4	5.7	6.0

Table 5.8: MRS for three numbers of categories based on CU (ALM)

CAT	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
3	10.0	10.9	9.2	3.3	8.6	4.0	15.9	11.4	7.9	5.6	11.4	9.2	9.3	6.0	6.8	6.5
5	13.2	12.0	7.3	3.7	8.0	4.7	15.6	12.2	6.3	5.2	12.0	10.7	7.2	5.5	6.0	6.3
7	14.7	11.5	6.6	4.3	7.4	5.0	15.4	12.0	5.8	5.9	12.8	10.8	6.8	5.5	5.6	5.9

Figure 5.8 and Figure 5.9 presents the assessment using the boxplots. This time, the PSFE and CU criteria results are in accordance. The obtained results are not surprising. The best clusters for all the similarity measures are created if the clustered variables contain three categories representing the simple dataset structure. The clustering with five categories always provides higher evaluation criteria scores than the clustering with seven categories. However, the differences are much smaller compared to the clustering with three categories.

Regarding the particular similarity measures, one can notice the inferior performance of the G4 measure that creates the worst clusters among all similarity measures by distance. Table IX and Table X in Appendix C contain the numeric characteristics of both the boxplots.

Figure 5.8: Boxplots of the PSFE values for three numbers of categories (ALM)

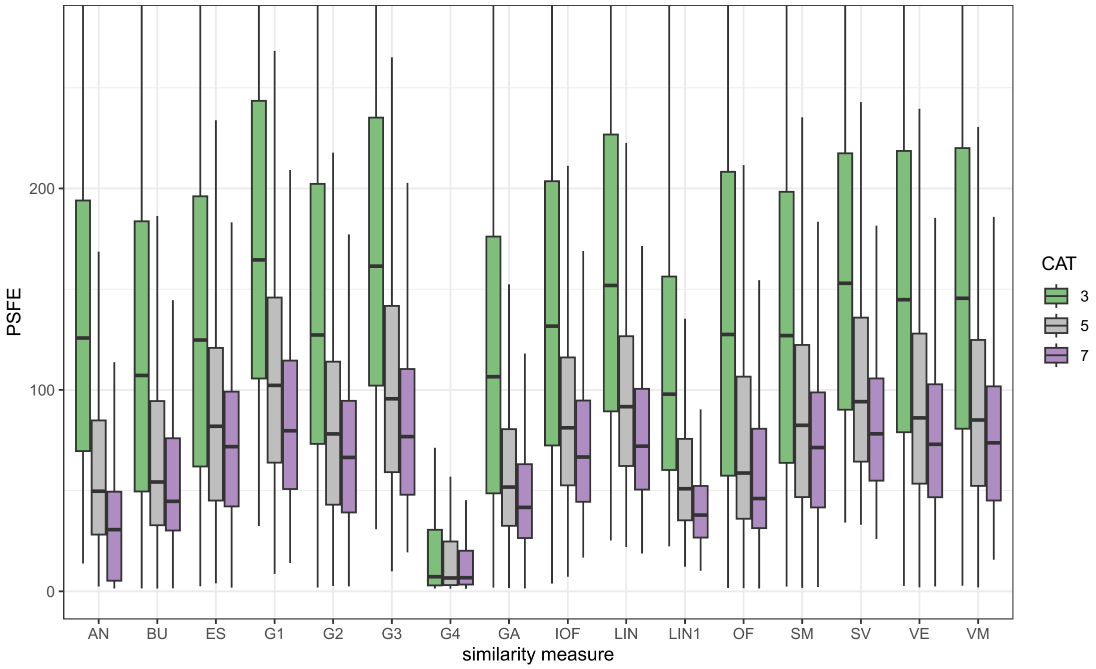
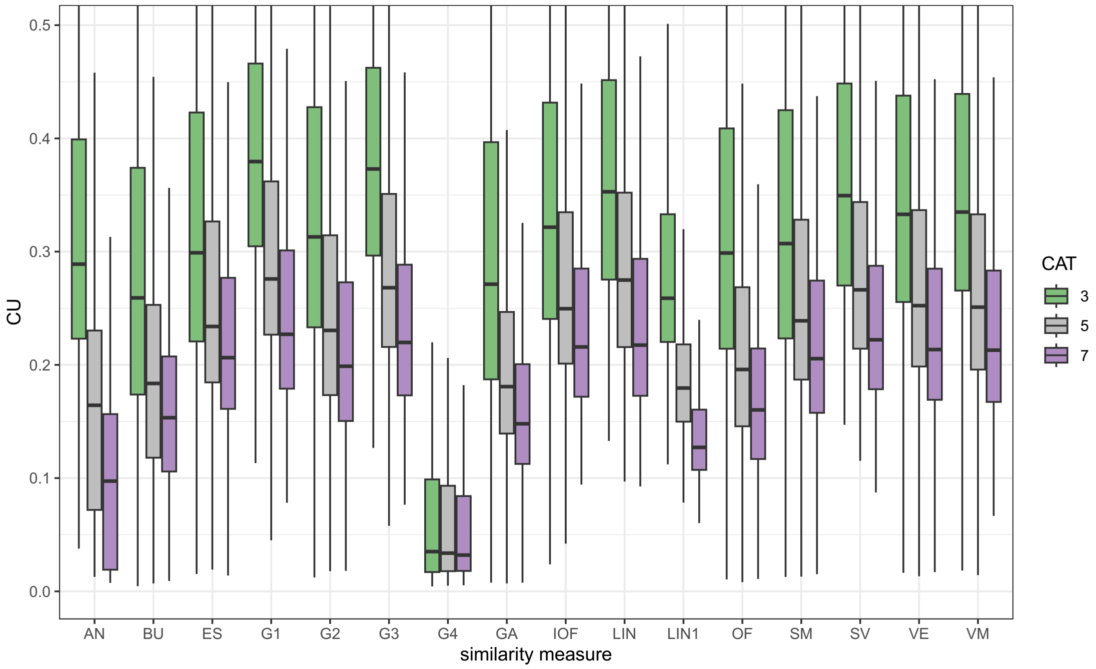


Figure 5.9: Boxplots of the CU values for three numbers of categories (ALM)



5.3.3 Complete linkage method

In CLM, a different set of similarity measures produces good-quality clusters than in ALM, as was demonstrated in Subsection 5.3.1. BU, GA, IOF, and G4 measures perform substantially better than in ALM. On the contrary, some clustering performances of some measures got worse, such as by ES, SM, G1, and G3. Still, there are some well-performing measures in both the linkages, namely LIN, VE, and VM.

Table 5.9 and Table 5.10 show MRS for 16 similarity measures broken down by three DIST levels according to the criteria PSFE and CU. Both criteria provide very similar results that differ only a little. Generally, the relative clustering performance of the examined similarity measures is not affected to a great extent by the decreasing DIST in CLM, but there are some differences among them. The measures G4, GA, and G2 perform relatively better by the small DIST but still do not belong to the best similarity measures. Smaller improvements also occur by the measures VE, VM, and IOF (according to CU) that generally construct good clusters in CLM. On the other hand, the measures ES, SM, G1, and G3 worsen their relative clustering performance with the decreasing DIST.

Table 5.9: MRS for three minimal between-cluster distances based on PSFE (CLM)

DIST	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
0.50	12.3	6.4	10.6	9.1	8.8	9.3	11.3	10.4	6.0	5.8	8.7	5.2	11.0	6.7	7.1	7.2
0.34	12.2	5.8	12.0	10.3	8.3	10.5	9.8	9.9	6.4	6.4	7.6	4.9	12.0	7.0	6.4	6.4
0.21	11.5	6.8	12.2	11.0	7.9	10.8	8.6	9.1	6.4	6.8	6.9	6.1	12.4	7.1	6.1	6.2

Table 5.10: MRS for three minimal between-cluster distances based on CU (CLM)

DIST	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
0.50	12.5	7.4	10.3	9.0	8.4	9.2	10.4	9.7	5.5	6.0	9.5	6.2	10.8	7.1	7.0	7.1
0.34	12.8	7.3	11.4	10.2	7.8	10.4	8.6	8.7	5.3	6.6	8.8	6.2	11.5	8.0	6.2	6.2
0.21	12.8	8.9	11.4	10.7	7.1	10.5	7.4	7.6	4.8	6.9	8.3	7.7	11.6	8.8	5.8	5.9

Figure 5.10 and Figure 5.11 express the absolute differences between the similarity measures based on PSFE and CU using the boxplots. Compared to ALM, see Figure 5.4 and Figure 5.5, there are generally smaller differences between the three DIST levels in CLM. When taking a closer look, one can observe that the cause lies in the usually lower clustering performance of CLM with lower values of both evaluation criteria. There are also generally smaller differences between the similarity measures than by ALM, so there is no completely unsuitable similarity measure for clustering. Both the charts express the relationships similarly, but the chart with the CU values presents the differences more clearly. The best-performing measures in CLM are IOF, LIN, and BU. They all have high median values with relatively small IQRs in all three examined DIST levels. The worst clusters create AN, ES, and a reference measure SM. Thus, researchers should avoid using the SM measure with CLM when clustering the categorical data. Table XI and Table XII in Appendix C contain the numeric characteristics of both the boxplots.

Figure 5.10: Boxplots of the PSFE values for three minimal between-cluster distances (CLM)

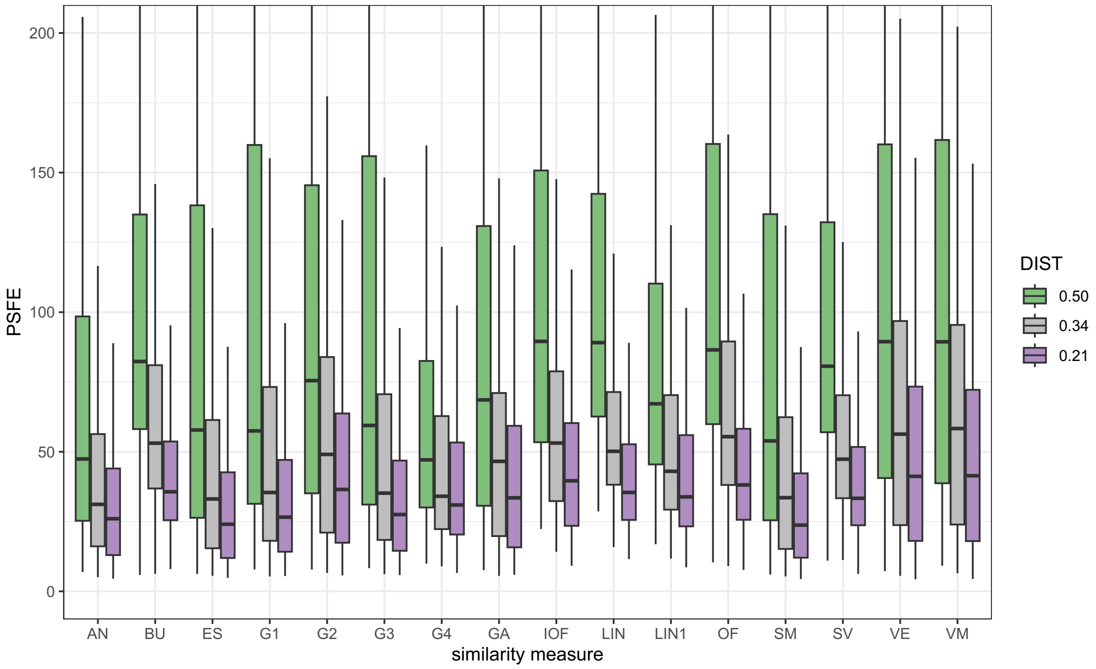


Figure 5.11: Boxplots of the CU values for three minimal between-cluster distances (CLM)

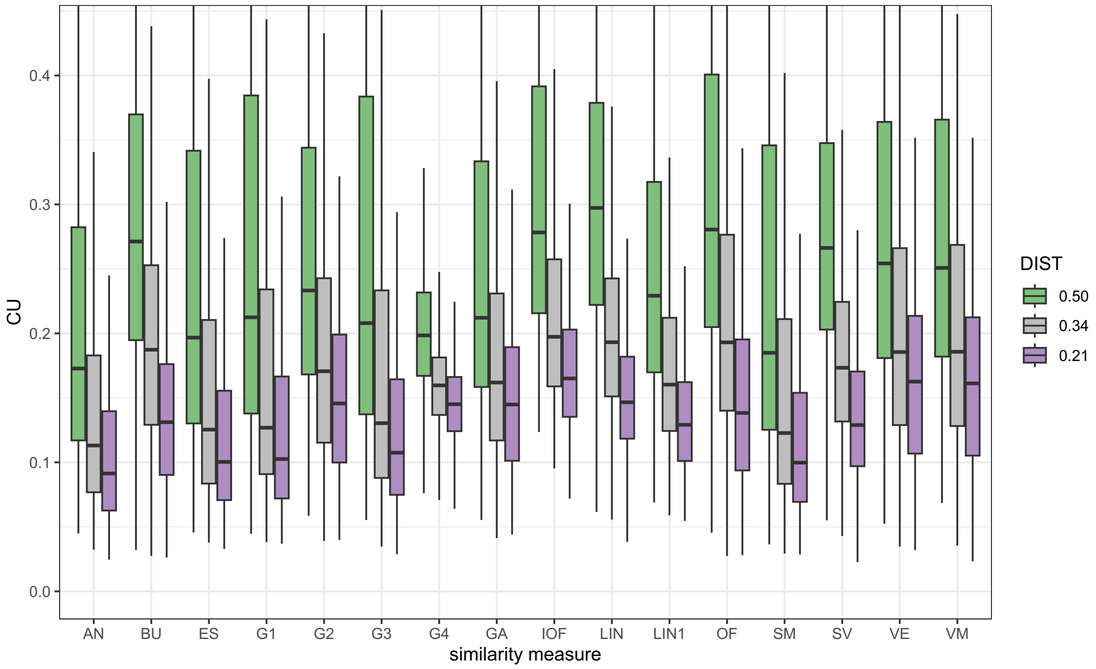


Table 5.11 and Table 5.12 present MRS based on PSFE and CU for the 16 similarity measures broken down by the VAR levels in CLM. Compared to ALM, where the clustering performance of the similarity measures was mostly constant (see Table 5.5 and Table 5.6), CLM shows interesting dependencies of the examined measures on the number of the clustered variables. Again, both the table outputs are in accordance, suggesting the relevance of the results.

Overall, the universal similarity measure in CLM is IOF which performs well across different numbers of variables. The measures LIN, OF, BU, and SV improve their relative clustering performance with the increasing number of variables, most of them to a large extent. Thus, they are suitable for datasets with many categorical variables.

On the contrary, the measures VE and VM create good-quality clusters when the number of clustered variables is four. When VAR increases, their relative clustering performance continuously deteriorates. Still, they provide the average-quality clusters by datasets with ten variables. The measures G2, GA, G4, and AM also deteriorate the created clusters' quality with the increasing VAR levels, but they perform poorly with lower VAR levels. The reference measure SM and further ES form constantly inferior clusters in CLM.

Table 5.11: MRS for three variable numbers based on PSFE (CLM)

VAR	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
4	11.4	9.3	11.1	11.0	6.9	11.0	9.3	8.5	6.6	7.9	7.5	8.1	11.4	8.4	3.9	3.9
7	12.0	5.2	12.5	10.4	8.6	10.5	9.4	9.9	6.7	6.0	7.2	4.4	12.7	6.4	6.9	7.1
10	12.5	4.6	11.2	9.0	9.5	9.1	11.0	11.0	5.5	5.3	8.5	3.8	11.3	6.0	8.8	8.8

Table 5.12: MRS for three variable numbers based on CU (CLM)

VAR	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
4	11.5	11.1	10.0	10.7	6.3	10.7	8.2	7.3	5.6	8.4	9.2	9.9	10.3	9.3	3.8	3.8
7	12.8	6.7	12.1	10.2	8.0	10.4	8.3	8.7	5.5	6.0	8.3	5.6	12.4	7.6	6.6	6.8
10	13.8	5.7	10.9	8.9	9.0	9.0	9.9	10.1	4.6	5.0	9.0	4.6	11.1	7.0	8.7	8.6

Figure 5.12 and Figure 5.13 show the absolute differences between evaluation criteria values broken down by the VAR levels for all the examined similarity measures in CLM. Similarly, as by ALM in Figure 5.6 and Figure 5.7, the evaluation criteria differ in the way they assess the cluster quality. The PSFE criterion always prefers the four-variable solution to the solutions with seven and ten variables that are more similar regarding their medians and IQRs. One can observe a superior performance of the VE and VM measures in datasets with four clustered variables compared to the rest of the similarity measures and also competitively good cluster quality by OF and BU in the higher VAR levels. Still, most of the PSFE values are lower than by ALM, and thus, ALM should be preferred to CLM unless there is a specific reason to use CLM.

The CU criterion generally expresses the exact relationships between the similarity measures as PSFE, but it proceeds differently with different numbers of variables. Then, the similarity measures performing well by higher VAR levels have a higher value of CU. This way, the outputs correspond more to the relative comparison presented in Table 5.11 and Table 5.12.

The numeric characteristics of both the boxplots are shown in Table XIII and Table XIV in Appendix C.

Figure 5.12: Boxplots of the PSFE values for three variable numbers (CLM)

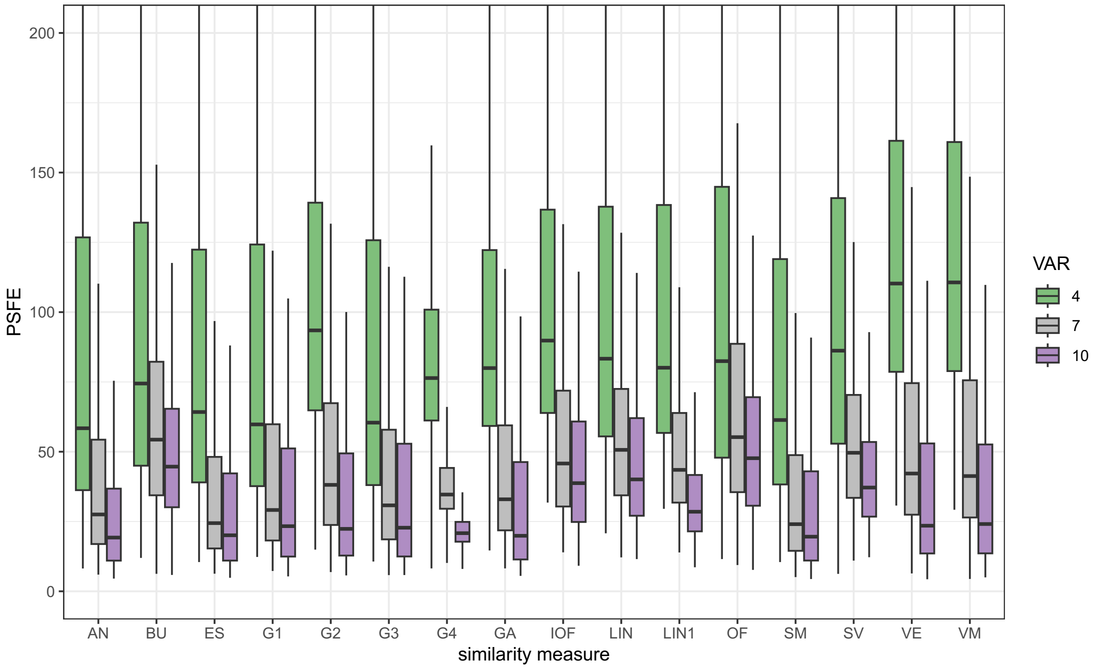


Figure 5.13: Boxplots of the CU values for three variable numbers (CLM)

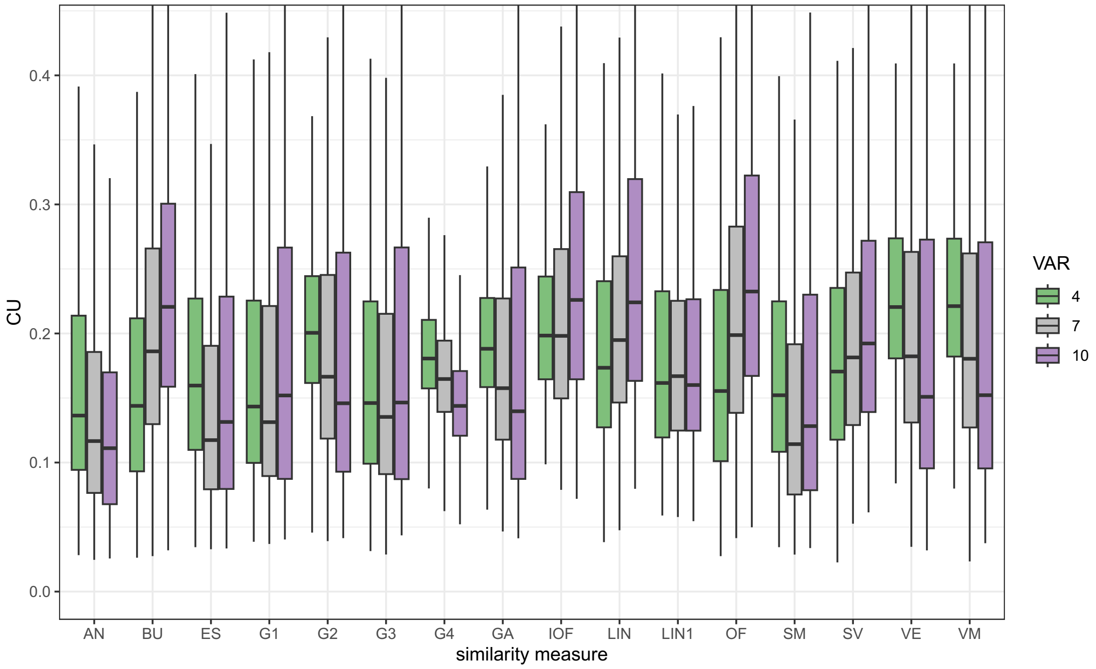


Table 5.13 and Table 5.14 contain MRS based on PSFE and CU for the examined similarity measures broken by the CAT levels in CLM. The similar MRS values in both tables indicate that the PSFE and CU criteria assess the created clusters similarly.

The results show that the number of categories is an influential factor for the cluster quality in CLM. The measures OF, VE, and VM create generally good clusters in CLM. However, their relative clustering performance slightly decreases with the increasing complexity of a dataset represented by increasing CAT levels.

The measures IOF, LIN, SV, LIN1, BU, and G4 improve their relative clustering performance with the increasing number of categories. While most of the mentioned measures generally perform well in CLM, the behavior of the G4 measure is surprising. It performs poorly in datasets with three categories by the clustered variables, but it creates clusters of outstanding quality in complex datasets containing variables with many categories.

The rest of the measures do not provide good-quality clusters. The G1, G2, and G3 measures perform relatively well in datasets with a simple structure represented by three categories. The ES, SM, and AN measures generally create poor-quality clusters in CLM.

Table 5.13: MRS for three numbers of categories based on PSFE (CLM)

CAT	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
3	10.3	7.4	9.8	7.1	7.6	7.4	15.1	9.2	7.7	8.9	9.9	5.2	9.8	9.2	5.8	5.7
5	12.5	5.7	12.5	11.1	8.5	11.2	8.5	10.8	5.7	5.6	7.2	5.2	12.5	5.7	6.5	6.8
7	13.1	6.0	12.5	12.3	8.8	12.1	6.1	9.5	5.3	4.7	6.0	5.8	13.1	5.9	7.4	7.4

Table 5.14: MRS for three numbers of categories based on CU (CLM)

CAT	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
3	11.5	9.0	9.2	7.0	7.2	7.3	14.9	8.5	7.0	8.7	10.0	6.2	9.3	9.7	5.5	5.4
5	13.2	7.1	12.0	11.0	8.0	11.0	7.1	9.5	4.5	5.7	8.4	6.5	12.1	7.0	6.4	6.6
7	13.5	7.5	11.8	12.0	8.2	11.8	4.3	8.1	4.2	5.1	8.1	7.4	12.5	7.1	7.2	7.2

Figure 5.14 and Figure 5.15 display the absolute differences between evaluation criteria values broken down by the CAT levels for all the examined similarity measures. The outputs of both charts are mostly in accordance. The best clusters are almost always[4] obtained if the variables with three categories are used, followed by the medium and complex dataset structure represented by five and seven categories. The differences between the CAT levels are much more substantial compared to VAR or DIST levels in the previous analyses. Both charts show that the created clusters are of the highest quality when the clustered variables contain only a small number of categories. With an increasing number of categories, the cluster quality decreases tremendously. Table XV and Table XVI in Appendix C contain the numeric characteristics for both the boxplots.

[4] The exception is the G4 measure by the CU criterion.

Figure 5.14: Boxplots of the PSFE values for three numbers of categories (CLM)

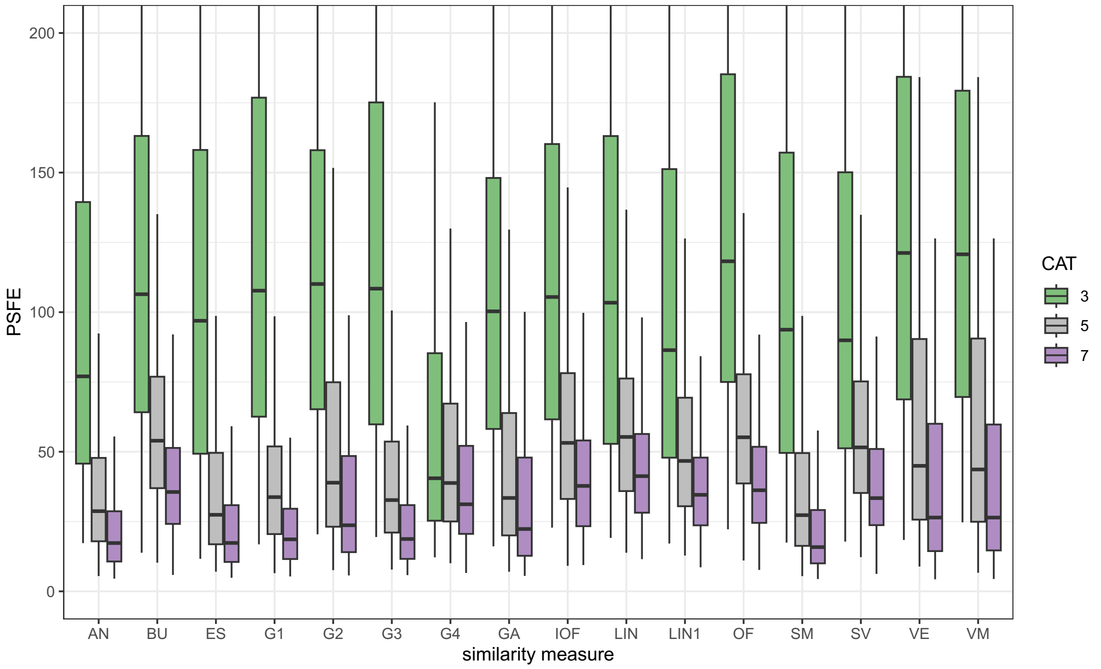
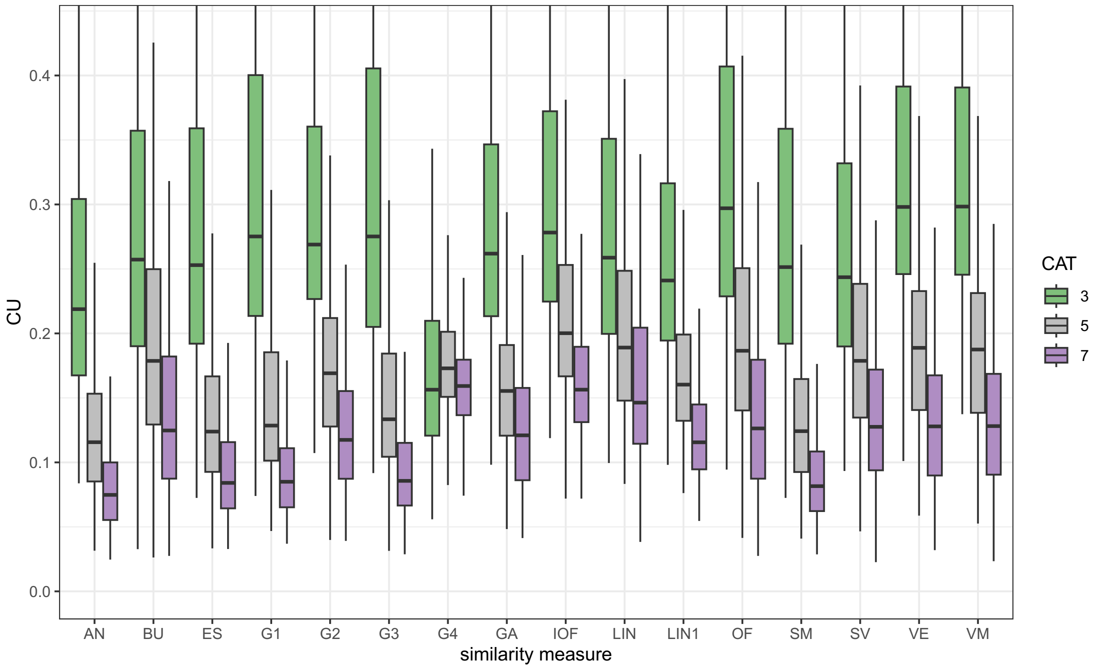


Figure 5.15: Boxplots of the CU values for three numbers of categories (CLM)



5.3.4 Single linkage method

Table 5.15 and Table 5.16 comprise MRS based on PSFE and CU for 16 similarity measures broken down by the DIST levels in SLM. Both evaluation criteria provide very similar MRS outputs. Some measures, whose relative clustering performance was poor in the other two linkages, create good clusters, e.g., AN. On the other hand, the SV measure, with a good clustering performance in ALM and CLM, generally performs poorly in SLM.

Most similarity measures provide a constant relative clustering performance in the three DIST levels; thus, they do not depend on it. An improvement with the decreasing minimal between-cluster distance can be observed by the measures AN, GA, and G4. The opposite direction of dependence occurs by the IOF measure and then further by ES and SM measures.

Table 5.15: MRS for three minimal between-cluster distances based on PSFE (SLM)

DIST	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
0.50	8.0	7.6	5.8	9.0	9.5	8.9	11.4	10.4	7.6	8.9	12.3	7.5	5.5	12.6	5.4	5.5
0.34	6.1	6.3	6.8	11.0	9.9	11.5	8.2	8.0	9.2	10.0	10.9	6.9	6.2	11.6	6.6	6.7
0.21	5.4	6.0	7.2	10.2	9.9	11.0	8.9	8.3	10.2	9.7	10.5	6.9	6.7	11.2	6.9	6.9

Table 5.16: MRS for three minimal between-cluster distances based on CU (SLM)

DIST	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
0.50	8.3	8.1	5.6	9.4	9.1	9.4	10.8	10.3	7.1	9.6	12.2	7.8	5.3	12.3	5.3	5.4
0.34	6.8	7.2	6.6	11.5	9.4	11.9	7.1	7.8	8.3	10.9	10.5	7.6	6.0	11.4	6.4	6.5
0.21	6.1	7.3	6.9	10.6	9.4	11.5	7.4	8.0	9.2	10.7	10.1	7.9	6.4	11.0	6.8	6.9

Figure 5.16 and Figure 5.17 show the absolute differences for the PSFE and CU evaluation criteria values in three DIST levels using the boxplots. Both charts show that the outputs of the examined criteria are in accordance.

The results show that most of the examined similarity measures fail to create satisfactory clusters by the medium DIST (0.34) representing partly intersecting clusters or the low DIST (0.21) representing intersecting clusters since the PSFE and CU values are close to zero. Thus, SLM can be used only in datasets with non-intersecting clusters (large DIST of 0.50) and only with specific similarity measures, such as VE and VM, which are the best-performing measures in SLM. IOF, ES, and SM also create relatively good clusters at the large DIST level. On the contrary, G4, GA, SV, and LIN1 form poor-quality clusters in all situations.

When comparing the absolute values of the evaluation criteria in SLM with the other two linkages, they are much lower than in ALM and CLM by most of the examined similarity measures, making SLM unsuitable for HCA of categorical data in most practical situations. Table XVII and Table XVIII in Appendix C contain the numeric characteristics for both the boxplots.

Figure 5.16: Boxplots of the PSFE values for three minimal between-cluster distances (SLM)

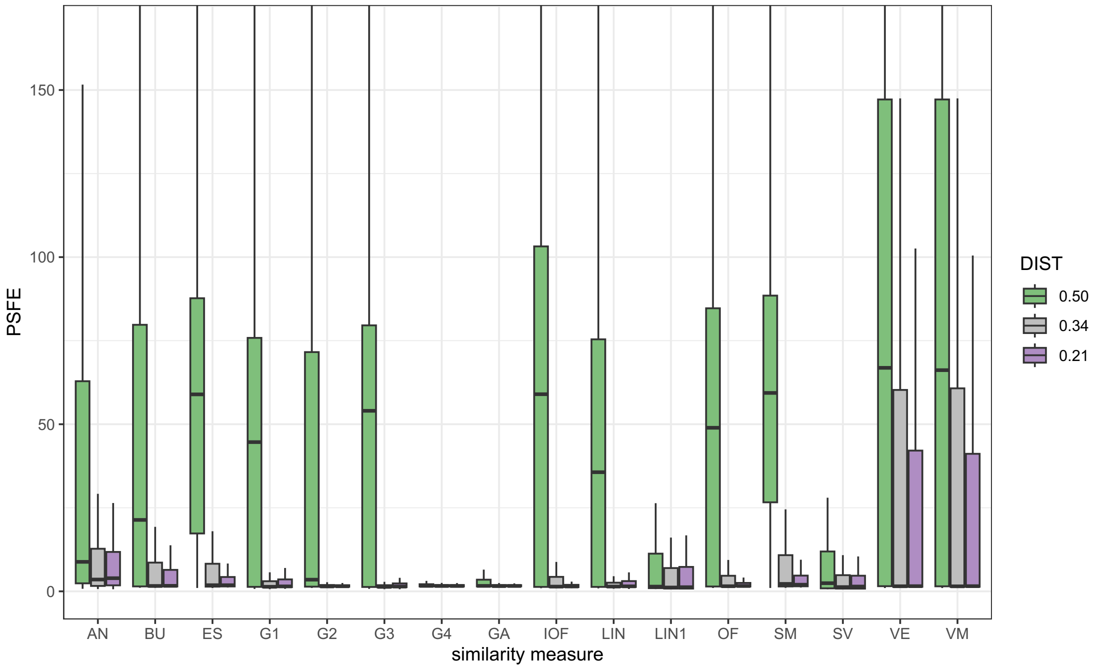


Figure 5.17: Boxplots of the CU values for three minimal between-cluster distances (SLM)

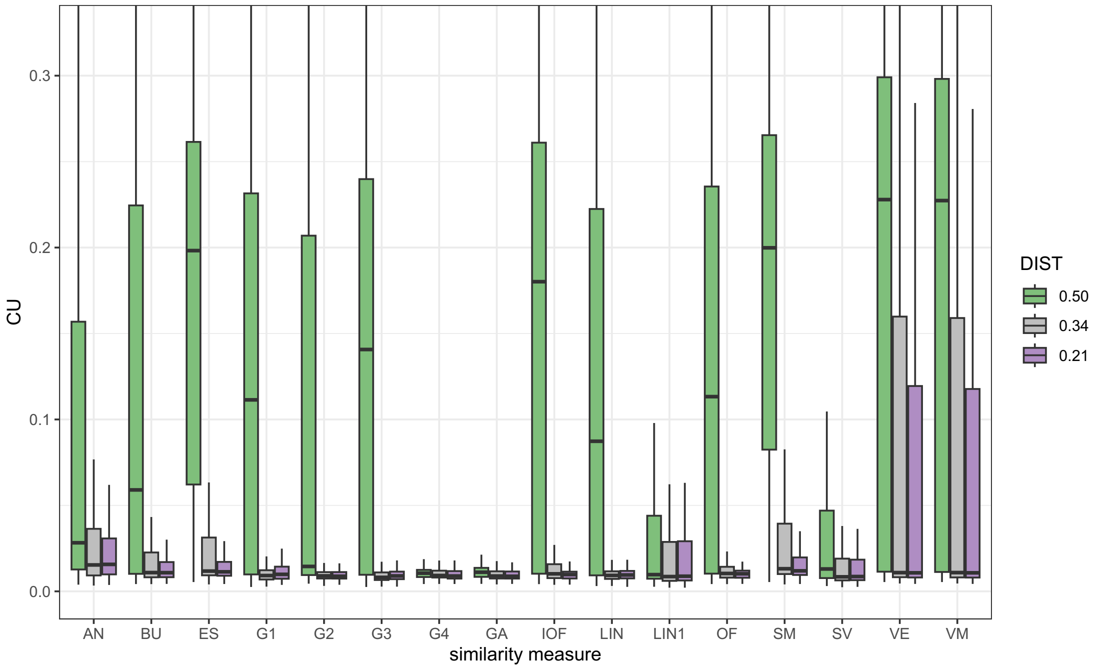


Table 5.17 and Table 5.18 contain MRS based on PSFE and CU criteria for the examined similarity measures in SLM that are broken down by three VAR levels. The results indicate that the relative clustering performance of the majority similarity measures remains constant. For instance, the AN, BU, and OF measures, whose clusters are of good quality overall. When increasing VAR, the highest improvement of the relative clustering performance is achieved by the G4 measure, which is closely followed by GA. Smaller increases in the relative cluster quality can also be observed by the measures ES and SM.

On the contrary, the largest worsening of the cluster quality occurs by the VE and VM measures. They perform outstandingly in datasets with four variables. However, their clusters are of average quality in the higher VAR levels. A smaller extent of cluster quality deterioration can also be observed by the IOF and SV measures which generally perform worse than VE and VM.

Table 5.17: MRS for three numbers of variables based on PSFE (SLM)

VAR	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
4	6.6	7.1	7.3	9.8	11.8	10.6	13.7	12.9	6.7	9.7	10.8	6.8	7.4	10.1	2.3	2.3
7	5.5	6.5	6.6	10.1	9.2	10.5	8.3	8.0	10.3	9.1	10.2	7.8	5.5	12.0	8.2	8.3
10	7.5	6.3	5.9	10.4	8.2	10.4	6.4	5.8	10.0	9.9	12.6	6.8	5.5	13.3	8.5	8.5

Table 5.18: MRS for three numbers of variables based on CU (SLM)

VAR	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
4	7.0	7.9	7.0	9.6	11.7	10.5	13.5	12.9	6.5	9.9	10.6	7.5	7.1	9.7	2.3	2.3
7	5.9	7.3	6.5	10.7	8.6	11.1	7.2	7.9	9.4	10.2	9.7	8.5	5.3	11.5	8.0	8.2
10	8.3	7.3	5.7	11.2	7.5	11.1	4.6	5.4	8.7	11.1	12.6	7.4	5.3	13.4	8.2	8.3

Figure 5.18 and Figure 5.19 show the absolute differences in the internal evaluation criteria values using the boxplots in three VAR levels. Both charts provide similar outputs in SLM, which contradicts the other two linkage methods, where the CU criterion usually prefers higher VAR levels.

The results show that decent clusters are provided only in datasets with four variables using specific similarity measures. In particular, the best clusters are provided by VE and VM measures, followed by IOF. In datasets with seven or ten variables, the PSFE and CU evaluation criteria medians are close to zero with low IQR; thus, the created clusters are of poor quality. Since the absolute differences between the evaluation criteria values by the similarity measures are small by higher VAR levels, some measures, such as G4 and GA, achieved good MRS in Table 5.17 and Table 5.18 although they created poor clusters in all situations. Table XIX and Table XX in Appendix C contain the numeric characteristics for both the boxplots.

Figure 5.18: Boxplots of the PSFE values for three variable numbers (SLM)

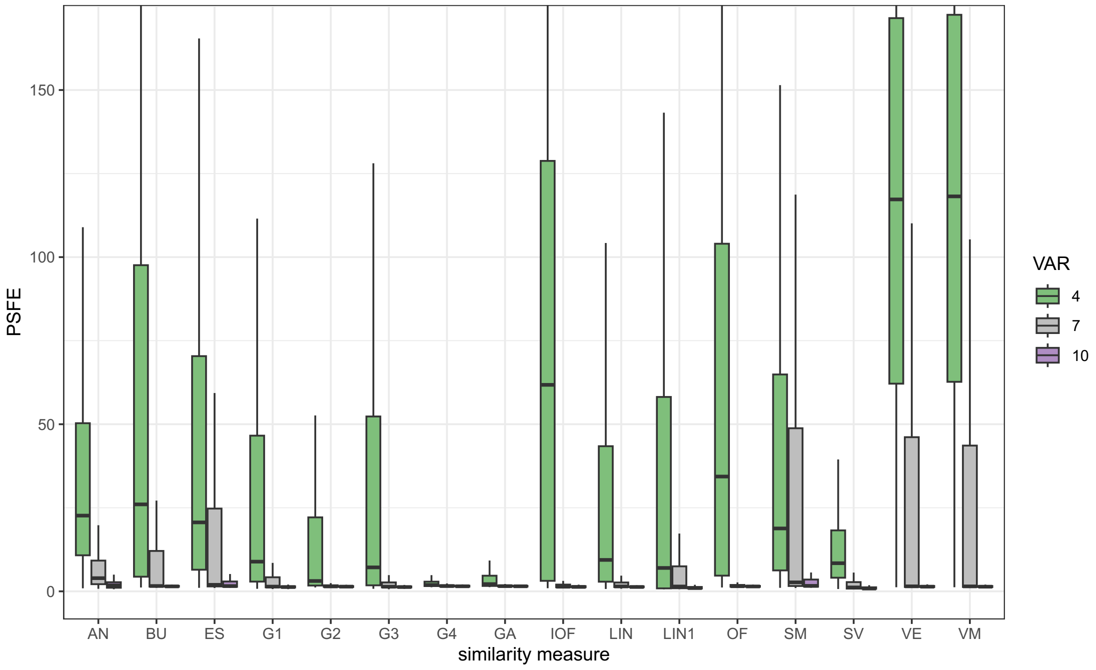


Figure 5.19: Boxplots of the CU values for three variable numbers (SLM)

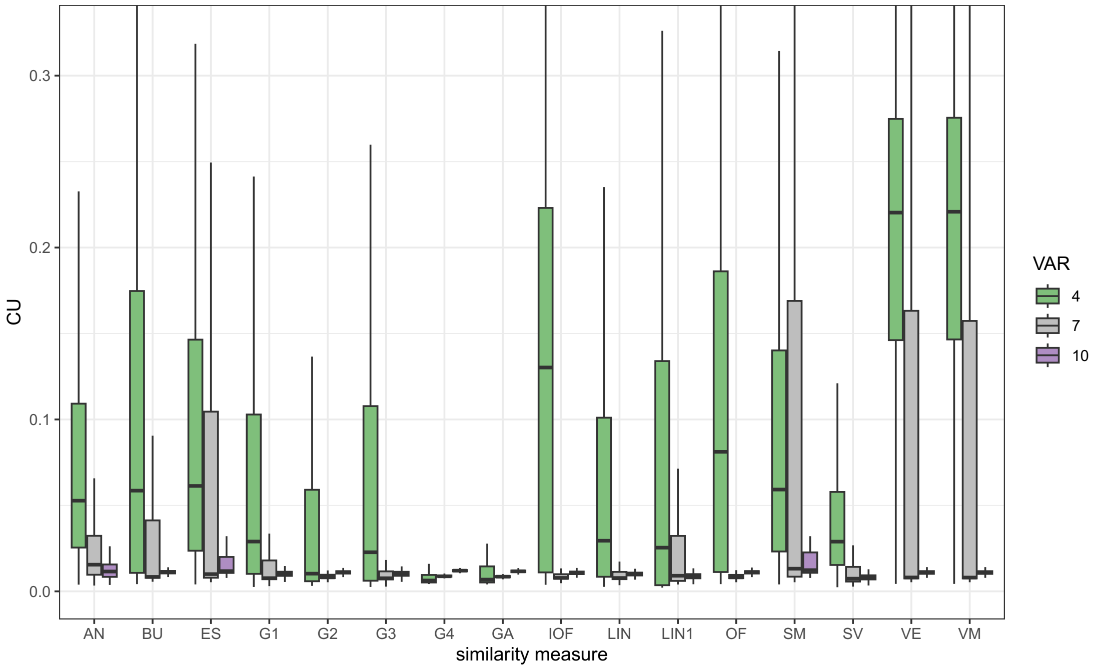


Table 5.19 and Table 5.20 show MRS based on PSFE and CU for 16 similarity measures broken down by the three different numbers of categories in SLM. Similarly, as in the previous analyses, both tables present similar outputs.

Most of the examined similarity measures do not substantially change their relative clustering performance with the increasing CAT levels, such as BU, VE, and VM, which perform generally well, or the measures G1, G2, and G3, whose relative clustering performance is below the average. The most considerable improvement of MRS with the increasing CAT occurs by the similarity measure AN, followed by ES and SM. Further improvements can also be observed by the measures GA and G4 with below-average MRS. On the other hand, the LIN1 measure performs relatively well in datasets containing variables with three categories. However, its clustering performance decreases vastly with more complex datasets containing variables with five or seven categories. Similar behavior can be observed in the IOF and SV measures.

Table 5.19: MRS for three numbers of categories based on PSFE (SLM)

CAT	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
3	8.5	6.4	7.4	10.2	10.7	11.1	10.2	10.1	7.8	10.2	6.8	7.5	7.4	9.7	6.1	6.1
5	6.4	6.6	6.8	10.5	9.5	10.4	9.5	8.7	9.1	9.1	12.3	6.8	5.5	12.5	6.1	6.1
7	4.7	6.8	5.6	9.6	9.1	10.0	8.7	8.0	10.1	9.2	14.5	7.1	5.6	13.2	6.8	6.9

Table 5.20: MRS for three numbers of categories based on CU (SLM)

CAT	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
3	8.7	6.9	7.1	10.6	10.3	11.4	9.8	10.0	7.3	10.3	6.9	7.8	7.1	10.1	5.9	5.9
5	6.9	7.7	6.7	10.9	9.1	10.8	8.3	8.6	8.3	10.1	11.8	7.6	5.2	12.1	6.0	6.0
7	5.6	7.8	5.4	10.0	8.4	10.4	7.2	7.7	9.1	10.9	14.1	7.9	5.4	12.5	6.7	6.8

Figure 5.20 and Figure 5.21 display the absolute differences of the PSFE and CU values in three different dataset complexities represented by CAT levels in SLM. Both charts provide similar outputs for the PSFE and CU evaluation criteria.

In SLM, meaningful clusters are obtained primarily on datasets with a simple structure. In more complex datasets containing five or seven categories, the clustering performance is deficient, represented by the median values close to zero and low IQR values by most similarity measures (except for VE and VM). In all dataset complexities, VE and VM provide the best clusters. In the simple dataset structure, good clusters also occur by the measures BU, LIN1, ES, and SM. Still, the evaluation criteria values are much lower than in ALM or CLM. The numeric characteristics of both the boxplots are presented in Table XXI and Table XXII in Appendix C.

Figure 5.20: Boxplots of the PSFE values for three numbers of categories (SLM)

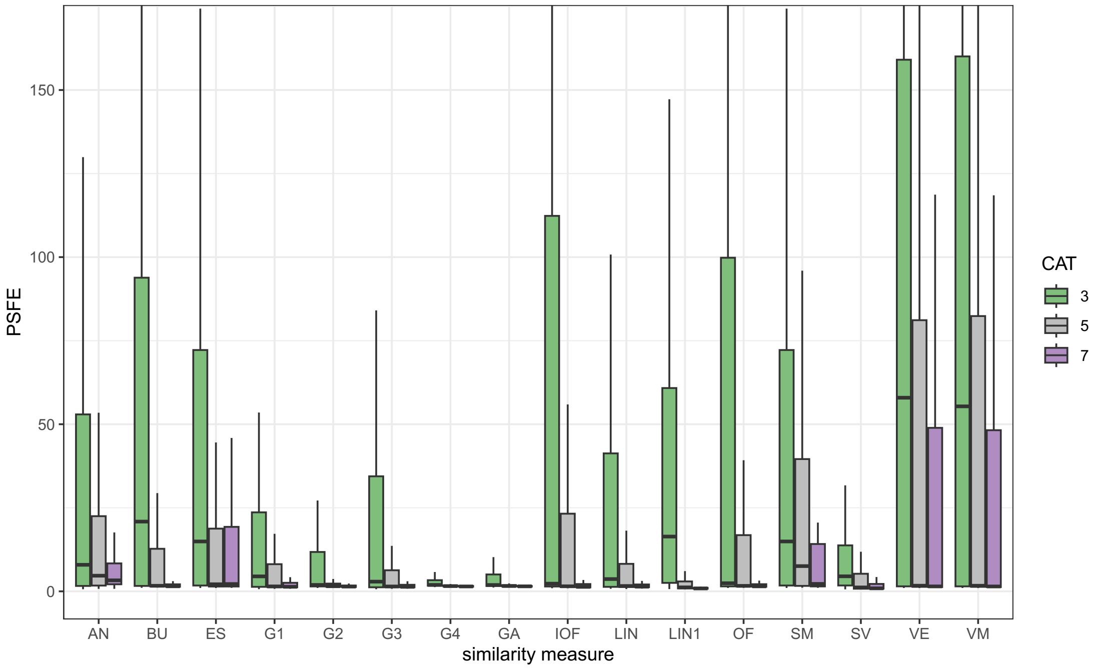
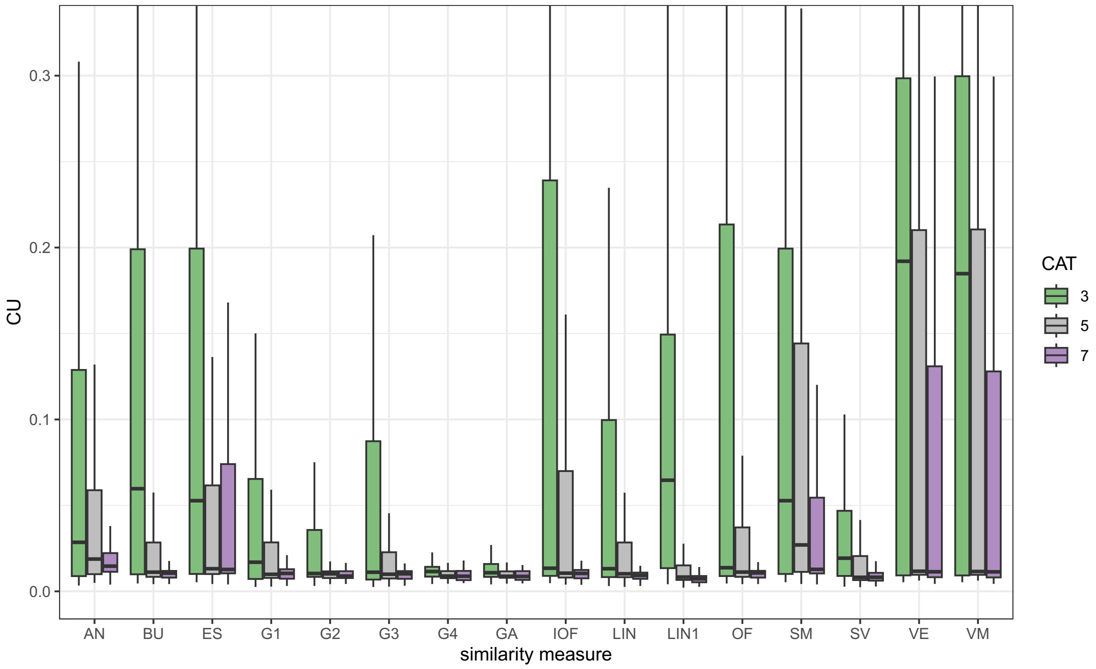


Figure 5.21: Boxplots of the CU values for three numbers of categories (SLM)



5.3.5 Experiment summary

The last part of the experiment focuses on summarizing the results obtained in the previous subsections. Since the linkage method proved to be a very influential analysis setting, mutual interactions of the similarity measures and the linkage methods will be examined in this subsection. Consequently, the most appropriate similarity measures based on the selected dataset properties and the used linkage will be recommended.

To determine which combinations of similarity measures and linkage methods create the best possible clusters, mean PSFE and CU scores of 48 combinations of 16 similarity measures and three linkage methods were ordered and ranked in decreasing order. The results are presented in Table 5.21 and Table 5.22.

Table 5.21: Ordered combinations of similarity measures and linkage methods based on PSFE

order	1	2	3	4	5	6	7	8	9	10	11	12
measure	G1	G3	SV	LIN	VE	VM	SM	IOF	ES	G2	OF	BU
linkage	ALM	ALM	ALM	ALM	ALM	ALM	ALM	ALM	ALM	ALM	ALM	ALM
mean	135.0	130.6	126.4	122.8	120.1	120.1	111.7	111.7	110.9	109.6	99.5	89.7
order	13	14	15	16	17	18	19	20	21	22	23	24
measure	AN	OF	GA	VE	VM	BU	LIN	IOF	LIN1	SV	G2	LIN1
linkage	ALM	CLM	ALM	CLM	CLM	CLM	CLM	CLM	ALM	CLM	CLM	CLM
mean	86.9	84.0	83.8	80.0	79.7	76.8	76.7	76.7	73.4	73.0	69.4	67.7
order	25	26	27	28	29	30	31	32	33	34	35	36
measure	G1	G3	GA	ES	SM	AN	VE	VM	G4	IOF	OF	SM
linkage	CLM	CLM	CLM	CLM	CLM	CLM	SLM	SLM	CLM	SLM	SLM	SLM
mean	67.3	66.9	63.4	59.4	58.5	56.6	55.2	55.1	49.8	34.3	30.1	28.8
order	37	38	39	40	41	42	43	44	45	46	47	48
measure	BU	ES	G4	AN	G3	G1	LIN	LIN1	G2	SV	GA	G4
linkage	SLM	SLM	ALM	SLM	SLM	SLM	SLM	SLM	SLM	SLM	SLM	SLM
mean	27.8	27.5	21.6	21.6	20.7	19.9	18.6	15.8	15.0	7.7	7.2	2.6

Both tables assess the combinations of similarity measures and the linkage methods similarly. However, there are some exceptions, e.g., the BU measure with ALM, which is on the 12th rank (row *order*) according to PSFE, and the 20th rank based on CU. The results show that the three linkage methods (row *linkage*) are well separated by the mean evaluation criteria values (row *mean*), which are averaged across all the examined dataset properties, and that there are only slight overlaps by either well- or poorly-performing similarity measures (row *measure*), e.g., the OF measure in CLM or the G4 measure in ALM. It is apparent that ALM mostly outperforms CLM and SLM (in this order) according to both evaluation criteria. Therefore, ALM should be preferred unless there is a specific reason for using other linkage methods.

The G1 measure with ALM creates the best clusters. They are followed by G3 with ALM and further by SV, LIN, VE, and VM (all with ALM). In CLM, the orders of the combinations based on PSFE and CU are not as unambiguous as in ALM, but the measures OF, IOF, VE, LIN, VM, and BU perform well with this linkage. The measures VE and VM work well with SLM. However,

Table 5.22: Ordered combinations of similarity measures and linkage methods based on CU

order	1	2	3	4	5	6	7	8	9	10	11	12
measure	G1	G3	LIN	SV	VE	VM	IOF	SM	ES	G2	IOF	OF
linkage	ALM	ALM	ALM	ALM	ALM	ALM	ALM	ALM	ALM	ALM	CLM	ALM
mean	0.32	0.31	0.30	0.30	0.28	0.28	0.28	0.27	0.27	0.27	0.23	0.23
order	13	14	15	16	17	18	19	20	21	22	23	24
measure	LIN	OF	VE	VM	GA	BU	LIN1	BU	G2	SV	GA	AN
linkage	CLM	CLM	CLM	CLM	ALM	CLM	ALM	ALM	CLM	CLM	CLM	ALM
mean	0.22	0.22	0.22	0.22	0.21	0.21	0.21	0.20	0.20	0.20	0.19	0.19
order	25	26	27	28	29	30	31	32	33	34	35	36
measure	LIN1	G1	G3	ES	SM	G4	AN	VE	VM	SM	ES	IOF
linkage	CLM	CLM	CLM	CLM	CLM	CLM	CLM	SLM	SLM	SLM	SLM	SLM
mean	0.19	0.19	0.19	0.17	0.17	0.17	0.15	0.11	0.11	0.08	0.08	0.08
order	37	38	39	40	41	42	43	44	45	46	47	48
measure	G4	OF	BU	G3	G1	AN	LIN	G2	LIN1	SV	GA	G4
linkage	ALM	SLM	SLM	SLM	SLM	SLM	SLM	SLM	SLM	SLM	SLM	SLM
mean	0.07	0.07	0.06	0.06	0.05	0.05	0.05	0.04	0.04	0.02	0.02	0.01

these two measures create good-quality clusters in all the examined linkages.

The newly examined similarity measures in this thesis, namely AN, BU, GA, and SV, differ substantially in their clustering performance, which is further enhanced by the linkage method used. The AN measure generally performs poorly. It creates average-quality clusters in SLM and poor clusters in the other two linkages. The BU measure creates relatively good clusters in CLM and SLM. Thus, it is a good option when ALM cannot be used. The clustering performance of the GA measure is average in CLM and SLM and poor in ALM. Therefore, it cannot be recommended for common use. The SV measure belongs to the best similarity measures in ALM, and it performs nicely in CLM as well. It does not work well in SLM, which is not a big drawback since this linkage generally provides poor clusters in HCA. Thus, only SV can be expected to create outstanding clusters from the newly examined similarity measures, especially in ALM.

The reference similarity measure, SM, creates good clusters in SLM. However, these clusters are usually useless in practice since SLM performs poorly overall. In ALM, the measure creates decent-quality clusters, which cannot be compared to clusters produced by the best similarity measures, such as G1 or LIN, but they are still good enough. In CLM, the SM measure produces poor-quality clusters; thus, it should not be used with this linkage method.

A special comment is dedicated to the ES measure, which creates similar clusters to the SM measure in the conducted experiment. The reason for this behavior is that with the fixed number of categories in the clustered variables, ES has no advantage over SM. In fact, in the previous research performed by Šulc (2016) and Šulc and Řezanková (2019) on datasets with different numbers of categories in generated variables, ES belonged among the best similarity measures in CLM and SLM. Thus, the ES measure could not unleash its full potential in the current experiment design. On the other hand, the current experiment design with fixed numbers of variables enables a better assessment of the different levels of complexity represented by different numbers of categories of the clustered variables.

There does not exist one universal similarity measure. Still, the main outputs of this experiment can be summarized in Table 5.23, which recommends several similarity measures for each linkage method. The similarity measures are divided into universal ones, which perform well no matter the clustered dataset properties, and specific ones, which create exceptional clusters if the particular dataset's properties are satisfied.

Table 5.23: Recommended similarity measures based on the dataset properties and the used linkage method

Linkage	Measure(s)	Use
ALM	G1, G3	universal
	LIN, SV	higher number of variables
CLM	IOF	universal
	OF, BU, LIN	higher number of variables and categories
	VE, VM	lower number of variables and categories
SLM	ES, SM	universal
	VE, VM	lower number of variables and categories

6 Comparison of Evaluation Criteria for Categorical Data Clustering

This chapter deals with the second goal of this thesis, the assessment of evaluation criteria for categorical data, which was presented in Introduction. The aim is to compare selected internal evaluation criteria for categorical data presented in Chapter 3 and to analyze their mutual relationships from different perspectives. The experiment excludes the WCM and WCE criteria, also presented in Chapter 3, since they are not designed to recommend the optimal number of clusters. Thus, 11 internal evaluation criteria are analyzed in this research. The experiment conclusions should help a researcher decide which evaluation criterion is suitable for a particular situation or inform which criteria assess the cluster quality almost identically. Another objective is to examine the relationship between the discussed internal criteria and the adjusted Rand index, which is a typical representative of the external criteria.

The chapter is divided into three sections. The first one describes the generated data and the similarity measures used in the experiment. The second one presents the methods used for evaluation criteria assessment, and the third one contains the conducted experiment.

6.1 Data Generation and Choice of Similarity Measures

The generated datasets for the experiment are obtained using the updated `gen_object()` function introduced by Šulc (2016), in the same way as it was described in Section 5.1.

In the experiment, 81 different dataset settings were used; see Figure 6.1. The datasets were generated with two, four, and six clusters. Three minimal between-cluster distances (0.21, 0.34, 0.50) were used, representing intersecting, partly intersecting, and almost non-intersecting clusters. Next, the datasets were generated with three different numbers of variables (4, 7, 10) covering the typical range of clustering of categorical datasets. There are three numbers of categories (3, 5, 7) by the generated variables representing simple, medium, and complex dataset structures. The number of objects in generated datasets was firmly set to 600 cases. Each dataset setting combination was replicated one hundred times to ensure the robustness of the obtained results. In total, this makes 8,100 generated datasets used for the analysis.

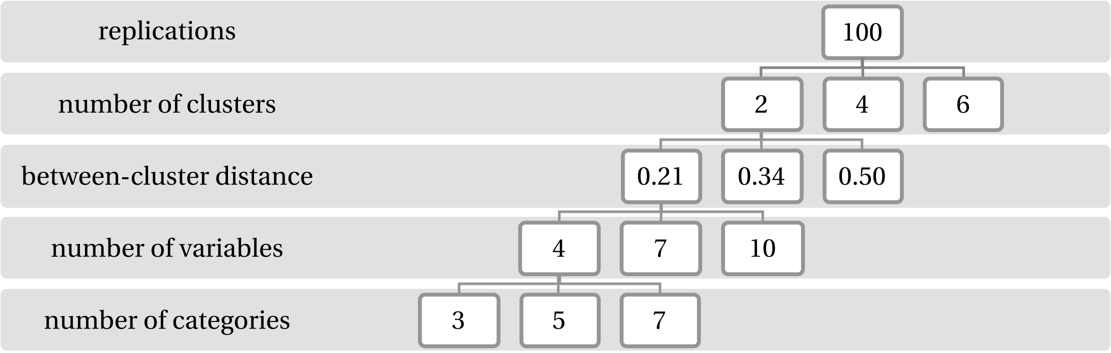


Figure 6.1: Dataset generation scheme for the second experiment

Chapter 5, and also the older research presented by Šulc (2016), proved that choosing a similarity measure can substantially influence the created clusters' quality in HCA of categorical data. Since this experiment does not aim to further analyze the similarity measures for categorical data but to analyze and compare the selected internal evaluation criteria for categorical data, six well-performing but different similarity measures for categorical data were chosen for the analysis. The ES measure uses the number of categories for the similarity definition; IOF uses the absolute frequencies of categories, G1 and LIN are based on relative frequencies, and VE uses the variable's entropy. The SM measure, known as the simple matching approach, is considered a reference similarity measure. Selecting only the well-performing similarity measures should ensure that the experiment results will not be biased by poorly performing measures, such as G4 or GA.

6.2 Methods for Evaluation Criteria Assessment

The internal evaluation criterion values can be either compared with values of other evaluation criteria using *correlation analysis* or analyzed across the values of dataset properties using *analysis of variance* (ANOVA). When comparing the created cluster partition with a vector with the known cluster membership, the *adjusted Rand index* (ARI) can be used. Since ARI was already presented in Section 3.1, this section briefly describes the remaining two methods used in the experiment.

The main output of the correlation analysis is the *correlation coefficient*, which is defined as the ratio between the covariance of two variables with two evaluation criteria values, e.g., PSFE and CU, and the product of their standard deviations. The correlation coefficient takes on values from -1 to 1, where the absolute values close to one indicate strong linear dependence and values close to zero linear independence. Nonlinear relationships cannot be expressed by this coefficient.

ANOVA is usually performed when analyzing relationships between quantitative and qualitative variables. The quantitative variable contains the values of a given evaluation criteria, and the qualitative one expresses a particular property of the clustered dataset, e.g., the number of clusters or variables. The method compares the between-group sum of squares (SS_B) with the within-group sum of squares (SS_W). The higher the SS_B variability, the more strongly the quantitative variable depends on the categorical one and vice versa. The strength of dependence can be expressed using the *eta-squared coefficient* η^2, which is defined as a ratio of SS_B and the total sum of squares in a dataset ($SS_T = SS_B + SS_W$). The eta-squared coefficient can take on values from zero to one. The values close to one indicate a high dependence, and close to zero show a low dependence of the evaluation criterion values on a given dataset's property.

6.3 Experiment

The experimental section consists of four subsections. The first one explores dependencies and differences between the examined evaluation criteria, the second one assesses the relationships between the internal and external criteria, and the third one investigates the dependences of the internal evaluation criteria on the properties of the clustered datasets and the used similarity measures. The results are summarized in the fourth subsection.

The analysis was performed on 81 types of datasets whose generation process was explained in Section 6.1. A series of HCAs for two to seven clusters with six selected similarity measures listed in section 6.1 and the average linkage method (ALM) were applied to each dataset. ALM was chosen since it generally provides the best cluster quality. In total, $81 \times 100 \times 6 = 48,600$ HCA outputs were obtained. Each was evaluated by 11 examined internal criteria presented in Section 3.2 (except for WCM and WCE). The evaluation criteria were analyzed mainly by the correlation and eta-squared coefficients and ARI. The experiments' scripts are discussed in Table II in Appendix B.

6.3.1 Similarity of evaluation criteria

Correlation analysis and multidimensional scaling are performed to determine if the studied evaluation criteria assess the cluster quality similarly. The analyzed data consists of evaluation criteria values for the number of clusters for which the datasets were generated, e.g., the value for PSFM in the two-cluster solution for the dataset generated with two original clusters and the value for PSFM in the four-cluster solution for the dataset generated with four original clusters. Thus, each correlation coefficient value is based on 8,100 evaluation criteria values. The resulting correlation matrix is consequently used as an input for multidimensional scaling.

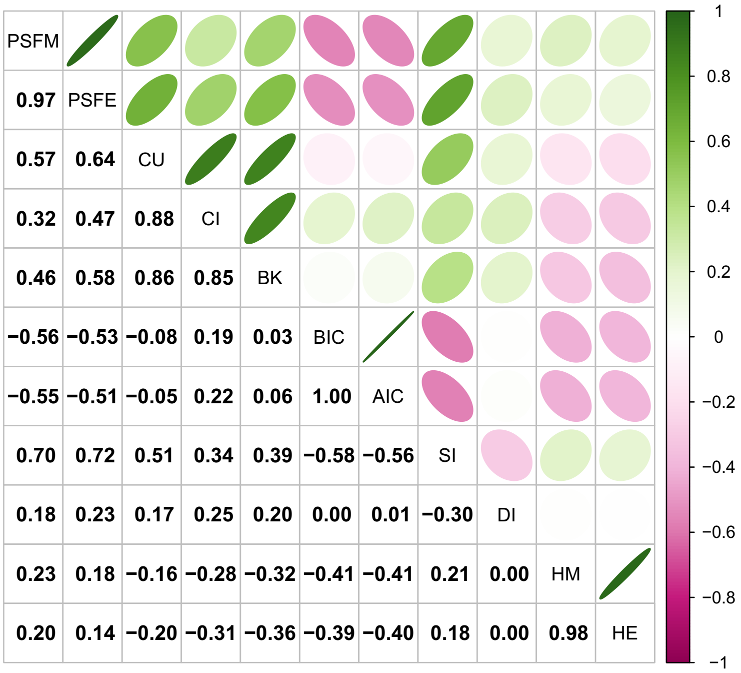


Figure 6.2: Correlations between pairs of internal evaluation criteria

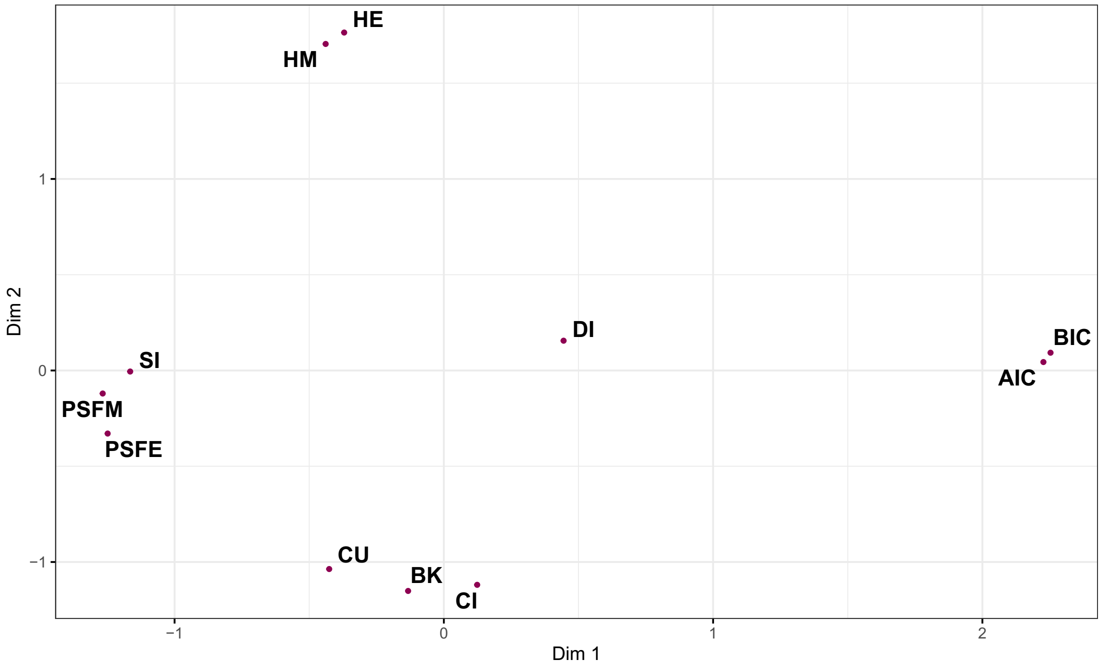


Figure 6.3: Multidimensional scaling of internal evaluation criteria

Figure 6.2 shows the results of correlation analysis and Figure 6.3 the outputs of multidimensional scaling. Both charts show that evaluation criteria based on the same principle but on a different variability measure (PSFM and PSFE, CU and CI, AIC and BIC, HM and HE) provide almost identical results. This fact is illustrated by correlations close or equal to one or close positions of the labels in Figure 6.3. Therefore, in this thesis, only one representative for each pair, namely PSFE, CU, BIC, and HE, will be used for the following analyses.

Regarding the non-paired criteria, the highest positive correlation ($r = 0.86$) occurs between the variability-based evaluation criteria CU and BK. Since the BK criterion utilizes the second order of the expected entropy to determine the optimal number of clusters, see Eq. (3.18), it often yields negative values, which decrease its correlation with CU, which could have been even higher otherwise.

The distance-based SI criterion's values correlate strongly with PSFE ($r = 0.72$) and BIC ($r = –0.58$). The negative correlation by BIC is all right since BIC (together with HE) prefers low values of a criterion. Although these three criteria are based on different principles, they are all highly related. By taking a closer look at the criteria's formulas in Eq. (3.10) and Eq. (3.19), the PSFE and BIC criteria utilize the clusters' variability expressed by entropy. They only differ in the way they transform it and penalize the higher number of clusters. A different situation occurs by SI, which uses distances (dissimilarities) between objects in the original dataset. Since every similarity measure for the dissimilarity calculation used in the experiment has a different range of possible values (their values are incomparable), there arises a question of whether the obtained correlation is relevant. Fortunately, when analyzing the correlation coefficients for individual similarity measures separately, even higher correlations are obtained, suggesting that the outputs presented in Fig. 6.2 can be considered valid.

There is a medium negative correlation between the newly proposed HE criterion and BIC ($r = –0.39$). Since both criteria look for the minimal value to identify the optimal number of clusters, this negative result suggests that these criteria often assess the cluster quality oppositely.

Interestingly, outputs of the distance-based criteria, SI and DI, are moderately negatively correlated ($r = -0.30$), suggesting that these distance-based criteria often provide contradictory results. While SI correlates moderately or strongly with the other examined criteria, DI shows only weak correlations with these criteria. However, when analyzing the correlations for individual similarity measures separately, substantial differences in the obtained correlations occur, suggesting that DI's performance depends strongly on the used similarity measure in HCA. This issue will be examined in the following analyses.

6.3.2 Relationships with the external evaluation

The quality of the created clusters in generated datasets can be easily assessed by external evaluation criteria, e.g., accuracy or the adjusted Rand index (ARI), since the cluster memberships are known by these datasets. In real-world datasets, the cluster memberships are unknown, so researchers need to rely on internal evaluation indices whose relationship with the external criteria was not analyzed in categorical data clustering. Therefore, the analyses in this subsection examine the relationships between external and internal evaluation criteria through two objectives. First, to determine how well the internal criteria can recognize the original number of clusters in datasets. Second, to determine to what extent they provide comparable results to ARI (as a representative of external criteria) and, thus, how well they represent the clustering quality.

The first objective is achieved through Table 6.1, which shows the proportions of cases (accuracies) when the internal criteria correctly identify the number of clusters in a dataset. Each calculated accuracy is based on 2,700 datasets according to the known number of clusters in generated datasets (two, four, six). For the analysis, HCA solutions with two to seven clusters are considered. Thus, a situation when a specific evaluation criterion often recommends the highest possible number of clusters does not positively influence the accuracies in the six-cluster solution.

Table 6.1: Criteria's ability to detect the original number of clusters measured by accuracy

Criterion	2 clusters	4 clusters	6 clusters	Total
PSFE	0.968	0.131	0.093	0.398
CU	0.954	0.165	0.033	0.384
BK	0.967	0.127	0.045	0.380
BIC	0.455	0.229	0.297	0.327
SI	0.917	0.243	0.193	0.451
DI	0.798	0.041	0.080	0.306
HE	0.246	0.182	0.243	0.224

The results indicate that the ability to determine the optimal number of clusters depends strongly on the original number of clusters in the datasets. The criteria can be divided into two groups. The first one, consisting of PSFE, CU, BK, and DI, performs well in the two-cluster solution. However, its classification performance drops drastically with the increasing number of natural clusters in a dataset. In this analysis, they are even lower than a random guess, 16.7%.

The second group comprises the SI, BIC, and HE criteria that show more balanced performance across the examined cluster solutions and perform better than a random guess on average (especially the former two). The criteria in this group perform worse in the two-cluster solution (except for SI), but their ability to correctly determine the original number of clusters is much better in four- and six-cluster solutions. Although the proportions are not high, especially compared to quantitative data with typical accuracies of around 80%, they do not depend to such a large extent on the number of clusters in datasets so that a researcher can expect relatively unbiased results. Still, one should not strictly rely on the provided results in practical tasks and should examine at least one lower- and one higher-cluster solution than the recommended one.

The second objective is carried out using Table 6.2, which describes the relationships between the internal criteria and ARI by correlation coefficient values. Similarly, as in the first objective, each calculated coefficient is based on 2,700 datasets. High correlation coefficient values indicate that internal criterion values are linearly dependent on the external ARI criterion values. Thus, a researcher can expect that the internal criteria provide comparable results with ARI in such a situation. Finally, since the low values of the BIC and HE criteria indicate good clusters, negative correlations are expected by these criteria.

Table 6.2: Correlations between the internal criteria and ARI

Criterion	2 clusters	4 clusters	6 clusters	Total
PSFE	0.441	0.598	0.536	0.640
CU	0.579	0.754	0.815	0.739
BK	0.685	0.543	0.342	0.687
BIC	0.045	−0.354	−0.508	−0.229
SI	0.220	0.487	0.597	0.514
DI	0.159	0.160	0.158	0.232
HE	−0.599	−0.133	0.094	−0.211

Overall, the highest correlation with ARI is achieved by the CU criterion ($r = 0.739$), whose values are increasingly more similar to ARI with the increasing number of clusters. It is followed by BK and PSFE. The BK criterion shows a high correlation in the two-cluster solution, but its values correspond with ARI less with the increasing number of clusters. On the contrary, PSFE performs consistently across the examined cluster solutions. The SI and BIC criteria values correspond more to ARI with a higher number of clusters. In total, the BIC criterion shows a moderate negative correlation with ARI ($r = -0.229$). A closer look reveals the high correlations in the four- and six-cluster solutions and the linear independence with ARI in the two-cluster solution. The DI criterion usually shows a weak positive correlation, meaning that it is only partly related to the ARI values. The HE criterion performs well in the two-cluster solution. However, its values in the cluster solutions with four and six clusters are not related to ARI.

When analyzing the correlations of the BIC criterion values with ARI in greater detail, the actual dependence of BIC and ARI is, in fact, higher than the correlation coefficient values show. Figure 6.4 demonstrates a non-linear relationship between BIC and ARI caused by the strong dependence of BIC on the number of variables and categories. When these relationships are analyzed separately for each number of variables and categories, stronger linear dependences occur in each group. Still, they are weaker than dependencies by the CU criterion. The dependence of the BIC values on the number of variables and categories is further examined in the following subsection.

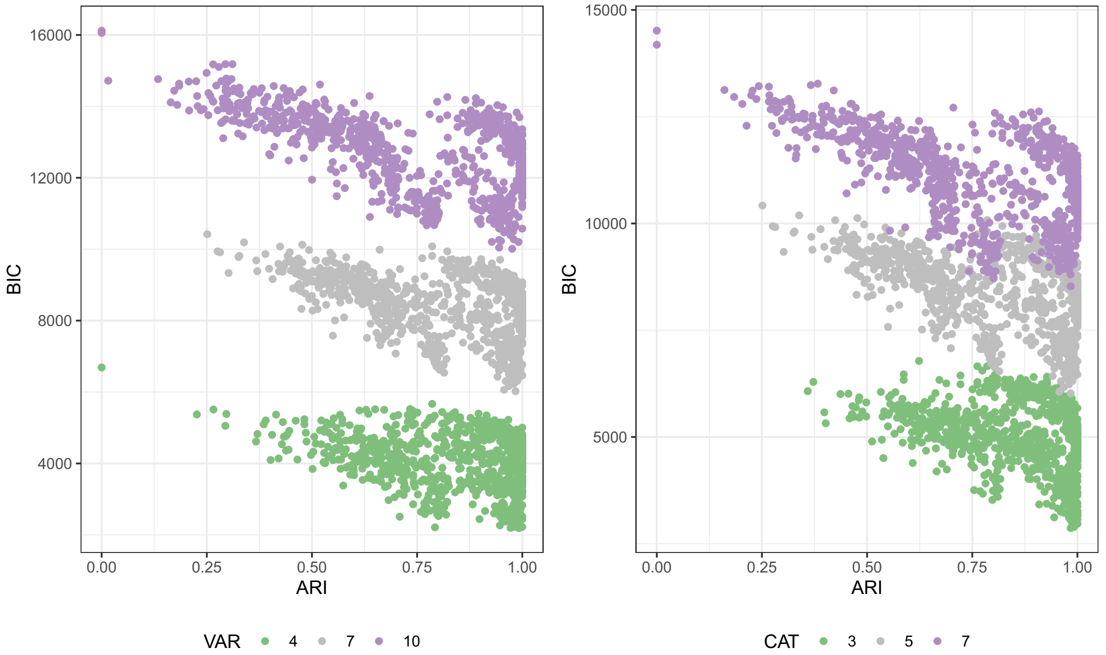


Figure 6.4: Relationship between the BIC criterion and ARI by the G1 similarity measure by the number of categories restricted to five (left) and the relationship between the BIC criterion and ARI by the G1 similarity measure by the fixed number of variables to seven (right)

6.3.3 Dependence on similarity measures and dataset properties

One of the aims of the experiment is to determine the extent to which the examined evaluation criteria depend on the analyzed dataset properties. The obtained strength of dependence can reveal what information the value of a given criterion truly expresses. Therefore, the dependencies of the criteria's values on the number of clusters (CLU), the number of variables (VAR), the number of categories (CAT), and the minimal between-cluster distance (DIST). In this research, the dependence of the similarity measure (SIM) used in HCA is also examined, mainly to determine how the used similarity measure influences the values of the distance-based evaluation criteria.

The ideal evaluation criterion depends only on CLU; this dataset's property is utilized in the optimal number of clusters determination task. Theoretically, the evaluation criteria values should not be influenced by DIST. In practice, however, lower DIST leads to clusters of poorer quality; see Chapter 5. Thus, dependence on DIST can be used as an indirect indicator of a criterion's ability to evaluate the cluster quality correctly. Some evaluation criteria also depend on VAR and CAT, which is considered a negative property. Fortunately, the impact of VAR and CAT plays a minor role if one does not compare evaluation criteria values in datasets with different numbers of variables and categories.

Table 6.3: Dependences of criteria values on the datasets' properties and the used similarity measure expressed by the eta-squared statistic

Criterion	CLU	DIST	VAR	CAT	SIM
PSFE	0.239	0.223	0.118	0.089	0.001
CU	0.289	0.256	0.070	0.061	0.002
BK	0.380	0.102	0.069	0.010	0.001
BIC	0.011	0.012	0.620	0.305	0.000
SI	0.091	0.114	0.053	0.106	0.453
DI	0.003	0.035	0.000	0.034	0.739
HE	0.139	0.071	0.213	0.075	0.004

Table 6.3 contains the eta-squared values expressing the strength of dependence between quantitative internal criteria values for the optimal number of clusters (two, four, and six clusters, depending on a dataset) and categorical datasets' properties and the used similarity measure. The results show that the BK, CU, and PSFE criteria values are moderately influenced by the original number of clusters in a dataset. The strongest dependence occurs by the BK criterion ($\eta^2 = 0.380$), which seems positive since this criterion was initially determined for the optimal number of clusters determination task. However, Table 6.1 shows that all three criteria perform well only in datasets with two original clusters. The CU and PSFE criteria's values depend moderately on DIST ($\eta^2 \approx 0.250$). Thus, these criteria better reflect the changes in cluster quality (indirectly expressed) compared to the other criteria. Still, all the examined evaluation criteria reveal the cluster quality to some extent, as shown in Table 6.2.

When analyzing the influence of VAR on the evaluation criteria values, the strongest dependence occurs by BIC ($\eta^2 = 0.620$). This criterion also depends moderately on the number of categories (CAT) ($\eta^2 = 0.305$), as shown in Fig. 6.4, so its values represent mainly the number of variables in a dataset and only partially the cluster quality. Thus, researchers may find difficult to recognize, e.g., subtle changes in the criterion's values in different cluster solutions when looking for the optimal number of clusters or judging the cluster quality. Apart from that, these dependencies do not influence the criterion's performance.

The distance-based criteria DI and SI show the high eta-squared values with the used similarity measure (SIM) representing strong ($\eta^2 = 0.739$) and moderate ($\eta^2 = 0.453$) dependences. Fig. 6.5 illustrates the issues associated with using the distance-based criteria to compare several similarity measures. Each chart represents the examined evaluation criterion values broken down by the number of clusters and six used similarity measures. Since only the well-performing similarity measures in ALM were chosen for the analysis, the cluster quality of the produced clusters should be comparable across the measures. This happens by the variability- and likelihood-based criteria but not by the distance-based ones, where different levels of evaluation criteria values among the similarity measures can be observed. The situation is especially problematic by the DI criterion, where the criterion's value for good clusters by one measure, e.g., LIN, can mean poor clusters by another measure, e.g., G1. Thus, the

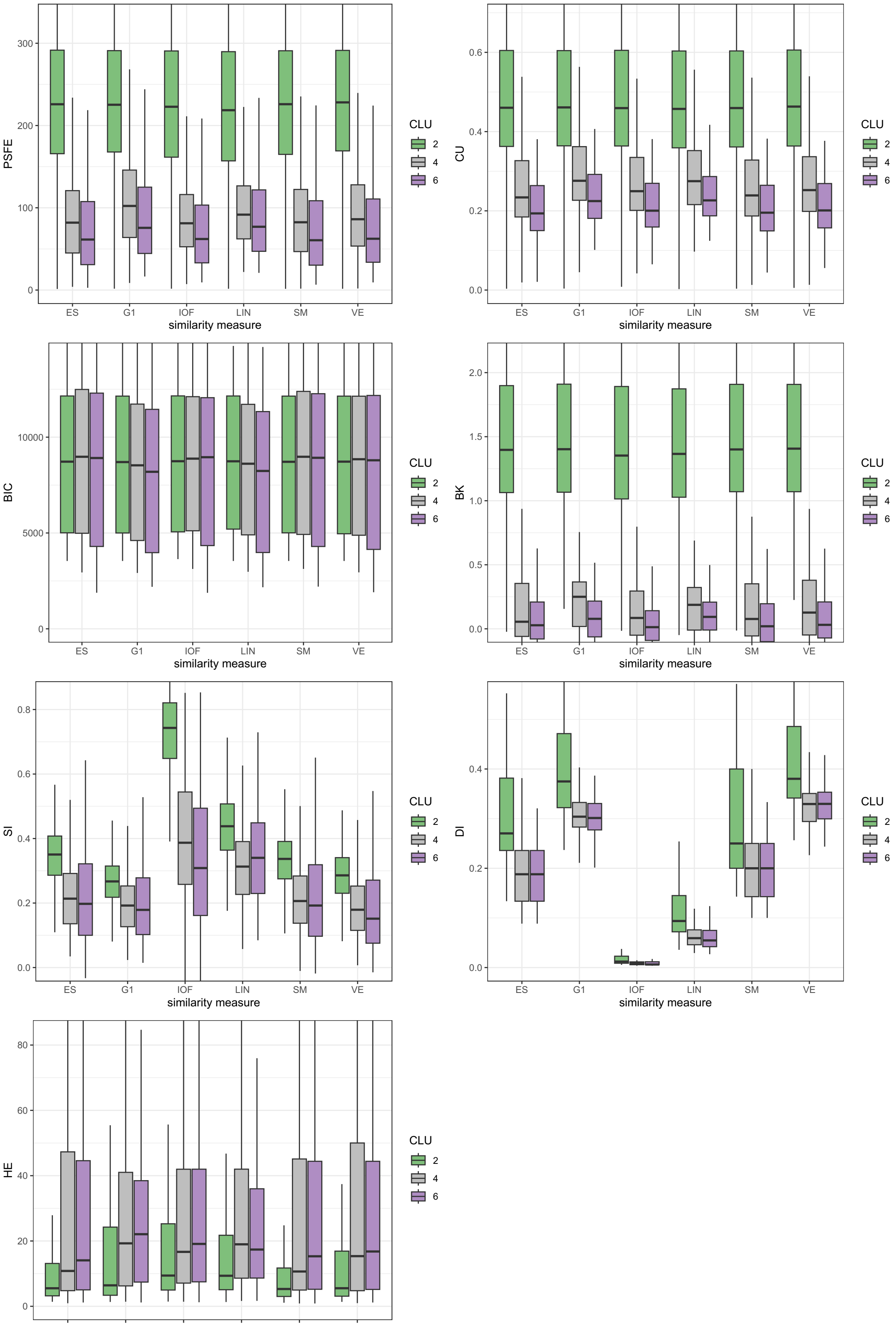


Figure 6.5: Comparison of evaluation criteria values across different similarity measures

distance-based criteria values should not be used to compare clusters created by different similarity measures. However, they can still be used for cluster quality evaluation if only one similarity measure is used, e.g., when comparing different numbers of clusters in a dataset.

6.3.4 Experiment summary

The experiment compared 11 internal evaluation criteria suitable for categorical data. The criteria were divided according to the principle they are based on, i.e., variability, likelihood, and distance. Additionally, two new variability-based criteria introduced in Subsection 3.2.4, HM and HE, were analyzed. The comparison focused mainly on the cluster quality assessment, but the criteria's ability to determine the optimal number of clusters was also examined.

The experiment showed that the values of the examined evaluation criteria differing only in the measure of variability (mutability vs. entropy), namely PSFM and PSFE, and CU and CI, are highly correlated. Hence, they are interchangeable for a researcher. Also, both likelihood-based criteria, BIC and AIC, provide almost identical outputs, and using them both does not offer additional insight into the cluster quality.

Table 6.4: Recommended evaluation criteria based on the intended task

Task	Criteria
Quality of clusters	CU (CI), PSFE (PSFM)
Quality of clusters (with one similarity measure)	CU (CI), PSFE (PSFM), SI
Optimal number of clusters	BIC (AIC)
Optimal number of clusters (with one similarity measure)	SI, BIC (AIC)
Standardized outputs (with one similarity measure)	SI

The experimental results indicate that no ideal evaluation criterion serves well in all situations. Therefore, Table 6.4 recommends the most suitable evaluation criteria for a given task (ordered by relevance).

The criteria CU (CI) and PSFE (PSFM) can be recommended for the cluster quality evaluation since their values mainly correlate with values of the adjusted Rand index (ARI), which is a commonly used external criterion for cluster quality assessment. Both criteria's values also depend on the minimal between-cluster distance, which is a necessary condition to judge the cluster quality properly.

The criteria SI and BIC (AIC) perform better than the rest of the examined evaluation criteria in the task of determining the optimal number of clusters. They can find the optimal number of clusters in 18 – 30% of cases in solutions with four and six clusters, which is a considerably worse result than a typical accuracy of criteria for quantitative data. It is most likely caused by the nature of categorical data, where clusters are far more challenging to recognize.

The distance-based criterion SI works well in cluster quality determination and the optimal number of cluster determination tasks. Moreover, its outputs are standardized, so they are easy to interpret. However, since the criterion's values depend on the used similarity measure in HCA, the criterion is unsuitable for comparison of the cluster quality achieved by several similarity measures. Otherwise, it is a versatile evaluation criterion.

BK and DI criteria do not outperform the criteria in Table 6.4 since they perform worse in all the observed situations (BK performs worse than CU, DI much worse than SI). Especially, DI performs poorly overall in both the cluster quality and optimal cluster determination tasks.

The newly proposed criteria, HE and HM, performed around the average regarding the optimal number of clusters determination. However, their performance in the cluster quality evaluation was inferior. Thus, these criteria can be recommended only for specific tasks.

Conclusion

The habilitation thesis aimed to thoroughly cover the topic of hierarchical cluster analysis (HCA) of categorical data, which consisted of dissimilarity matrix calculation, application of a given HCA algorithm, and cluster quality evaluation. Although the thesis dealt with all three steps, it mainly focused on similarity measures for dissimilarity matrix calculation and assessing the internal evaluation criteria for categorical data.

In Introduction, three main research goals were stated. The first one dealt with comparing similarity measures for categorical data, the second one with assessing internal evaluation criteria for categorical data, and the third one with developing a second generation of the `nomclust` package. In the following paragraphs, these goals are evaluated in detail.

The **first goal** was to compare and evaluate the clustering performance of the selected similarity measures for categorical data, including those not examined before. The objective was also to explore the combinations of similarity measures and three different linkage methods that might be useful in practical applications. To accomplish this goal, Chapter 2 was written, where 16 examined similarity measures for categorical data were presented. The experiment was carried out on 2,700 generated datasets in Chapter 5, where the similarity measures' ability to create good-quality clusters was examined based on the datasets' properties and the linkage method used. Compared to the previous research in this area performed by Šulc (2016), a new way of cluster quality assessment based on boxplots was used. Moreover, the influence of different minimal between-cluster distances and linkage methods was explored. Finally, more similarity measures, a new evaluation criterion, and a substantially higher number of datasets were used for the experiment.

The experimental results showed that there were considerable differences between the linkage methods. By far, the best clusters were created by the average linkage method (ALM), so it should be preferred unless there is a specific need for another linkage method. The method shows the lowest dependence on the clustered datasets' properties, namely the number of variables, categories, and, especially, the minimal between-cluster distances, which are impossible to determine in real applications. The complete linkage method (CLM) usually creates worse clusters than ALM. This method proved to be the most sensitive to the clustered datasets' properties, especially the number of variables and their categories. The single linkage method (SLM) generally created poor-quality clusters. Theoretically, the method can be suitable for simple datasets with few variables and their categories. However, it cannot be recommended for common use.

Regarding the examined similarity measures for categorical data, generally, the best measures are G1 and G3, which usually create the best clusters in ALM. Good clustering performance can also be expected when using LIN, SV, and VE measures. CLM created the best clusters with the IOF and OF similarity measures and SLM with the ES and SM measures. The VE and VM similarity measures, proposed by Šulc (2016), performed well in all the examined linkages. Such a consistent behavior did not occur by any other studied similarity measure. The SM measure served as a reference similarity measure in the thesis since it is commonly used by most researchers in the HCA of categorical data. The measure provided average-quality clusters in ALM, poor clusters in CLM, and excellent clusters in SLM.

The research also confirmed that some of the examined similarity measures perform better by specific dataset properties. For instance, LIN and SV improve their relative clustering performance in ALM compared to the other measures when the number of variables is higher. Another example is the outstanding performance of the VE and VM measures in CLM when the number of clustered variables is low. Thus, the knowledge of the clustered dataset's properties can help a researcher select the best suitable similarity measures and thus maximally increase the chance of getting the best possible clusters.

The newly examined similarity measures AN, BU, GA, and SV differ in their clustering performance substantially. AN creates relatively good clusters only in SLM, which is useless in practice since SLM generally performs poorly. BU works well in CLM and SLM. Thus, this measure can be a preferred choice if CLM needs to be used. The GA measure constantly provides below-average clusters, so it cannot be recommended for practical applications. On the contrary, the SV measure performs excellently in ALM and well in CLM. Thus, this measure can be recommended for regular use in ALM.

The **second goal** was to compare and assess 11 internal evaluation criteria for categorical data from different perspectives, such as their mutual similarity, coherence with the adjusted Rand index, or dependence on the clustered dataset's properties. This goal was achieved through Chapter 3 and Chapter 6. In Chapter 3, the examined criteria were described, presented, and divided according to the principles they were built on. Moreover, two new variability-based criteria were proposed there. Chapter 6 contains the experiment conducted on 8,100 generated datasets, where the evaluation criteria were mainly assessed by well-known statistical methods, such as correlation analysis or ANOVA.

The experiment identified evaluation criteria that assess the cluster quality almost identically, namely PSFM and PSFE, CU and CI, BIC and AIC, and HM and HE. Thus, using both criteria from a given pair to analyze the created clusters is redundant. Therefore, only a representative from each group was analyzed in the consequent analyses.

First, the internal evaluation criteria were examined concerning the ability to identify the optimal number of clusters. The criteria can be divided into two groups. The first one, consisting of the SI, BIC (AIC) and HE (HM) criteria, is able to predict the optimal number of clusters approximately in 20–30% of cases. The second group, containing the remaining evaluation criteria, primarily recommends the two-cluster solution, and thus, it is unsuitable for this task. To sum it up, all the examined criteria performed poorly. Therefore, one cannot rely on the recommended number of clusters by any of these criteria, and it is always good to try out solutions with at least one less and one more cluster.

Although the ability to determine the optimal number of clusters is overall low, most of the evaluation criteria correlate moderately or strongly with ARI, suggesting that they can identify the clusters in a dataset well. The most consistent results with ARI are provided by the CU (CI) criterion. PSFE (PSFM), SI, and BIC (AIC) also perform well. Thus, even though the true cluster membership is unknown in practical situations, a researcher can mostly trust the aforementioned internal criteria results. The newly proposed HE criterion and DI do not correlate with ARI to a large extent, which corresponds to their overall poor performance.

The analysis of evaluation criteria dependence on the clustered datasets' properties revealed several interesting conclusions: First, although the values of the BK, CU, and PSFE criteria depend moderately on the original number of clusters in a dataset, they are unable to predict the correct number of clusters if there are more than two clusters in a dataset. Second, the outstanding performance of the CU and PSFE criteria in the cluster quality assessment, which was indirectly expressed by a moderate dependence of the criteria on the minimal between-cluster distance, has been proved. Third, the distance-based criteria SI and DI are influenced by the similarity measure used. Thus, these criteria are unsuitable for comparing clusters created by different similarity measures.

The **third goal** was to present and improve the second generation of the `nomclust` package and to illustrate its use. It was accomplished in Chapter 4, where the methods in the package are described, and the package functionalities are clearly demonstrated. Thus, researchers from various fields can use it as a single tool for complex hierarchical clustering of categorical data, enabling them to choose from many similarity measures for categorical data and evaluation criteria.

The package was essential when performing both experiments in this thesis. Due to C++ optimizations, the analyses could be performed on a much larger number of datasets than in the previous studies, increasing their validity. Moreover, the large number of datasets enabled an easy visualization of the evaluation criteria values using the boxplots.

The `nomclust` package is under continuous development. Compared to the package presented by Šulc et al. (2022), the version described in this thesis (version 2.8) contains the DI criterion and the variable weighting procedure theoretically introduced in Section 2.3. In future releases, handling the missing observations and the approaches for mixed-type data are planned.

In conclusion, all three primary goals were satisfied. The thesis provided a complex analysis of categorical data clustering, mainly focused on not much-explored topics of similarity measures and evaluation criteria. Additionally, it offers a software application that enables convenient adoption of the researched methods. The obtained results can be considered legitimate thanks to a robust simulation study, where each dataset combination was 100 times replicated. Thus, the thesis can help researchers to get oriented in categorical data HCA and recommend the best combination of similarity measure and linkage method for a given situation. The conducted research on evaluation criteria can also be beneficial in other classification studies that deal with categorical data, including those that exceed the area of HCA.

Bibliography

AHMAD, A. AND KHAN, S. S. 2019. Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access 7,* 31883–31902.

AKAIKE, H. 1973. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory,* B. N. Petrov and F. Csaki, Eds. Akadémiai Kiado, Budapest, 267–281.

ANDERBERG, M. R. 1973. *Cluster Analysis for Applications.* Probability and Mathematical Statistics. Academic Press.

ANDERLUCCI, L. AND HENNIG, C. 2014. The clustering of categorical data: A comparison of a model-based and a distance-based approach. *Communication in Statistics – Theory and Methods 43,* 4, 1–16.

ANSELIN, L., SYABRI, I., AND KHO, Y. 2006. GeoDa: An introduction to spatial data analysis. *Geographical Analysis 38,* 1, 5–22.

ARBELAITZ, O., GURRUTXAGA, I., MUGUERZA, J., PÉREZ, J. M., AND PERONA, I. 2013. An extensive comparative study of cluster validity indices. *Pattern Recognition 46,* 1, 243–256.

BACHER, J., WENZIG, K., AND VOGLER, M. 2004. *SPSS TwoStep Cluster – a First Evaluation.* Arbeits- und Diskussionspapiere / Universität Erlangen-Nürnberg, Sozialwissenschaftliches Institut, Lehrstuhl für Soziologie. Universität Erlangen-Nürnberg, Wirtschafts- und Sozialwissenschaftliche Fakultät, Sozialwissenschaftliches Institut Lehrstuhl für Soziologie, Nürnberg.

BAI, L. AND LIANG, J. 2015. Cluster validity functions for categorical data: A solution-space perspective. *Data Mining and Knowledge Discovery 29,* 6, 1560–1597.

BARBARÁ, D., LI, Y., AND COUTO, J. 2002. COOLCAT: An entropy-based algorithm for categorical clustering. In *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 4-9, 2002.* ACM, 582–589.

BERGMAN, L. R. AND EL-KHOURI, B. M. 1999. Studying individual patterns of development using i-states as objects analysis (ISOA). *Biometrical Journal 41,* 6, 753–770.

BIEM, A. 2003. A model selection criterion for classification: application to hmm topology optimization. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition.* 104–108.

BONTEMPS, D. AND TOUSSILE, W. 2013. Clustering and variable selection for categorical multivariate data. *Electronic Journal of Statistics 7,* 2344 – 2371.

BORIAH, S., CHANDOLA, V., AND KUMAR, V. 2008. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the eighth SIAM International Conference on Data Mining.* 243–254.

BRUN, M., SIMA, C., HUA, J., LOWEY, J., CARROLL, B., SUH, E., AND DOUGHERTY, E. R. 2007. Model-based evaluation of clustering validation measures. *Pattern Recognition 40,* 3, 807–824.

BURNABY, T. P. 1970. On a method for character weighting a similarity coefficient, employing the concept of information. *Journal of the International Association for Mathematical Geology 2,* 25–38.

CALIŃSKI, T. AND HARABASZ, J. 1974. A dendrite method for cluster analysis. *Communications in Statistics 3,* 1, 1–27.

CHANDOLA, V., BORIAH, S., AND KUMAR, V. 2009. A framework for exploring categorical data. In *SIAM International Conference on Data Mining.* SIAM 2009, 187–198.

CHATURVEDI, A., GREEN, P. E., AND CAROLL, J. D. 2001. K-modes clustering. *Journal of Classification 18,* 1, 35–55.

CHEN, K. AND LIU, L. 2009. "Best K": Critical clustering structures in categorical datasets. *Knowledge and Information Systems 20,* 1, 1–33.

CHEN, L. AND GUO, G. 2014. Centroid-based classification of categorical data. In *Web-Age Information Management.* Springer International Publishing, 472–475.

CHIANG, M. M.-T. AND MIRKIN, B. G. 2010. Intelligent choice of the number of clusters in k-means clustering: An experimental study with different cluster spreads. *Journal of Classification 27,* 1, 3–40.

CIBULKOVÁ, J., ŠULC, Z., ŘEZANKOVÁ, H., AND SIROTA, S. 2020. Associations among similarity and distance measures for binary data in cluster analysis. *Metodoloski zvezki 17,* 1, 33–54.

COOMBES, K. R. AND COOMBES, C. E. 2022. *Mercator: Clustering and visualizing distance matrices.* R package version 1.1.2.

CORTER, J. E. AND GLUCK, M. A. 1992. Explaining basic categories: Feature predictability and information. *Psychological Bulletin 111,* 291–303.

de Souto, M. C., Coelho, A. L., Faceli, K., Sakata, T. C., Bonadia, V., and Costa, I. G. 2012. A comparison of external clustering evaluation indices in the context of imbalanced data sets. In *2012 Brazilian Symposium on Neural Networks.* 49–54.

Desai, A., Singh, H., and Pudi, V. 2011. *DISC: Data-intensive similarity measure for categorical data.* Springer Berlin Heidelberg, Berlin, Heidelberg, 469–481.

Deza, M. M. and Deza, E. 2009. *Encyclopedia of Distances.* Springer Berlin Heidelberg.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology 26,* 3, 297–302.

Dimitriadou, E., Dolničar, S., and Weingessel, A. 2002. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika 67,* 1, 137–159.

Draszawka, K. and Szymański, J. 2011. *External validation measures for nested clustering of text documents.* Springer Berlin Heidelberg, Berlin, Heidelberg, 207–225.

Dunn, J. C. 1973. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics 3,* 3, 32–57.

Eddelbuettel, D. and Francois, R. 2011. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software 40,* 8, 1–18.

Ellerman, D. 2013. An introduction to logical entropy and its relation to shannon entropy. *International Journal of Semantic Computing 7,* 2, 121–145.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S. 2002. *A geometric framework for unsupervised anomaly detection.* Springer US, Boston, MA, 77–101.

Everitt, B. S., Landau, S., and Leese, M. 2009. *Cluster Analysis,* 4th ed. Wiley Publishing.

Fisher, D. H. 1987. Knowledge acquisition via incremental conceptual clustering. *Machine Learning 2,* 2, 139–172.

Gambaryan, P. 1964. A mathematical model of taxonomy. *Izvestiia Akademii nauk Armianskoĭ SSR 17,* 12, 47–53.

Github. 2020. Github.

Goodall, D. W. 1966. A new similarity index based on probability. *Biometrics 22,* 4, 882–907.

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

Guha, S., Rastogi, R., and Shim, K. 1999. Rock: A robust clustering algorithm for categorical attributes. In *Proceedings 15th International Conference on Data Engineering.* 512–521.

Hagenaars, J. and McCutcheon, A. 2002. *Applied Latent Class Analysis.* Cambridge University Press.

HAHSLER, M., BUCHTA, C., GRUEN, B., AND HORNIK, K. 2015. *arules: Mining association rules and frequent itemsets.* R package version 1.7-6.

HALKIDI, M., BATISTAKIS, Y., AND VAZIRGIANNIS, M. 2001. On clustering validation techniques. *Journal of Intelligent Information Systems 17,* 2-3, 107–145.

HAMANN, U. 1961. Merkmale bestand und verwandtschaftsbeziehungen der farinose. Ein beitrag zum system der monokotyledonen. *Willdenowia 2,* 639–768.

HARTIGAN, J. 1975. *Clustering algorithms.* John Wiley and Sons, New York.

HENNIG, C. 2022. An empirical comparison and characterisation of nine popular clustering methods. *Advances in Data Analysis and Classification 16,* 1, 201–229.

HENNIG, C., MEILA, M., MURTAGH, F., AND ROCCI, R. 2015. *Handbook of Cluster Analysis.* Chapman & Hall / CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton.

HUBERT, L. AND ARABIE, P. 1985. Comparing partitions. *Journal of Classification 2,* 1, 193–218.

JACCARD, P. 1912. The distribution of the flora in the alpine zone. *New Phytologist 11,* 2, 37–50.

KARGAR, M., IZADKHAH, H., AND ISAZADEH, A. 2019. Tarimliq: A new internal metric for software clustering analysis. In *2019 27th Iranian Conference on Electrical Engineering (ICEE).* 1879–1883.

LIN, D. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning.* Morgan Kaufmann, 296–304.

LINZER, D. A. AND LEWIS, J. B. 2011. poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software 42,* 10, 1–29.

LIU, Y., LI, Z., XIONG, H., GAO, X., AND WU, J. 2010. Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining.* 911–916.

MAECHLER, M., ROUSSEEUW, P., STRUYF, A., HUBERT, M., AND HORNIK, K. 2022. *cluster: Cluster analysis basics and extensions.* R package version 2.1.4.

MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. 2008. *Introduction to Information Retrieval.* Cambridge University Press, New York, USA.

MILIGAN, G. W. AND COOPER, M. C. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika 50,* 159–179.

MORLINI, I. AND ZANI, S. 2012. A new class of weighted similarity indices using polytomous variables. *Journal of Classification 29,* 2, 199–226.

MURAKOSHI, K. AND FUJIKAWA, S. 2016. Growing hierarchical self-organizing map using category utility. *International Journal of Software Engineering and Knowledge Engineering 26,* 2, 217–237.

QIU, W. AND JOE, H. 2006. Generation of random clusters with specified degree of separation. *Journal of Classification 23,* 2, 315–334.

R CORE TEAM. 2021. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

RAND, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66,* 336, 846–850.

RENDÓN, E., ABUNDEZ, I. M., GUTIERREZ, C., ZAGAL, S. D., ARIZMENDI, A., QUIROZ, E. M., AND ARZATE, H. E. 2011. A comparison of internal and external cluster validation indexes. In *Proceedings of the 2011 American Conference on Applied Mathematics and the 5th WSEAS International Conference on Computer Engineering and Applications.* World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 158–163.

ŘEZANKOVÁ, H., LÖSTER, T., AND HÚSEK, D. 2011. Evaluation of categorical data clustering. *Advances in Intelligent Web Mastering 3,* 173–182.

ROGERS, D. J. AND TANIMOTO, T. T. 1960. A computer program for classifying plants. *Science 132,* 3434, 1115–1118.

ROSENBERG, J. M., SCHMIDT, J. A., AND BEYMER, P. N. 2020. *prcr: Person-centered analysis.* R package version 0.2.1.

ROUSSEEUW, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20,* 53–65.

RUSSEL, P. F. AND RAO, T. R. 1940. On habitat and association of species of anopheline larvae in south-eastern madras. *Journal of Malaria Institute India 3,* 153–178.

SCHWARZ, G. 1978. Estimating the dimension of a model. *The Annals of Statistics 6,* 2, 461–464.

SHANNON, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal 27,* 379–423.

SMIRNOV, E. S. 1968. On exact methods in systematics. *Systematic Biology 17,* 1, 1–13.

SOKAL, R. AND SNEATH, P. H. A. 1963. *Principles of numerical taxonomy.* W. H. Freeman, London.

SOKAL, R. R. AND MICHENER, C. D. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin 28,* 1409–1438.

SPSS, Inc. 2001. *The SPSS TwoStep Cluster component.* SPSS, Inc.

SPÄRCK JONES, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation 28,* 11–21.

STRAUSS, T. AND VON MALTITZ, M. J. 2017. Generalising Ward's method for use with Manhattan distances. *PLOS ONE 12,* 1, 1–21.

ŠULC, Z. 2016. *Similarity Measures for Nominal Data in Hierarchical Clustering.* Ph.D. thesis, Prague University of Economics and Business.

ŠULC, Z., CIBULKOVÁ, J., PROCHÁZKA, J., AND ŘEZANKOVÁ, H. 2018. Internal evaluation criteria for categorical data in hierarchical clustering: Optimal number of clusters determination. *Metodoloski zvezki 15,* 2, 1–20.

ŠULC, Z., CIBULKOVÁ, J., AND ŘEZANKOVÁ, H. 2022. Nomclust 2.0: An R package for hierarchical clustering of objects characterized by nominal variables. *Computational Statistics 37,* 5, 2161–2184.

ŠULC, Z. AND ŘEZANKOVÁ, H. 2015. nomclust: An R package for hierarchical clustering of objects characterized by nominal variables. In *Proceedings of the 9th International Days of Statistics and Economics.* Melandrium, Slaný, 1581–1590.

ŠULC, Z. AND ŘEZANKOVÁ, H. 2019. Comparison of similarity measures for categorical data in hierarchical clustering. *Journal of Classification 36,* 1, 58–72.

SZEPANNEK, G. 2018. clustMixType: User-friendly clustering of mixed-type data in R. *The R Journal,* 200–208.

THORNDIKE, R. L. 1953. Who belongs in the family? *Psychometrika 18,* 267–276.

TODESCHINI, R., CONSONNI, V., XIANG, H., HOLLIDAY, J., BUSCEMA, M., AND WILLETT, P. 2012. Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling 52,* 11, 2884–2901.

TOMASINI., C., N. BORGES., E., MACHADO., K., AND EMMENDORFER., L. 2017. A study on the relationship between internal and external validity indices applied to partitioning and density-based clustering algorithms. In *Proceedings of the 19th International Conference on Enterprise Information Systems – Volume 1: ICEIS,.* INSTICC, SciTePress, 89–98.

VENDRAMIN, L., CAMPELLO, R. J. G. B., AND HRUSCHKA, E. R. 2010. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining: The ASA Data Science Journal 3,* 4, 209–235.

VERMUNT, J. AND MAGIDSON, J. 2016. *Technical Guide for Latent GOLD 5.1: Basic, Advanced, and Syntax.* Technical report, Belmont, MA: Statistical Innovations Inc.

WARRENS, M. J. 2008. *Similarity Coefficients for Binary Data.* Ph.D. thesis, University of Leiden.

WARRENS, M. J. 2016. Inequalities between similarities for numerical data. *Journal of Classification 33,* 2, 141–148.

WEIHS, C., LIGGES, U., LUEBKE, K., AND RAABE, N. 2005. klaR analyzing german business cycles. In *Data Analysis and Decision Support*, D. Baier, R. Decker, and L. Schmidt-Thieme, Eds. Springer-Verlag, Berlin, 335–343.

WITTEN, I., FRANK, E., HALL, M., AND PAL, C. 2016. *Data Mining: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

XAVIER, J. C., CANUTO, A. M. P., ALMEIDA, N. D., AND GONÇALVES, L. M. G. 2013. A comparative analysis of dissimilarity measures for clustering categorical data. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*. 1–8.

YI, J., YANG, G., AND WAN, J. 2016. Category discrimination based feature selection algorithm in chinese text classification. *Journal of Information Science and Engineering 32,* 5, 1145–1159.

YULE, G. U. 1900. On the association of attributes in statistics: With illustrations from the material of the childhood society. *Philosophical Transactions of the Royal Society of London 194*, 257–319.

YULE, G. U. 1912. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society 75,* 6, 579–652.

ZHAO, Y., KARYPIS, G., AND FAYYAD, U. 2005. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery 10*, 141–168.

Appendices

A Dissimilarity Matrix Calculations Steps of the Single Linkage Method

Table I: Dissimilarity matrix calculation steps of the single linkage method

step 1	1	2	3	4	5	6
1	0	0.617	1.000	0.796	0.889	0.907
2		0	0.841	0.637	0.730	0.748
3			0	0.841	0.466	0.654
4				0	0.730	0.594
5					0	0.841
6						0

step 2	1	2	3+5	4	6
1	0	0.617	0.889	0.796	0.907
2		0	0.730	0.637	0.748
3+5			0	0.730	0.654
4				0	0.594
6					0

step 3	1	2	3+5	4+6
1	0	0.617	0.889	0.796
2		0	0.730	0.637
3+5			0	0.654
4+6				0

step 4	1+2	3+5	4+6
1+2	0	0.730	0.637
3+5		0	0.654
4+6			0

step 5	1+2+4+6	3+5
1+2+4+6	0	0.654
3+5		0

B Scripts Used in the Experiments

Table II: Files used for the conducted experiments

Data file	The file is used to ...
`00_run.R`	run the whole analysis.
`01_data_settings.csv`	provide properties of generated datasets.
`02_data_generation.R`	generate the datasets.
`03_calculations.R`	get and save the outputs of the performed HCAs.
`04_evaluation.R`	extract evaluation criteria values from the HCA outputs.
`05_experiment_1.R`	calculate MRS and produce boxplots in Experiment I.
`06_experiment_2.R`	run the evaluation criteria assessment in Experiment II.

The file `01_data_settings.csv` contains properties for 8,100 datasets with different numbers of clusters (2, 4, 6). Experiment I, performed in Chapter 5, is based on the datasets with four original clusters, i.e., on a subset of 2,700 datasets. Experiment II, conducted in Chapter 6, uses all 8,100 datasets.

C Characteristics of the Boxplots

Table III: Boxplot characteristics for three linkage methods (PSFE)

LINK	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
ALM	Q1	28.8	34.9	47.7	67.7	46.4	63.3	3.2	33.5	52.5	62.4	35.2	37.3	48.6	66.3	55.2	55.4
	median	55.3	57.2	83.5	106.7	82.6	101.7	6.8	54.4	83.2	94.4	54.0	65.6	84.1	97.3	90.3	90.6
	Q3	112.6	115.2	137.8	166.7	133.4	163.6	24.2	101.1	134.5	149.6	86.4	128.9	137.6	155.5	149.1	148.4
	IQR	83.7	80.3	90.1	99.0	86.9	100.3	21.0	67.6	82.1	87.1	51.2	91.6	89.0	89.2	93.9	93.0
CLM	Q1	17.6	35.0	17.0	19.4	23.2	19.7	23.3	20.9	33.8	36.3	29.8	36.8	16.1	32.7	25.9	25.7
	median	33.7	55.4	33.9	38.2	49.9	37.8	35.9	46.9	55.8	55.0	47.6	58.3	33.9	51.4	58.1	57.9
	Q3	61.7	88.9	73.4	78.1	92.6	76.5	65.2	81.7	93.9	90.3	78.2	96.6	69.8	85.3	107.1	106.8
	IQR	44.1	53.9	56.4	58.7	69.4	56.7	42.0	60.8	60.1	54.0	48.4	59.8	53.7	52.6	81.3	81.1
SLM	Q1	1.9	1.4	1.5	1.2	1.4	1.2	1.5	1.5	1.3	1.3	0.9	1.4	1.6	0.8	1.4	1.4
	median	4.5	1.7	2.8	1.6	1.5	1.5	1.6	1.6	1.5	1.6	1.3	1.6	4.0	1.5	1.8	1.7
	Q3	19.0	31.8	48.0	10.6	3.0	10.3	1.9	2.0	57.3	10.8	8.0	38.8	49.2	6.4	86.0	86.4
	IQR	17.1	30.4	46.5	9.3	1.6	9.1	0.5	0.5	56.1	9.5	7.2	37.4	47.6	5.6	84.7	85.0

$Q1$ stands for the first quartile (25% quantile) corresponding to the lower border of the box in a boxplot, median (50% quantile) expresses the median value representing the line in the box, $Q3$ stands for the third quartile (75% quantile) corresponding the upper border of the box. IQR can be expressed as $Q3-Q1$, and it represents the width of the box containing 50% of the middle evaluation criterion values.

Table IV: Boxplot characteristics for three linkage methods (CU)

LINK	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
ALM	Q1	0.09	0.13	0.18	0.23	0.18	0.22	0.02	0.13	0.20	0.21	0.14	0.14	0.18	0.21	0.20	0.20
	median	0.18	0.19	0.25	0.29	0.24	0.28	0.03	0.19	0.26	0.28	0.19	0.21	0.25	0.27	0.26	0.26
	Q3	0.27	0.27	0.34	0.39	0.34	0.38	0.09	0.27	0.35	0.37	0.25	0.30	0.34	0.36	0.36	0.35
	IQR	0.17	0.14	0.16	0.16	0.16	0.16	0.07	0.14	0.16	0.16	0.11	0.15	0.16	0.15	0.16	0.16
CLM	Q1	0.08	0.12	0.09	0.09	0.12	0.09	0.14	0.12	0.16	0.14	0.12	0.13	0.09	0.13	0.13	0.13
	median	0.12	0.18	0.13	0.14	0.18	0.14	0.16	0.17	0.20	0.20	0.16	0.20	0.13	0.18	0.20	0.20
	Q3	0.19	0.27	0.22	0.24	0.25	0.23	0.20	0.23	0.27	0.27	0.23	0.28	0.22	0.25	0.27	0.27
	IQR	0.11	0.14	0.13	0.14	0.13	0.14	0.06	0.11	0.11	0.13	0.11	0.15	0.13	0.12	0.14	0.14
SLM	Q1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	median	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01
	Q3	0.05	0.09	0.14	0.04	0.01	0.03	0.01	0.01	0.16	0.03	0.03	0.09	0.15	0.02	0.23	0.23
	IQR	0.04	0.08	0.12	0.03	0.00	0.02	0.00	0.00	0.15	0.03	0.03	0.08	0.14	0.02	0.22	0.22

Table V: Boxplot characteristics for three minimal between-cluster distances in ALM (PSFE)

DIST	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
0.50	Q1	51.8	61.1	97.7	114.5	95.0	109.5	5.0	63.8	101.6	110.0	47.7	68.4	98.0	107.7	102.7	100.2
	median	92.4	100.3	140.2	168.7	136.6	163.6	17.8	98.6	138.0	151.5	73.5	120.3	141.0	156.2	147.6	146.1
	Q3	162.5	181.4	201.6	244.8	205.9	235.9	54.9	165.5	196.1	220.2	119.8	192.0	205.2	218.1	223.3	220.9
	IQR	110.7	120.3	103.8	130.3	111.0	126.4	49.8	101.6	94.5	110.3	72.1	123.6	107.2	110.4	120.7	120.8
0.34	Q1	32.5	36.6	50.3	69.3	49.3	63.4	3.0	35.8	55.3	66.3	34.5	38.8	51.7	67.7	58.4	58.6
	median	54.2	50.5	77.6	102.3	77.5	98.0	6.1	48.1	78.3	89.0	50.6	55.7	78.7	92.1	84.6	84.6
	Q3	101.2	107.1	119.2	145.9	114.8	143.7	16.1	79.6	112.0	126.2	79.3	112.8	119.3	134.7	126.7	129.3
	IQR	68.7	70.5	68.9	76.6	65.6	80.3	13.1	43.7	56.7	59.9	44.8	74.0	67.5	67.0	68.2	70.7
0.21	Q1	7.8	5.5	31.9	45.0	30.3	41.7	2.6	16.1	34.7	44.9	28.0	24.4	30.7	49.4	35.2	33.8
	median	31.5	33.0	45.8	68.8	45.0	64.0	4.5	30.8	48.7	59.8	40.6	35.9	46.4	66.4	53.5	52.5
	Q3	63.9	62.9	76.6	102.4	75.8	98.8	11.1	50.6	72.4	82.7	65.2	74.0	77.5	95.3	84.0	84.9
	IQR	56.2	57.4	44.7	57.4	45.6	57.0	8.5	34.5	37.8	37.9	37.2	49.7	46.8	45.8	48.8	51.1

Table VI: Boxplot characteristics for three minimal between-cluster distances in ALM (CU)

DIST	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
0.50	Q1	0.19	0.23	0.31	0.34	0.30	0.33	0.03	0.23	0.32	0.32	0.18	0.24	0.31	0.32	0.32	0.31
	median	0.26	0.28	0.37	0.41	0.37	0.40	0.08	0.29	0.39	0.41	0.24	0.31	0.37	0.39	0.39	0.38
	Q3	0.40	0.37	0.45	0.50	0.45	0.49	0.18	0.40	0.46	0.49	0.33	0.41	0.45	0.48	0.46	0.46
	IQR	0.20	0.15	0.14	0.16	0.15	0.16	0.15	0.17	0.14	0.17	0.15	0.17	0.14	0.16	0.14	0.14
0.34	Q1	0.11	0.14	0.20	0.24	0.19	0.23	0.02	0.14	0.21	0.23	0.14	0.16	0.20	0.23	0.21	0.22
	median	0.17	0.18	0.24	0.28	0.24	0.27	0.03	0.18	0.25	0.28	0.18	0.20	0.24	0.27	0.26	0.26
	Q3	0.26	0.24	0.28	0.36	0.30	0.35	0.07	0.23	0.31	0.34	0.24	0.27	0.29	0.33	0.31	0.32
	IQR	0.15	0.09	0.09	0.12	0.10	0.12	0.05	0.09	0.09	0.12	0.10	0.11	0.09	0.10	0.10	0.10
0.21	Q1	0.02	0.03	0.14	0.18	0.14	0.17	0.02	0.06	0.16	0.17	0.12	0.09	0.14	0.18	0.15	0.15
	median	0.10	0.12	0.17	0.22	0.17	0.21	0.02	0.13	0.18	0.21	0.15	0.14	0.17	0.21	0.19	0.18
	Q3	0.18	0.16	0.21	0.27	0.21	0.27	0.04	0.17	0.22	0.25	0.19	0.18	0.21	0.24	0.23	0.23
	IQR	0.15	0.13	0.07	0.09	0.07	0.10	0.03	0.11	0.06	0.08	0.08	0.09	0.07	0.07	0.07	0.07

Table VII: Boxplot characteristics for three variable numbers in ALM (PSFE)

VAR	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
4	Q1	56.6	80.7	97.1	120.5	90.7	115.8	8.6	58.8	85.9	95.4	62.7	87.6	99.2	112.5	101.7	106.8
	median	117.4	128.1	140.4	176.2	138.7	171.4	25.2	94.6	130.7	143.8	87.5	137.8	140.9	159.9	157.3	158.1
	Q3	185.9	198.4	208.8	245.3	212.2	237.2	60.4	169.2	206.8	217.8	140.3	213.8	214.2	218.6	226.5	228.3
	IQR	129.3	117.7	111.7	124.8	121.6	121.3	51.8	110.5	120.9	122.4	77.7	126.2	115.1	106.1	124.8	121.5
7	Q1	32.1	34.2	49.8	68.1	48.1	63.0	3.5	33.7	54.6	62.5	36.1	36.9	51.8	69.1	56.8	56.0
	median	52.8	50.5	76.5	100.6	77.2	96.1	6.7	49.9	80.1	90.5	50.0	56.7	78.2	92.0	83.1	83.3
	Q3	91.6	83.7	118.5	144.3	114.1	141.0	15.7	87.2	123.9	135.5	73.3	99.6	116.7	131.2	125.4	122.4
	IQR	59.5	49.4	68.7	76.2	66.0	78.0	12.2	53.5	69.3	73.0	37.2	62.6	64.8	62.2	68.6	66.4
10	Q1	6.2	12.8	32.4	45.2	31.1	42.5	2.2	22.6	35.9	46.9	24.5	27.0	31.9	48.1	35.5	34.6
	median	37.9	37.9	46.7	70.4	46.3	65.9	3.2	37.5	53.1	67.4	34.3	39.3	46.7	65.5	54.6	57.1
	Q3	56.0	54.6	80.8	103.6	81.7	97.0	5.6	55.6	85.1	97.8	48.0	60.3	81.4	92.3	85.9	84.7
	IQR	49.8	41.8	48.4	58.4	50.5	54.5	3.4	32.9	49.3	50.9	23.5	33.3	49.5	44.1	50.5	50.1

Table VIII: Boxplot characteristics for three variable numbers in ALM (CU)

VAR	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
4	Q1	0.11	0.15	0.19	0.21	0.18	0.21	0.03	0.14	0.19	0.18	0.13	0.16	0.19	0.19	0.20	0.20
	median	0.20	0.21	0.24	0.27	0.24	0.26	0.07	0.20	0.24	0.24	0.18	0.23	0.24	0.25	0.25	0.25
	Q3	0.27	0.28	0.30	0.33	0.30	0.32	0.14	0.27	0.31	0.31	0.23	0.30	0.31	0.31	0.31	0.31
	IQR	0.17	0.13	0.11	0.12	0.12	0.12	0.11	0.13	0.12	0.12	0.10	0.14	0.11	0.12	0.12	0.11
7	Q1	0.11	0.12	0.18	0.24	0.18	0.22	0.02	0.13	0.21	0.22	0.14	0.13	0.19	0.22	0.20	0.20
	median	0.17	0.17	0.25	0.31	0.25	0.30	0.03	0.18	0.27	0.29	0.19	0.19	0.26	0.29	0.27	0.27
	Q3	0.27	0.27	0.35	0.40	0.34	0.39	0.08	0.28	0.36	0.38	0.25	0.30	0.35	0.38	0.37	0.36
	IQR	0.17	0.14	0.17	0.16	0.17	0.17	0.06	0.15	0.16	0.16	0.11	0.17	0.16	0.15	0.16	0.16
10	Q1	0.03	0.06	0.17	0.24	0.17	0.22	0.02	0.13	0.19	0.24	0.14	0.14	0.17	0.23	0.19	0.18
	median	0.17	0.18	0.24	0.33	0.23	0.31	0.02	0.19	0.27	0.32	0.19	0.19	0.24	0.30	0.27	0.27
	Q3	0.25	0.25	0.36	0.43	0.37	0.42	0.04	0.27	0.39	0.43	0.25	0.28	0.37	0.41	0.39	0.38
	IQR	0.22	0.20	0.19	0.20	0.20	0.20	0.02	0.14	0.20	0.18	0.11	0.14	0.20	0.18	0.20	0.20

Table IX: Boxplot characteristics for three numbers of categories in ALM (PSFE)

CAT	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
3	Q1	69.6	49.6	62.0	105.7	73.2	102.1	2.9	48.6	72.4	89.4	60.2	57.4	63.8	90.2	79.0	80.7
	median	125.8	107.2	124.8	164.5	127.3	161.4	7.3	106.5	131.7	151.9	97.9	127.5	126.9	152.9	144.8	145.5
	Q3	194.0	183.7	196.1	243.5	202.3	235.2	30.5	176.1	203.6	226.8	156.3	208.3	198.3	217.4	218.6	220.0
	IQR	124.4	134.1	134.1	137.8	129.1	133.0	27.6	127.5	131.2	137.4	96.0	150.9	134.6	127.3	139.6	139.3
5	Q1	28.2	32.8	45.1	63.9	43.0	59.2	3.1	32.5	52.6	62.2	35.3	36.0	46.8	64.4	53.5	52.3
	median	49.7	54.3	81.9	102.2	78.2	95.6	6.6	51.8	81.2	91.7	50.9	58.7	82.4	94.2	86.1	85.1
	Q3	84.8	94.5	120.9	145.8	114.0	141.7	24.7	80.5	116.1	126.7	75.7	106.6	122.3	135.9	128.0	124.8
	IQR	56.7	61.6	75.8	81.9	71.0	82.5	21.7	48.0	63.5	64.4	40.4	70.5	75.5	71.5	74.5	72.4
7	Q1	5.3	30.2	42.2	50.8	39.2	48.0	3.4	26.5	44.4	50.5	26.7	31.4	41.7	54.9	46.7	45.1
	median	30.6	44.7	71.8	79.7	66.5	76.8	6.8	41.7	66.7	72.1	37.9	46.1	71.4	78.2	73.0	73.7
	Q3	49.5	76.0	99.1	114.5	94.6	110.4	20.1	63.1	94.7	100.5	52.3	80.7	98.8	105.7	102.8	101.8
	IQR	44.2	45.8	57.0	63.8	55.4	62.4	16.8	36.7	50.3	50.0	25.6	49.3	57.1	50.7	56.1	56.7

Table X: Boxplot characteristics for three numbers of categories in ALM (CU)

CAT	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
3	Q1	0.22	0.17	0.22	0.30	0.23	0.30	0.02	0.19	0.24	0.28	0.22	0.21	0.22	0.27	0.26	0.27
	median	0.29	0.26	0.30	0.38	0.31	0.37	0.04	0.27	0.32	0.35	0.26	0.30	0.31	0.35	0.33	0.33
	Q3	0.40	0.37	0.42	0.47	0.43	0.46	0.10	0.40	0.43	0.45	0.33	0.41	0.42	0.45	0.44	0.44
	IQR	0.18	0.20	0.20	0.16	0.19	0.17	0.08	0.21	0.19	0.18	0.11	0.19	0.20	0.18	0.18	0.17
5	Q1	0.07	0.12	0.18	0.23	0.17	0.22	0.02	0.14	0.20	0.22	0.15	0.15	0.19	0.21	0.20	0.20
	median	0.16	0.18	0.23	0.28	0.23	0.27	0.03	0.18	0.25	0.27	0.18	0.20	0.24	0.27	0.25	0.25
	Q3	0.23	0.25	0.33	0.36	0.31	0.35	0.09	0.25	0.33	0.35	0.22	0.27	0.33	0.34	0.34	0.33
	IQR	0.16	0.13	0.14	0.14	0.14	0.14	0.08	0.11	0.13	0.14	0.07	0.12	0.14	0.13	0.14	0.14
7	Q1	0.02	0.11	0.16	0.18	0.15	0.17	0.02	0.11	0.17	0.17	0.11	0.12	0.16	0.18	0.17	0.17
	median	0.10	0.15	0.21	0.23	0.20	0.22	0.03	0.15	0.22	0.22	0.13	0.16	0.21	0.22	0.21	0.21
	Q3	0.16	0.21	0.28	0.30	0.27	0.29	0.08	0.20	0.29	0.29	0.16	0.21	0.27	0.29	0.28	0.28
	IQR	0.14	0.10	0.12	0.12	0.12	0.12	0.07	0.09	0.11	0.12	0.05	0.10	0.12	0.11	0.12	0.12

Table XI: Boxplot characteristics for three minimal between-cluster distances in CLM (PSFE)

DIST	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
0.50	Q1	25.3	58.1	26.3	31.3	35.1	31.1	30.0	30.6	53.4	62.6	45.5	59.9	25.5	57.0	40.6	38.7
	median	47.4	82.3	57.8	57.5	75.5	59.4	47.1	68.6	89.5	89.1	67.2	86.5	53.9	80.7	89.4	89.4
	Q3	98.5	135.0	138.3	159.9	145.5	155.9	82.5	130.9	150.7	142.4	110.2	160.2	135.1	132.2	160.1	161.7
	IQR	73.2	76.9	111.9	128.5	110.4	124.8	52.5	100.2	97.3	79.8	64.8	100.4	109.6	75.2	119.5	123.0
0.34	Q1	16.1	36.8	15.4	18.1	21.0	18.4	22.3	19.8	32.3	38.2	29.3	38.1	15.2	33.4	23.7	23.9
	median	31.2	53.1	33.1	35.4	49.1	35.2	34.1	46.6	53.1	50.2	43.0	55.4	33.5	47.3	56.3	58.3
	Q3	56.3	81.0	61.4	73.2	83.9	70.6	62.8	71.0	78.8	71.4	70.3	89.5	62.4	70.2	96.8	95.4
	IQR	40.2	44.2	45.9	55.1	62.9	52.2	40.5	51.3	46.5	33.2	41.0	51.4	47.2	36.9	73.1	71.5
0.21	Q1	13.0	25.5	11.9	14.2	17.4	14.5	20.3	15.8	23.5	25.6	23.3	25.6	12.1	23.7	18.1	18.0
	median	26.0	35.7	24.0	26.6	36.5	27.5	30.9	33.5	39.6	35.4	33.8	38.1	23.7	33.3	41.2	41.4
	Q3	44.0	53.7	42.6	47.1	63.7	46.8	53.3	59.3	60.3	52.7	56.0	58.2	42.3	51.7	73.3	72.2
	IQR	31.0	28.2	30.7	32.9	46.3	32.3	33.0	43.5	36.8	27.1	32.7	32.6	30.2	28.0	55.3	54.2

Table XII: Boxplot characteristics for three minimal between-cluster distances in CLM (CU)

DIST	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
0.50	Q1	0.12	0.19	0.13	0.14	0.17	0.14	0.17	0.16	0.22	0.22	0.17	0.20	0.13	0.20	0.18	0.18
	median	0.17	0.27	0.20	0.21	0.23	0.21	0.20	0.21	0.28	0.30	0.23	0.28	0.18	0.27	0.25	0.25
	Q3	0.28	0.37	0.34	0.38	0.34	0.38	0.23	0.33	0.39	0.38	0.32	0.40	0.35	0.35	0.36	0.37
	IQR	0.17	0.18	0.21	0.25	0.18	0.25	0.06	0.18	0.18	0.16	0.15	0.20	0.22	0.14	0.18	0.18
0.34	Q1	0.08	0.13	0.08	0.09	0.12	0.09	0.14	0.12	0.16	0.15	0.12	0.14	0.08	0.13	0.13	0.13
	median	0.11	0.19	0.13	0.13	0.17	0.13	0.16	0.16	0.20	0.19	0.16	0.19	0.12	0.17	0.19	0.19
	Q3	0.18	0.25	0.21	0.23	0.24	0.23	0.18	0.23	0.26	0.24	0.21	0.28	0.21	0.22	0.27	0.27
	IQR	0.11	0.12	0.13	0.14	0.13	0.15	0.04	0.11	0.10	0.09	0.09	0.14	0.13	0.09	0.14	0.14
0.21	Q1	0.06	0.09	0.07	0.07	0.10	0.07	0.12	0.10	0.14	0.12	0.10	0.09	0.07	0.10	0.11	0.11
	median	0.09	0.13	0.10	0.10	0.15	0.11	0.15	0.14	0.17	0.15	0.13	0.14	0.10	0.13	0.16	0.16
	Q3	0.14	0.18	0.16	0.17	0.20	0.16	0.17	0.19	0.20	0.18	0.16	0.20	0.15	0.17	0.21	0.21
	IQR	0.08	0.09	0.08	0.09	0.10	0.09	0.04	0.09	0.07	0.06	0.06	0.10	0.08	0.07	0.11	0.11

Table XIII: Boxplot characteristics for three variable numbers in CLM (PSFE)

VAR	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
4	Q1	36.2	45.0	39.0	37.7	64.8	38.0	61.2	59.2	63.9	55.5	56.7	47.9	38.3	52.9	78.6	78.9
	median	58.4	74.4	64.2	59.8	93.5	60.4	76.4	80.0	89.8	83.3	80.1	82.4	61.4	86.2	110.2	110.6
	Q3	126.8	132.0	122.4	124.2	139.2	125.8	100.9	122.2	136.7	137.8	138.4	144.9	119.0	140.8	161.4	160.9
	IQR	90.6	87.1	83.4	86.6	74.4	87.7	39.7	63.0	72.8	82.3	81.6	97.0	80.7	88.0	82.8	82.0
7	Q1	16.9	34.4	15.3	18.2	23.8	18.6	29.6	21.8	30.4	34.4	31.8	35.5	14.5	33.5	27.4	26.4
	median	27.5	54.3	24.4	29.1	38.1	30.8	34.7	33.0	45.8	50.7	43.5	55.2	24.0	49.6	42.2	41.3
	Q3	54.3	82.2	48.1	59.8	67.4	57.9	44.2	59.4	71.9	72.5	63.9	88.7	48.8	70.3	74.6	75.6
	IQR	37.4	47.8	32.8	41.6	43.6	39.3	14.6	37.6	41.5	38.1	32.1	53.1	34.3	36.9	47.1	49.2
10	Q1	11.0	30.1	11.0	12.4	12.8	12.5	17.8	11.4	24.8	27.1	21.5	30.7	11.0	26.7	13.6	13.6
	median	19.2	44.7	20.1	23.3	22.4	22.8	20.8	19.9	38.7	40.1	28.5	47.7	19.6	37.2	23.5	24.1
	Q3	36.8	65.4	42.2	51.2	49.4	52.9	24.9	46.3	60.8	62.0	41.7	69.5	43.0	53.5	53.0	52.6
	IQR	25.8	35.3	31.2	38.7	36.6	40.4	7.1	34.9	36.0	35.0	20.2	38.9	32.0	26.7	39.4	39.0

Table XIV: Boxplot characteristics for three variable numbers in CLM (CU)

VAR	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
4	Q1	0.09	0.09	0.11	0.10	0.16	0.10	0.16	0.16	0.16	0.13	0.12	0.10	0.11	0.12	0.18	0.18
	median	0.14	0.14	0.16	0.14	0.20	0.15	0.18	0.19	0.20	0.17	0.16	0.16	0.15	0.17	0.22	0.22
	Q3	0.21	0.21	0.23	0.23	0.24	0.22	0.21	0.23	0.24	0.24	0.23	0.23	0.22	0.24	0.27	0.27
	IQR	0.12	0.12	0.12	0.13	0.08	0.13	0.05	0.07	0.08	0.11	0.11	0.13	0.12	0.12	0.09	0.09
7	Q1	0.08	0.13	0.08	0.09	0.12	0.09	0.14	0.12	0.15	0.15	0.12	0.14	0.08	0.13	0.13	0.13
	median	0.12	0.19	0.12	0.13	0.17	0.14	0.16	0.16	0.20	0.19	0.17	0.20	0.11	0.18	0.18	0.18
	Q3	0.19	0.27	0.19	0.22	0.25	0.22	0.19	0.23	0.27	0.26	0.23	0.28	0.19	0.25	0.26	0.26
	IQR	0.11	0.14	0.11	0.13	0.13	0.12	0.06	0.11	0.12	0.11	0.10	0.14	0.12	0.12	0.13	0.13
10	Q1	0.07	0.16	0.08	0.09	0.09	0.09	0.12	0.09	0.16	0.16	0.12	0.17	0.08	0.14	0.10	0.10
	median	0.11	0.22	0.13	0.15	0.15	0.15	0.14	0.14	0.23	0.22	0.16	0.23	0.13	0.19	0.15	0.15
	Q3	0.17	0.30	0.23	0.27	0.26	0.27	0.17	0.25	0.31	0.32	0.23	0.32	0.23	0.27	0.27	0.27
	IQR	0.10	0.14	0.15	0.18	0.17	0.18	0.05	0.16	0.15	0.16	0.10	0.16	0.15	0.13	0.18	0.18

Table XV: Boxplot characteristics for three numbers of categories in CLM (PSFE)

CAT	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
3	Q1	45.7	64.2	49.3	62.6	65.2	59.8	25.3	58.2	61.6	52.8	47.9	75.0	49.6	51.2	68.8	69.6
	median	77.0	106.4	96.9	107.7	110.1	108.4	40.5	100.3	105.4	103.4	86.4	118.2	93.7	89.9	121.2	120.7
	Q3	139.4	163.1	158.1	176.8	158.0	175.1	85.3	148.1	160.2	163.1	151.2	185.2	157.1	150.1	184.3	179.3
	IQR	93.7	99.0	108.8	114.3	92.8	115.3	60.0	89.9	98.6	110.3	103.4	110.2	107.6	98.9	115.5	109.7
5	Q1	17.9	37.0	16.9	20.5	23.1	21.1	25.0	20.0	33.1	35.9	30.5	38.7	16.3	35.2	25.7	24.9
	median	28.7	54.0	27.4	33.7	38.9	32.7	38.8	33.4	53.2	55.3	46.7	55.2	27.3	51.6	45.0	43.7
	Q3	47.8	76.9	49.6	51.9	74.9	53.6	67.2	63.9	78.2	76.2	69.4	77.8	49.5	75.2	90.4	90.6
	IQR	29.9	39.9	32.7	31.4	51.8	32.6	42.2	43.8	45.1	40.3	38.9	39.1	33.2	40.0	64.7	65.6
7	Q1	10.7	24.2	10.5	11.6	14.0	11.6	20.6	12.7	23.3	28.2	23.6	24.5	10.0	23.7	14.4	14.7
	median	17.3	35.6	17.3	18.6	23.7	18.7	31.2	22.3	37.8	41.3	34.6	36.2	15.8	33.4	26.4	26.4
	Q3	28.7	51.4	30.9	29.6	48.5	30.9	52.1	47.9	54.0	56.4	47.9	51.8	29.1	51.0	60.0	59.8
	IQR	18.0	27.2	20.3	18.0	34.4	19.3	31.5	35.2	30.7	28.2	24.3	27.2	19.1	27.3	45.6	45.1

Table XVI: Boxplot characteristics for three numbers of categories in CLM (CU)

CAT	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
3	Q1	0.17	0.19	0.19	0.21	0.23	0.21	0.12	0.21	0.22	0.20	0.19	0.23	0.19	0.19	0.25	0.25
	median	0.22	0.26	0.25	0.28	0.27	0.28	0.16	0.26	0.28	0.26	0.24	0.30	0.25	0.24	0.30	0.30
	Q3	0.30	0.36	0.36	0.40	0.36	0.41	0.21	0.35	0.37	0.35	0.32	0.41	0.36	0.33	0.39	0.39
	IQR	0.14	0.17	0.17	0.19	0.13	0.20	0.09	0.13	0.15	0.15	0.12	0.18	0.17	0.14	0.15	0.15
5	Q1	0.09	0.13	0.09	0.10	0.13	0.10	0.15	0.12	0.17	0.15	0.13	0.14	0.09	0.13	0.14	0.14
	median	0.12	0.18	0.12	0.13	0.17	0.13	0.17	0.16	0.20	0.19	0.16	0.19	0.12	0.18	0.19	0.19
	Q3	0.15	0.25	0.17	0.19	0.21	0.18	0.20	0.19	0.25	0.25	0.20	0.25	0.16	0.24	0.23	0.23
	IQR	0.07	0.12	0.07	0.08	0.08	0.08	0.05	0.07	0.09	0.10	0.07	0.11	0.07	0.10	0.09	0.09
7	Q1	0.06	0.09	0.06	0.07	0.09	0.07	0.14	0.09	0.13	0.11	0.09	0.09	0.06	0.09	0.09	0.09
	median	0.07	0.12	0.08	0.08	0.12	0.09	0.16	0.12	0.16	0.15	0.12	0.13	0.08	0.13	0.13	0.13
	Q3	0.10	0.18	0.12	0.11	0.16	0.12	0.18	0.16	0.19	0.20	0.14	0.18	0.11	0.17	0.17	0.17
	IQR	0.04	0.09	0.05	0.05	0.07	0.05	0.04	0.07	0.06	0.09	0.05	0.09	0.05	0.08	0.08	0.08

Table XVII: Boxplot characteristics for three minimal between-cluster distances in SLM (PSFE)

DIST	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
0.50	Q1	2.4	1.5	17.3	1.3	1.4	1.3	1.5	1.5	1.3	1.3	0.9	1.4	26.6	0.9	1.5	1.5
	median	8.8	21.4	58.9	44.6	3.5	54.0	1.6	1.6	59.0	35.6	1.4	48.9	59.4	2.4	66.9	66.2
	Q3	62.9	79.8	87.7	75.8	71.6	79.6	2.1	3.5	103.2	75.4	11.3	84.7	88.5	11.9	147.2	147.2
	IQR	60.5	78.3	70.4	74.5	70.2	78.3	0.7	2.0	101.9	74.1	10.3	83.2	61.8	11.0	145.6	145.7
0.34	Q1	1.6	1.4	1.4	1.2	1.3	1.2	1.5	1.4	1.2	1.3	0.8	1.4	1.5	0.8	1.3	1.3
	median	3.5	1.6	1.8	1.4	1.5	1.3	1.6	1.6	1.4	1.4	1.1	1.5	2.2	1.3	1.5	1.5
	Q3	12.7	8.6	8.2	3.0	1.9	1.8	1.9	1.9	4.3	2.6	7.0	4.7	10.8	4.8	60.3	60.7
	IQR	11.1	7.2	6.8	1.9	0.6	0.7	0.4	0.4	3.1	1.3	6.1	3.3	9.4	4.0	59.0	59.4
0.21	Q1	1.8	1.5	1.5	1.3	1.4	1.2	1.5	1.5	1.3	1.3	0.8	1.4	1.5	0.8	1.4	1.4
	median	3.9	1.6	1.8	1.5	1.5	1.4	1.6	1.6	1.5	1.5	1.2	1.6	2.0	1.4	1.5	1.5
	Q3	11.8	6.4	4.3	3.6	1.8	2.4	1.9	1.9	1.9	3.1	7.3	2.5	4.7	4.7	42.1	41.1
	IQR	10.0	5.0	2.8	2.3	0.4	1.1	0.4	0.4	0.7	1.8	6.5	1.1	3.2	3.9	40.7	39.8

Table XVIII: Boxplot characteristics for three minimal between-cluster distances in SLM (CU)

DIST	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
0.50	Q1	0.01	0.01	0.06	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.08	0.01	0.01	0.01
	median	0.03	0.06	0.20	0.11	0.01	0.14	0.01	0.01	0.18	0.09	0.01	0.11	0.20	0.01	0.23	0.23
	Q3	0.16	0.22	0.26	0.23	0.21	0.24	0.01	0.01	0.26	0.22	0.04	0.24	0.27	0.05	0.30	0.30
	IQR	0.14	0.21	0.20	0.22	0.20	0.23	0.00	0.01	0.25	0.21	0.04	0.23	0.18	0.04	0.29	0.29
0.34	Q1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	median	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Q3	0.04	0.02	0.03	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.03	0.01	0.04	0.02	0.16	0.16
	IQR	0.03	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.01	0.03	0.01	0.15	0.15
0.21	Q1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	median	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Q3	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.01	0.02	0.02	0.12	0.12
	IQR	0.02	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.01	0.11	0.11

Table XIX: Boxplot characteristics for three variable numbers in SLM (PSFE)

VAR	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
4	Q1	10.8	4.4	6.5	2.9	1.7	1.8	1.6	1.6	3.1	2.9	0.9	4.7	6.3	4.1	62.1	62.7
	median	22.7	26.0	20.6	8.9	3.1	7.2	1.9	2.2	61.8	9.4	7.0	34.3	18.8	8.4	117.3	118.2
	Q3	50.3	97.6	70.4	46.6	22.1	52.3	2.9	4.7	128.8	43.4	58.2	104.0	64.9	18.2	171.5	172.5
	IQR	39.5	93.2	63.9	43.7	20.4	50.5	1.3	3.1	125.6	40.5	57.3	99.3	58.6	14.2	109.4	109.8
7	Q1	2.1	1.4	1.5	1.2	1.3	1.2	1.5	1.4	1.2	1.3	0.9	1.4	1.6	0.8	1.3	1.3
	median	3.9	1.6	2.0	1.4	1.5	1.4	1.6	1.6	1.4	1.5	1.5	1.5	2.7	1.2	1.5	1.5
	Q3	9.2	12.1	24.8	4.2	1.8	2.7	1.8	1.8	2.0	2.6	7.5	1.9	48.8	2.8	46.1	43.6
	IQR	7.1	10.7	23.3	3.0	0.5	1.5	0.4	0.3	0.8	1.4	6.6	0.5	47.2	1.9	44.8	42.3
10	Q1	1.1	1.4	1.4	1.1	1.3	1.2	1.4	1.4	1.2	1.2	0.8	1.4	1.4	0.8	1.3	1.3
	median	1.8	1.5	1.6	1.3	1.4	1.3	1.5	1.5	1.3	1.3	1.0	1.5	1.6	0.8	1.4	1.4
	Q3	2.7	1.6	2.9	1.5	1.6	1.5	1.7	1.6	1.5	1.5	1.3	1.6	3.5	1.2	1.6	1.6
	IQR	1.7	0.3	1.6	0.4	0.3	0.3	0.3	0.2	0.3	0.3	0.4	0.3	2.1	0.4	0.3	0.3

Table XX: Boxplot characteristics for three variable numbers in SLM (CU)

VAR	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
4	Q1	0.03	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.02	0.02	0.15	0.15
	median	0.05	0.06	0.06	0.03	0.01	0.02	0.01	0.01	0.13	0.03	0.03	0.08	0.06	0.03	0.22	0.22
	Q3	0.11	0.17	0.15	0.10	0.06	0.11	0.01	0.01	0.22	0.10	0.13	0.19	0.14	0.06	0.27	0.28
	IQR	0.08	0.16	0.12	0.09	0.05	0.10	0.00	0.01	0.21	0.09	0.13	0.17	0.12	0.04	0.13	0.13
7	Q1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	median	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Q3	0.03	0.04	0.10	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.01	0.17	0.01	0.16	0.16
	IQR	0.02	0.03	0.10	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.16	0.01	0.16	0.15
10	Q1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	median	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Q3	0.02	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01
	IQR	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00

Table XXI: Boxplot characteristics for three numbers of categories in SLM (PSFE)

CAT	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
3	Q1	1.6	1.6	1.7	1.3	1.5	1.2	1.7	1.6	1.4	1.3	2.5	1.5	1.7	1.8	1.5	1.5
	median	7.9	20.9	14.9	4.5	1.9	2.9	1.9	1.9	2.3	3.7	16.4	2.4	14.9	4.5	57.9	55.3
	Q3	52.9	93.8	72.2	23.6	11.8	34.4	3.3	5.1	112.3	41.3	60.8	99.8	72.2	13.8	159.1	160.1
	IQR	51.3	92.2	70.5	22.3	10.3	33.2	1.6	3.5	110.9	39.9	58.3	98.3	70.5	12.0	157.6	158.6
5	Q1	1.7	1.5	1.4	1.2	1.4	1.2	1.5	1.5	1.3	1.3	0.9	1.4	1.6	0.8	1.3	1.4
	median	4.7	1.7	2.1	1.5	1.5	1.5	1.6	1.6	1.5	1.6	1.2	1.6	7.5	1.2	1.7	1.7
	Q3	22.5	12.7	18.8	8.1	2.3	6.3	1.8	1.8	23.2	8.2	3.0	16.8	39.6	5.3	81.1	82.4
	IQR	20.8	11.3	17.3	6.9	0.9	5.1	0.3	0.4	22.0	6.9	2.1	15.4	37.9	4.5	79.8	81.0
7	Q1	2.1	1.4	1.5	1.2	1.3	1.2	1.4	1.4	1.2	1.3	0.8	1.4	1.5	0.8	1.3	1.3
	median	3.3	1.5	2.2	1.4	1.4	1.4	1.5	1.5	1.3	1.4	0.9	1.5	2.2	0.9	1.5	1.5
	Q3	8.4	2.1	19.3	2.6	1.7	2.0	1.6	1.6	2.2	2.1	1.0	2.1	14.2	2.2	48.9	48.2
	IQR	6.3	0.7	17.8	1.3	0.4	0.7	0.2	0.2	1.0	0.8	0.3	0.8	12.7	1.4	47.6	46.9

Table XXII: Boxplot characteristics for three numbers of categories in SLM (CU)

CAT	char.	AN	BU	ES	G1	G2	G3	G4	GA	IOF	LIN	LIN1	OF	SM	SV	VE	VM
3	Q1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	median	0.03	0.06	0.05	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.06	0.01	0.05	0.02	0.19	0.18
	Q3	0.13	0.20	0.20	0.07	0.04	0.09	0.01	0.02	0.24	0.10	0.15	0.21	0.20	0.05	0.30	0.30
	IQR	0.12	0.19	0.19	0.06	0.03	0.08	0.01	0.01	0.23	0.09	0.14	0.20	0.19	0.04	0.29	0.29
5	Q1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	median	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.01	0.01	0.01
	Q3	0.06	0.03	0.06	0.03	0.01	0.02	0.01	0.01	0.07	0.03	0.02	0.04	0.14	0.02	0.21	0.21
	IQR	0.05	0.02	0.05	0.02	0.00	0.02	0.00	0.00	0.06	0.02	0.01	0.03	0.13	0.01	0.20	0.20
7	Q1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	median	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Q3	0.02	0.01	0.07	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.05	0.01	0.13	0.13
	IQR	0.01	0.00	0.06	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.04	0.00	0.12	0.12