LIVES Working Paper 2026/108

What Are We Looking For?

A Comparative Review of Clustering Algorithms and Cluster Quality Indices For Sequence Analysis

LEONHARD UNTERLERCHNER, MATTHIAS STUDER

RESEARCH PAPER
http://dx.doi.org/10.12682/lives.2296-1658.2026.108
ISSN 2296-1658

UNIL | Université de Lausanne
UNIVERSITÉ DE GENÈVE
LIVES
SWISS CENTRE OF EXPERTISE IN LIFE COURSE RESEARCH

Abstract

Sequence Analysis (SA) aims to provide a holistic view of life trajectories by creating a typology. Technically, it involves three steps: comparing the trajectories using a dissimilarity measure, regrouping similar pathways into types using a clustering algorithm before evaluating its quality using a cluster quality index.
This article aims to provide a comprehensive overview of the different clustering algorithms and cluster quality indices and to draw guidelines on their choice. The different methods are evaluated using simulations designed to reflect the different kinds of research questions addressed with sequence analysis as well as common longitudinal data characteristics.
The results highlight the good performance of Consensus Clustering with Ward linkage and Partition Around Medoids depending on the research aims and data characteristics. It also highlights the need to further define the required level of detail of a typology before choosing a cluster quality index. While the Calinsky-Harabasz Index with squared distances can be advocated for when aiming to create a parsimonious typology, the Hubert C index is recommended otherwise.

Authors
Leonhard Unterlerchner, Matthias Studer

Authors' Affiliation
Université de Genève

Keywords
Sequence Analysis, Cluster Analysis , Clustering Algorithms, Cluster Quality Indices, Typology Creation

What Are We Looking For? A Comparative Review of Clustering Algorithms and Cluster Quality Indices For Sequence Analysis

Leonhard Unterlerchner Matthias Studer

Abstract

Sequence Analysis (SA) aims to provide a holistic view on life trajectories by creating a typology. Technically, it involves three steps: comparing the trajectories using a dissimilarity measure, regrouping similar pathways into types using a clustering algorithm, and then evaluating its quality using a cluster quality index.

This article provides a comprehensive overview of different clustering algorithms and cluster quality indices, before drawing guidelines on their choice. The methods are evaluated using simulations designed to reflect the different kinds of research questions addressed with sequence analysis as well as common longitudinal data characteristics.

The results highlight the good performance of Consensus Clustering with Ward linkage and Partition Around Medoids depending on the research aims and data characteristics. It also highlights the need to further define the required level of detail of a typology before choosing a cluster quality index. While the Calinski-Harabasz Index with squared distances

can be advocated for when aiming to create a parsimonious typology, the Hubert C index is recommended otherwise.

1 Introduction

The life-course paradigm has gained increasing prominence in recent decades, contributing to a wide range of disciplines, including sociology, demography, gerontology, medicine and psychology (Elder et al., 2003; Bernardi et al., 2019). In this context, sequence analysis (SA) is regarded as one of the key longitudinal approaches (Shanahan, 2003; Brüderl et al., 2019; Mayer, 2009; Brzinsky-Fay, 2007; Liefbroer, 2019). It encompasses a broad set of methods aimed at analysing trajectories from a holistic perspective, including visualization and explanatory methods.

The standard use of SA involves four steps (Gauthier et al., 2014). First, the trajectories are coded as state sequences. Second, trajectories are compared to one another using a dissimilarity measure. Studer and Ritschard (2016) provide a comprehensive review of these developments and their usefulness for life-course research. Third, a typology of trajectories is created using cluster analysis. This typology aims to describe the various patterns observed in the data without any assumption on the data-generating models. This exploratory approach can capture complex and potentially unexpected dynamics in trajectories, which is useful for understanding the many life-course interdependencies.

Finally, subsequent analyses may use the typology to understand how the different types of trajectories relate to key social determinants or outcomes (Gauthier et al., 2014). Technically, it can be included in any statistical method handling categorical data.

Within this standard SA framework, cluster analysis plays a critical role, as it determines how the typology is created. However, this key methodological step has been only scantly discussed. This article aims to overcome this limitation through a comparative review of available methods, by assessing their relevance using simulations before providing practical guidelines to SA users.

The creation of a typology typically involves several steps (Hennig et al., 2015; Piccarreta and Studer, 2019). First, an appropriate clustering algorithm (CA) is chosen to create typologies with different numbers of groups. Then, the quality of these clusterings is evaluated using cluster quality indices (CQIs) (see Studer, 2013, for a review). Finally, the solution maximizing the CQI values is kept.

To date, most SA studies rely either on Partition Around Medoids (PAM) or the agglomerative Ward algorithms (Liao et al., 2022). However, many other algorithms have been proposed in the data-mining literature (Hennig et al., 2015). They were developed to identify different kinds of structure in the data, follow different aims and logic, and are optimized for various situations or data characteristics. Some of them might, therefore, be especially relevant for SA. The same applies to the CQIs.

According to the constructivist perspective proposed by Hennig (2015), which is highly relevant for the social sciences, the CA and the CQIs should be chosen by first defining "what kind of truth they [the researchers] are interested in and what should constitute a 'real' cluster." In other words, before trying to uncover different types of trajectories, one should first define what constitutes a type in the social sciences or in the life-course literature. Following this perspective, we start by reviewing and discussing recurrent research questions that could be addressed using SA. We also highlight common data characteristics that might affect typology creation. Then, relying on this discussion, we define a simulation

framework, which allows us to systematically evaluate the suitability of 15 CAs and seven CQIs for these research aims and data characteristics.

The remainder of the paper is structured as follows. Section 2 develops theoretical considerations on typology creation in the social sciences. The design of the simulation study is presented in Section 3. Section 4 is dedicated to the CA evaluation, first presenting the algorithms under review, then discussing the simulation results in subsection 4.2. Section 5 focuses on the use of CQIs. It follows the same structure as the CA section. Evaluated CQIs are presented first, and CQI-related findings are then formulated in Subsection 5.4. Section 6 presents the guidelines on CAs and CQIs. Finally, Section 7 concludes the paper.

2 Typologies in Social Sciences

To evaluate CAs and CQIs, we should first define what constitutes a good typology in life-course research, i.e. what we are looking for. In SA, the typology generally aims to describe the *relevant* types of trajectories observed in the data. The definition of what is relevant depends on the research question and the characteristics of the data (Kaufman and Rousseeuw, 1990; Hennig, 2015). In this article, we emphasize four key aspects which might be important for a typology in life-course research: what should be considered as a *type*, the required *level of detail* of the typology, the *level of the clustering structure* and the *sample size*.

2.1 Types' Definition

The definition of what should be considered as a type of trajectory typically depends on the research question. We identify three different *type definitions* that are relevant for choosing among CAs and CQIs: *common*, *atypical* and

hybrid types of trajectories.

Most life-course studies are interested in identifying the *common* types of trajectories. Here, the aim is to describe the most recurrent and frequent types of trajectories to provide an overview. These *common* types are then often related to key covariates to understand the profiles associated with each type. For instance, several studies have tried to understand how *common* types of family formation or professional trajectories have changed across cohorts to highlight the emergence of new *common* types (Widmer and Ritschard, 2009; Billari et al., 2006; Studer et al., 2018; Bras et al., 2010; Struffolino et al., 2016).

Some studies further aim to identify types of *atypical* trajectories, for instance to understand their social consequences (Lelièvre, 2019). These *atypical* trajectories are (implicitly) defined as infrequent trajectories that differ strongly from *common* types. However, there is a tension here as *atypical* trajectories should be sufficiently frequent to be of interest, and not only anecdotal cases. Many life-course researches are interested in identifying *atypical* trajectories, which are often associated with adverse life conditions. For instance, McVicar and Anyadike-Danes (2002) aim to identify the profiles of atypical, named "at-risk", trajectories of school-to-work transitions marked by joblessness. Similarly, Sacchi and Meyer (2016) showed that pupils undertaking atypical paths in secondary vocational education are more likely to encounter difficulties when entering the labour market.

Third, some studies are interested in describing and identifying *hybrid* types of trajectories. *Hybrid* trajectories are infrequent and lie in between *common* types of trajectories. Contrary to *atypical* trajectories, these trajectories share many similarities with the *common* types. In some studies, these infrequent *hybrid* cases might reveal particularly vulnerable or resilient patterns requiring further inves-

tigation. However, in other studies, one would prefer assigning these trajectories to the closest types in order to avoid overloading the description, as they are relatively infrequent. For instance, if we study professional trajectories described by the two *common* types lifelong employment and lifelong unemployment, there might be an infrequent pattern characterized by back-and-forth movement between work and unemployment. In some studies, one might want to describe this uncommon situation as a *hybrid* type, as it reveals a vulnerable and/or peculiar situation. In other studies, one might prefer to assign this *hybrid* type to one of the *common* types to obtain a more parsimonious typology. In a study on changes in family formation and employment trajectories in Switzerland during the 20$^{\text{th}}$ century, Widmer and Ritschard (2009) identify a type of employment trajectories lying in between the pattern of full-time employment —typical of men— and the at-home pattern —typical of women. The association of this hybrid pattern with younger cohorts of parents is interpreted as a sign of changes in family-formation behaviours. As another example, Unterlerchner et al. (2023) identified an infrequent school-to-work type of trajectory mixing vocational and academic education, which is associated with one of the highest average incomes in later life.

This discussion highlights that there might be different definitions of what should be considered as a *type* of trajectories. These definitions are not mutually exclusive, as one might be interested in uncovering all of them. However, in others, *hybrid* and/or *atypical* trajectories might be less relevant. The choice of a CA is thus expected to depend on these *type definitions*, as each algorithm has a different underlying definition of *what* is worth being described as a type. One aim of our simulations is thus to measure the ability of each CA and CQI to uncover *common*, *atypical* or *hybrid* types.

2.2 Level of Details

Apart from the definition of what should be considered a type, longitudinal studies using SA differ according to the *level of detail* they provide on trajectories. While many studies use between 3 and 8 types to describe trajectories, others go into more detail. For instance, Mattijssen et al. (2020) proposed a detailed view of non-standard employment by describing seventeen types of trajectories.

The level of detail required by a study mostly depends on the research question. Prior research might also inform the minimal number of types required to describe a specific domain. In the sociology of education, for instance, a specific educational system might require a minimal number of types to describe the main pathways.

While the required *level of detail* can be partially set theoretically, it also depends on the data itself. Indeed, some *common* or *atypical* patterns might have not been foreseen. Conversely, some planned paths might only be followed by a few individuals. Similarly, a small sample size might not allow describing the trajectories in too much detail.

As stated by Hennig (2015), some CAs and CQIs might perform better when the aim is to provide a detailed description, while others might be more suited to deriving of a parsimonious typology. This is especially expected regarding the behaviour of CQIs, as some of them include an explicit penalization for high numbers of groups.

2.3 Level of Clustering Structure

The *level of the clustering structure* describes the extent to which the trajectories are grouped into types. In data mining, this is often defined by relating the

intragroup homogeneity — the similarity of trajectories within the same type — with the intergroup separation — the dissimilarity of trajectories in different types. When types are homogeneous and highly separated from one another, there is a *strong clustering structure*. Conversely, there is a *weak clustering structure* when the types are either highly heterogeneous or not clearly separated from one another. Generally speaking, the identification of the typology is easier when there is a *strong clustering structure* in the data.

The *level of the clustering structure* depends both on the study domain and on the data characteristics, but to a lesser extent on the research question. In some domains, a stronger clustering structure can be expected if trajectories are deeply shaped by constraints, such as strong social norms, laws, institutions or economic limitations. For instance, educational pathways are strongly shaped by institutions, laws and regulations, which determine the accepted duration of various educational spells and their order. Indeed, tertiary education cannot take place before primary education. In contrast, professional trajectories can be much more volatile, as virtually any states' ordering can be observed.

Previously presented aspects also affect the *level of the clustering structure*. Indeed, the presence of *hybrid* types generally results in a weaker clustering structure. Furthermore, a higher *level of detail* also tends to result in a weaker clustering structure, as it is generally linked with a decrease in type separation.

The *level of the clustering structure* is also linked with several data properties, such as the length of the sequences or the number of states in the alphabet. A trajectory described in more detail, with a higher number of time points or using a larger number of states, might be less structured, as it leaves more room for variations.

As a result, prior domain knowledge and data characteristics might lead us

to expect a stronger or weaker clustering structure. At the same time, some CAs were specifically developed for weak or strong clustering structures. For this reason, we expect some methods to perform better in one context or the other.

2.4 Sample Size

Finally, the sample size must be considered. Indeed, some clustering techniques may suffer from instability when applied to small samples, while others may be intractable when applied to large samples. In addition, an interaction between sample size and the number of groups is expected. With a small sample size, identifying a high number of groups can become difficult due to their reduced size.

2.5 Synthesis

We identified four key aspects to take into account when creating a typology of trajectories. Some of these aspects should be defined by the researcher prior to the creation of a typology, including the *type definitions* of interest, the required *level of detail* or the expected *level of the clustering structure*. The *sample size*, the *level of detail* and the *level of the clustering structure* also depend on the data characteristics. Some aspects should therefore also be derived from the analysis.

As already stated, we expect CAs and CQIs to perform better in specific settings, as they were developed for different goals. We now turn to the presentation of the simulation design built upon the theoretical considerations of this section.

3 Simulation Design

This paper aims to systematically evaluate the presented clustering methods to create a typology with SA. We do so by relying on the simulation framework presented in Van Mechelen et al. (2023) involving the following three steps:

1. Generate sequences clustered into types, i.e. generate trajectories with an a priori known partition.

2. Cluster the data with a CA to be evaluated.

3. Evaluate the CA:

    - By comparing the estimated typology with the underlying partition defined in step one.
    - By evaluating the recovery of a particular type in the underlying partition.

The data-generation step is the most challenging. It requires generating sequences organized into types reflecting the different data characteristics and research questions identified in Section 2.

Van Mechelen et al. (2023) distinguish two strategies for generating data with a known partition: model-based and empirical simulations. We further distinguish between two empirical simulation approaches. On the one hand, simulations may rely on *observed* characteristics of the data. In this context, the underlying partition is actually realized. Consequently, the clustering structure cannot be weak. On the other hand, simulations can be driven by a *probabilistic* categorization of the data. In this case, an underlying type might potentially generate an observed sequence that strongly resembles another type, resulting in a weaker clustering structure.

Characteristic	Modalities				Total	
Approach	Model-based	Observed	Probabilistic		3	
Type's Definition	Common	Atypical	Hybrid		×3	
Clustering Structure Strength	Weak	Moderate	Strong		×3	
Level of Detail	Three	Four	Six	Nine	×4	
Sample Size	50	500			×2	= 216
Replications					×50	= 10,800

Table 1: Summary of the Simulations Generation Characteristics

Each approach has its pros and cons. The model-based approach enables the entire data-generation process to be controlled and allows the performance of a clustering method to be directly attributed to specific data properties or model specifications. As such, it allows a greater generalization of the results. However, models might struggle to reproduce the complexity of empirical data, and therefore move away from typical applications of the evaluated methods. On the other hand, empirical approaches can provide clearer recommendations for practical applications, as they rely on real-life data. However, the results might be specific to the dataset used for the simulation, and depend on uncontrolled data characteristics. Therefore, it might provide less general conclusions. Given the advantages and complementarities of each approach, we use all three of them to provide robust results.

Following Van Mechelen et al. (2023) we followed a full-factorial design resulting in the 216 simulations summarized in Table 1.

Even if the data-generation process differs between the three approaches, several characteristics can be operationalized in the same way to improve the comparability of the results. We considered four *levels of detail* with three, four, six or nine types. This reflects numbers of clusters commonly used in SA applications. The *sample size* was operationalized by considering either 50 or 500 observations per type. Following the discussions in Section 2, we considered three *type definitions* using evenly sized groups when focusing on *common* types. The *hybrid* and

atypical types are obtained by adding a group with a smaller size, i.e., one-fifth of the *common* types size. The *atypical* type is characterized by a state that is infrequent in the *common* groups. It is therefore infrequent and different from all other groups. The *hybrid* type contains trajectories with states frequent in at least two other groups. This infrequent group can therefore be seen as a mixture of *common* types.

The operationalization of the *level of the clustering structure* differs in the three simulation approaches, as it is tightly linked to the data-generating process. However, in each case, we considered the same three qualitative levels of *strong*, *moderate* or *weak clustering structure*.

To ensure the stability of the results, each simulation is replicated 50 times, resulting in the generation of 10,800 simulations.

In step 2, the trajectories are clustered using various CAs, which are all based on the computation of a distance measure. Studer and Ritschard (2016) provide an extensive review of the available options. In this article, we use Optimal Matching with constant costs for two reasons. First, it is by far the most widely used distance measure and is particularly versatile (Zimmermann and Seiler, 2019; Dlouhy and Biemann, 2015; Liao et al., 2022). Second, using the same distance with all the CAs is required to ensure consistency in the results.

In step 3, using an a priori known typology is common in CA benchmarkings. By knowing which structure an algorithm should find, the researcher is thus able to know whether the algorithm succeeds or fails in retrieving such structure (Van Mechelen et al., 2023). We rely on two complementary measures. First, we use the Adjusted Rand Index (ARI) for simulations involving common types (Rand, 1971; Hubert and Arabie, 1985). This index measures the similarity of the underlying true clustering with the one obtained with a given clustering method.

This index equals zero for similarity obtained "by chance" and 1 for two identical clusterings. Negative values can occur for highly dissimilar clusterings. Second, to specifically evaluate the algorithm's ability to identify hybrid and atypical types, we rely on the Jaccard Index (Jaccard, 1901). Ranging from 0 to 1, it measures the recovery of a particular cluster, with 1 indicating perfect recovery (Hennig, 2007). We consider a type as identified when the Jaccard value is greater than or equal to 0.7.

We now turn to the data generation according to empirical and model-based approaches, described in sections 3.1 and 3.2.

3.1 Model-Based Simulations

The first set of simulations relies on Markov models, which are widely used in benchmarking studies (Van Mechelen et al., 2023; Milligan and Cooper, 1987). These models can reproduce some key characteristics of categorical processes, such as their organization in spells, which is very common in SA applications (Studer, 2021). This section starts by presenting the models, before discussing their parametrization to capture the different aspects presented in Section 2.

3.1.1 Markov Models

In the present application, we consider Markov models described by three states, A, B, C, and sequence length $\ell = 20$. From these states, nine models are constructed that can be divided in two types: *dominated* and *transition* models. The three *dominated* models produce sequences characterized by a high chance of being in a given state, while still allowing small chances of experiencing spells in other states. The six *transition* models describe a transition from one state to another. Model specifications and a data-generation example are available in the

Appendix A.1.

3.1.2 Simulation Characteristics

The presented Markov models can simulate sequences for the various data characteristics and research aims discussed in Section 2. The input parameter p controls the *strength of the clustering structure*. We used $p = 0.85$ for weakly, $p = 0.92$ for moderately and $p = 0.99$ for strongly structured typologies. Lowering p results in less homogeneous sequences that might be closer to other types, depending on the other types included in the simulations (see Figure A.1 in the Appendix) .

The *level of detail* required by the typology is controlled by including different Markov models in the simulations, as described in Table 2. For three groups, models AA (for *dominant* in state A), AB (transition from A to B) and BB are included. For conciseness, the table does not repeat previously included types. The CC model (dominant in state C) is added for simulations in 4 groups and so on.

	Level of Detail	**Atypical Type**	**Hybrid Type**
3 Groups	AA & AB & BB	BB	AB
4 Groups	& CC	CC	AB
6 Groups	& BA & CB	CC	AB
9 Groups	& AC & BC & CA	CC	AB

Table 2: Groups present for each level of detail (left), the clusters of the previous rows are added to the corresponding row to obtain the desired number of groups. Models selected when atypical or hybrid types are simulated (right).

To simulate a situation where the interest lies only in the identification of *common* types, each model is used to generate the same number of sequences. For the identification of *atypical* and *hybrid* types, the size of one of the types is reduced to one-fifth of the others. For the *atypical* type, we used the CC model,

as the C state is less frequent in most of the other included types. It is, however, more frequent in simulations with more groups. On the other hand, we used the AB model (transition from state A to B) to operationalize the *hybrid* type. This *hybrid* group always lies between the types AA and BB.

3.2 Empirical Simulations

Our second and third sets of simulations are based on empirical data, the first cohort of TREE (TREE, 2016). In this longitudinal survey, Swiss compulsory school leavers are followed from the end of secondary lower education in 2000 until 2014. It allows reconstructing educational trajectories for 22 semesters, distinguishing between five states: general secondary education (G); secondary vocational education and training (V); tertiary education (T); transitional solutions (TS); and out of education (O).

Apart from the sequences, the simulations require us to specify the associated, a priori known, typology. This typology cannot be defined using cluster analysis as we would typically do in practice, as it would irremediably favour the used CA. We therefore define types theoretically according to expected curricula and previous studies on the Swiss educational system (Scharenberg et al., 2014).

The types are defined using ideal-typical sequences and presented in Table 3. These ideal-typical sequences are described using their state followed by the duration in semesters in this state. The column "description" provides a detailed version of the type.

The logic to account for different *levels of detail* is the same as for model-based simulations. We start with a solution in 3 groups, representing the *common* patterns expected within the Swiss education system. In 4 groups, the trajectories of individuals struggling to find an apprenticeship is added. This is an unconven-

	Sequence	Description
3 groups	V/8 - O/14	Four years secondary vocational education
	V/8 - T/6 - O/8	Vocational followed by tertiary education
	G/6 - T/6 - O/10	High school and bachelor
4 groups	TS/2 - V/6 - O/14	Transitional followed by 3-year vocational education
6 groups	V/8 - O/8 - T/6	Vocational followed by tertiary education with a break
	G/6 - O/16	High school
9 groups	V/6 - O/16	Three years secondary vocational education
	G/6 - T/10 - /6	High school and Master
	G/6 - T/16	High school and PhD

Table 3: Ideal-Types Trajectories

tional trajectory followed by one-fifth of the pupils, which might translate into further difficulties in later-life professional careers (Sacchi and Meyer, 2016). The 6-groups typology brings *common* trajectories that are not explicitly expected in the educational system. The 9-groups one further details the typology by distinguishing varying durations in each educational spell, which results in different diplomas.

Once the types have been defined, the sequences attributed to a type are randomly sampled (with replacement) from the real sequences. The difference between the *observed* and *probabilistic* approaches lies in the sampling procedure used in the simulation.

The *observed* approach proceeds by sampling sequences lying within a given radius r around the ideal-typical sequence. This radius is defined using the same distance measure as for the clustering, i.e. Optimal Matching with constant costs in our case. As a result, a value of two represents a difference of one semester from the ideal type.

The radius r controls the *clustering structure strength*, a higher value results in less homogeneous and less clearly separated clusters. We used the following r values: 10 (5 semesters) for weak, 6 (3 semesters) for moderate and 4 (2 semesters) for strong clustering structure.

The *probabilistic* approach relies on the logic of fuzzy clustering, sequences

are sampled according to their probability p_i of belonging to a given type using the following formula $p_i = \frac{d_i^{-\frac{1}{m-1}}}{\sum d_i^{-\frac{1}{m-1}}}$. A fuzzifier m allow blurring the probabilities p and thus decreasing the clustering structure, d_i being the distance to the ideal-typical sequence. We used the following m values: 1.6 for weak, 1.3 for intermediate and 1.01 for strong clustering structure.

Atypical groups are formed around the sequence *three years of secondary vocational education preceded by transitional solutions* (TS/2 - V/6 - O/14). This type being extremely close to other ones, we restricted the sampling to sequences beginning with the spell TS/2. This ensures the actual presence of the type in the simulations. Hybrid groups are formed around the sequence *vocational education followed up to the tertiary level* (V/8 - T/6 - O/8). It shares a spell in common with two well-separated clusters: *four-year secondary vocational education* (V/8 - O/14) and *high school diploma and bachelor* (G/6 - T/6 - O/10). As for model-based data, the size of the atypical and hybrid groups is reduced to one-fifth relative to the other groups.

3.3 Softwares

The *R* programming language (Core and Team, 2020) is used for the computations along with the following *R* packages: `TraMineR` (Gabadinho et al., 2011) for SA, `seqHMM` (Helske and Helske, 2023) for Markov models, `viridis` (Garnier et al., 2024) for colour-blind-friendly palettes, `WeightedCluster` (Studer, 2013), `fastcluster` (Müllner and Inc, 2024), `mclust` (Fraley et al., 2024), `dbscan` (Hahsler et al., 2023), `apcluster` (Bodenhofer et al., 2024) and `clue` (Hornik and Böhm, 2023) for clustering.

4 Clustering Algorithms

The aim of the following section is to provide guidelines for the use of CAs applied to SA. In consequence, we evaluate CAs performance according to the research aims and data characteristics discussed in Section 2. We first introduce the evaluated CAs in Section 4.1, before turning to the result presentation in Section 4.2. Guidelines are formulated in Section 6.

4.1 Evaluated Clustering Algorithms

Many different clustering algorithms have been proposed in the data-mining literature (Hennig et al., 2015). In this study, we focus on algorithms that can be used within the SA framework, i.e. that rely on dissimilarities or distances. These algorithms can be classified into five families according to their inner functioning: agglomerative, divisive, partitioning, density-based and consensus-based. In this section, we present the evaluated algorithms, along with their respective parameterizations.

Table 4 summarizes the evaluated algorithms. The first column shows their family and names. The next three columns provide information about the algorithms, their aims and important remarks, followed by the tested values of the parameters. The last two columns indicate the *R* libraries used and whether the results are presented within the article. Indeed, some clustering algorithms showed extremely poor results. We decided not to discuss them to shorten our presentation. Further details on these discarded algorithms are provided in Section 4.2.5.

Hierarchical algorithms are either agglomerative or divisive. The agglomerative approach forms groups by merging observations into clusters, which are

	Name	Description	Remarks & Param. Values	Incl.	R library
Agglomerative	Single	Merges the clusters having the smallest minimal between-cluster dissimilarity.	Might produce chained clustering.	No	fastcluster
	Complete	Merges clusters by maximizing the between-cluster dissimilarity.	Sensitive to outliers	Yes	fastcluster
	UPGMA	Merges the clusters with the smallest average between dissimilarity.	Compromise between single and complete linkage.	Yes	fastcluster
	β-Flexible UPGMA	UPGMA extension, β controls space distortions with positive values leading to space contraction and negative values space dilatation.	Tested β values: -1, -0.625, -0.3125, 0, 0.3125, 0.625	No	cluster
	WPGMA	Similar to UPGMA, but with higher penalization when merging small groups.	Merging of small groups is penalized.	Yes	fastcluster
	β-Flexible WPGMA	WPGMA extension, β controls space distortions as in β-Flexible UPGMA.	Tested β values: -1, -0.625, -0.3125, 0, 0.3125, 0.625	Yes	cluster
	Ward (squared diss.)	Minimize average within-cluster dissimilarities to their centroid.	Aim to produce homogenous clusters.	Yes	fastcluster
	Ward (non-squared)	Ward with non-squared dissimilarities.	Idem	Yes	fastcluster
	Ward (SQ module)	Identical sequences are agglomerated prior to Ward clustering.	Frequent sequences tend to have lower weights.	Yes	fastcluster
Divisive	Property Based	Recusively splits the sequences by looking at the share of the sequence discrepancy explained by sequence properties at each split.	Rules of cluster identification are directly linked to sequence properties.	Yes	WeightedCluster
	DIANA	Splits the cluster with the largest average dissimilarity.	Less affected by previous merges than agglomerative algorithms.	Yes	cluster
Partitioning	PAM	Identify a medoid for each cluster that minimizes the sum of dissimilarities to other sequences belonging to this cluster.	The medoids allow a clear definition of the clusters. Clusters tend to be compact.	Yes	WeightedCluster
	PAM (Ward)	PAM initialized using Ward clustering.	Might improve PAM results.	Yes	WeightedCluster
	PAM (UPGMA)	PAM initialized using UPGMA clustering.	Might improve PAM clustering and small-subgroup detection (hybrid or atypical).	Yes	WeightedCluster
	Affinity Propagation	Identify cluster exemplars by simultaneously scanning the data space, so that exemplars maximize the within-cluster similarity.	Might better explore the data space than PAM.	Yes	apcluster
Density	DBSCAN	Identify areas of high density as clusters.	Non-exhaustive clustering, outliers and noise can be identified. Tested parameter values: eps = $(2, 3, 4, 6)$, min. points = $(2, 4, 6, 8, 12)$	No	dbscan
	HDBSCAN	Finds areas of high density as clusters and produce a hierarchy.	Can capture clusters of various densities. Tested parameter values: min. points = $(2, 4, 6, 10, 14)$	No	dbscan
Consensus	Consensus (PAM)	Finds a consensus between multiple partitions generated by Bayesian resampling and PAM.	Soft Euclidian consensus functions.	Yes	clue
	Consensus (Ward D)	Finds a consensus between multiple partitions generated by Bayesian resampling and Ward (D)	Same consensus functions.	Yes	clue
	Consensus (UPGMA)	Multiple partitions generated with UPGMA.	Same consensus functions.	Yes	clue
	Consensus (PAM+UPGMA)	Multiple partitions generated with UPGMA and PAM.	Same consensus functions.	Yes	clue

Table 4: Summary of evaluated clustering algorithms

then merged into larger clusters until all observations belong to a single cluster. Once merged, observations cannot be separated. This means that previous merges affect the subsequent ones. This strategy might lead to poor results, particularly when aiming to identify atypical or hybrid types (i.e. the clusters are unbalanced), when there are many ties in the distance matrix (which is often the case in SA) or when the clustering structure is weak (Balcan et al., 2014; Martin et al., 2008). We review four agglomerative algorithms.

First, single linkage creates clusters by merging the clusters having the smallest minimal between-cluster dissimilarity (Florek et al., 1951; Sneath, 1957). This strategy is known to produce "chainings" when the clustering structure is weak (Milligan and Cooper, 1987). Chaining is a phenomenon occurring when clusters are merged because two data points are close to each cluster. This produces an elongated cluster with a low level of homogeneity. This property is not useful in social sciences.

Second, complete linkage takes the opposite strategy by merging the clusters maximizing the between-cluster dissimilarity (McQuitty, 1960). However, this might make complete linkage too sensitive to outliers.

Third, the average linkage algorithm merges the clusters with the smallest between-cluster average dissimilarity. In this study, we consider four versions of this algorithm. UPGMA, which stands for Unweighted Pair Group Method with Arithmetic mean (UPGMA) (Sokal et al., 1958), is expected to be efficient in identifying hybrid and atypical types (Lesnard, 2006), but Milligan and Cooper (1987) reported a lack of consistency in their simulation study. Sokal (1963) proposed an adaptation to penalize the merging of small groups named Weighted Pair Group Method with Arithmetic mean (WPGMA). In both algorithms, the clusters tend to get closer or farther apart, a phenomenon called space distortion.

The β-flexible versions of these two algorithms allow controlling this distortion with the β parameter. A positive β leads to space contraction, whereas negative values lead to space dilatation (Belbin et al., 2010).

Finally, the Ward algorithm is probably the most used in SA (Liao et al., 2022). It aims to minimize the within-cluster sum-of-squares inertia, which can be directly linked to the concept of residual variance when applied to Euclidean spaces (Ward, 1963). This goal is particularly useful when aiming to find common patterns in the data. In their simulation study, Dlouhy and Biemann (2015) identified Ward Algorithm as the best performing agglomerative algorithm. We included three versions of the algorithm. The original version uses squared dissimilarities. However, there is an ongoing debate on whether this is suitable to analyse non-Euclidean data (Batagelj, 1988; Murtagh and Legendre, 2014; Studer et al., 2011). We therefore included the version using raw dissimilarities. Furthermore, the SQ Stata's module makes use of a third version of the algorithm (Kohler et al., 2006). It starts by agglomerating identical sequences before clustering the data. While it considerably speeds up computations, this strategy might strongly reduce the relative weight of frequent sequences.

Divisive algorithms proceed in the opposite way. They start with all the observations grouped together, before splitting the data until each observation forms a cluster by itself. The algorithm, therefore, starts by maximizing a global criterion, but the process is more computationally demanding (Studer, 2018; Kaufman and Rousseeuw, 1990). Two algorithms are reviewed. First, Divisive Analysis (DIANA) splits clusters with the largest average dissimilarity (Macnaughton-Smith et al., 1964; Kaufman and Rousseeuw, 1990). The behaviour of this algorithm is close to UPGMA but should be more accurate at identifying typologies with a small number of groups, as it maximizes a global criterion at this hierarchical level.

Second, property-based clustering divides the data according to a comprehensive set of sequence properties to minimize the sum-of-square inertia (Piccarreta and Billari, 2007; Studer, 2018). This makes splitting rules explicit and thus the interpretation of the clustering easier.

Partitioning algorithms follow a different strategy. They start with an initial solution, in a pre-defined number of groups, that is recursively improved according to a global criterion. Partitioning Around Medoids (PAM) creates typologies by identifying a medoid for each cluster and minimizing the sum of dissimilarities of the cluster members to their medoid (Kaufman and Rousseeuw, 1990). This optimization strategy is probably well suited to find common types in data as it tends to find clusters of similar sizes. Some authors have advocated initializing the algorithm with the results of a prior hierarchical clustering (Lebart et al., 2006; Studer, 2013). We therefore also initialized PAM with UPGMA and Ward Linkage, which are expected to behave differently. Ward aims to maximize a criterion similar to PAM, while UPGMA might be better suited at identifying small subgroups.

Affinity Propagation is another partitioning technique which aims to identify clusters by maximizing the within-cluster similarity (i.e. minimizing the average residual squared error of the cluster members to their representative sequence, coined exemplar). However, rather than selecting a set of exemplars and later iteratively refining the partition, the Affinity Propagation technique initially considers all data points as potential exemplars. Then messages are conveyed between the data points in accordance with their probability of being a superior exemplar (Frey and Dueck, 2007). According to its authors, this approach facilitates a more comprehensive examination of the data space and yields a more robust clustering outcome than alternative partitioning algorithms.

Density-based clustering groups data by finding areas of high density. Areas of low density are labelled as noise in the resulting typology. Two algorithms are considered: DBSCAN and HDBSCAN. DBSCAN requires clusters to have the same density. Two input parameters are needed. The radius around a given point where the density is computed and the minimal number of points to find within the radius to consider the area as dense (Schubert et al., 2017; Ester et al., 1996). HDBSCAN is an adaptation of DBSCAN that produces a hierarchy and allows the clusters to have different densities, it requires only the minimal number of points parameter (Campello et al., 2015). Liao et al. (2022) highlight two potential strengths of these two algorithms for SA, although we found no application of these algorithms in SA. First, by identifying unclassifiable sequences, they might better distinguish anecdotal cases from core ones. Second, these algorithms might be applied to large databases, while most of the previously presented ones are more limited.

Finally, **consensus clustering** aims to increase the robustness of the clustering results, by first generating several clusterings before finding a consensus partition. A variety of algorithms have been proposed for this purpose, differing in the generation of prior clusterings and in the approach to consensus formation. Monti et al. (2003) proposed to generate alternate clustering using bootstraps to improve the robustness of the results. Following Hornik and Böhm (2023), we relied on Bayesian resampling followed by a weighted cluster analysis using Ward (unsquared), PAM, UPGMA or a combination of PAM and UPGMA. The number of clusters to be obtained from bootstrapped clusterings is set to produce a known number of clusters. This ensures the comparison with non-consensus CA. In the second step, the consensus function can deviate from the expected number of clusters but only to a smaller number of clusters. As proposed by

Monti et al. (2003) this allows the algorithm to deviate from a suboptimal known number of clusters. It is anticipated that Ward, UPGMA and PAM-based consensus clusterings will be more stable than their non-consensus variants. The combination of both UPGMA and PAM-based clustering might benefit from the strengths of each method, including the ability to identify atypical, hybrid, and common types. In this case, the objective of consensus clustering is to achieve greater flexibility in the clustering process (Hennig et al., 2015).

4.2 Results

We now turn to the presentation of the simulation results regarding CA performances. To do so, we assume that the number of clusters is known beforehand. However, in real-world applications, the number of groups is typically unknown, making its selection a significant challenge. We address this in Section 5 by comparing the effectiveness of cluster quality indices in determining the ideal number of groups and in identifying the most suitable algorithm for each scenario.

CA results are presented in three parts. First, we report the overall findings across all simulations concerning common types. Second, results concerning atypical and hybrid types are presented in a dedicated subsection. Third, an evaluation of each algorithm and —if applicable— its parametrization is provided. We formulate guidelines on which CA to use according to each research aim and data characteristic in Section 6.

Results are reported as the proportion of simulations in which each algorithm ranked among the best performers according to the Adjusted Rand Index (ARI) between the estimated clustering and the clustering known by design. In each experiment, all CAs with an ARI close to the highest (within a 0.01 deviation) are counted as top performers. This ensures that well-performing algorithms are

included even if not strictly the best. To ease the reading of the plots, poorly performing algorithms were discarded; this includes single linkage, β-flexible UPGMA, β-flexible WPGMA with positive β values, DBSCAN, and HDBSCAN (see Figure A.2 in the Appendix). Additionally, clusterings including a cluster constituted by fewer than five observations are discarded as well.

Figure 1 presents the summarized results from all experiments, with model-based simulations displayed in the left panel, probabilistic empirical simulations in the central one and observed empirical simulations in the left one. Each panel represents a total of 1,200 experiments.

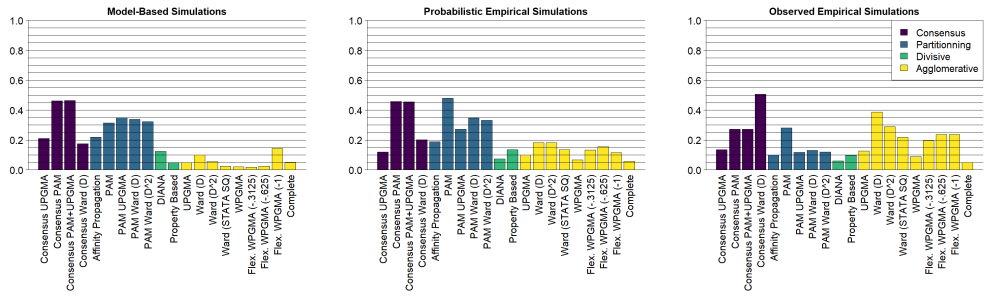

Figure 1: Proportion of simulations where each CA features among the best-performing ones.
Reading note: Bars display the proportion of simulations in which a given algorithm is associated with the highest ARI (a deviation of 0.01 is allowed). Colours represent the algorithm families.

These overall results reveal that there is no single best-performing algorithm across all scenarios. In the model-based simulations, consensus partitioning algorithms are the best performers, followed by the four PAM. Initializing PAM does not bring added value here. In probabilistic empirical simulations, PAM and consensus partitioning algorithms perform best, followed by the other PAM parameterizations. In observed empirical simulations, Consensus Ward and Ward (D) feature among the best, followed by consensus partitioning algorithms, other

Ward variants and β-flexible WPGMA.

With the objective of deriving guidelines on the use of CA, we now examine how CA performances varies according to the previously discussed data characteristics and research aims, beginning with the level of detail. As smaller datasets are less frequent and the results for small and large cluster sizes are similar, the results presented here focus on simulations with 500 observations per cluster. A dedicated subsection addresses the challenges associated with smaller cluster sizes.

4.2.1 Level of Detail

As a recall, the level of detail refers to the underlying number of groups in the *known* typology (see Figure A.3 in Appendix).

In model-based simulations, consensus (PAM and PAM+UPGMA) and PAM variants consistently rank among the top performers across all detail levels. Consensus UPGMA performs well for clusterings in nine groups. Probabilistic empirical simulations indicate good performances of consensus (PAM and PAM + UPGMA) and PAM (original) for all levels of detail. The other PAM variants do not outperform its original formulation but can be considered for clustering in more than 6 groups.

The picture is more varied for observed empirical simulations. For scenarios with three groups, consensus Ward and agglomerative algorithms perform well. With four groups, consensus (PAM and PAM+UPGMA) and PAM (original) are the best performers. For partitions in 6 or nine groups, consensus Ward outperforms, by far, the other algorithms.

4.2.2 Level of Clustering Structure

As discussed in Section 2, the level of clustering structure refers to the extent to which trajectories are grouped into types. Figure A.4 in the Appendix presents the best-performing algorithms for each experiment categorized by the level of clustering structure.

In model-based simulations, consensus Ward and WPGMA (-1) show the best performance for strong clustering structures. For moderated ones, consensus (PAM and PAM+UPGMA) and PAM are the best options. As expected, consensus clustering improves the quality of these algorithms at lower levels of clustering structure. Consensus UPGMA outperforms the other CAs for the weak-clustering case. In probabilistic empirical simulations, consensus (PAM and PAM+UPGMA) and PAM show the best performances for strong and moderate cases, but fall behind PAM and PAM initialized with Ward for the weak clustering case. In the observed empirical simulations, many algorithms perform well for the strong cases, indicating that the choice of algorithms matters less in this case. While for moderate and weak structures, consensus Ward and Ward (D) are the best CAs.

4.2.3 Sample Size

All the simulations were conducted with a sample size of either 50 or 500 sequences, resulting in different overall sample sizes according to the number of groups. The results of each algorithm are quite similar according to the cluster size (see Figure A.5 in Appendix). Consequently, this typology dimension does not influence CA choice, except regarding the computational burden associated with greater sample sizes.

4.2.4 Hybrid and Atypical Types

As a reminder, we identified three definitions of what constitutes a type in Section 2: *common*, *atypical*, and *hybrid* types. The goal is to evaluate the CA ability to identify these two latter types. We thus analyse the results concerning this typology dimension using a specific logic. As a recall, we rely on the Jaccard index, indicating the recovery of a type of interest.

Results indicate that identifying atypical or hybrid types is a major struggle for evaluated CAs. These uncommon types are rarely identified. Simulations with strong clustering structures are the exception where atypical or hybrid types might be successfully identified (see Figures A.6 and A.7 in the Appendix).

4.2.5 Algorithms

Several implementations and parameterizations of the CAs were included in the simulations. In this section, we summarize our findings and explain why some algorithms were discarded from our results.

PAM featured among the best CAs in model-based and probabilistic empirical simulations, but fell behind Ward in the observed empirical ones. This could be explained by the differences in the simulation design. PAM relies on medoids, and therefore requires good *observed* medoid candidates. This is more likely to occur in model-based simulations, where many small variations of the sequences can be generated. In the empirical ones, such fine-tuning might not exist, as some sequences are too unlikely, such as changing school in the middle of the year. Ward does not have such a requirement, as it relies on *gravity centres* (Batagelj, 1988). This discussion highlights that PAM might perform better when many variations of the trajectories are present. This is in line with the discussion of

Hennig (2022) on the differences and similarities between PAM and Ward.

Several PAM initializations were considered. Its original version regularly overshadowed the initialized ones. Using UPGMA stood out only in model-based simulations, while using Ward provided no clear additional value. This may be explained by the fact that PAM and Ward aim to minimize within-cluster error variance in two incompatible ways. The best initialization therefore varied between simulations. Consensus PAM was consistently close to the best variant, and can, therefore, be used as a *safe* choice.

Ward stood out as one of the best in the observed empirical simulations and should, therefore, be regularly considered. Milligan and Cooper (1987) already pointed out Ward as a versatile CA providing good results in various situations. The unsquared-distance version almost always provided the best results. The squared version often provides similar performances but lags behind in some situations. The Stata SQ version, based only on unique sequences, almost always provided poorer results. The consensus version of Ward (D) almost always outperformed the other Ward variants.

β-flexible WPGMA provided interesting results in the three sets of simulations, especially for strong clustering structure. We found better results with negative β values, between -0.625 and -1, with lower β increasing the performance for higher number of groups. β should, therefore, be tuned according to the underlying number of groups. This is expected, as Belbin et al. (2010) pointed out that negative values tend to space dilatation, avoiding clusters being merged when getting closer to each other —which occurs when the number of groups increases.

Consensus Algorithms often improved the results of the original version of each algorithm. This confirms the adequacy of Monti et al. (2003) approach. With PAM, it consistently equalled or superseded the best initialization method. Consensus UPGMA and Ward were always better than their non-consensus version. However, consensus UPGMA is still found to fail in situations where UPGMA fails as well. We did not observe improvement by combining different CAs with consensus. Consensus PAM+UPGMA provided almost the same results as consensus PAM.

Other Algorithms

Among the other algorithms, some provided unstable results, while others were only relevant in specific cases. While they might be considered in those situations, there were often alternatives performing better in the other simulations.

Affinity propagation proved to be interesting with small sample size, but is computationally intensive and often fails to converge with larger sample sizes. It still has the interesting property to be suited to the clustering of non-metric dissimilarities (Frey and Dueck, 2007).

DIANA has been found to perform relatively well for higher number of groups in the model-based simulations, but it has only small added value on top of the previously discussed algorithms. This relative performance at higher number of groups was not expected, since divisive CA optimizes the clustering at the lower levels of the hierarchy. Following this logic, divisive CA should perform best for solutions in few groups due to the increased dependence on previous merges —and thus on errors in clustering estimation— as the number of clusters increases.

Property-based clustering leads to explicit clustering rules based on sequence properties, which might be of interest where features of the sequences are at the

core of the research question (Studer, 2018). However, it provided globally poor results and especially at higher levels of detail.

UPGMA logic is designed to identify clusters of various sizes, which can be useful when looking for uneven-sized clusters. However, it regularly produces clusters containing very few observations. This strongly reduces the relevance of UPGMA typologies. Poor performance is observed with large numbers of groups or weakly structured data. These two conclusions are in line with the one of Milligan and Cooper (1987) methodological review.

Among the algorithms tested in the present study, some were found to be irrelevant to SA. They can be consequently avoided. Single linkage proved too sensitive to chaining, leading to the creation of very large and heterogeneous clusters. This behaviour was expected, since several reviews already pointed it out (Hennig, 2022; Milligan and Cooper, 1987). β-flexible WPGMA clustering with positive β showed a similar behaviour. β-flexible UPGMA consistently provided poor performances, except when $\beta = 0$, which is equivalent to standard UPGMA.

DBSCAN produced poor results as well, but for different reasons. By design, DBSCAN assumes that all clusters have the same density, leading to either grouping all the sequences into a single noise cluster, or to generate a very high number of groups. In both cases the resulting typology is useless.

HDBSCAN, a DBSCAN extension, is designed to find clusters of various densities. However, it does not improve the results. Moreover, in these two algorithms, the number of clusters is directly estimated and cannot be specified, which deteriorates the ability of the researchers to choose a meaningful typology among several candidates. Its expected benefits were to identify unclassifiable observations and to be applied to large databases (Liao et al., 2022). The case of large databases was recently addressed by Studer (2024).

The study also tested a hierarchical extension of HDBSCAN, which allows the user to choose the number of clusters to keep within a hierarchy. This approach provided slightly better results, but still far below the performances of the above presented algorithms.

5 Cluster Quality Indices

In practice, the true clustering structure is unknown. In consequence, the quality of the obtained clusterings has to be evaluated for two reasons (Studer, 2021). First, it provides an evaluation of the underlying strength of the clustering structure, which can then be used to tone the interpretation accordingly. Second, it can guide the identification of the adequate number of groups and the most suitable CA. To do so, CQIs aim to measure the statistical quality of a partition. Several of them have been proposed (see Studer, 2013, for a review).

In this section, we aim to formulate recommendations for this second use according to the research aim and data characteristics. We start by presenting the evaluated CQIs. We then discuss the results regarding CA selection, before moving to their use to decide on the number of clusters. We then evaluate the added value of standardized CQIs using parametric bootstraps (Studer, 2021). Findings are summarized in Section 5.4. Guidelines are provided in Section 6.

5.1 Reviewed CQIs

CQIs generally balance intragroup homogeneity and / or between-group separations. However, they differ in the measurement of each aspect and their relative weights. In the following lines, we present the CQIs under review.

A first set of CQIs the clustering's ability to reproduce the information of the

original dissimilarity matrix. This is generally achieved by computing a correlation measure between the original dissimilarities and a binary matrix representing the clustering. The "Point Biserial Correlation" (PBC) uses the Pearson correlation (Milligan and Cooper, 1985; Hennig and Liao, 2010), the "Hubert's Gamma" (HG) relies on the Gamma coefficient (Hubert and Arabie, 1985), while the "Hubert Somers' D" (HGSD) uses the Somers' D association, which accounts for ties in dissimilarities. In their simulation study, based on multivariate normally distributed data, Milligan and Cooper (1985) showed that PBC index tends to underestimate the number of clusters, while the HG index overestimates it.

Second, the "Hubert's C" (HC) index conceptually aims to measure a gap between an ideal partition and the obtained one for a given distance matrix and number of groups. Doing so, the HC is primarily focused on intragroup homogeneity. While it provided good results in Milligan and Cooper (1985), the authors recommended not taking into account small variations with the increasing number of groups, as it tends to overestimate the number of groups otherwise.

Third, the "Average Silhouette Width" (ASW) and "Calinski-Harabasz" (CH) index explicitly relate between-group separation with intragroup homogeneity. While the ASW index measures it at the observation level, the CH index can be seen as an extension of the F statistic in ANOVA (Studer et al., 2011) relating within and between sums of squares. CH can be computed by using the raw distance matrix or the squared distances, denoted "CHsq". The ASW index is one of the most used CQI, but tends to favour two or three group solutions (Hennig and Liao, 2013). The CH index was the best performing index in Milligan and Cooper (1985), but for normally distributed data.

Fourth, Studer (2013) proposed to look at agreement between different CQIs to select the appropriate number of groups. Here, we operationalize this idea

using majority voting, which is a common data mining technique. We do it in two steps. We start by ranking the different solutions according to each CQI, before summing up these ranks over all the considered CQIs. This overall rank is then used as a composite CQI measure. We consider four combinations of CQIs. MV^{all} aggregates all available CQIs. $MV^{HcCHsqHg}$ is computed on the basis of one CQIs per logic presented above, namely: HC, CHsq and HG. It aims to take advantage of each logic to lead to more stable results. MV^{HcCHsq} is based on HC and CHsq. These two CQIs are based on two opposing objectives: maximizing cluster homogeneity and balancing separation and homogeneity. Finally, MV^{HcHg} is constructed using HC and HG.

Finally, Studer (2021) proposed to standardize the CQI using a parametric bootstrap procedure specifically designed for SA. The procedure compares the quality of the obtained clustering with the quality of clustering of non-clustered but similar data. It is computationally intensive and can, therefore, only be used with some clustering algorithms. The method is mainly designed to provide a test-like interpretation of an underlying clustering structure of the data. However, we include it to test its extension to the choice of the underlying number of groups.

5.2 Results on Cluster Algorithm Selection

This section investigates the use of CQIs to identify the best CA in each experiment. As for the CA evaluation, we discarded the results of poor performing CAs and selected all experiments on common types of 500 observations. In each simulation, only the clusterings where the number of requested clusters is equal to the number of known clusters are considered. For each CQI, its optimal value per experiment is pinpointed and associated with the corresponding clustering. Keeping the number of groups constant compels CQIs to choose clusterings only

according to their algorithms and leads to the selection of clusterings obtained with different CAs for the same experiment. Finally, the number of times each CQI leads to the best ARI in each experiment is counted and reported, regardless of the CA it is obtained with. To avoid penalizing CQIs close to the best, we allowed a deviation of 0.01 from the best ARI per experiment.

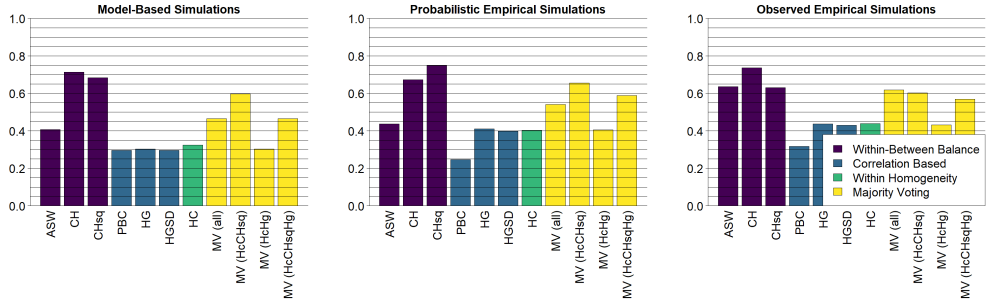

Figure 2: Proportion of experiments where each CQI selects the best performing algorithm among all simulations.

Figure 2 illustrates the results, model-based simulations on the left panel, probabilistic empirical ones in the middle and observed empirical simulations on the right panel. CH, CHsq, $\text{MV}^{\text{HcCHsq}}$ and $\text{MV}^{\text{HcCHsqHg}}$ feature at first glance among the best CQIs.

Interesting variation exists according to the expected level of detail (see Figure A.8 in the appendix). As expected, the ASW shows much worse results for larger number of groups. In this case, CH and CHsq are safe choices according to all three simulation sets. MVs do not outperform the other CQIs in any configuration, but appear to be quite stable. However, their gains in stability only partially compensate for these mediocre results. Furthermore, they fail in some of the simulations even when some of the CQIs they rely on perform well. Results on the level of clustering structure are contrasted (see Figure A.9 in the Appendix). Model-based simulations show good performance of CH and

CHsq when the structure is strong or moderate. When the structure turns up to be weak, correlation-based CQIs, HC and MV^{HcHg} are the best performers. Probabilistic and observed empirical simulations do not discriminate CQIs when the structure is strong. For moderate or weak clustering structures, CH, CHsq, MV^{HcCHsq} and $MV^{HcCHsqHg}$ are the best options.

Regarding the other simulation dimensions, cluster size did not discriminate CQIs and discussing uncommon types identification does not make sense, since evaluated CAs are not found to be suited for this task.

In a nutshell, CH and CHsq are the best CQI for selecting among CAs. However, they might lead to poor results when the clustering structure is weak. In this latter case, HC provides better results.

5.3 Number of Groups Selection

We now turn to the use of CQIs to establish the optimal number of groups to keep. We do so by relying on the same set of simulations as before. However, the number of groups is not fixed anymore, but estimated through the CQIs. More precisely, all the solutions between 2 and 12 groups are kept as well as the results from all CAs discussed in Section 4.2. We then let each CQI "decide" on the best number of groups to keep separately for each CA. Since results differ the most according to the expected level of detail, we begin by discussing this aspect, before briefly addressing the other ones.

Figure 3 presents the proportion of the experiments and algorithms in which each CQI selected a clustering close to the highest ARI with the known cluster. Again, a variation of 0.01 is allowed. In this Figure, panels are separated by number of a priori known clusters. In model-based simulations, two groups of CQIs can be identified. The first one includes CQIs based on the balance between

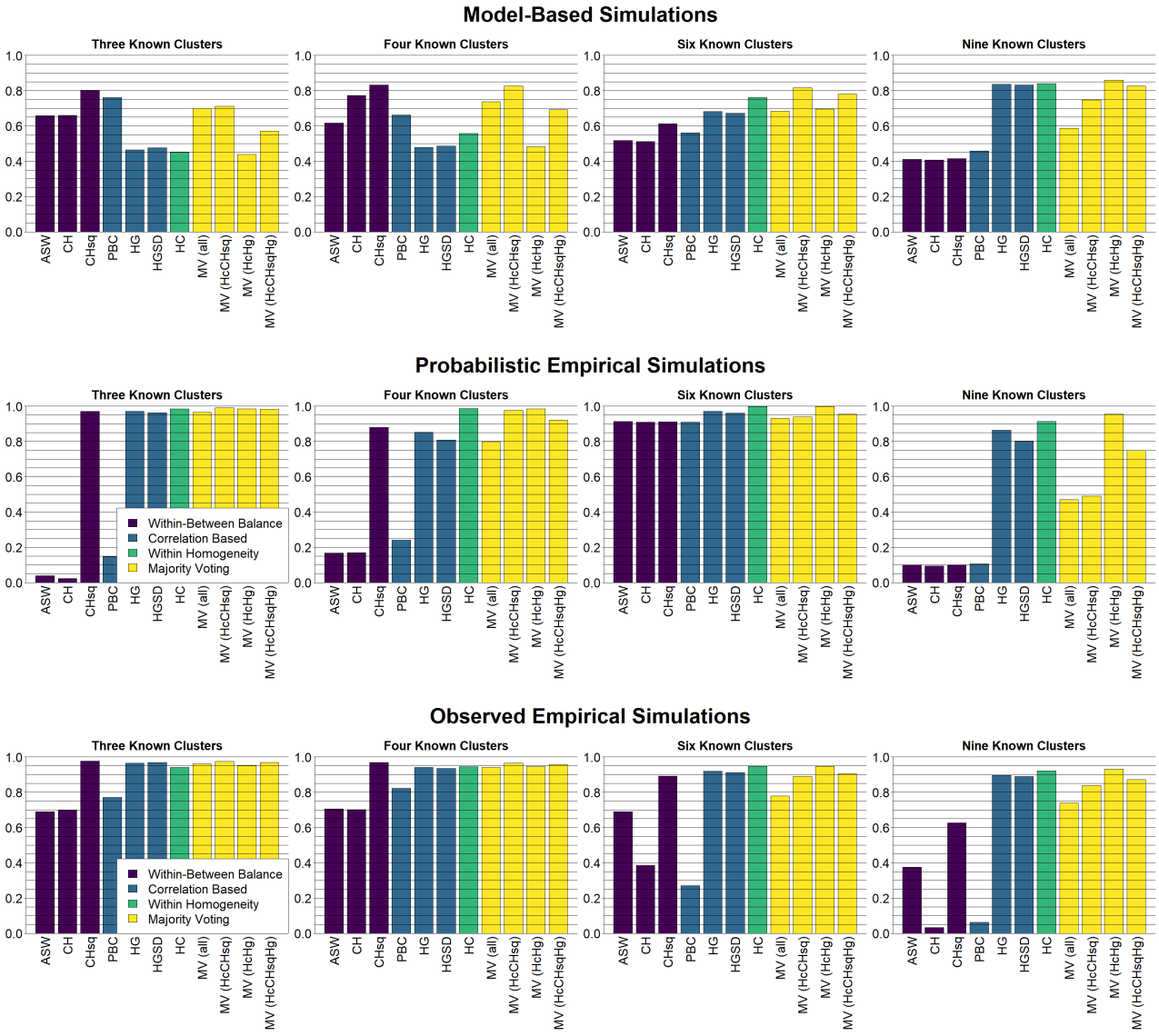

Figure 3: Proportion of experiments where each CQI selects the optimal number of groups among all simulations, separated by levels of detail

cluster homogeneity and separation and PBC, which are the best at identifying solutions in fewer than 6 groups. The second group is made up of HC, HG and HGSD, which perform best for higher levels of detail. Majority voting performs well in any configuration. The results of observed empirical simulations are generally very good, with less variation between CQIs. Correlation-based CQIs and HC perform well at every level of detail and CHsq fails only for nine groups. ASW, CH and PBC underperform in all settings, but particularly for the six and nine-group solutions. Probabilistic empirical simulations provide a similar picture to observed ones but ASW, CH and PBC show poor results also at lower levels of detail.

A complementary analysis investigates the difference between the number of groups advocated for by each CQI and the number of groups expected by design. Findings indicate a tendency to underestimate the number of clusters. All CQIs exhibit this behaviour, but especially the balance-based ones. In model-based simulations, correlation-based CQIs tend to overestimate this number (see Figure A.10 in the Appendix). Two solutions are available to mitigate this. First, one can ignore the two-cluster solutions (see Figures A.12 in the Appendix). The second approach is to consider also the clustering associated with the second (or even the third) best CQI. This is particularly beneficial for HG, HGSD and HC (see Figure A.11). The good performance of ASW, CH and CHsq for a lower number of groups is expected, as they take into account the between-cluster separation, which automatically decreases for a higher number of groups. Finally, the relative performance of CHsq for higher numbers of groups was not expected and makes it particularly versatile. The performance of HC at higher levels of detail was also expected, as it aims to minimize within-cluster homogeneity. Indeed, the within-cluster homogeneity mechanically tends to increase with the

number of groups. MVs are associated with average results. Although the results provided by MVs often lag behind other CQIs, they appear to be more stable than those provided by individual CQIs.

Regarding the other dimensions, sample size literally makes no difference. The level of clustering structure further highlights three minor points. First, HC is less affected by the clustering structure than the other CQIs. Second, correlation-based CQIs perform poorly in model-based simulations at low levels of clustering structure. They tend to overestimate the number of clusters in such cases (see Figure A.13 in the appendix). Finally, MVs were associated with particularly good results for weakly structured data in model-based simulations, but not in the empirical ones.

5.3.1 Standardized CQIs

Studer (2021) proposed a parametric bootstrap procedure providing standardized CQIs values. We now discuss its relevance in selecting the number of groups. This standardization is conducted according to an underlying model, which defines the structure searched for. Due to its computational burden, the standardized CQIs were only computed for three CAs: β-flexible WPGMA (with β = -1), PAM and Ward D. In these simulations, we tested three null models focusing on differences either in sequencings, durations or combining both trajectory aspects.

Figure 4 presents, as for the other CQIs, the proportion of experiments in which a standardized CQI selects a clustering close to the underlying known clustering. Darker coloured bars represent standardized CQIs, while the lighter ones show the results of the corresponding raw CQIs.

Firstly, the bootstrap procedure seems to equalize the performances of the CQIs, and therefore achieves its goal. For instance, in model-based simulations,

Figure 4: Share of times each CQI selects the optimal clustering according to three null models

the standardized version of correlation-based CQIs is improved for low number of clusters, i.e. when they were underperforming. Conversely, they stayed constant for nine groups, where they already featured among the best-performing ones (see Figure A.14 in the Appendix).

Second, the actual performance depends on the adequacy between the true and the parametric bootstrap models. In the model-based simulations, the sequencing null model is in line with the data-generating model, leading to good results of bootstrapped CQIs. This is not the case for the observed empirical simulations, where the duration model is slightly more suited.

Summarizing, the procedure can be recommended when there is strong correspondence, but should probably be avoided otherwise. These conclusions are, however, only relevant when it is used to select the number of groups. The procedure also provides further information, such as a test-like method and interpretation threshold to estimate the strength of the underlying clustering structure of the data, which are not evaluated here and still provide useful information.

5.4 Conclusion

CQIs can serve two purposes: evaluating the CAs or deciding on the number of groups to keep in a clustering. This latter aim is the most common one (Hennig et al., 2015). Results are summarized in the following lines. We propose guidelines for the use of CQIs in Section 6.

ASW provided rather poor results for selecting the best CA among several candidates or to select the appropriate number of clusters. It works best when the aim is to find a low number of clusters. Because ASW has a strong tendency to underestimate the number of clusters, it regularly fails at higher levels of detail. Ignoring the two-cluster solutions improved the results, as discussed by Hennig

and Liao (2013).

CH is effective at identifying the best CA, but shows mixed results when confronted with weakly structured data. Its performance to select the optimal number of groups follows ASW's trend.

CHsq has proved to be a versatile and generally good CQI. It provides similar performance to CH to select the best CA, and is the most relevant among the balance-based CQIs to identify the appropriate number of groups. However, like the other balance-based CQIs, it is better at identifying solutions in lower numbers of groups and might fail at higher levels of detail.

PBC is based on the correlation between the partition and the original dissimilarity matrix. It shows similar results to balance-based CQIs as it is effective at low levels of detail but regularly fails when confronted with weakly structured data or highly detailed typologies.

HG and HGSD are also correlation-based. They diverge on the measure of the correlation. Their relevance for CA selection is limited. However, these two indices feature among the best performers to identify highly detailed typologies. They might, however, fail to identify more parsimonious typologies, as they tend to overestimate the optimal number of clusters.

HC explicitly focuses on cluster homogeneity and shows diverging behaviour from the previous CQIs. It shows particularly stable results. It is a good CQI for CA selection, albeit not as good as CH and CHsq, and it performs very well in identifying the appropriate number of clusters at higher levels of detail, without failing at lower levels.

Majority Votings were found to be particularly stable. However, they are top performers only when the CQIs they are based on are already among the best ones. This limits the usefulness of this approach.

Standardized CQIs using parametric bootstraps was found to improve the results when some of their non-standardized version is underperforming. For instance, correlation-based CQIs were improved when looking for few groups and balanced CQIs when looking for more detailed typologies. However, the improvement highly depends on the suitability of the underlying null model used for standardization.

6 Guidelines

The following section is devoted to the recommendations regarding first the CA choice and second the CQI one. Regarding CAs, there is no absolute winner. However, Consensus PAM and Consensus Ward are the most versatile and clearly the most efficient to identify common types, but close to the top in many other simulations. The additional computing time remaining limited, one can, therefore, recommend them as a default choice. Consensus Ward clustering provided good performances in the empirical simulations but not in the model-based ones, and consensus PAM appeared as one of the best in model-based but lagged behind in empirical ones. In doubt, one could use both algorithms and compare the results. These conclusions also apply to PAM and Ward, although their performances are increased by their consensus versions. Additionally, using consensus PAM avoids the need to decide on the appropriate initialization process.

Other CAs can be considered to further explore the data. β-flexible WPGMA should be considered for highly detailed typologies —especially with lower β values— or when a strong clustering structure is expected. Finally, property-based clustering can be considered to take advantage of the explicit clustering criteria it provides, although it rarely ranks among the best CAs.

The choice among the clustering algorithms can, therefore, be grounded in the aims of the study by clarifying the research question and the expectations related to the structure of the data. However, the choice can also be made by using the four above-mentioned algorithms and using cluster quality indices.

We now turn to the guidelines concerning CQIs. We first discuss their use to choose among CAs and in second place on their ability to identify the optimal number of groups.

CH and CHsq featured among the best CQI for selecting between different CAs and can serve as default choices. However, they might lead to poor results in some situations, such as in cases of weak clustering structure. In such a context, we recommend using the HC index.

To select the number of groups, we recommend three CQIs: CHsq, HG and HC. Their performance differs according to the required detail level. CHsq provides better performances for parsimonious typologies, but might fail for detailed ones. Similar results are reported in Sugar and James (2003) and Milligan and Cooper (1985), even if they used a very different simulation framework. Conversely, the HG and HC indexes provided the best performances when interested in detailed typologies. but might fail for parsimonious ones. This is consistent with Milligan and Cooper (1985) findings. However, HC and HG performances at lower levels of detail exceed CHsq's at higher levels. HC is a particularly versatile index and can be considered in combination with other more specific CQIs in different situations.

Using MV strategies might limit the drawbacks of individual CQIs, but it seldom leads to the best results. In consequence, we recommend using this approach as an exploratory tool before relying on specific CQIs for the final decision. The added value of standardized CQIs to select the number of groups strongly depends

on the appropriateness of the chosen null model. Consequently, we recommend using it to choose the number of groups when there is a strong correspondence between the chosen null model and the data structure. This consideration does not affect the test-like feature of the procedure as advocated for in Studer (2021).

Evaluated CQIs are also found to regularly underestimate the optimal number of groups. This is more detrimental than overestimating the optimal number of groups because the clustering might overlook information (Sugar and James, 2003). Two strategies allow mitigating this bias. First, one can exclude the two-clusters solutions as advocated by Hennig and Liao (2013). Second, we strongly recommend considering as well the second- or third-best clusterings indicated by the CQI value. In definitive, CQIs should be participating in the clustering choice rather than imposing it.

7 Conclusion

This study reviewed and evaluated several cluster algorithms and cluster quality indices (CQI) to create a typology of trajectories in SA. We further introduced and discussed the use of consensus clustering in SA. This method is found to regularly enhance the performance of the CA it is based on.

In line with Hennig (2022) constructivist approach, we proposed a simulation framework grounded in the different kinds of uses and aims of the resulting typology in life-course research. Furthermore, common variations in data characteristics were incorporated into the framework.

The present study does not proclaim any absolute winners, but rather highlights that the choice of a CA or a CQI should be grounded in the research aim and data characteristics. As a result, SA users are invited to further specify

their research questions and investigate some of their data characteristics. Our results highlight that the choice of a CA should be made according to the expected strength of the clustering structure. The choice of a CQI should rather be grounded in the expected level of detail of the resulting typology.

The proposed guidelines on CAs do not contradict current practices in SA as PAM and Ward appear to be the best-performing algorithms (Liao et al., 2022). Nonetheless our results point out that using consensus clustering improves the performance and stability of these two algorithms. Conversely, findings on CQIs shed new light on their use in SA. The ASW, which is a widely used CQI, appears to be a rather poor indicator compared to CHsq, HG and HC.

The originality of this paper is that we evaluate CAs applied to categorical data, which is unusual in the cluster analysis literature. For this reason, the results might differ with other studies mostly based on numerical data. Furthermore, dimensionality tends to also be higher in our case. Life-course trajectories often exhibit a rather weak clustering structure. All these differences further explain the need, within the SA community, for the presented review to make informed choices when creating typologies. However, our results still converge with previous studies in many cases.

Finally, the scope of this paper is limited to crisp CAs and CQIs. A systematic evaluation of robust clustering techniques, including fuzzy and noise clustering, would be an evident follow-up of the present work and would find strong interest in the research community, notably for the identification of atypical or hybrid types.

8 Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship and publication of this article.

9 Funding

The authors gracefully acknowledge the grant support of the Swiss National Science Foundation (project "Strengthening Sequence Analysis," grant numbers: 10001A_204740).

10 Declaration of Generative AI Use

During the preparation of this manuscript, the authors used DeepL translator in order to correct the text and reformulate sentences. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

Balcan, Maria-Florina, Yingyu Liang, and Pramod Gupta. 2014. "Robust Hierarchical Clustering." *Journal of Machine Learning Research* 15:4011–4051.

Batagelj, Vladimir. 1988. "Generalized Ward and Related Clustering Problems." *Classification and Related Methods of Data Analysis* .

Belbin, Lee, Daniel P. Faith, and Glenn W. Milligan. 2010. "A Comparison of Two Approaches to Beta-Flexible Clustering." *Multivariate Behavioral Research* .

Bernardi, Laura, Johannes Huinink, and Richard A. Settersten. 2019. "The Life Course Cube: A Tool for Studying Lives." *Advances in Life Course Research* 41:100258.

Billari, Francesco C., Johannes Fürnkranz, and Alexia Prskawetz. 2006. "Timing, Sequencing, and Quantum of Life Course Events: A Machine Learning Approach." *European Journal of Population / Revue européenne de Démographie* 22:37–65.

Bodenhofer, Ulrich, Johannes Palme, Chrats Melkonian, Andreas Kothmeier, and Nikola Kostic. 2024. "Apcluster: Affinity Propagation Clustering."

Bras, Hilde, Aart C. Liefbroer, and Cees H. Elzinga. 2010. "Standardization of Pathways to Adulthood? An Analysis of Dutch Cohorts Born between 1850 and 1900." *Demography* 47:1013–1034.

Brüderl, Josef, Fabian Kratz, and Gerrit Bauer. 2019. "Life Course Research with Panel Data: An Analysis of the Reproduction of Social Inequality." *Advances in Life Course Research* 41:100247.

Brzinsky-Fay, Christian. 2007. "Lost in Transition? Labour Market Entry Sequences of School Leavers in Europe." *European Sociological Review* 23:409–422.

Campello, Ricardo J. G. B., Davoud Moulavi, Arthur Zimek, and Jörg Sander. 2015. "Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection." *ACM Transactions on Knowledge Discovery from Data* 10:5:1–5:51.

Core, R and Team. 2020. *A Language and Environment for Statistical Computing, R Foundation for Statistical Computing; 2013.*

Dlouhy, Katja and Torsten Biemann. 2015. "Optimal Matching Analysis in Career Research: A Review and Some Best-Practice Recommendations." *Journal of Vocational Behavior* 90:163–173.

Elder, Glen H., Monica Kirkpatrick Johnson, and Robert Crosnoe. 2003. "The Emergence and Development of Life Course Theory." In *Handbook of the Life Course*, edited by Jeylan T. Mortimer and Michael J. Shanahan, Handbooks of Sociology and Social Research, pp. 3–19. Boston, MA: Springer US.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 226–231, Portland, Oregon. AAAI Press.

Florek, K., J. Łukaszewicz, J. Perkal, Hugo Steinhaus, and S. Zubrzycki. 1951. "Sur La Liaison et La Division Des Points d'un Ensemble Fini." *Colloquium Mathematicum* 2:282–285.

Fraley, Chris, Adrian E. Raftery, Luca Scrucca, Thomas Brendan Murphy, and Michael Fop. 2024. "Mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation."

Frey, Brendan J. and Delbert Dueck. 2007. "Clustering by Passing Messages Between Data Points." *Science* 315:972–976.

Gabadinho, Alexis, Gilbert Ritschard, Nicolas S. Müller, and Matthias Studer. 2011. "Analyzing and Visualizing State Sequences in R with TraMineR." *Journal of Statistical Software* 40:1–37.

Garnier, Simon, Noam Ross, Bob Rudis, Marco Sciaini, Antônio Pedro Camargo, and Cédric Scherer. 2024. "Viridis: Colorblind-Friendly Color Maps for R."

Gauthier, Jacques-Antoine, Felix Bühlmann, and Philippe Blanchard. 2014. "Introduction: Sequence Analysis in 2014." In *Advances in Sequence Analysis: Theory, Method, Applications*, edited by Philippe Blanchard, Felix Bühlmann, and Jacques-Antoine Gauthier, Life Course Research and Social Policies, pp. 1–17. Cham: Springer International Publishing.

Hahsler, Michael, Matthew Piekenbrock, Sunil Arya, and David Mount. 2023. "Dbscan: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Related Algorithms."

Helske, Jouni and Satu Helske. 2023. "seqHMM: Mixture Hidden Markov Models for Social Sequence Data and Other Multivariate, Multichannel Categorical Time Series."

Hennig, Christian. 2007. "Cluster-Wise Assessment of Cluster Stability." *Computational Statistics & Data Analysis* 52:258–271.

Hennig, Christian. 2015. "What Are the True Clusters?" *arXiv:1502.02555 [stat]* .

Hennig, Christian. 2022. "An Empirical Comparison and Characterisation of Nine Popular Clustering Methods." *Advances in Data Analysis and Classification* 16:201–229.

Hennig, C. and T. Liao. 2010. "Comparing Latent Class and Dissimilarity Based Clustering for Mixed Type Variables with Application to Social Stratification."

Hennig, Christian and Tim F. Liao. 2013. "How to Find an Appropriate Clustering for Mixed-Type Variables with Application to Socio-Economic Stratification." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 62:309–369.

Hennig, Christian, Marina Meila, Fionn Murtagh, and Roberto Rocci (eds.). 2015. *Handbook of Cluster Analysis*. New York: Chapman and Hall/CRC.

Hornik, Kurt and Walter Böhm. 2023. "Clue: Cluster Ensembles."

Hubert, Lawrence and Phipps Arabie. 1985. "Comparing Partitions." *Journal of Classification* 2:193–218.

Jaccard, Paul. 1901. "Distribution de La Flore Alpine Dans Le Bassin Des Dranses et Dans Quelques Regions Voisines." *Bulletin de Société Vaudoise Sciences Naturelles* 37:241–272.

Kaufman, Leonard and Peter J. Rousseeuw (eds.). 1990. *Finding Groups in Data*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Kohler, Ulrich, Magdalena Luniak, and Christian Brzinsky-Fay. 2006. "SQ: Stata Module for Sequence Analysis." .

Lebart, L., Marie Piron, and A. Morineau. 2006. "Statistique Exploratoire Multidimensionnelle : Visualisations et Inférences En Fouille de Données."

Lelièvre, Eva. 2019. "Setting Research Priorities in Developing Approaches for the Life Course." *Advances in Life Course Research* 41:100275.

Lesnard, Laurent. 2006. "Optimal Matching and Social Sciences." Working Paper 2006-01, Center for Research in Economics and Statistics.

Liao, Tim F., Danilo Bolano, Christian Brzinsky-Fay, Benjamin Cornwell, Anette Eva Fasang, Satu Helske, Raffaella Piccarreta, Marcel Raab, Gilbert Ritschard, Emanuela Struffolino, and Matthias Studer. 2022. "Sequence Analysis: Its Past, Present, and Future." *Social Science Research* 107:102772.

Liefbroer, Aart C. 2019. "Methodological Diversity in Life Course Research: Blessing or Curse?" *Advances in Life Course Research* 41:100276.

Macnaughton-Smith, P., W. T. Williams, M. B. Dale, and L. G. Mockett. 1964. "Dissimilarity Analysis: A New Technique of Hierarchical Sub-division." *Nature* 202:1034–1035.

Martin, Peter, Ingrid Schoon, and Andy Ross. 2008. "Beyond Transitions: Applying Optimal Matching Analysis to Life Course Research." *International Journal of Social Research Methodology* 11:179–199.

Mattijssen, Lucille, Dimitris Pavlopoulos, and Wendy Smits. 2020. "Occupations and the Non-Standard Employment Career: How the Occupational Skill Level and Task Types Influence the Career Outcomes of Non-Standard Employment." *Work, Employment and Society* 34:495–513.

Mayer, Karl Ulrich. 2009. "New Directions in Life Course Research." *Annual Review of Sociology* 35:413–433.

McQuitty, Louis L. 1960. "Hierarchical Linkage Analysis for the Isolation of Types." *Educational and Psychological Measurement* 20:55–67.

McVicar, Duncan and Michael Anyadike-Danes. 2002. "Predicting Successful and Unsuccessful Transitions from School to Work by Using Sequence Methods." *Journal of the Royal Statistical Society Series A: Statistics in Society* 165:317–334.

Milligan, Glenn W. and Martha C. Cooper. 1985. "An Examination of Procedures for Determining the Number of Clusters in a Data Set." *Psychometrika* 50:159–179.

Milligan, Glenn W. and Martha C. Cooper. 1987. "Methodology Review: Clustering Methods." *Applied Psychological Measurement* 11:329–354.

Monti, Stefano, Pablo Tamayo, Jill Mesirov, and Todd Golub. 2003. "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data." *Machine Learning* 52:91–118.

Müllner, Daniel and Google Inc. 2024. "Fastcluster: Fast Hierarchical Clustering Routines for R and 'Python'."

Murtagh, Fionn and Pierre Legendre. 2014. "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?" *Journal of Classification* 31:274–295.

Piccarreta, Raffaella and Francesco C. Billari. 2007. "Clustering Work and Family Trajectories by Using a Divisive Algorithm." *Journal of the Royal Statistical Society Series A: Statistics in Society* 170:1061–1078.

Piccarreta, R. and M. Studer. 2019. "Holistic Analysis of the Life Course: Methodological Challenges and New Perspectives." *Advances in Life Course Research* .

Rand, William M. 1971. "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association* 66:846–850.

Sacchi, Stefan and Thomas Meyer. 2016. "Übergangslösungen Beim Eintritt in

Die Schweizer Berufsbildung: Brückenschlag Oder Sackgasse?" *Swiss Journal of Sociology* 42.

Scharenberg, Katja, Melania Rudin, Barbara Mueller, Thomas Meyer, and Sandra Hupka-Brunner. 2014. *Education Pathways from Compulsory School to Young Adulthood: The First Ten Years. Results of the Swiss Panel Survey TREE, Part I*.

Schubert, Erich, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. 2017. "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN." *ACM Transactions on Database Systems* 42:19:1–19:21.

Shanahan, Michael. 2003. "Pathways to Adulthood in Changing Societies: Variability and Mechanisms in Life Course Perspective." *Annual Review of Sociology* 26:667–692.

Sneath, P. H. A. 1957. "The Application of Computers to Taxonomy." *Microbiology* 17:201–226.

Sokal, Robert Reuven. 1963. *Principles of Numerical Taxonomy*. A Series of Books in Biology. San Francisco: W.H. Freeman.

Sokal, Robert R., Robert R. Sokal, and Charles D. Michener. 1958. "A Statistical Method for Evaluating Systematic Relationships." *The University of Kansas science bulletin* 38:1409–1438.

Struffolino, Emanuela, Matthias Studer, and Anette Eva Fasang. 2016. "Gender, Education, and Family Life Courses in East and West Germany: Insights from New Sequence Analysis Techniques." *Advances in Life Course Research* 29:66–79.

Studer, Matthias. 2013. "WeightedCluster Library Manual: A Practical Guide to Creating Typologies of Trajectories in the Social Sciences with R." *LIVES Working Papers* 2013:1–32.

Studer, Matthias. 2018. "Divisive Property-Based and Fuzzy Clustering for Sequence Analysis." In *Sequence Analysis and Related Approaches: Innovative Methods and Applications*, edited by Gilbert Ritschard and Matthias Studer, Life Course Research and Social Policies, pp. 223–239. Cham: Springer International Publishing.

Studer, Matthias. 2021. "Validating Sequence Analysis Typologies Using Parametric Bootstrap." *Sociological Methodology* 51:290–318.

Studer, Matthias. 2024. "Seqclararange: Sequence Analysis for Large Databases."

Studer, Matthias, Aart Liefbroer, and Jarl Mooyaart. 2018. "Understanding Trends in Family Formation Trajectories: An Application of Competing Trajectories Analysis." *Advances in Life Course Research* 36.

Studer, Matthias and Gilbert Ritschard. 2016. "What Matters in Differences between Life Trajectories: A Comparative Review of Sequence Dissimilarity Measures." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 179:481–511.

Studer, Matthias, Gilbert Ritschard, Alexis Gabadinho, and Nicolas S. Müller. 2011. "Discrepancy Analysis of State Sequences." *Sociological Methods & Research* 40:471–510.

Sugar, Catherine A and Gareth M James. 2003. "Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach." *Journal of the American Statistical Association* 98:750–763.

TREE. 2016. "TRansition de l'École à l'Emploi, Documentation de La Première Cohorte de TREE (TREE1). 2000-2016." Technical report.

Unterlerchner, Leonhard, Matthias Studer, and Andres Gomensoro. 2023. "Back to the Features. Investigating the Relationship Between Educational Pathways and Income Using Sequence Analysis and Feature Extraction and Selection Approach." *Swiss Journal of Sociology* 49:417–446.

Van Mechelen, Iven, Anne-Laure Boulesteix, Rainer Dangl, Nema Dean, Christian Hennig, Friedrich Leisch, Douglas Steinley, and Matthijs J. Warrens. 2023. "A White Paper on Good Research Practices in Benchmarking: The Case of Cluster Analysis." *WIREs Data Mining and Knowledge Discovery* 13:e1511.

Ward, Joe H. 1963. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association* 58:236–244.

Widmer, Eric D. and Gilbert Ritschard. 2009. "The De-Standardization of the Life Course: Are Men and Women Equal?" *Advances in Life Course Research* 14:28–39.

Zimmermann, Barbara and Simon Seiler. 2019. "The Relationship between Educational Pathways and Occupational Outcomes at the Intersection of Gender and Social Origin." *Social Inclusion* 7:79–94.

A Appendix

A.1 Model-Based Simulations

Markov models are based on the assumption that the state observed at a given time point depends (solely) on the previous one. Practically, these models require specifying the vector **B** of the state probabilities at time 1 and the single matrix **T** describing the transition probabilities between the states at each time point; we kept these probabilities constant over time. A sequence is generated by first randomly selecting a state from **B**. Then, the next states are randomly drawn position by position according to the transition matrix **T** and the state at the previous time point. The sequence is complete when the expected length ℓ is reached. Repeating the procedure n times results in the generation of n sequences according to the same model.

Practically, the transition matrix **T** can be summarized as follows. When in the dominant state, the probability of staying in this state is high, while the probability of moving to any other state is low. When in a non-dominant state, the probability of moving to the other non-dominant state is null and the probability of returning to the dominant state is high.

	A	B	C
A	p	$\frac{1-p}{2}$	$\frac{1-p}{2}$
B	$\frac{1}{3}$	$\frac{2}{3}$	0
C	$\frac{1}{3}$	0	$\frac{2}{3}$

	A	B	C
A	$\frac{5p}{6}$	$\frac{3-2p}{6}$	$\frac{1-p}{2}$
B	$\frac{1-p}{2}$	p	$\frac{1-p}{2}$
C	$\frac{p}{6}$	$\frac{p}{6}$	$\frac{3-p}{3}$

Table 5: Transition matrices **T** of Markov models for sequences dominated by state A (left) or transiting from state A to B (right)

The transition matrix **T** presented on the left-hand side of Table 5 describes

the model for sequences *dominated by state A*. The *dominated* models for state B or C are constructed similarly. The vector **B**—the initial probabilities to be in each state—is defined as the first row of the matrix **T**. The model depends on a single parameter p ranging from 0.85 to 0.99, which is used to describe different *levels of clustering strength* (see subsection 3.1.2).

Let us illustrate the generation of a sequence for this model using $p = 0.99$. The probability of starting the sequence in state A equals 0.99 and 0.005 for B and C. If the first state is A, the distribution of probabilities for the second one is the same. If the first state is B, the probability distribution is given by the second row of the transition matrix **T**, i.e. $\frac{1}{3}$ to move to the dominant state A and $\frac{2}{3}$ to stay in B. If C is the first state, the probability of moving to A equals $\frac{1}{3}$, and to $\frac{2}{3}$ to stay in C. The same procedure is applied until the 20^{th} state of the sequence is generated. The three left panels of Figure A.1 display sequences generated by this model for three p value.

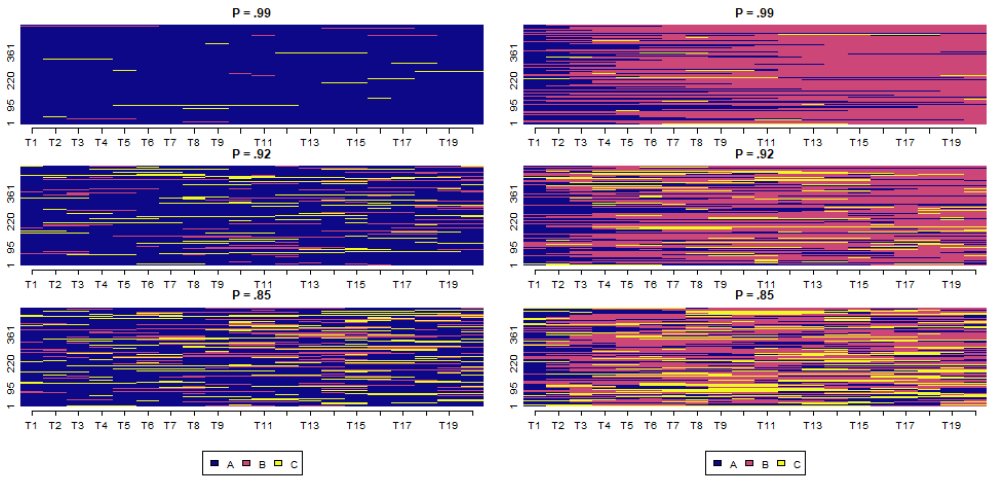

Figure A.1: Index plots (unordered) of sequences generated with Markov models, dominated by state A (left) and transiting from state A to B (right)

The transition matrix **T** for the model transiting from A to B is presented on the right-hand side of Table 5. Here again, the model depends on a single param-

eter p and the first row further describes the initial state distribution. Overall, the model start with a higher chance of starting the sequence with state A and staying in it. However, the chance to move to state B is higher than in the A *state dominated* model. Sequences generated by this model using various p values are shown in the right-hand panels of Figure A.1.

A.2 Supplementary Figures

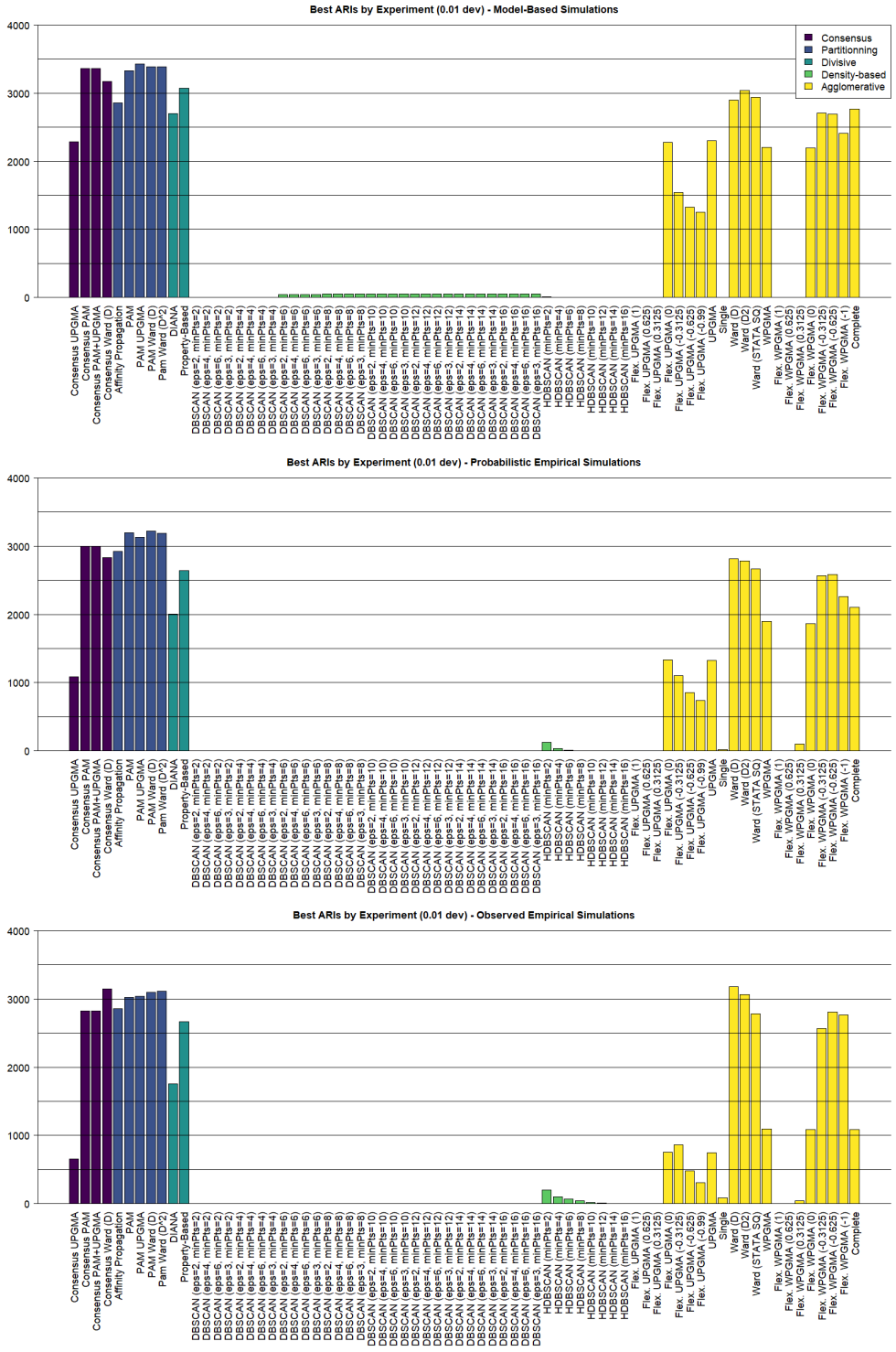

Figure A.2: Number of times each algorithm features among the best-performing ones in each experiment. All algorithms included.

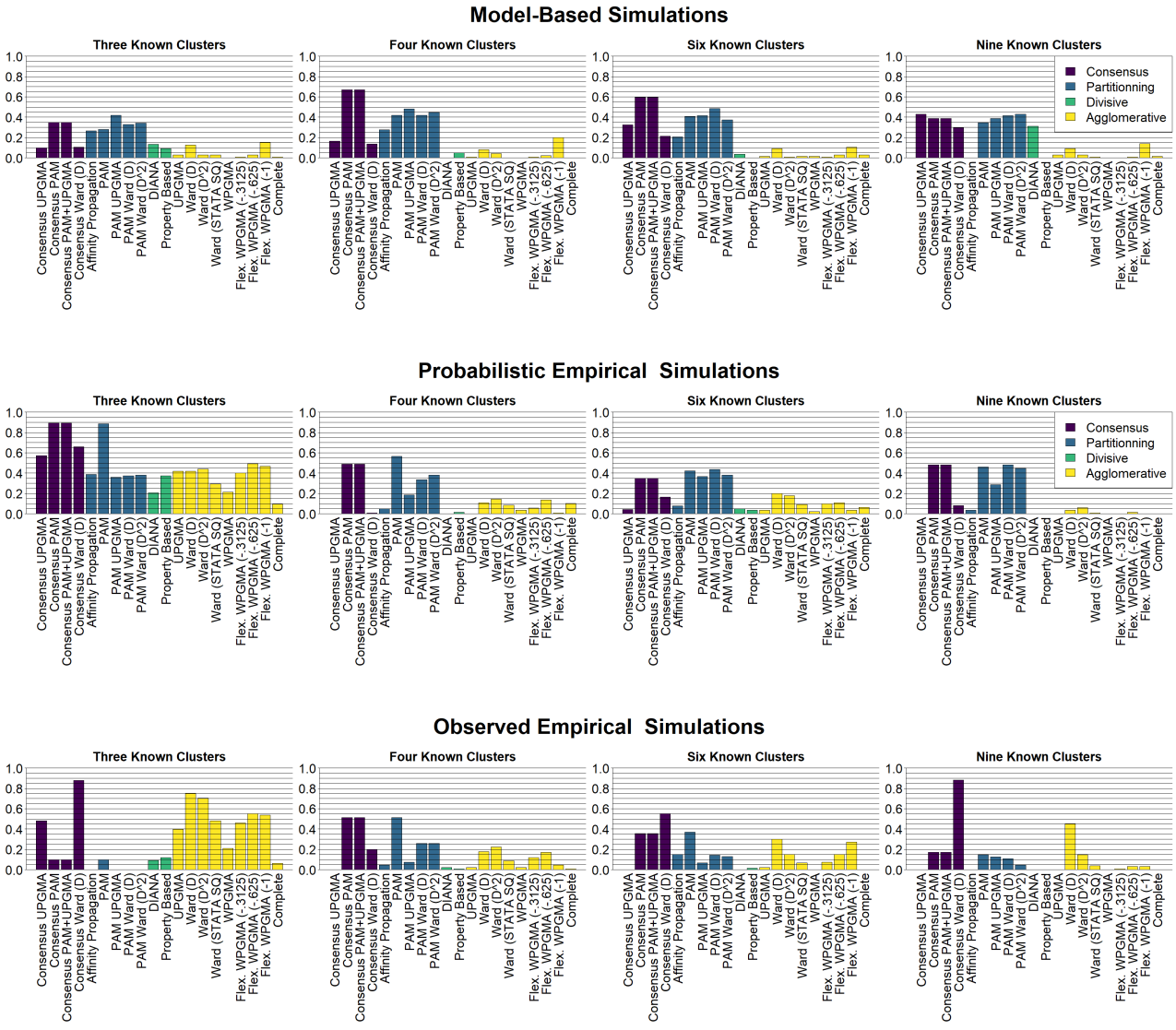

Figure A.3: Proportion of simulations where each CA features among the best-performing ones, separated by levels of detail.

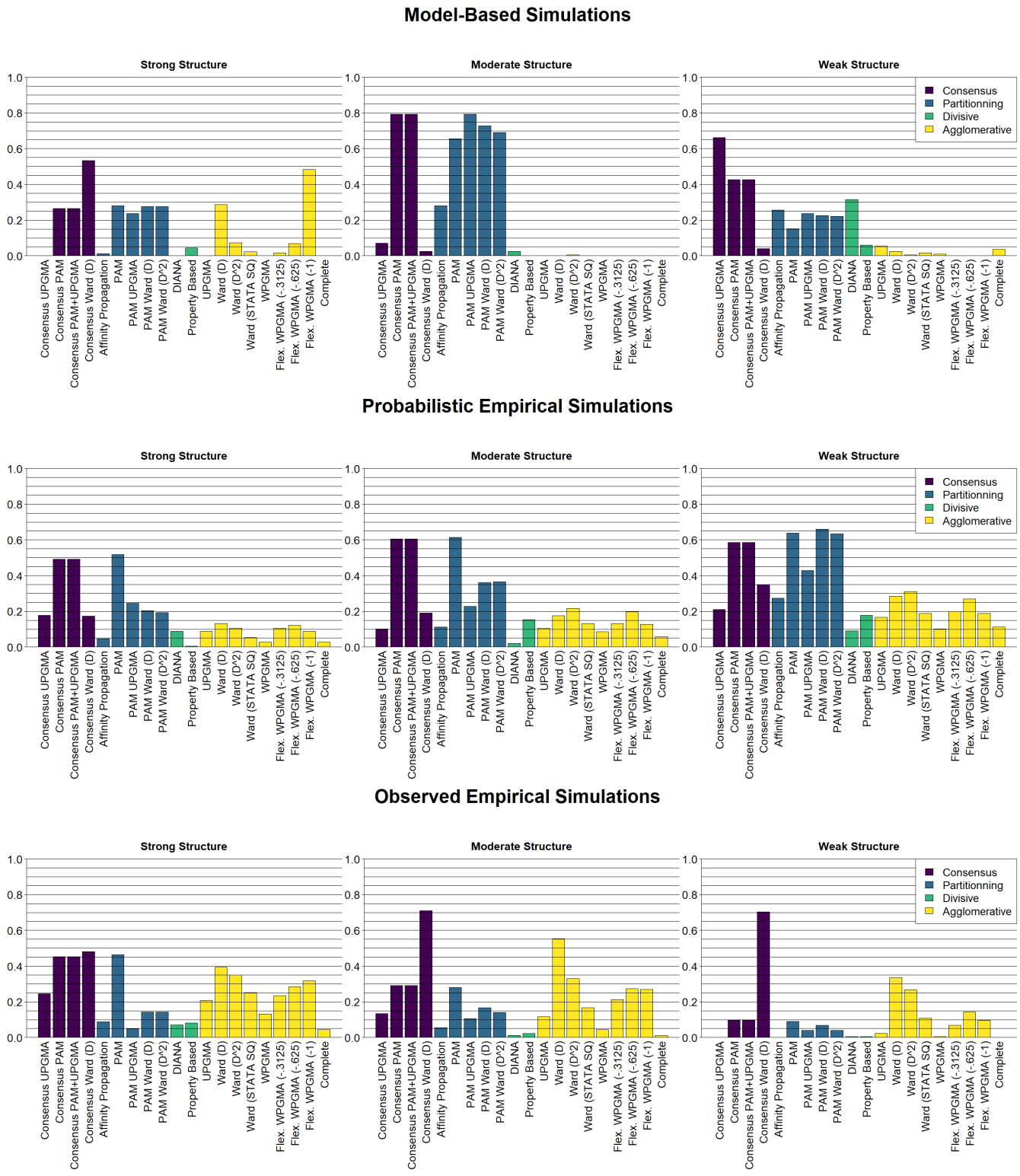

Figure A.4: Proportion of simulations where each CA features among the best-performing ones, separated by levels of clustering structure

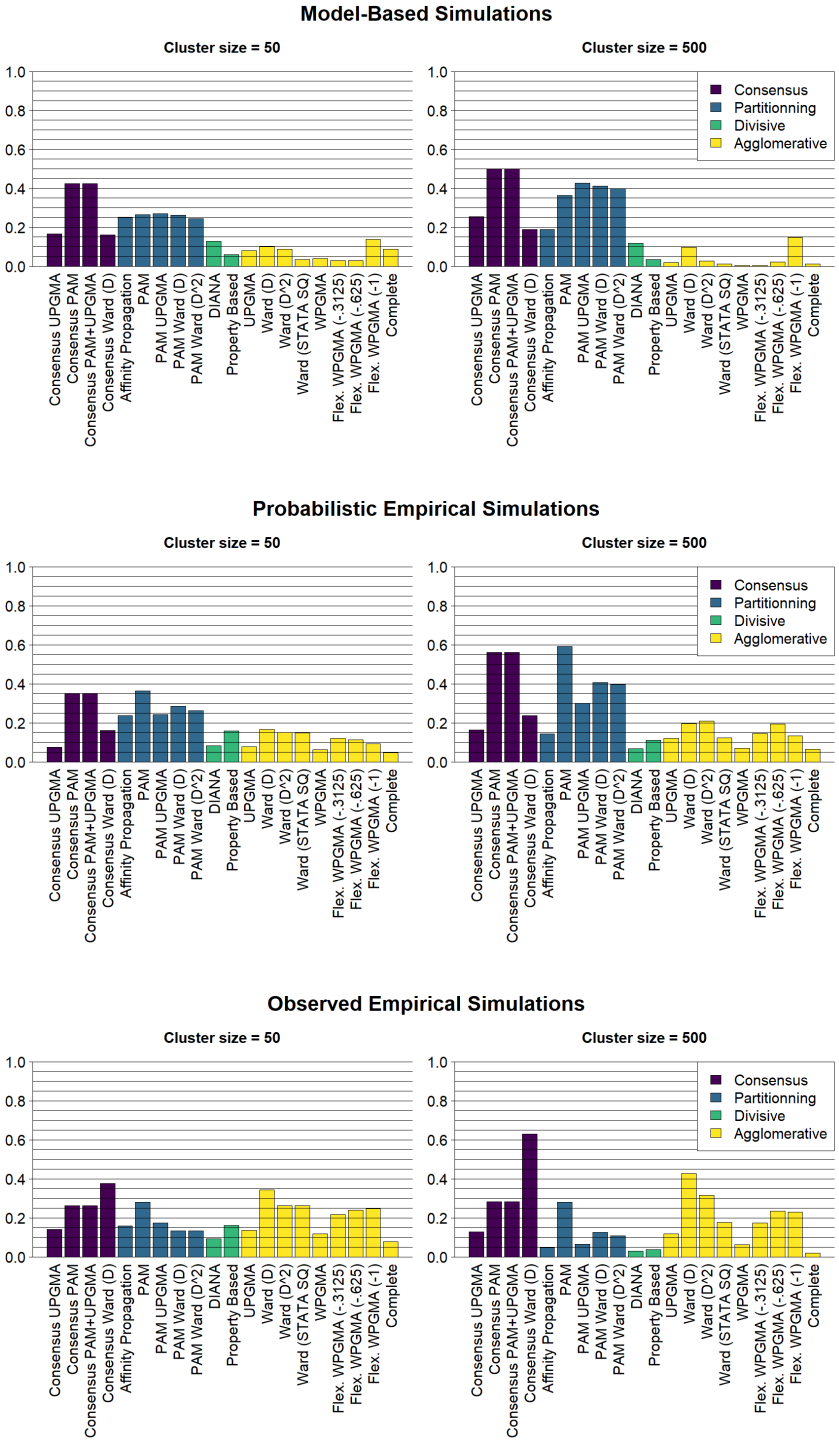

Figure A.5: Proportion of simulations where each CA features among the best-performing ones, separated by cluster sizes.

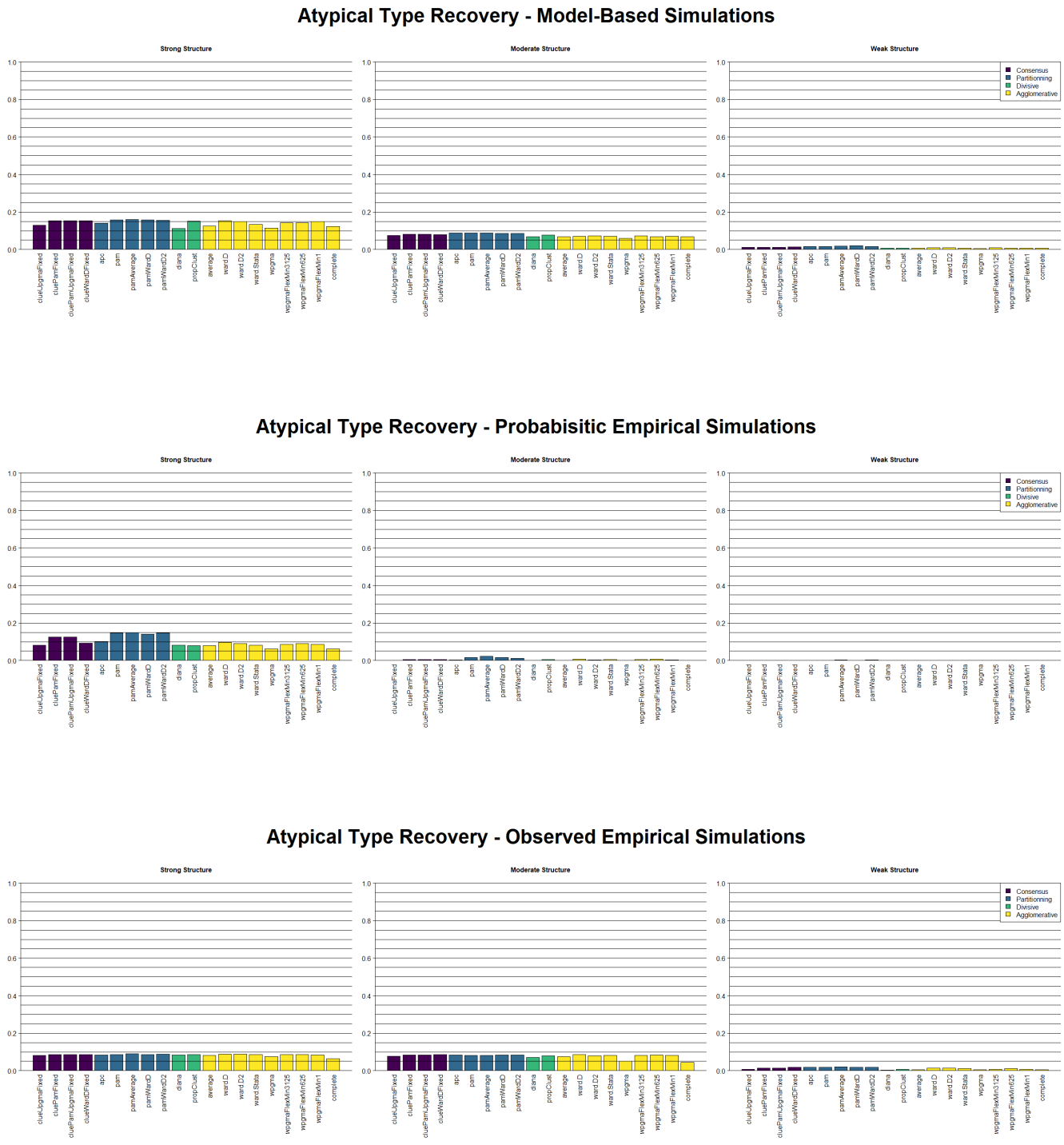

Figure A.6: Proportion of simulations where atypical types are identified, Jaccard index ≥ 0.7, sepatated by levels of clutering stucture

Figure A.7: Proportion of simulations where hybrid types are identified, Jaccard index ≥ 0.7, sepatated by levels of clutering stucture

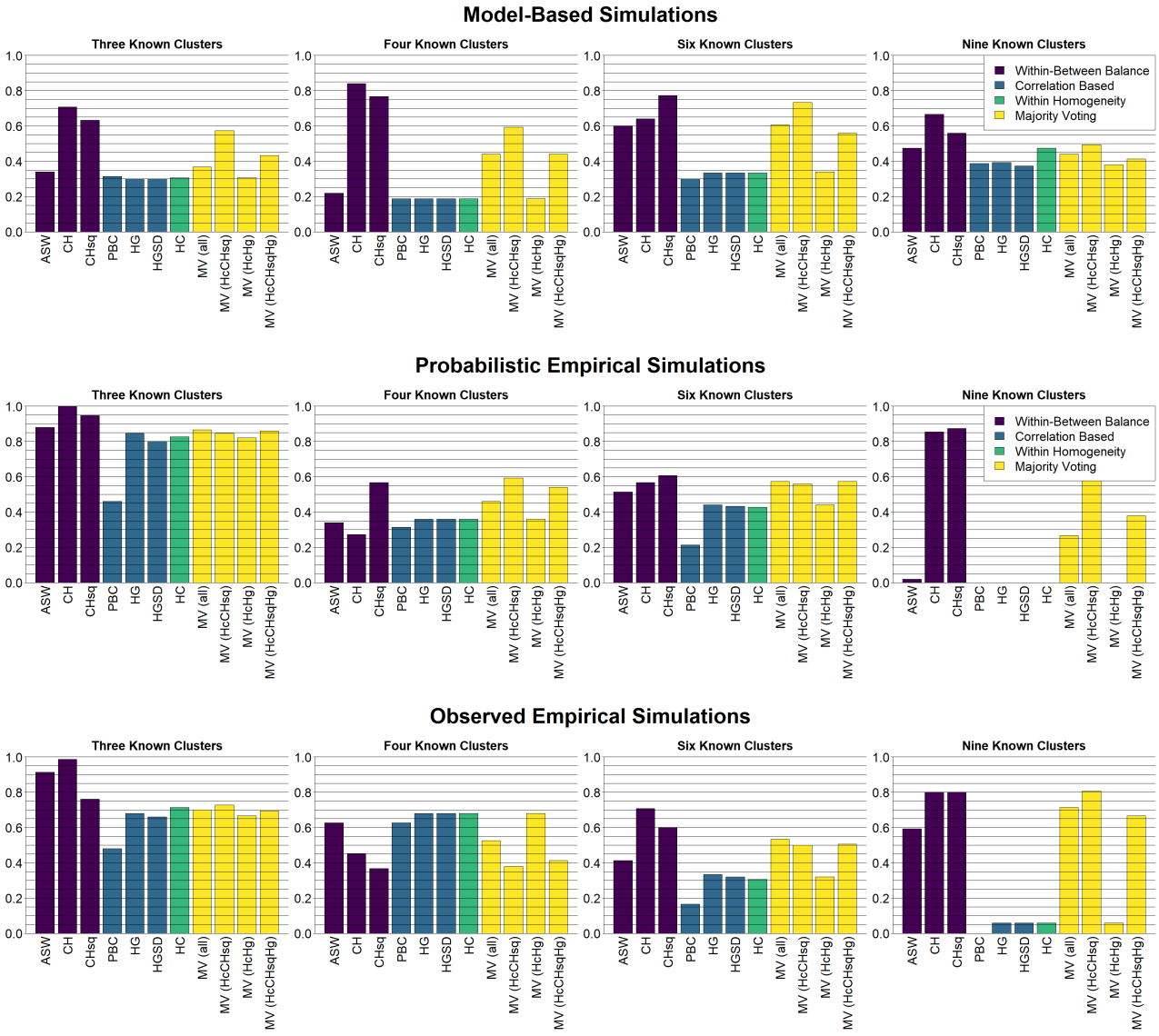

Figure A.8: Proportion of experiments where each CQI selects the best performing algorithm among all simulations, separated by number of known clusters.

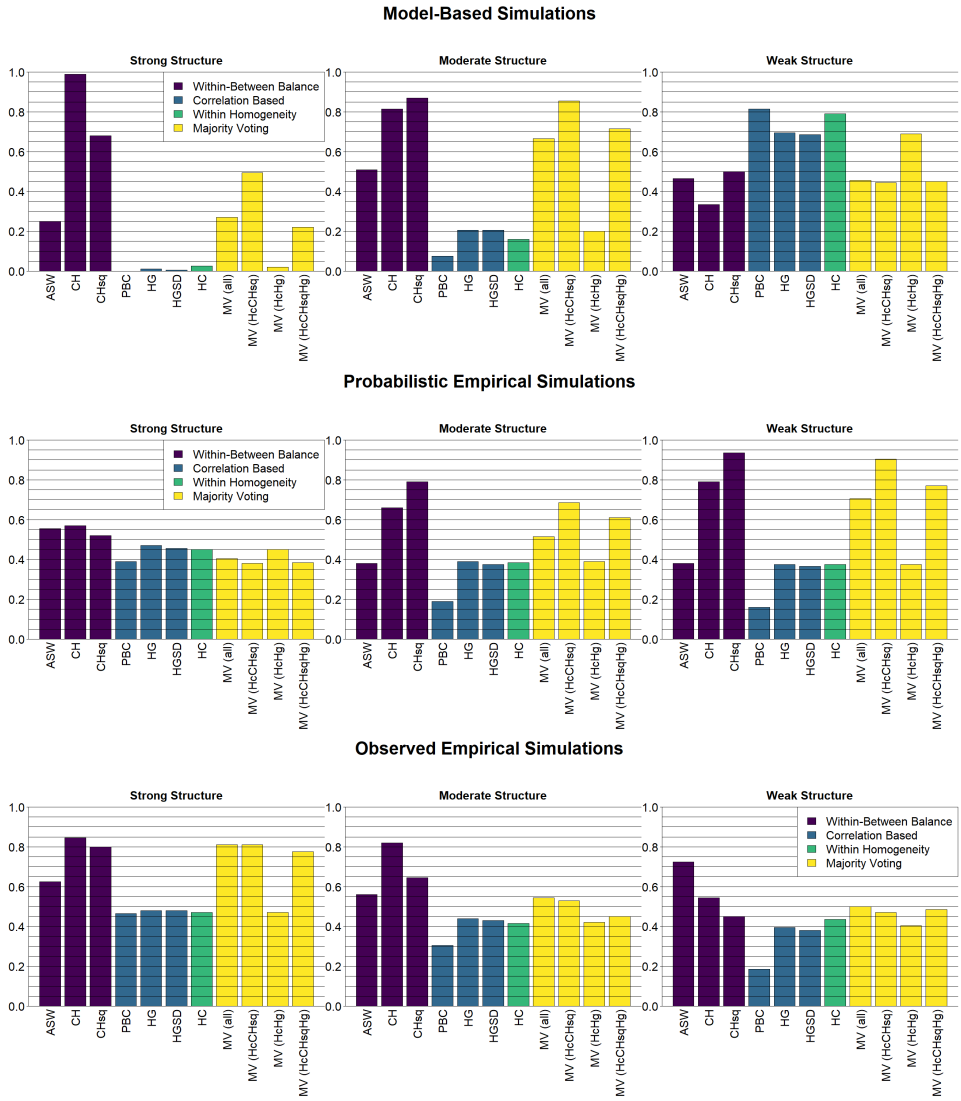

Figure A.9: Proportion of experiments where each CQI selects the best performing algorithm among all simulations, separated by clustering structure levels.

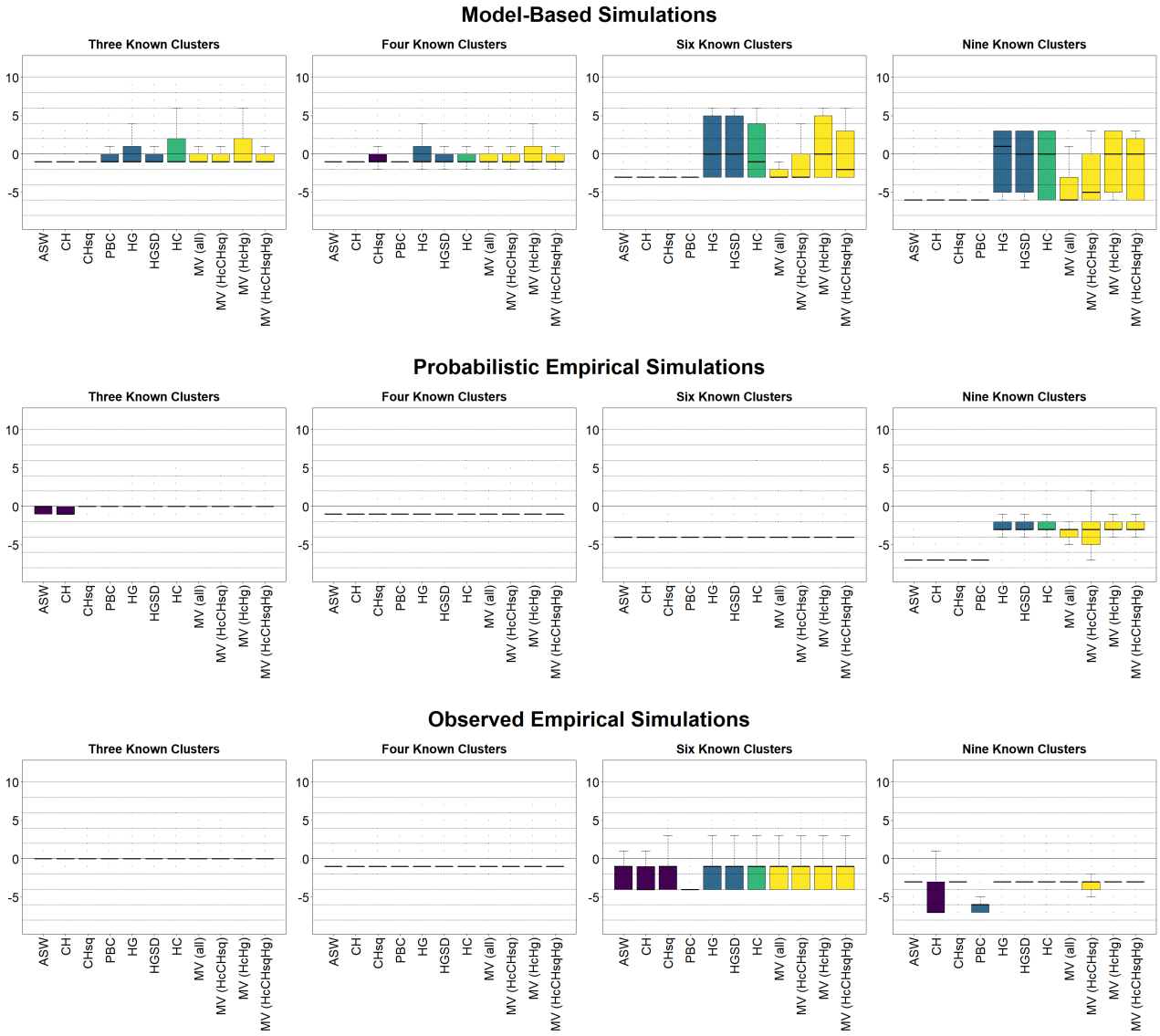

Figure A.10: Gap between expected and selected number of clusters, separated by typology level of detail.

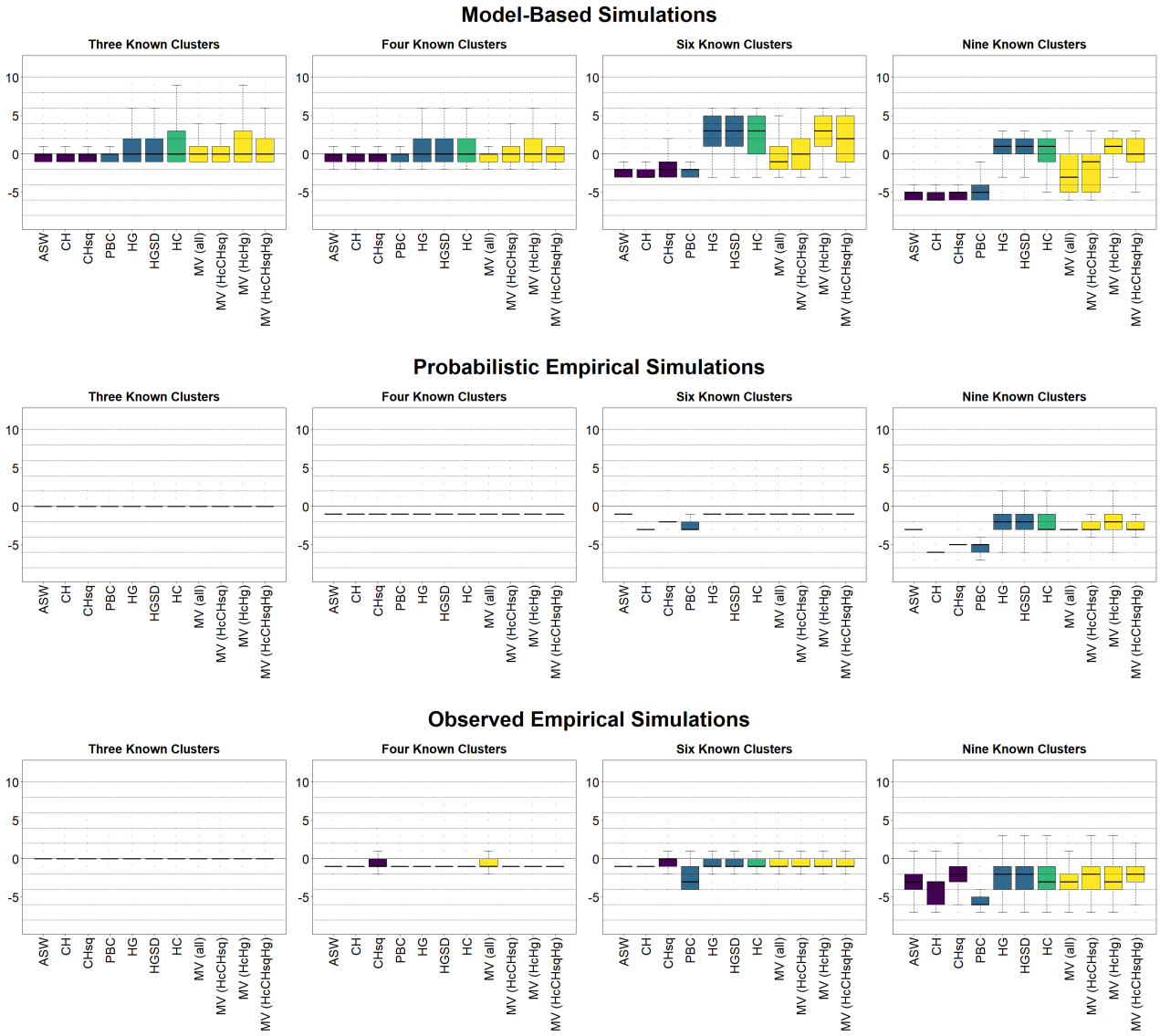

Figure A.11: Gap between expected and selected number of clusters, separated by typology level of detail. The two best CQIs are considered

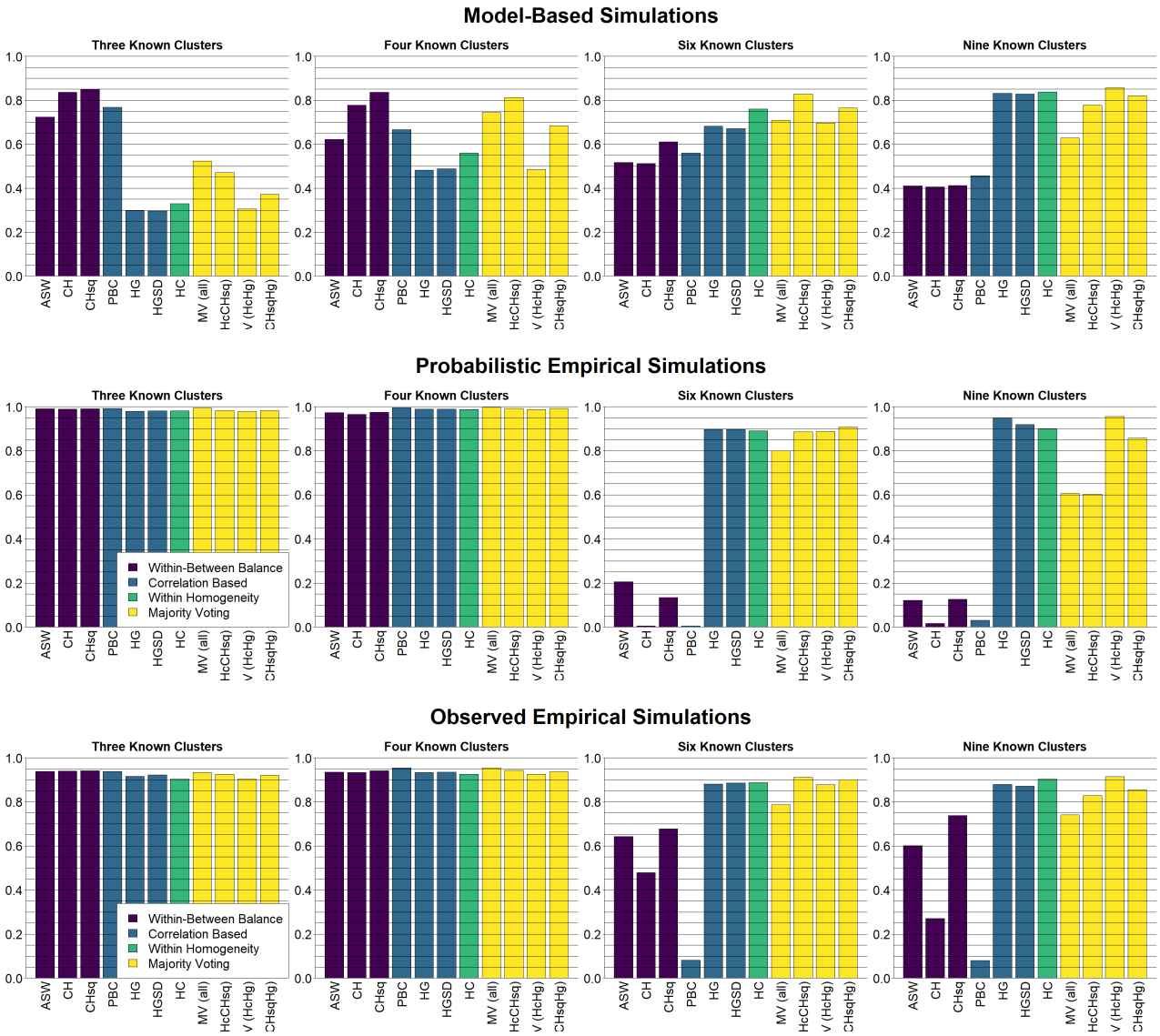

Figure A.12: Proportion of experiments where each CQI selects the optimal number of groups, separated by level of detail. Clusterings in two groups are excluded

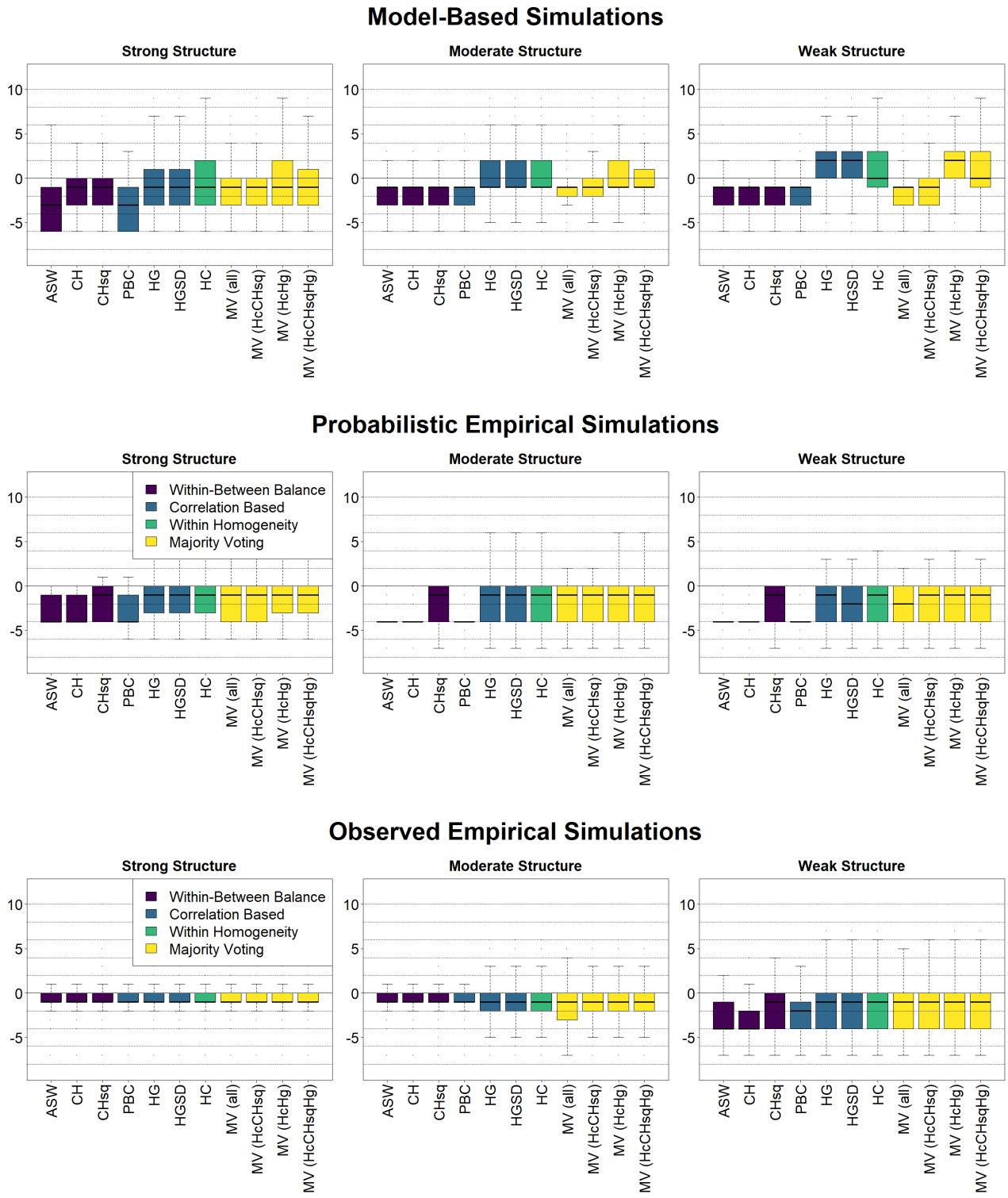

Figure A.13: Gap between expected and selected number of clusters, separated by typology level of detail. Separated by levels of clustering structure

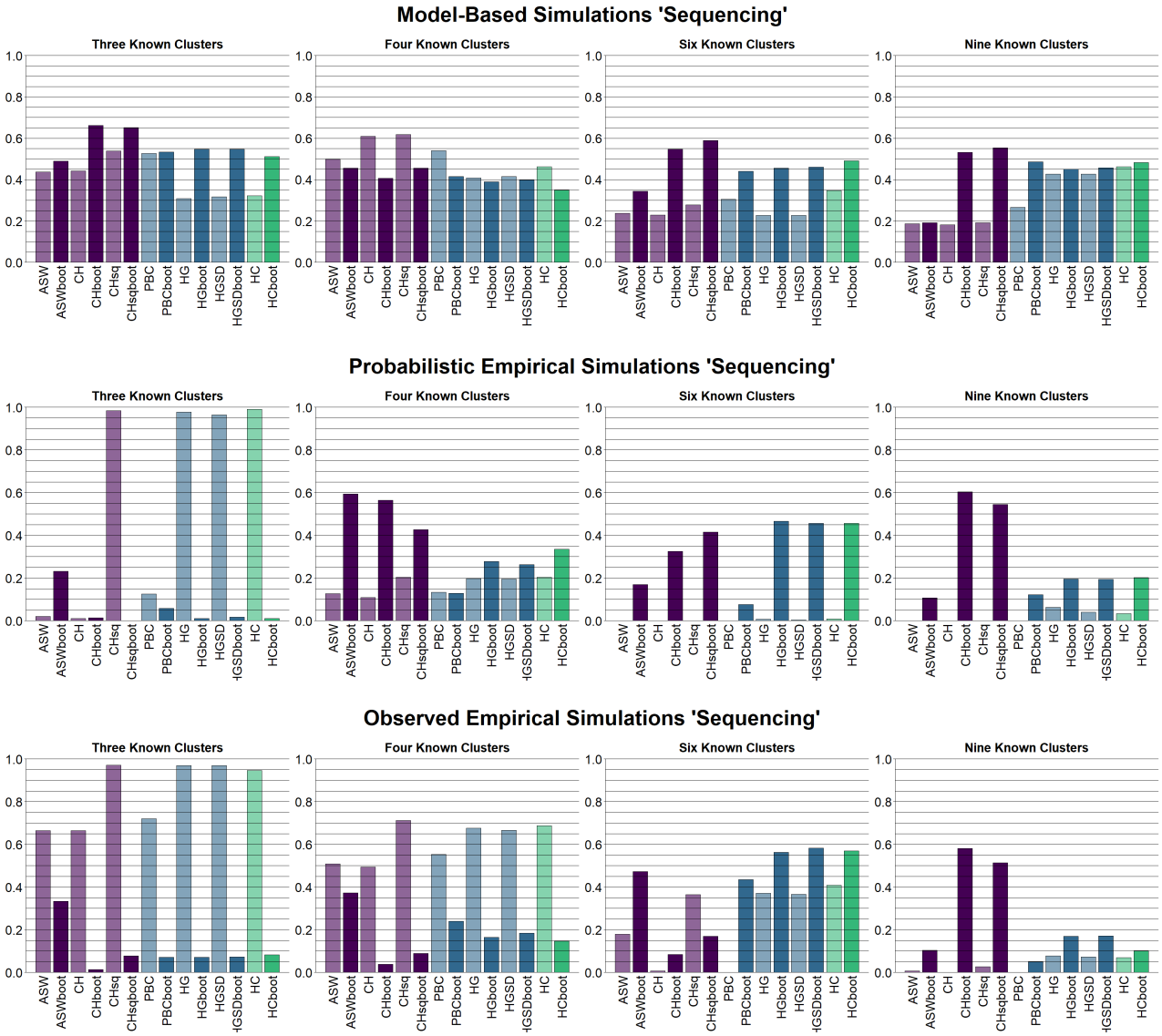

Figure A.14: Proportion of experiments where each CQI selects the optimal number of groups, separated by levels of detail.