Received: 17 April 2023 | Revised: 29 June 2023 | Accepted: 1 July 2023

DOI: 10.1002/widm.1511

OVERVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY

WILEY

A white paper on good research practices in benchmarking: The case of cluster analysis

Iven Van Mechelen[1] | **Anne-Laure Boulesteix**[2] | **Rainer Dangl**[3] | **Nema Dean**[4] | **Christian Hennig**[5] | **Friedrich Leisch**[3] | **Douglas Steinley**[6] | **Matthijs J. Warrens**[7]

[1]Quantitative Psychology and Individual Differences, University of Leuven, Leuven, Belgium

[2]Faculty of Medicine, LMU Munich and Munich Center of Machine Learning, Munich, Germany

[3]Institute of Statistics, University of Natural Resources and Life Sciences, Vienna, Austria

[4]School of Mathematics & Statistics, University of Glasgow, Glasgow, UK

[5]Department of Statistical Sciences "Paolo Fortunati", University of Bologna, Bologna, Italy

[6]Psychological Sciences, University of Missouri, Columbia, Missouri, USA

[7]Department of Pedagogical and Educational Sciences, University of Groningen, Groningen, The Netherlands

Correspondence
Iven Van Mechelen, University of Leuven, Leuven, Belgium.
Email: iven.vanmechelen@kuleuven.be

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Number: 01IS18036A; Engineering and Physical Sciences Research Council, Grant/Award Number: EP/K033972/1; German Research Foundation, Grant/Award Numbers: BO3139/7-1, 9-1; Research Foundation - Flanders, Grant/Award Numbers: G080219N, K802822N; Research Fund of KU Leuven, Grant/Award Number: C14/19/054

Edited by: Mehmed Kantardzic, Associate Editor and Witold Pedrycz, Editor-in-Chief

Abstract

To achieve scientific progress in terms of building a cumulative body of knowledge, careful attention to benchmarking is of the utmost importance, requiring that proposals of new methods are extensively and carefully compared with their best predecessors, and existing methods subjected to neutral comparison studies. Answers to benchmarking questions should be evidence-based, with the relevant evidence being collected through well-thought-out procedures, in reproducible and replicable ways. In the present paper, we review good research practices in benchmarking from the perspective of the area of cluster analysis. Discussion is given to the theoretical, conceptual underpinnings of benchmarking based on simulated and empirical data in this context. Subsequently, the practicalities of how to address benchmarking questions in clustering are dealt with, and foundational recommendations are made based on existing literature.

This article is categorized under:

Fundamental Concepts of Data and Knowledge > Data Concepts
Fundamental Concepts of Data and Knowledge > Key Design Issues in Data Mining
Technologies > Structure Discovery and Clustering

KEYWORDS

conceptual underpinnings, foundational recommendations, method comparison

1 | INTRODUCTION

Presently, many domains of science are shaken to their foundations by powerful calls for scientific integrity, for buttressing up claims in evidence-based ways, for setting up studies in reproducible and replicable manners, and for adhering in all possible respects to good research practices (see, e.g., Haibe-Kains et al., 2020; Hutson, 2018). This tremor has also hit the fields of statistics, machine learning, and data mining.

Within these fields, both users and methodological researchers face a major challenge when having to choose among a wealth of new and old methods. Obviously, a proper choice requires addressing the question of which alternatives are optimal in some respect, necessitating comparative evaluation or *benchmarking*. Linking up with the calls referred to above, answers to benchmarking questions should be evidence-based, with the relevant evidence being collected through well-thought-out research practices, in reproducible and replicable ways.

In the present article, we review good research practices in method benchmarking from the perspective of the area of cluster analysis. The reason for this choice is that, whereas there is an initial impetus to a benchmarking tradition in bioinformatics (Weber et al., 2019) as well as in many subdomains of statistics and machine learning, including the supervised part of the domain of statistical learning (see, e.g., Eugster, 2011; Hoffmann et al., 2019; Stone, 1974), the situation is rather different for unsupervised learning, and in particular cluster analysis. Indeed, from the early days on, several prominent researchers in the clustering area have been emphasizing that very many choices have to been made in the selection of a proper clustering strategy (e.g., Anderberg, 1973; Dubes & Jain, 1976; Hartigan, 1985). Moreover, only very few theoretical guidelines appeared to be available for this selection, in spite of praiseworthy attempts to list formal characteristics of useful clustering methods (Jardine & Sibson, 1971) as well as so-called "admissibility criteria" (Fisher & Van Ness, 1971; Rubin, 1967). As a way out, one necessarily had to turn to benchmarking studies, with instances of seminal benchmarking attempts being reported by, for example, Baker (1974), Hubert (1974), and especially Milligan (Milligan, 1980, 1985; Milligan et al., 1983; Milligan & Cooper, 1985; for overviews of earlier benchmarking work in the area, see, e.g., Jain & Dubes, 1988; Milligan, 1981a, 1981b). Later on, there has been some follow-up to this seminal work (e.g., Anderlucci & Hennig, 2014; Arbelaitz et al., 2013; Costa et al., 2022; Hennig, 2022; Rossbroich et al., 2022; Schepers et al., 2006; Shireman et al., 2017; Steinley, 2003; Steinley & Brusco, 2008; Šulc & Řezanková, 2019; Wilderjans et al., 2013). Nevertheless, there is much less of a benchmarking tradition in the clustering area than in the field of supervised learning. This is evidenced by the fact that not infrequently new methods are proposed without a sound comparison with their best predecessors, by a broad lack of neutral comparisons of existing methods, by a dearth of recommendations and guidelines for benchmarking, and even by a lack of philosophical scrutiny of the issue (Watson, 2023). At least part of all this goes with the fact that in supervised classification, classes and class membership are known a priori, which naturally leads to a concrete metric for evaluating performance (viz., the error rate or proportion of objects misclassified: see, e.g., Schiavo & Hand, 2000). In an unsupervised clustering context, however, classes and class memberships are unknown, and must be "discovered" by the clustering algorithm, which hampers determining how well a specific approach is performing in real data.

In this article, we present an in-depth review of the fundamental conceptual underpinnings of good research practices for benchmarking in the cluster analysis domain, along with a set of foundational guidelines based on existing literature. The latter will include guidelines that are relevant for method benchmarking in general as well as guidelines that are specific for clustering.—That being said, the specific guidelines may also serve as a source of inspiration for areas of statistics, machine learning, and data mining other than clustering. Indeed, because of the absence of straightforward evaluation criteria, the clustering area is obliged to reflect more deeply on principles of benchmarking, unlike other areas that may be inclined to go too quickly into autopilot ways of comparative evaluation.—Due to the tendency of established recommendations to be often ignored, the guidelines will reiterate some well-wrought best practices as well as supplement them with new advice. All in all, we hope that this endeavor will foster scientific debate on benchmarking, and that this will subsequently lead to improved research practices.

The remainder of this article is structured as follows: Section 2 presents an in-depth review of the conceptual underpinnings of benchmarking in the clustering area; Section 3 continues with a discussion of the practicalities of how to address benchmarking questions in clustering along with foundational guidelines; Section 4 presents a handful of illustrative examples.

2 | CONCEPTUAL UNDERPINNINGS

In this section, we outline a conceptual framework for benchmarking studies of clustering. Four considerations are central to inform the key decisions to be drawn in such studies: (1) features of the clustering methods to be compared in them, (2) the data to be used for benchmarking, (3) the quality or performance criteria focused on, and (4) the type of answers benchmarking studies may lead to. We will deal with these successively.

2.1 | Features of clustering methods to be compared in benchmarking studies

Benchmarking in statistical data analysis pertains to a comparison of different data-analytic methods. Importantly, one should pay ample attention to the fundamental features that characterize these methods and that differentiate between them, as those features may have major relevance both for understanding the methods and for informing decisions in benchmarking. In this subsection we will do so for clustering methods, which may be characterized in terms of features of the points of departure for the clustering process (Section 2.1.1) and features of that process itself (Section 2.1.2).

2.1.1 | Points of departure for clustering process

Point of departure 1: Research goal for clustering

The goal of any clustering process is to learn or induce unknown groups from the data, with the groups being defined in terms of their memberships, that is, the clusters' *extension*, and/or their characteristics, definitions, and membership rules, that is, the clusters' *intension* (Leibniz, 1764). The aim underlying this can be the search for the "true" clustering (see Hennig (2015c) for a discussion of that concept), which can be linked to the Socratic target of "carving nature at its joints" (Plato, n.d., approximately 370 BC) and can be qualified as fundamental or "realistic." Alternatively, the underlying goal can be more pragmatic or "constructive" (Hennig, 2015b), with the ideal clustering being optimal in terms of some extrinsic goal such as information compression, communication, or providing a suitable basis for some kind of action. Apart from the realistic–constructive distinction, one should also consider the question whether induction of a clustering is the only research goal at issue, or rather whether the clustering also goes with an alternative goal, such as prediction, dimension reduction, or (biological, psychological, …) process modeling. In the latter cases, the primary concern may be either in the clustering or in the alternative goal, even for one and the same clustering method. As an example, a mixture of factor analyzers (McLachlan & Peel, 2000) can sometimes be used with finding groups as primary concern (with the factor analysis part being invoked only to cope with technical complications resulting from ill-conditioned data), and sometimes with the retrieval of latent dimensions underlying the data as primary goal (with the mixture part being invoked only to deal with a secondary issue of population heterogeneity).

Point of departure 2: Data for clustering

An important second point of departure for a clustering process is the type of data that are to be subjected to the clustering method. Importantly, the type of data structure is discussed in this subsection as a characteristic of a clustering method, and not from the viewpoint of the type of data to be used for the benchmarking (although the type of data that is required by the clustering method obviously also impacts the type of data to be used for the benchmarking).

Two basic data structures are most frequently encountered in the clustering domain: object by variable data and object by object proximity (similarity or dissimilarity) data with unstructured object or variable data modes. Apart from these, quite a few more complex data structures may be encountered as well, including graph or network data, multiway data, multiblock data (of which multilevel data are a special case), and data with structured object or variable data modes (e.g., time or space) such as sequence or functional data.

Leaving aside semi-supervised cases, data for a cluster analysis normally do not include a priori known cluster memberships. If such information were nonetheless available, it would be kept apart from the "internal" part of the data that is to be used in the cluster analysis, and could optionally be used afterwards as "external" information to validate clustering results.

2.1.2 | Clustering process

The clustering process comprises three steps: (1) pre-processing the data, (2) the actual cluster analysis, and (3) post-processing the output of the second step.

Step 1: Pre-processing the data

More modest forms of pre-processing include various ways for handling missing data and various kinds of data selection. Data selection can refer to the experimental units or objects with, for example, a removal of outliers before the actual cluster analysis. Alternatively, selection can also refer to the variables included in the data set. Otherwise, data selection can take place both before and during the actual cluster analysis, with only the former qualifying as data pre-processing.

A somewhat more involved type of pre-processing is that of transformations of variables or even entire data blocks. Classic examples in the clustering domain are various combinations of centering and scaling (e.g., Milligan & Cooper, 1988; Steinley, 2004). Data transformations as a way of data-pre-processing have to be distinguished from data transformations that are part of the actual cluster analysis (see, e.g., van Buuren & Heiser, 1989).

The most comprehensive types of pre-processing imply a conversion of the data structure into a different type of structure before the actual cluster analysis. A classic instance of this is the conversion of object by variable data into object-by-object proximities, for which an unlimited wealth of proximity measures could be invoked (see, e.g., Gower & Legendre, 1986). Note that this type of conversion implies that, in benchmarking studies, clustering methods based on proximity data can be compared with methods based on object by variable data.

Step 2: Actual cluster analysis

Aspects of the actual cluster analysis include: (a) the nature of the clusters aimed at, (b) the criterion or objective function that is to be optimized, and (c) the choice of a suitable algorithm or computational procedure.

Nature of the clusters. The clustering aimed at can be typified in terms of set-theoretic characteristics of the membership of the clusters of interest, that is, their extension, and of organizing principles on the level of the clusters' definition, associated features or attributes, that is, their intension. The following five questions can be helpful to characterize the clusters' **extension**:

1. *What is the nature of the elements of the clusters*? The elements of the clusters can be objects, variables, or some other type of entities involved in the data (for example, sources). Alternatively, in case of data that involve two or more types of entities, such as object by variable or source by object by variable data, the elements to be clustered could be n-tuples such as ordered pairs (of, for example, an object and a variable) in two-mode, bi- or co-clustering (see, further, e.g., Hartigan, 1975; Van Mechelen et al., 2004).
2. *Is cluster membership crisp or fuzzy?* The standard form of cluster membership is crisp in nature, with elements not belonging versus belonging to a cluster under study, formalized by membership values of 0 versus 1. Alternatively, graded forms of membership can be considered, with membership values varying from 0 to 1, which optionally could be formalized in terms of some fuzzy set-theoretic framework (e.g., Bezdek et al., 2005).
3. *A single cluster* versus *a comprehensive clustering?* Certain clustering methods aim at retrieving a single cluster that in itself is optimal in some sense. Another, and far more common alternative is to look for a collection of clusters (i.e., a clustering) that is to be assessed as a whole in some respect; in the latter case, the number of clusters in the collection can be either pre-specified or chosen during or after the actual cluster analysis.
4. *Are all elements to be clustered?* This question is relevant if one is looking for a clustering, which, as a whole, is optimal in some sense.
5. *Is cluster overlap allowed, and, if yes, is it restricted to nested clusters?* Here overlap pertains to extensional overlap. Nestedness further means that for every pair of overlapping clusters it holds that one of the clusters is a subset of the other. Some authors call nested clusterings hierarchical.

To typify the structure of the clusters' **intension**, both the within- and the between-cluster organization can be considered. Regarding the *within-cluster organization*, a first issue is what should be the unifying or common ground for elements to belong to the same cluster. Possible answers include a common pattern of values for a subset of the variables, small within-cluster dissimilarities, and large similarities of elements to the centroid(s) of the cluster(s) to which they belong. A second issue refers to constraints on the form of the within-cluster heterogeneity, such as topological or geometric constraints (for example, connectedness or convexity), and constraints on the within-cluster dependence structure of variables, such as, for example, the shape implied by the within-cluster covariance matrices. As to the *between-cluster organization*, one may first wonder what should be the discriminating ground for elements to belong to different clusters. Possible answers include large between-cluster dissimilarities, (linear or quadratic—Coraggio & Coretto, 2023) separability, and clearly distinguishable distributional shapes. Second, one may wonder whether there are constraints on the form of the between-cluster differences, such as that all cluster centroids should lie in a low-dimensional space.

The criterion or objective function to be optimized. Any clustering method involves some optimization, yet the nature of the criterion that is optimized may vary considerably between methods. For example, a number of methods proceed in a stepwise fashion, with some specific criterion being optimized in each step, but without the whole of all steps targeting the optimization of an overall objective function. For many other methods, however, an overall objective function is in place. This may or may not refer to some kind of goodness of fit, as in, for example, a likelihood or a least squares loss function.

The algorithm or computational procedure. For the actual optimization, clustering methods rely on computing routines. In absence of an overall objective function, these are mere computational procedures. In presence of such a function, they can be considered algorithms. To be sure, in almost all cases, there is no guarantee that the algorithms in question will yield the globally optimal solutions looked for, with different algorithms, and even different runs of the same algorithm, often leading to different clusterings. Major reasons for this are that the optimization problems at hand are typically nonconvex and at least partially discrete in nature, with a ubiquitous presence of local optima (see, e.g., Steinley, 2003).

Apart from optimization-related aspects, Ackerman et al. (2021) emphasize the importance in benchmarking of how computational procedures react to object duplication, or, slightly more general, object weights. In particular, they draw a distinction between: (1) procedures that are affected by weights on all data sets, (2) procedures that respond to weights on some data sets only, and (3) procedures that ignore weights.

Step 3: Post-processing the output of step 2

A final possible step in the data-analytic process concerns post-processing the output of the chosen clustering method. This often involves a form of model selection, including the choice of the number of clusters in the final model, insofar as this has not been chosen during the actual cluster analysis. This step could boil down to a classical selection among different models, but also to a selection of one part within a model. As an example of the latter, one may think of the selection of a single partition within a hierarchical clustering, conceived as a collection of nested partitions. Other forms of post-processing imply some kind of simplification of the output. As an example, one may think of converting estimated posterior cluster membership probabilities in mixture models as obtained from Bayes' rule into 0/1 membership values.

2.2 | Data to be used for benchmarking

In the present article, we focus on dealing with benchmarking questions by means of analyses of data sets. Different types of relevant data sets can be distinguished.

A first type is **data obtained from Monte Carlo simulations**. If these have been set up on the basis of a priori, technical or artificial assumptions, we will further call them *synthetic simulations*, in contrast with the realistic simulations to be described below.

A second type is that of **empirical data** (from one or more substantive domains of interest). These may or may not include "external" true group information. Data that include such information imply an easy possible validation criterion (albeit not necessarily the best or the only legitimate one: see further Section 3.3.1), yet data that do not include it are more representative of real clustering problems in which groups are not known in advance. Empirical data sets may be conceived as elements sampled from some population of data sets (Boulesteix, Hable, et al., 2015). Questions at this point include how the latter population can be characterized, whether a sample at hand can be considered representative, and which sampling mechanism can be assumed to be in place. Given two or more empirical data sets, one may further also wonder whether they can be considered exchangeable members of some class of data sets for which, for example, the nature of the clusters aimed at is the same. Note that whether or not this is the case will depend on the data set features one cares about, and, hence, is "in the eye of the beholder."

Three variants of using empirical data sets in benchmarking can further be considered. First, the data sets may be used as given. Second, they may be modified, for example, by adding outliers or noise variables, or by removing objects or variables. Third, they may be used as input for so-called *realistic simulations*, to be distinguished from the synthetic simulations discussed above; these involve the use in the simulations of structures, parameter settings, or distributions found in analyses of the empirical data (for an illustration, see El Abbassi et al., 2021). Examples include informing simulations to make the range of the studied factors realistic, and creating simulated data sets by applying a resampling procedure to the residuals that result from an analysis of the empirical data.

2.3 | Quality criteria for benchmarking

For the comparative evaluation of methods that is inherently implied by any benchmarking study, numerous quality criteria can be considered. Below, we will outline a few conceptual distinctions between possible criteria, with the primary distinction being the one between omnibus or overall formal criterion types that may be used for the evaluation of a broad range of methods versus criteria that relate to the particular goal of the data analysis under study.

2.3.1 | Omnibus formal criteria

Optimization performance

In the presence of an objective or loss function that all to be compared methods intend to optimize, a first type of criterion relates to the value of that objective function.

Stability or replicability

A second type of criteria pertains to the stability of the output of the cluster analysis, in terms of its extensional or intensional aspects. Stability may be investigated in itself, for instance, by means of bootstrap samples. Alternatively, it may also be investigated with regard to several choices in the clustering process. As to the latter, on the level of the data preprocessing, these choices may pertain to the entry/removal of outliers, the type of transformation of variables or data blocks, and the type of conversion of the data structure (e.g., the choice of a dissimilarity measure). On the level of the actual data analysis, they may include algorithm initialization and the values of tuning parameters (Mishra et al., 2022). On the level of the post-processing of the output, the choice of a type of method for model selection may be considered.

Computational cost

A third type of criteria is the computational cost implied by the data analysis, in terms of number of operations, computational complexity, or computing time (often as a function of data size).

2.3.2 | Criteria related to the goal of the data analysis

A second group of criteria refers to the specific primary goal of the data analysis. Successively, we will consider the case in which this goal concerns the identification of some unknown clustering, and the case of other primary goals.

Primary goal is identification of clustering

Within this subsection we draw a distinction between three types of criteria: (a) recovery performance, (b) other external validation, and (c) internal validation (for an early reference to the distinction between external and internal validation, see Dubes & Jain, 1979).

Recovery performance. In case external validation information on the truth underlying the data is available, recovery of some aspect of this truth is a possible criterion. In a clustering context, relevant aspects may be the true number of clusters, true cluster membership, true cluster characteristics (for example, the values of the cluster centroids), and the truly relevant variables for the clustering. Note, though, that even if data go with true grouping information, this is not necessarily the best or only legitimate validation criterion, as in cluster analysis the "truth" depends on the context and the goals of the analysis (Hennig, 2015c), and therefore reasonable clusterings may or may not correspond with the given grouping information.

Other external validation. Other external validation criteria may be relevant if the goal underlying the cluster analysis is not the search for the "true" clustering, but is more pragmatic in nature. In the latter case, criteria that capture the pragmatic goal can be considered, such as, for example, response to treatment to validate a patient clustering induced from patient by biomarker data.

Internal validation. Internal validation criteria may reflect what constitutes a "good" or "desirable" clustering, in terms of set-theoretic characteristics of the clusters' extension as well as organizing principles for the clusters' intension (see Section 2.1.2). As examples, one may think of degree of extensional overlap, small within-cluster dissimilarities, and degree of between-cluster separability.

Other primary goal

If the primary concern behind the data analysis is not the induction of a clustering but something else such as prediction, dimension reduction, or modeling of the (biological, psychological, …) processes through which the data came about, criteria that capture this concern can be used, such as measures of predictive quality, or of assumed underlying processes.

2.4 | Possible answers resulting from benchmarking studies

Benchmarking questions imply by definition a comparative evaluation of different methods in terms of their performance. Answers to such questions on the basis of some benchmarking study may or may not refer to differences between conditions in that study, with conditions pertaining to subsets of the data types or performance criteria included in it. Answers that do not refer to such differences are called here "unconditional"; examples include statements about performance rank orders of methods or about an overall winner. Answers that do refer to such differences are called here "conditional"; such statements typically look like: "Method A outperformed method B for such and such data types or with regard to such and such criterion types, whereas for other data and/or criterion types this rank order does not hold or is even reversed."

Conditional answers imply that concepts such as global optimality are to be replaced by more shaded counterparts. If, for a given criterion, the nature of the optimal method further differs across data types, one ends up in a "one size does not fit al" type of scenario (Strobl & Leisch, 2022), which calls for a reporting in terms of proper conditional statements. A formally very similar scenario is well known in clinical trial research and occurs if there is no therapeutic alternative that outperforms all other alternatives for all (types of) patients; this further calls for a precision medicine approach. Likewise, within a benchmarking context, one may look for so-called optimal data-analytic regimes, that is to say, decision rules that specify, given some criterion, for which data types which methods are to be preferred (Boulesteix & Hatz, 2016; Doove et al., 2017). Otherwise, conditional answers may not only pertain to conditional optimality of some method, but also to aspects such as telling apart conditions where different state-of-the-art clustering methods tend to give similar results, and conditions where this is less the case and where some kind of expert intervention to choose some particular result may be needed (Fütterer & Augustin, 2021).

3 | PRACTICALITIES AND FOUNDATIONAL GUIDELINES

3.1 | Choice of methods to be compared

3.1.1 | Issues

For benchmarking, two or more methods have to be chosen, which differ from one another with regard to at least one of the three steps of the data-analytic process: (1) the input pre-processing, (2) the actual data analysis, and (3) the output post-processing.—For a survey of popular R packages for cluster analysis see Flynt and Dean (2016), and for a software library of 50 clustering methods: see Thrun and Stier (2021).—Moreover, several specific aspects have to be decided as well, including the choice of implementations, initializations, tuning parameters, convergence criteria, and time limits.

3.1.2 | Recommendations

We propose three sets of recommendations:

- Make a suitable choice of competing methods, meaning that:
    - The intended scope of the methods should be well-defined, well-justified from the point of view of the aims of the clustering, and well-reported. Moreover, the intended scope should be covered properly. For example, within the intended scope, the choice should be sufficiently broad and should not be limited to close neighbors or variants of the same method, unless the intended scope would be limited to such variants.
    - Do include known strong competitors, that is to say, methods that have been shown in previous studies to outperform others for the type(s) of problems and criteria under study (Boulesteix, 2015). Furthermore, in the search for strong competitors, take into account that clustering methods are being developed and studied in a broad range of research domains (including discrete mathematics, statistics, data mining, computer science, machine learning, and bioinformatics), and therefore do not limit this search to a single research domain only.
    - If methods are to be compared that differ from one another in terms of different aspects (e.g., initialization method and type of iteration scheme, or type of variable selection and type of algorithm), consider, if meaningful and possible, to orthogonally manipulate these aspects according to a full factorial design, to enhance inferential capacity for disentangling the main effects of each of the aspects and their interactions.
- Make full information on the competing methods available, in view of reproducibility (Donoho, 2010; Hofner et al., 2016; Peng, 2011) and of enabling follow-up research:
    - Make the code of the methods under study fully available, whenever possible via platforms such as GitHub or Gitlab.
    - Disclose full information on aspects like initialization, values of tuning parameters, convergence criteria, time limits, random number generators and seeds.
- Organize a fair comparison in terms of the choices of specifics, with fairness meaning that particular methods are neither wrongfully favored nor put at a disadvantage. This implies:
    - that the same amount of prior information is made available to the different methods
    - fair choices with regard to the specification of tuning parameters (such as the number of starts)
    - fairness in terms of the number of operations/amount of computing time apportioned to the different methods under study, with the obvious problem that the latter may depend fairly strongly on the particular implementation that has been chosen (insofar the involved algorithms allow to tune this).
    - that the same amount of time should be spent with a similar amount of expertise to implement the different methods, handle bugs, choose tuning parameters, and so on.

3.2 | Data sets used for benchmarking

3.2.1 | Issues

In case of data that are obtained from **synthetic simulations**, choices have to be made regarding several aspects. The first of these pertains to the factors that are to be manipulated in the simulations, along with the levels of each chosen factor. Examples include the number of objects, the number of variables, the proportion of variables that are relevant for the clustering, the number of clusters, (in)equality of cluster sizes, the cluster-generating distribution, the amount of distributional overlap, and so on. A second aspect pertains to the experimental design of the simulation setup. The most simple option at this point is a full factorial design. A third aspect pertains to the number of replications in each cell of the design. See further Hennig (2018) for some considerations regarding synthetic simulations for benchmarking, partly related to the present article.

For the actual simulations, either own code or existing data generators can be used. Regarding the latter, over the past decades quite a few generators have been proposed, including the Milligan (1985) algorithm for generating artificial test clusters, OCLUS (Steinley & Henson, 2005), the Qiu and Joe (2006) random cluster generation algorithm, and MixSim (Melnykov et al., 2012). In the justification of these generators, quite some emphasis has been put on the aspects of separability and overlap, where overlap refers to *intensional* rather than to extensional overlap, that is to say, overlap in terms of variables or component distributions, with all generated clusterings being partitions. Advantages of existing data generators include their ease of use, and comparability with results of other studies that have used the same generators; disadvantages include that they produce a somewhat limited scope of data sets.

A blueprint of a novel device for simulating data for benchmarking in unsupervised learning has been designed by Dangl and Leisch (2017). This blueprint comprises the plan of a web repository and an accompanying R (R Core Team, 2017) package for the actual production of metadata objects and for the subsequent generation of data sets on a local computer.

A promising framework for a flexible generation of simulated benchmark data sets was recently proposed by Shand et al. (2022). Within this framework, an instance space of data sets is created on the basis of a set of features that quantify the difficulty of a clustering problem from a range of perspectives. By varying the relative importance of the features, different collections of benchmarking data sets can be obtained. A possible target in the use of the framework could be to generate data sets that elicit performance differences between pairs of clustering methods, including relative strengths and weaknesses of each method.

Alternatively, a collection of **empirical data sets** can be used, which could, for example, be retrieved from various (general or specific, e.g., Javed et al., 2020) repositories (with recently also a few mixed empirical-simulated data repositories being proposed for cluster benchmarking: Gagolewski, 2022; Thrun & Ultsch, 2020). Such a collection may be considered a sample from the population of all empirical data sets. Although it can be hardly anything better than a convenience sample, it is worthwhile to try to make it as representative as possible. To draw a sample of empirical data, either a top-down or a bottom-up approach can be followed. In a top-down approach, one starts from a profile of data characteristics, along with common criteria of what would constitute a "good" clustering, to subsequently look for a collection of data sets that sufficiently covers this profile. Alternatively, a bottom-up approach can be taken, in terms of starting from a single data set and an in-depth analysis of what would constitute a "good" clustering for it, to subsequently look for a collection of data sets with similar profiles. Irrespective of the approach that is followed, one should take into account that, within the sample of data sets, typically empirical correlations between data characteristics show up, unlike in simulated setups in which data characteristics can usually be orthogonally manipulated. Such correlations can hamper inferences from benchmarking studies using empirical data.

3.2.2 | Recommendations

We propose three groups of recommendations:

- Make a suitable choice of data sets and give an explicit justification of the choices made. This means that:
    - the intended scope of generalization for the data sets should be well-defined, well-justified, and well-reported;
    - the population of data sets of interest should be well covered;
    - the data sets should be sufficient in number to allow for reliable inferences, also taking into account power issues (see further Section 3.4);
    - whenever feasible, one should consider to carefully combine synthetic simulations and empirical data as these may yield complementary information (Friedrich & Friede, 2023; Zimmermann, 2020);
    - in case of simulated data sets that differ from one another in terms of multiple aspects (e.g., sample size, degree of cluster separability), consider, if meaningful and possible, to orthogonally manipulate these aspects according to a full factorial design to enhance inferential capacity (see, e.g., Costa et al., 2022);
    - whereas, when proposing a new method, it is typically desirable to show that this method does something useful on some data sets that is not covered by already existing methods, report explicitly and honestly about how the data sets in question were selected, for example, because they yielded favorable results for the newly proposed method (Boulesteix et al., 2017); moreover, when proposing a new method, the inclusion of cases in which this method does *not* work, that is, "foils," is particularly informative to clarify the method's actual scope.

- In case of simulated data, organize a fair comparison in terms of the relation between the methods under study and the data-generating mechanisms of the simulations, with fair meaning that one should not exclusively rely on mechanisms that unilaterally favor methods which explicitly or implicitly assume that these mechanisms are in place.
- Disclose full information on the data sets that are used (making use, whenever meaningful, of platforms such as GitHub or Gitlab) in view of reproducibility (Dangl & Leisch, 2017; Donoho, 2010; Hofner et al., 2016; Peng, 2011) and of enabling follow-up research. This means that:
  - for simulated data sets, provide implementable data-generating code with full information on cluster-specific parameters, the data-generating function, random seeds, the type and version of the random number generator, and so on;
  - for empirical data sets, provide the full data sets, with sufficient detail on format, codes used to denote missing values, pre-processing, and so on.

3.3 | Evaluation measures

3.3.1 | Issues

Regarding *formal/technical criteria*, optimization performance can be directly measured in terms of the objective function that is optimized in the data analysis. Examples include the trace of the pooled within-cluster variance–covariance matrix $\mathbf{W}$ that is minimized in *k*-means (MacQueen, 1967), as well as least squares and likelihood objective functions. Specific challenges include that their global optimum is usually not known and that their values are often not comparable across different benchmarking setups or data sets. The latter comparability problems can be occasionally solved by some kind of normalization procedure (see, e.g., Brusco & Steinley, 2007). Regarding formal criteria referring to stability and replicability, a broad range of (dis)similarity measures could be considered (see, e.g., Breckenridge, 1989; Masoero et al., 2023; Ullmann et al., 2022).

Regarding *criteria related to the goal of the data analysis*, we first consider the case in which the primary goal is the identification of a clustering:

- Various indices of recovery performance may be considered. To measure recovery of the subset of variables that is truly relevant for the clustering of interest, measures of recall and precision are available (Steinley & Brusco, 2008). When choosing a measure of cluster membership recovery, one should take the nature of the clustering into account, with different indices being relevant in a partitioning context (e.g., the adjusted Rand index: Hubert & Arabie, 1985), in that of hierarchical clusterings (e.g., the cophenetic correlation: Sokal & Rohlf, 1962), and in that of overlapping clusterings (e.g., the mean absolute difference between the overlapping cluster membership matrix and its true counterpart, minimized across all possible permutations of the clusters: Depril et al., 2012).

  More specifically in a partitioning context, cluster membership recovery indices can be roughly categorized into three approaches: (1) counting object pairs (Warrens & Van der Hoef, 2022), (2) information theory (van der Hoef & Warrens, 2019), and (3) matching sets (Fränti et al., 2014; Steinley et al., 2016). Most indices are of the pair-counting type, based on counting pairs of objects placed in identical and different clusters. Prototypical examples are the (adjusted) Rand index and the Jaccard index. Information-theoretic indices (such as the normalized mutual information) capture the difference in information between two partitions, based on concepts like mutual information, Shannon entropy, and joint entropy (Pfitzner et al., 2009; van der Hoef & Warrens, 2019). Finally, set-matching indices (such as the centroid index) are based on matching entire parts of clusters. One may note that indices based on counting object pairs or information theory are commonly affected by cluster size imbalance: If cluster sizes are unbalanced these measures will primarily reflect the degree of agreement between the large clusters.—Indices that are not susceptible to unbalanced cluster sizes are considered in Pfitzner et al. (2009), Fränti et al. (2014) and Warrens and Van der Hoef (2022).—Finally, the value of the Rand index is determined to a large extent by the number of object pairs that are not joined in either of the partitions (Warrens & Van der Hoef, 2022).

- Criteria that involve external validation information other than information on the truth may include effect sizes (such as R^2- and η^2-type measures) associated with regression analyses, analyses of variance, or discriminant analyses that relate cluster membership to one or more external validation variables.

- Alternatively, criteria of different aspects of internal validation can be considered, such as within-cluster homogeneity, between-cluster separation, and similarity of cluster members to their cluster centroid (Hennig, 2015b, 2015c; Milligan, 1981a). A broad range of measures exists for this purpose, which can optionally further be combined (Hennig, 2015a). Examples include the average silhouette width (Rousseeuw, 1987) or a fuzzy extension of it (Campello & Hruschka, 2006), which, in a partitioning context, measures how similar objects are to their own cluster compared with other clusters, and the cophenetic correlation coefficient (Sokal & Rohlf, 1962), which, in a hierarchical clustering context, may also be used to measure the relation between the ultrametric implied by the hierarchical clustering and the proximity data from which this clustering was derived. Admittedly, classical internal validity measures have recently been criticized with the claim that they often promote clusterings that match expert knowledge poorly (Gagolewski et al., 2021). This further also led to proposals of novel such indices (e.g., density-based measures that should also work well in case of clusters with arbitrary shapes: see Tavakkol et al., 2022).

A second group of criteria related to the goal of the data analysis shows up in case of a primary goal other than the identification of a clustering. Examples of measures include the proportion of explained criterion variance associated with prediction models and goodness-of-fit measures of (biological, psychological, …) process models that involve some kind of latent heterogeneity as captured by a clustering.

3.3.2 | Recommendations

- Think very carefully about the nature of the quality criterion/criteria that you choose, with different criteria possibly implying different optimal clusterings (Hennig, 2015b, 2015c). Choose the criteria in an a priori way, before any comparative data analysis. Make your choice explicit along with a justification of it. Consider multiple criteria if appropriate, optionally with an index of their relative importance.
- Select suitable measures for the chosen criteria, meaning that:
  - the measures should capture the chosen target criterion/criteria;
  - the application of the measures in question in the context under study should be technically correct (e.g., the use of the adjusted Rand index as a measure of cluster recovery only makes sense within the context of partitions);
  - the values of the chosen measures should be comparable across all benchmarking data sets, if necessary after some kind of normalization.
- Examine the interrelations between the performance regarding different criteria whenever appropriate. For example, it may be desirable to examine the relation between recovery of the true number of clusters and recovery of the true cluster memberships, with recovery of the number of clusters in some cases coming with a worse recovery of the memberships.

3.4 | Analysis, results, and discussion

3.4.1 | Issues

Unconditional conclusions on the performance rank order of the methods included in a benchmarking study with simulated data sets may result from: (1) a consistent rank order across all combinations of individual data sets and criteria included in that study, or (2) a consistent rank order across all data types and performance criteria, after averaging across all data sets (or replications) within each data type, or (3) a rank order after averaging across all data sets and criteria under study. Regarding the second and third option, averaging across data sets may in principle require a suitable framework for the population of data sets under study; averaging across criteria is even more challenging because of obvious commensurability problems with different criteria.—Unfortunately, "convenience averaging," just ignoring the raised issues is ubiquitous.—Optionally, the null hypothesis of no differences can be tested in terms of the main effect of the within-subject method factor in a, possibly factorial, analysis of variance, in which data sets act as experimental units. Alternatively, in case of comparability or commensurability obstacles, some type of consensus ranking of the methods could be made up, with consensus referring to an aggregation across data types and/or criteria, for example, based on Kemeny's axiomatic framework (see also Eugster, 2011).

All of the above can also be applied to draw unconditional conclusions from benchmarking studies with empirical data sets (Boulesteix, Stierle, & Hapfelmeier, 2015). Obviously, in the latter case, critical ANOVA assumptions such as random sampling from a population of real data sets of interest and independent and identically distributed data can be expected to be violated, with, for example, the sampling process most probably being subjected to several kinds of selection bias. This needs to be acknowledged. Still, classical power calculations may be meaningful in benchmarking studies based on real data sets, as they may provide a useful order of magnitude for the required sample sizes and inferential error rates (Boulesteix, Stierle, & Hapfelmeier, 2015).

With regard to *conditional conclusions*, in the absence of comparability and commensurability problems, benchmarking data can be subjected to (repeated measures) analyses of variance. Conditional conclusions then would be implied by method by data generation and/or criterion factor interactions. (At this point, one should take into account that testing interactions in a reliable way typically requires larger sample sizes than testing main effects.) In the presence of comparability or commensurability problems, one may look at the interactions in question in a more informal, descriptive way.

Finally, all types of conclusions drawn from a benchmarking study only hold conditional on the scope of that study in terms of the set of methods, data types, and performance criteria included in it. Moreover, whereas within a study, with a limited number of data sets, data types, and criteria, a consistent rank order of performance may show up, universal winning methods across all possible data sets, data types and criteria can be shown not to exist, as for a given data set different quality criteria may imply different optimal clustering solutions (Hennig, 2015b, 2015c).

3.4.2 | Recommendations

- Disclose full information on the code and results of the benchmarking analysis, making use, whenever possible, of platforms such as GitHub or Gitlab, for the purpose of reproducibility.
- Regarding unconditional statements about performance rank orders of methods:
  - Always investigate explicitly whether overall between-method differences are qualified by sizeable between-data type or between-criterion differences.
  - Whenever unconditional statements are reported, make clear whether or not they are based on some kind of averaging, and, if yes, on which one.
  - As means can be dominated by a few situations in which methods do very badly, consider, next to averages, more comprehensive representations of the distribution of results, for example, by means of boxplots.
- If possible, examine main and interaction effects by means of a (repeated measures) factorial analysis of variance, given its obvious inferential advantages. When doing so, take care of an adequate reporting of effects:
  - Go beyond significance levels, which are often meaningless because of large sample sizes in many simulation studies; report effect sizes (of, e.g., an η^2-type), and consider using a threshold on them to discuss only substantial effects.
  - Go beyond main effects and also look at interactions.
  - Inspect the contents of the found effects, test contrasts whenever needed to clarify the patterns involved in them, and provide an insightful reporting of such patterns; custom-made graphical displays can be helpful in the latter.
- Properly address developers of methods:
  - Provide insight into the reasons for the found effects in terms of underlying mechanisms, relations between characteristics of methods and data, and so on.
- Properly address practitioners:
  - Explain which method gives the best results for which cluster concepts, for which criteria, and for which types of data. Qualitative or disordinal method by data characteristic interactions should be highlighted insofar they imply that the best method differs across criteria and/or data types. Furthermore, if the latter type of interactions shows up, consider to look for an optimal data-analytic regime, that is to say, a decision rule that specifies, given some criterion measure, for which data types which methods are to be preferred (Doove et al., 2017).
  - From a pragmatic viewpoint, one should focus on recommendations in terms of data characteristics known to the researcher when carrying out a cluster analysis in practice, outside the context of Monte Carlo simulations (such as, e.g., data size); in case unknown characteristics (such as, e.g., the true number of clusters) appear to play a critical role in disordinal method by data characteristic interactions, one may try to replace them by proxies that are known in data-analytic practice (Doove et al., 2017).

- Do not overgeneralize obtained results. Never ever forget to mention in the final conclusion the restricted scope of the problem that has been looked at in the study in terms of selected methods, (types of) data sets, and specific performance criteria being used; highlight the limitations of this scope.

3.5 | Authorship

3.5.1 | Issues

The relationship between the author(s) of the benchmarking study and the authorship of the methods investigated in this study deserves special attention. Authors who propose new methods are in practice typically obliged to show that these methods outperform competitors in at least some situations, as a kind of existential justification of these methods. This, however, may also be looked at as an instance of publication bias, in that it is difficult to publish a paper on a new method that does not outperform competitors (Boulesteix, Stierle, & Hapfelmeier, 2015). Moreover, in practice, the need for evidence to justify the existence of a newly proposed method may degenerate into some kind of less warranted over-optimism, for example, in terms of reports of benchmarking evidence that this method uniformly outperforms a number of competitors for all criteria and all data types under study. This is further evidenced by a survey conducted by Boulesteix et al. (2013) of articles published in seven high-ranked computational science journals. From this survey it appears that papers in which new methods were proposed very often identify these methods as winner in benchmarking comparisons. In the same vein, Ullmann et al. (2022, 2023) illustrated several mechanisms that may lead to over-optimistic evaluations of newly introduced clustering methods (see also Nießl et al., 2022).

3.5.2 | Recommendations

- Disclose in benchmarking studies any possible conflicts of interest, such as, for example, a vested interest in some of the methods under study.
- "Neutral" comparison studies by authors without vested interests and, for the authorship as a collective, ideally with approximately the same level of familiarity with all methods under study (Boulesteix et al., 2017), should be especially encouraged. This further implies that journals should welcome reports of them and should beware of any possible publication biases against neutral comparison studies of existing methods (Boulesteix et al., 2013). In cases where foils, that is, cases in which a method does not work, were missing in earlier work, it would be helpful if neutral comparison studies could contribute some of these.

4 | A FEW ILLUSTRATIVE EXAMPLES

In this section we will briefly discuss a few illustrative examples of benchmarking studies in the clustering area. Rather than providing self-contained summaries of the studies in question, we will highlight a few distinctive features of each of them.

4.1 | Steinley (2003)

Steinley (2003) presented an impactful comparison of three widespread commercial software packages and one multistart MATLAB routine for k-means analysis, one of the most basic and prototypical methods for cluster analysis. In this study, he addressed a classical problem in the optimization of overall objective or loss functions in many methods of cluster analysis, namely the ubiquitous occurrence of locally optimal solutions. For this purpose, he made use of two empirical data sets and synthetically simulated data on the basis of a full factorial design. Criteria included the value of the (error sum of squares) loss function for the empirical data sets, and recovery of the true clustering for the simulated data sets (measured via the adjusted Rand index). The comparison suggested that the commercial packages most often return locally optimal solutions or solutions that imply a worse recovery of the underlying true clustering than those produced by a suitable multistart procedure.

4.2 | Schepers et al. (2006)

Schepers et al. (2006) set up a four-part benchmarking study to investigate the sensitivity of 49 different alternating least squares (ALS) partitioning algorithms for multiway data to local optima. The 49 algorithms were obtained by orthogonally combining seven types of initializations with seven combinations of an ALS-scheme and a procedure to deal with empty clusters. As primary criterion measure, the authors calculated for each data set whether or not an algorithm attained the optimal value of the least squares loss function that was found across all analyses of that data set (as proxy for the globally optimal loss value for that data set). Interestingly, the rank order of the algorithms found in a study with synthetically simulated data sets (Part 1) did not correspond to the rank order of the same algorithms when applied to empirical data (Part 2). Parametric bootstrap tests (Part 3) revealed that assumptions of the stochastic model underlying the data-generating mechanism used in Part 1 were violated in the empirical data of Part 2. In a new study with simulated data using a modified data-generating mechanism (Part 4), this hypothesis was confirmed.

4.3 | Steinley and Brusco (2008)

Steinley and Brusco (2008) set up an extensive comparison of eight prespecified combinations of a procedure for model-based or non-model-based clustering with a variable selection procedure (without orthogonally combining clustering and variable selection procedures). For this comparison, they made use of benchmarking data obtained from synthetic simulations with OCLUS according to a full factorial design with 6804 conditions and three replications per cell; this yielded 20,412 simulated data sets. Subsequently, they applied all eight methods under study to each data set (under the assumption that the true number of clusters was known), which resulted in 163,296 clustering runs. Recovery of the subset of truly relevant variables was evaluated in terms of precision and recall, and recovery of the true cluster membership in terms of the adjusted Rand index. Recovery data were further subjected to factorial multivariate analyses of variance, along with a careful monitoring of η^2-type effect sizes.

4.4 | Arbelaitz et al. (2013)

In a neutral benchmarking study, Arbelaitz et al. (2013) compared 30 internal validity indices. Given a collection of partitions with different numbers of clusters as obtained from subjecting a given data set (with a known underlying true partition) to some given clustering method, the key performance aspect focused on was whether or not an internal validity index under study attained its optimal value for the partition that was most similar to the true partition (in terms of some partition similarity index such as the adjusted Rand index). Unlike in many benchmarking studies that evaluate performance in terms of recovery of the true number of clusters, Arbelaitz et al. (2013) left room for the partition that is most similar to the true partition to comprise a number of clusters that differed from the true number of clusters. They further made use of both synthetically simulated benchmarking data (according to a full factorial design) and real benchmarking data sets with a known truth. Each data set was subjected to three different clustering algorithms, and partition similarity was evaluated with three different measures (two of which were of the object pair counting type and one of the information-theoretic type). In the discussion, the authors paid attention to various interactions that showed up in the results.

4.5 | Anderlucci and Hennig (2014)

Anderlucci and Hennig (2014) compared two approaches for clustering categorical data that are based on totally different conceptual frameworks: partitioning around medoids (Kaufman & Rousseeuw, 1990), which attempts to produce homogeneous clusters with low within-cluster dissimilarities, and latent class modeling (Vermunt & Magidson, 2002), which tries to estimate an underlying mixture model. In the study, which relied on synthetically simulated data according to a full factorial design, two different criteria were used to evaluate each analysis result, one corresponding to each approach: the average silhouette width for finding clusters that are homogeneous and well-separated from each other in terms of dissimilarities, and the adjusted Rand index to measure the recovery of the clusters that were true according to the underlying mixture model. One result was that the latent class approach was in most situations surprisingly competitive regarding the average silhouette width, which implies that the "home advantage" of partitioning around medoids for this criterion was less outspoken than expected.

4.6 | Rossbroich et al. (2022)

Rossbroich et al. (2022) addressed the (underinvestigated) issue of selecting the number of clusters in an overlapping clustering model (ADPROCLUS: Depril et al., 2008; Mirkin, 1987). For this purpose, they proposed and compared 13 model selection strategies, 11 of which were (minor or major) adaptations of similar strategies for partitioning models, and 2 of which were crossvalidation-based. For the comparison, the authors made use of benchmarking data obtained from synthetic simulations with a full factorial design. Critical criterion measures pertained to recovery of the true number of clusters in terms of accuracy (whether or not a given strategy yielded for a given dataset the true number of clusters) and precision (i.e., the absolute difference between the true and estimated numbers of clusters). Results were subjected to (repeated measures) analyses of variance. Particularly strong aspects of the study included a careful discussion and formulation of conclusions, with cautionary notes about the performance of all methods (including the marginally best one) in case of data sets with fewer variables and data sets with higher noise levels, about the occurrence of local minima, and about a possible lack of generalizability of the found results to larger data sets than the ones used in the benchmarking study.

5 | CONCLUSION

The primary take home message to investigators of clustering methods is that we think it is desirable to invest in benchmarking. This should not only be an absolute requisite when proposing new methods, but should likewise be taken care of in follow up research, particularly also by authors who were not involved in the development of the methods under study.

If we further analyze the specific recommendations for good research practices listed in the present paper, we may conclude that many of them are relevant for benchmarking in statistics, machine learning, and data mining in general. Examples include:

- Do include in any benchmark attempt known strong competitors for the target methods under study.
- Orthogonally manipulate aspects of methods and characteristics of simulated benchmark data whenever possible.
- Make full information available on methods (code, values of tuning parameters etc.) and benchmark data in view of full reproducibility.
- Make use of a suitable combination of simulations and empirical data when creating a collection of benchmark data sets.
- Take care of fairness towards all methods involved when setting up benchmark comparisons.
- Always examine whether overall between-method differences in a benchmarking study are qualified by sizeable between-data type or between-criterion differences.
- Disclose any possible conflicts of interest.
- Journals should especially welcome "neutral" comparisons by authors without vested interests in any of the methods under study.

Apart from this, we also listed several specific recommendations that are more typical for the clustering domain. Perhaps the most fundamental of these is to go for a deep reflection of what constitutes, within a particular context, and given particular research questions and aims, a good or desirable clustering. The result of this may have far-reaching consequences for the choice of relevant methods, benchmark data sets, and evaluation criteria. Much of this fundamental recommendation may be linked to the lack of an obvious metric for a performance evaluation of methods for unsupervised statistical learning in general and for clustering in particular. Such a lack could be perceived as a bottleneck. However, it could also be perceived as an opportunity and even as a blessing, which could finally lead to insights that may be inspiring for the clustering domain as well as for many other subdomains of statistics, machine learning, and data mining (Box 1).

BOX 1 Recommendations for method benchmarking in statistics, machine learning, carriage return and data mining

General recommendations:

- Do include in any benchmark attempt known strong competitors for the target methods under study.
- Orthogonally manipulate aspects of methods and characteristics of simulated benchmark data whenever possible.
- Make full information available on methods (code, values of tuning parameters, etc.) and benchmark data in view of full reproducibility.
- Make use of a suitable combination of simulations and empirical data when creating a collection of benchmark data sets.
- Take care of fairness towards all methods involved when setting up benchmark comparisons.
- Always examine whether overall between-method differences in a benchmarking study are qualified by sizeable between-data type or between-criterion differences.
- Disclose any possible conflicts of interest.
- Journals should especially welcome "neutral" comparisons by authors without vested interests in any of the methods under study.

Specific recommendation for benchmarking in cluster analysis:

- Go for a deep reflection of what constitutes, within a particular context, and given particular research questions and aims, a good or desirable clustering.

AUTHOR CONTRIBUTIONS

Iven Van Mechelen: Conceptualization (lead); funding acquisition (equal); methodology (equal); writing – original draft (lead); writing – review and editing (equal). **Anne-Laure Boulesteix:** Conceptualization (supporting); funding acquisition (equal); methodology (equal); writing – original draft (supporting); writing – review and editing (equal). **Rainer Dangl:** Conceptualization (supporting); methodology (equal); writing – original draft (supporting); writing – review and editing (equal). **Nema Dean:** Conceptualization (supporting); methodology (equal); writing – original draft (supporting); writing – review and editing (equal). **Christian Hennig:** Conceptualization (supporting); funding acquisition (equal); methodology (equal); writing – original draft (supporting); writing – review and editing (equal). **Friedrich Leisch:** Conceptualization (supporting); methodology (equal); writing – original draft (supporting); writing – review and editing (equal). **Douglas Steinley:** Conceptualization (supporting); methodology (equal); writing – original draft (supporting); writing – review and editing (equal). **Matthijs Warrens:** Conceptualization (supporting); methodology (equal); writing – original draft (supporting); writing – review and editing (equal).

ACKNOWLEDGMENTS

The authors thank Isabelle Guyon and Theresa Ullmann for helpful comments on earlier versions of this paper.

FUNDING INFORMATION

The work on this paper has been supported in part by the Research Foundation – Flanders (grants G080219N and K802822N to Iven Van Mechelen), by the Research Fund of KU Leuven (grant C14/19/054, Co-PI: Iven Van Mechelen), by the German Research Foundation (grants BO3139/7–1 and /9-1 to Anne-Laure Boulesteix), by the Bundesministerium für Bildung und Forschung (grant 01IS18036A, Co-PI: Anne-Laure Boulesteix) and by the Engineering and Physical Sciences Research Council (grant EP/K033972/1 to Christian Hennig).

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Iven Van Mechelen https://orcid.org/0000-0002-9917-8540
Anne-Laure Boulesteix https://orcid.org/0000-0002-2729-0947
Rainer Dangl https://orcid.org/0000-0001-8015-8083
Nema Dean https://orcid.org/0000-0002-5080-2517
Christian Hennig https://orcid.org/0000-0003-1550-5637
Friedrich Leisch https://orcid.org/0000-0001-7278-1983
Douglas Steinley https://orcid.org/0000-0001-9900-5028
Matthijs J. Warrens https://orcid.org/0000-0002-7302-640X

RELATED WIREs ARTICLES

Benchmarking in classification and regression
Method evaluation, parameterization, and result validation in unsupervised data mining: A critical survey
Validation of cluster analysis results on validation data:A systematic framework

REFERENCES

Ackerman, M., Ben-David, S., Brânzei, S., & Loker, D. (2021). Weighted clustering: Towards solving the user's dilemma. *Pattern Recognition*, *120*, 108–152. https://doi.org/10.1016/j.patcog.2021.108152

Anderberg, M. R. (1973). *Cluster analysis for applications*. Academic Press. https://doi.org/10.1016/C2013-0-06161-0

Anderlucci, L., & Hennig, C. (2014). Clustering of categorical data: A comparison of a model-based and a distance-based approach. *Communications in Statistics – Theory and Methods*, *43*, 704–721. https://doi.org/10.1080/03610926.2013.806665

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*(1), 243–256. https://doi.org/10.1016/j.patcog.2012.07.021

Baker, F. B. (1974). Stability of two hierarchical grouping techniques case I: Sensitivity to data errors. *Journal of the American Statistical Association*, *69*(346), 440–445. https://doi.org/10.1080/01621459.1974.10482971

Bezdek, J. C., Keller, J., Raghu, K., & Pal, N. R. (2005). *Fuzzy models and algorithms for pattern recognition and image processing*. Springer. https://doi.org/10.1007/b106267

Boulesteix, A.-L. (2015). Ten simple rules for reducing overoptimistic reporting in methodological computational research. *PLoS Computational Biology*, *11*(4), 1–6. https://doi.org/10.1371/journal.pcbi.1004191

Boulesteix, A.-L., Hable, R., Lauer, S., & Eugster, M. J. A. (2015). A statistical framework for hypothesis testing in real data comparison studies. *The American Statistician*, *69*, 201–212. https://doi.org/10.1080/00031305.2015.1005128

Boulesteix, A.-L., & Hatz, M. (2016). Benchmarking for clustering methods based on real data: A statistical view. In F. Palumbo, A. Montanari, & M. Vichi (Eds.), *Data science: Studies in classification, data analysis and knowledge organization* (pp. 73–82). Springer. https://doi.org/10.1007/978-3-319-55723-6_6

Boulesteix, A.-L., Lauer, S., & Eugster, M. J. A. (2013). A plea for neutral comparison studies in computational sciences. *PLoS One*, *8*(4), e61562. https://doi.org/10.1371/journal.pone.0061562

Boulesteix, A.-L., Stierle, V., & Hapfelmeier, A. (2015). Publication bias in methodological computational research. *Cancer Informatics*, *14*(S5), 11–19. https://doi.org/10.4137/CIN.S30747

Boulesteix, A.-L., Wilson, R., & Hapfelmeier, A. (2017). Towards evidence-based computational statistics: Lessons from clinical research on the role and design of real-data benchmark studies. *BMC Medical Research Methodology*, *17*, 138. https://doi.org/10.1186/s12874-017-0417-2

Breckenridge, J. N. (1989). Replicating cluster analysis: Method, consistency, and validity. *Multivariate Behavioral Research*, *24*, 147–161. https://doi.org/10.1207/s15327906mbr2402_1

Brusco, M., & Steinley, D. (2007). A comparison of heuristic procedures for minimum within-cluster sums of squares partitioning. *Psychometrika*, *72*, 583–600. https://doi.org/10.1007/s11336-007-9013-4

Campello, R. J. G. B., & Hruschka, E. R. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, *157*, 2858–2875. https://doi.org/10.1016/j.fss.2006.07.006

Coraggio, L., & Coretto, P. (2023). Selecting the number of clusters, clustering models, and algorithms: A unifying approach based on the quadratic discriminant score. *Journal of Multivariate Analysis*, *196*, 105181. https://doi.org/10.1016/j.jmva.2023.105181

Costa, E., Papatsouma, I., & Markos, A. (2022). Benchmarking distance-based partitioning methods for mixed-type data. *Advances in Data Analysis and Classification*. https://doi.org/10.1007/s11634-022-00521-7

Dangl, R., & Leisch, F. (2017). On a comprehensive metadata framework for artificial data in unsupervised learning. *Archives of Data Science Series A*, *2*(1), 16. https://doi.org/10.5445/KSP/1000058749/22

Depril, D., Van Mechelen, I., & Mirkin, B. G. (2008). Algorithms for additive clustering of rectangular data tables. *Computational Statistics & Data Analysis*, *52*, 4923–4938. https://doi.org/10.1016/j.csda.2008.04.014

Depril, D., Van Mechelen, I., & Wilderjans, T. F. (2012). Low dimensional additive overlapping clustering. *Journal of Classification*, *29*, 297–320. https://doi.org/10.1007/s00357-012-9112-5

Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics*, *11*, 385–388. https://doi.org/10.1093/biostatistics/kxq028

Doove, L. L., Wilderjans, T. F., Calcagnì, A., & Van Mechelen, I. (2017). Deriving optimal data-analytic regimes from benchmarking studies. *Computational Statistics & Data Analysis*, *107*, 81–91. https://doi.org/10.1016/j.csda.2016.10.016

Dubes, R., & Jain, A. K. (1976). Clustering techniques: The user's dilemma. *Pattern Recognition*, *8*(4), 247–260. https://doi.org/10.1016/0031-3203(76)90045-5

Dubes, R., & Jain, A. K. (1979). Validity studies in clustering methodologies. *Pattern Recognition*, *11*(4), 235–254. https://doi.org/10.1016/0031-3203(79)90034-7

El Abbassi, M., Overbeck, J., Braun, O., Calame, M., van der Zant, H. S. J., & Perrin, M. L. (2021). Benchmark and application of unsupervised classification approaches for univariate data. *Communications on Physics*, *4*, 50. https://doi.org/10.1038/s42005-021-00549-9

Eugster, M. J. A. (2011). Benchmark experiments: A tool for analyzing statistical learning algorithms [Unpublished doctoral dissertation]. Ludwig-Maximilians-Universität München.

Fisher, L., & van Ness, J. W. (1971). Admissible clustering procedures. *Biometrika*, *58*(1), 91–104. https://doi.org/10.1093/biomet/58.1.91

Flynt, A., & Dean, N. (2016). A survey of popular R packages for cluster analysis. *Journal of Educational and Behavioral Statistics*, *41*(2), 205–225. https://doi.org/10.3102/1076998616631743

Fränti, P., Rezaei, M., & Zhao, Q. (2014). Centroid index: Cluster level similarity measure. *Pattern Recognition*, *47*, 3034–3045. https://doi.org/10.1016/j.patcog.2014.03.017

Friedrich, S., & Friede, T. (2023). On the role of benchmarking data sets and simulations in method comparison studies. *Biometrical Journal*, *00*, e2200212. https://doi.org/10.1002/bimj.202200212

Fütterer, C., & Augustin, T. (2021). Internal validation of unsupervised clustering following an association accuracy heuristic. In 2021 IEEE international conference on bioinformatics and biomedicine (BIBM) (pp. 2201–2210). https://doi.org/10.1109/BIBM52615.2021.9669782

Gagolewski, M. (2022). A framework for benchmarking clustering algorithms. *SoftwareX*, *20*, 101270. https://doi.org/10.1016/j.softx.2022.101270

Gagolewski, M., Bartoszuk, M., & Cena, A. (2021). Are cluster validity measures (in)valid? *Information Sciences*, *581*, 620–636. https://doi.org/10.1016/j.ins.2021.10.004

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, *3*, 5–48. https://doi.org/10.1007/BF01896809

Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Massive Analysis Quality Control (MAQC) Society Board of Directors, Shraddha, T., Kusko, R., Sansone, S. A., Tong, W., Wolfinger, R. D., Mason, C. E., Jones, W., Dopazo, J., Furlanello, C., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., … Aerts, H. J. W. L. (2020). Transparency and reproducibility in artificial intelligence. *Nature*, *586*, E14–E16. https://doi.org/10.1038/s41586-020-2766-y

Hartigan, J. A. (1975). *Clustering algorithms*. Wiley.

Hartigan, J. A. (1985). Statistical theory in clustering. *Journal of Classification*, *2*(1), 63–76. https://doi.org/10.1007/BF01908064

Hennig, C. (2015a). Cluster validation by measurement of clustering characteristics relevant to the user. In C. H. Skiadas & J. R. Bozeman (Eds.), *Data analysis and applications 1: Clustering and regression, model-estimating, forecasting and data mining* (Vol. 2, pp. 1–24). Wiley. https://doi.org/10.1002/9781119597568.ch1

Hennig, C. (2015b). Clustering strategy and method selection. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 703–730). Chapman & Hall. https://doi.org/10.1201/b19706

Hennig, C. (2015c). What are the true clusters? *Pattern Recognition Letters*, *64*, 53–62. https://doi.org/10.1016/j.patrec.2015.04.009

Hennig, C. (2018). Some thoughts on simulation studies to compare clustering methods. *Archives of Data Science, Series A*, *5*(1), 1–21. https://doi.org/10.5445/KSP/1000087327/24

Hennig, C. (2022). An empirical comparison and characterisation of nine popular clustering methods. *Advances in Data Analysis and Classification*, *16*, 201–229. https://doi.org/10.1007/s11634-021-00478-z

Hoffmann, F., Bertram, T., Mikut, R., Reischl, M., & Nelles, O. (2019). Benchmarking in classification and regression. *WIREs Data Mining and Knowledge Discovery*, *9*, e1318. https://doi.org/10.1002/widm.1318

Hofner, B., Schmidt, M., & Edler, L. (2016). Reproducible research in statistics: A review and guidelines for the biometrical journal. *Biometrical Journal*, *58*(2), 416–427. https://doi.org/10.1002/bimj.201500156

Hubert, L. (1974). Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. *Journal of the American Statistical Association*, *69*(347), 698–704. https://doi.org/10.1080/01621459.1974.10480191

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218. https://doi.org/10.1007/BF01908075

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science*, *359*, 725–726. https://doi.org/10.1126/science.359.6377.725

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

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

Javed, A., Lee, B. S., & Rizzo, D. M. (2020). A benchmark study on time series clustering. *Machine Learning with Applications*, *1*, 100001. https://doi.org/10.1016/j.mlwa.2020.100001

Kaufman, L., & Rousseeuw, P. J. (1990). Finding groups in data: An introduction to cluster analysis. Wiley. https://doi.org/10.1002/9780470316801

Leibniz, G. W. (1764). Nouveaux essais sur l'entendement humain. livre iv, chap. xvii.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1: Statistics, pp. 281–297). University of California Press.

Masoero, L., Thomas, E., Parmigiani, G., Tyekucheva, S., & Trippa, L. (2023). Cross-study replicability in cluster analysis. *Statistical Science*, *38*, 303–316. https://doi.org/10.1214/22-STS871

McLachlan, G. J., & Peel, D. (2000). Mixtures of factor analyzers. In P. Langley (Ed.), *Proceedings of the seventeenth international conference on machine learning* (pp. 599–606). Morgan Kaufmann.

Melnykov, V., Chen, W.-C., & Maitra, R. (2012). MixSim: R package for simulating data sets with pre-specified clustering complexity. *Journal of Statistical Software*, *51*(12), 1–25. https://doi.org/10.18637/jss.v051.i12

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, *45*, 325–342. https://doi.org/10.1007/BF02293907

Milligan, G. W. (1981a). A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, *46*, 187–199. https://doi.org/10.1007/BF02293899

Milligan, G. W. (1981b). A review of Monte Carlo tests of cluster analysis. *Multivariate Behavioral Research*, *16*(3), 379–407. https://doi.org/10.1207/s15327906mbr1603_7

Milligan, G. W. (1985). An algorithm for generating artificial test clusters. *Psychometrika*, *50*, 123–127. https://doi.org/10.1007/BF02294153

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*, 159–179. https://doi.org/10.1007/BF02294245

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, *5*, 181–204. https://doi.org/10.1007/BF01897163

Milligan, G. W., Soon, S. C., & Sokol, L. M. (1983). The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *1*, 40–47. https://doi.org/10.1109/TPAMI.1983.4767342

Mirkin, B. G. (1987). The method of principal clusters. *Automation and Remote Control*, *48*(10), 1379–1388.

Mishra, S., Monath, N., Boratko, M., Kobren, A., & McCallum, A. (2022). An evaluative measure of clustering methods incorporating hyperparameter sensitivity. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(7), 7788–7796. https://doi.org/10.1609/aaai.v36i7.20747

Nießl, C., Herrmann, M., Wiedemann, C., Casalicchio, G., & Boulesteix, A.-L. (2022). Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(2), e1441. https://doi.org/10.1002/widm.1441

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Pfitzner, D., Leibbrandt, R., & Powers, D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, *19*, 361–394. https://doi.org/10.1007/s10115-008-0150-6

Plato. (n.d.). (approx. 370 BC). Phaedrus.

Qiu, W.-L., & Joe, H. (2006). Generation of random clusters with specified degree of separation. *Journal of Classification*, *23*, 315–334. https://doi.org/10.1007/s00357-006-0018-y

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Rossbroich, J., Durieux, J., & Wilderjans, T. F. (2022). Model selection strategies for determining the optimal number of overlapping clusters in additive overlapping partitional clustering. *Journal of Classification*, *39*(2), 264–301. https://doi.org/10.1007/s00357-021-09409-1

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rubin, J. (1967). Optimal classification into groups: An approach for solving the taxonomy problem. *Journal of Theoretical Biology*, *15*(1), 103–144. https://doi.org/10.1016/0022-5193(67)90046-X

Schepers, J., Van Mechelen, I., & Ceulemans, E. (2006). Three-mode partitioning. *Computational Statistics & Data Analysis*, *51*, 1623–1642. https://doi.org/10.1016/j.csda.2006.06.002

Schiavo, R. A., & Hand, D. J. (2000). Ten more years of error rate research. *International Statistical Review*, *68*, 295–310. https://doi.org/10.1111/j.1751-5823.2000.tb00332.x

Shand, C., Allmendinger, R., Handl, J., Webb, A., & Keane, J. (2022). HAWKS: Evolving challenging benchmark sets for cluster analysis. *IEEE Transactions on Evolutionary Computation*, *26*, 1206–2022. https://doi.org/10.1109/TEVC.2021.3137369

Shireman, E., Steinley, D., & Brusco, M. J. (2017). Examining the effect of initialization strategies on the performance of mixture modeling. *Behavior Research Methods*, *49*, 282–293. https://doi.org/10.3758/s13428-015-0697-6

Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, *11*, 30–40. https://doi.org/10.2307/1217208

Steinley, D. (2003). Local optima in K-means clustering: What you don't know may hurt you. *Psychological Methods*, *8*, 294–304. https://doi.org/10.1037/1082-989x.8.3.294

Steinley, D. (2004). Standardizing variables in K-means clustering. In D. Banks, L. House, F. McMorris, P. Arabie, & W. Gaul (Eds.), *Classification, clustering, and data mining applications* (pp. 53–60). Springer. https://doi.org/10.1007/978-3-642-17103-1_6

Steinley, D., & Brusco, M. J. (2008). Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika*, *73*, 125–144. https://doi.org/10.1007/s11336-007-9019-y

Steinley, D., Brusco, M. J., & Hubert, L. J. (2016). The variance of the adjusted Rand index. *Psychological Methods*, *21*, 261–272.

Steinley, D., & Henson, R. (2005). OCLUS: An analytic method for generating clusters with known overlap. *Journal of Classification*, *22*, 221–250. https://doi.org/10.1007/s00357-005-0015-6

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *36*, 111–147. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Strobl, C., & Leisch, F. (2022). Against the "one method fits all data sets" philosophy for comparison studies in methodological research. *Biometrical Journal*, *00*, 1–8. https://doi.org/10.1002/bimj.202200104

Šulc, Z., & Řezanková, H. (2019). Comparison of similarity measures for categorical data in hierarchical clustering. *Journal of Classification*, *36*, 58–72. https://doi.org/10.1007/s00357-019-09317-5

Tavakkol, B., Choi, J., Jeong, M. K., & Albin, S. L. (2022). Object-based cluster validation with densities. *Pattern Recognition*, *121*, 108223. https://doi.org/10.1016/j.patcog.2021.108223

Thrun, M. C., & Stier, Q. (2021). Fundamental clustering algorithms suite. *SoftwareX*, *13*, 100642. https://doi.org/10.1016/j.softx.2020.100642

Thrun, M. C., & Ultsch, A. (2020). Clustering benchmark datasets exploiting the fundamental clustering problems. *Data in Brief*, *30*, 105501. https://doi.org/10.1016/j.dib.2020.105501

Ullmann, T., Beer, A., Hünemörder, M., Seidl, T., & Boulesteix, A.-L. (2023). Over-optimistic evaluation and reporting of novel cluster algorithms: An illustrative study. *Advances in Data Analysis and Classification*, *17*(1), 211–238. https://doi.org/10.1007/s11634-022-00496-5

Ullmann, T., Hennig, C., & Boulesteix, A.-L. (2022). Validation of cluster analysis results on validation data: A systematic framework. *WIREs Data Mining and Knowledge Discovery*, *12*, e1444. https://doi.org/10.1002/widm.1444

van Buuren, S., & Heiser, W. J. (1989). Clustering n objects into k groups under optimal scaling of variables. *Psychometrika*, *54*, 699–706. https://doi.org/10.1007/BF02296404

van der Hoef, H., & Warrens, M. J. (2019). Understanding information theoretic measures for comparing clusterings. *Behaviormetrika*, *46*, 353–370. https://doi.org/10.1007/s41237-018-0075-7

Van Mechelen, I., Bock, H.-H., & De Boeck, P. (2004). Two-mode clustering methods: A structured overview. *Statistical Methods in Medical Research*, *13*, 363–394. https://doi.org/10.1191/0962280204sm373ra

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. Hagenaars & A. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge University Press. https://doi.org/10.1017/CBO9780511499531

Warrens, M. J., & van der Hoef, H. (2022). Understanding the adjusted Rand index and other partition comparison indices based on counting object pairs. *Journal of Classification*, *39*, 487–509. https://doi.org/10.1007/s00357-022-09413-z

Watson, D. S. (2023). On the philosophy of unsupervised learning. *Philosophy & Technology*, *36*(2), 28. https://doi.org/10.1007/s13347-023-00635-6

Weber, L. M., Saelens, W., Cannoodt, R., Soneson, C., Hapfelmeier, A., Gardner, P. P., Boulesteix, A.-L., Saeys, Y., & Robinson, M. D. (2019). Essential guidelines for computational method benchmarking. *Genome Biology*, *20*, 125. https://doi.org/10.1186/s13059-019-1738-8

Wilderjans, T. F., Depril, D., & Van Mechelen, I. (2013). Additive biclustering: A comparison of one new and two existing ALS algorithms. *Journal of Classification*, *30*, 56–74. https://doi.org/10.1007/s00357-013-9120-0

Zimmermann, A. (2020). Method evaluation, parameterization, and result validation in unsupervised data mining: A critical survey. *WIREs Data Mining and Knowledge Discovery*, *10*, e1330. https://doi.org/10.1002/widm.1330

How to cite this article: Van Mechelen, I., Boulesteix, A.-L., Dangl, R., Dean, N., Hennig, C., Leisch, F., Steinley, D., & Warrens, M. J. (2023). A white paper on good research practices in benchmarking: The case of cluster analysis. *WIREs Data Mining and Knowledge Discovery*, *13*(6), e1511. https://doi.org/10.1002/widm.1511