An automated process for supporting decisions in clustering-based data analysis

José Antonio Bernabé-Díaz[a], Manuel Franco[b], Juana-María Vivo[b], Manuel Quesada-Martínez[c], Jesualdo T. Fernández-Breis[a,*]

[a]*Dept. Informática y Sistemas, Universidad de Murcia, IMIB-Arrixaca, Spain*
[b]*Dept. Statistics and Operations Research, University of Murcia, IMIB-Arrixaca, Spain*
[c]*Center of Operations Research (CIO), Miguel Hernández University of Elche, Spain*

Abstract

Background and objective: Metrics are commonly used by biomedical researchers and practitioners to measure and evaluate properties of individuals, instruments, models, methods, or datasets. Due to the lack of a standardized validation procedure for a metric, it is assumed that an adequate metric should exhibit a similar stochastic behavior in different datasets. There is an implicit assumption of homogeneity in the sets of resources to be evaluated, so a metric is assumed to exhibit the same behavior in different scenarios. The study of such stochastic behavior of a metric is the objective of this paper, since it would allow for assessing its reliability before drawing any conclusion about biomedical datasets.
Methods: We present a method to support in evaluating the stochastic behavior of quantitative metrics on datasets. Our approach assesses a metric by using clustering-based data analysis, and enhancing the decision-making process in the optimal classification. Our method assesses the metrics by applying two important criteria of the unsupervised classification validation are calculated on the clusterings generated by the metric, namely stability and goodness of the clusters. The application of our method is facilitated to biomedical researchers by our *evaluomeR* tool.

*Corresponding author: jfernand@um.es, phone +34 868884613
Email addresses: `joseantonio.bernabe1@um.es` (José Antonio Bernabé-Díaz), `mfranco@um.es` (Manuel Franco), `jmvivomo@um.es` (Juana-María Vivo), `mquesada@umh.es` (Manuel Quesada-Martínez), `jfernand@um.es` (Jesualdo T. Fernández-Breis)

Preprint submitted to Computer Methods and Programs in Biomedicine *December 28, 2021*

Results: The analytical power of our methods is shown in the results of the application of our method to analyze (1) the behavior of the impact factor metric for a series of journal categories; (2) which structural metrics provide a better partitioning of the content of a repository of biomedical ontologies, and (3) the heterogeneity sources in effect size metrics of biomedical primary studies.
Conclusions: The use of statistical properties such as stability and goodness of classifications allows for a useful analysis of the behavior of quantitative metrics, which can be used for supporting decisions about which metrics to apply on a certain dataset.

Keywords: Evaluation metrics, Clustering-based data analysis, Unsupervised classification, Structural metrics, Meta-analysis

1. Introduction

Biomedical researchers usually measure and evaluate the properties of individuals, instruments, models, methods, or datasets through quantitative or qualitative metrics. Metrics are applied for different purposes such as analysis, classification and ranking. Examples of metrics can be RNA quality metrics for the assessment of gene expression difference [1], ontology metrics [2], variable blood prefusion [3], validation of electronic healthcare data [4] and for machine learning [5]. New metrics are continuously being proposed in order to make evaluation processes objective and reproducible, and an example is the current development of metrics for assessing the fairness of datasets [6]. However, the lack of systematic evaluation workflows has been considered an issue in biomedical domains [7, 8].

The validation of metrics is not a standarized process and, in most cases, the creators of the metrics apply them to a series of resources. In particular, when the gold standard associated to a classification is available, some measurements have been used to evaluate the performance and accuracy of a metric classifier, e.g., see Moccia et al. [7], Vivo et al. [9] and Franco and Vivo [10]). However, the gold standard might be unavailable, which is frequent in practice. Thus, if the results are satisfactory, the metric is then accepted as an appropriate measurement instrument for a certain feature. As a consequence, the metric is systematically applied to new resources. In most cases, such evaluations do not analyze how reliable the metric is for evaluating a new set of resources. There is an implicit assumption of homogeneity in the sets of resources to be evaluated, so a metric is assumed to exhibit the same behavior in different scenarios. Heterogeneity has also been identified as a limitation for comparative studies [11]. To the best of our knowledge, it has been not sufficiently studied whether such shared behavior really holds.

For instance, in the field of research synthesis, meta-analyses combine the results of different studies to draw conclusions by assuming homogeneity in the primary studies [12, 13, 14, 15]. In most cases, a meta-analysis summarizes the results provided by each individual study. The summary is obtained by using a set of dependent variables or summary metrics. A traditional criticism to meta-analysis is that such an average view may not be representative of the individual studies due to the presence of heterogeneity in the primary studies [12].

In this work we describe an approach that aims at supporting biomedical researchers in analyzing the stochastic behavior of quantitative metrics based on an automated process which combines two validity criteria of unsupervised classification. By proceeding in this way, researchers will know if the datasets are homogeneous from the perspective provided by such a metric. If the stochastic behavior of the metric is dissimilar in the datasets, then the metric might not be the optimal one for the study. We believe that the stochastic behavior of a metric should be studied and its optimal configuration justified before drawing any conclusion about datasets. In order to facilitate such knowledge studies to the research community we have developed *evaluomeR*, which implements our approach.

Starting with the pre-computed measurements of metrics for a set of resources, *evaluomeR* can be used for assessing each metric. In our work reliability is assessed by applying two important criteria of the unsupervised classification validation are calculated on the clusterings generated by the metric, namely stability and goodness of the clusters. The stability refers to whether a meaningful cluster is more or less influenced by small variations in the data, which may be analyzed by bootstrap clustering [16]. The goodness of the clustering is related to the cohesion and separation of the clusters [17]. In detail, both validation features are described in Section 2. The classification of the instances reported from a metric is the result of applying an unsupervised partition algorithm with a number k of clusters which is often unknown [18]. Thus, a range of k values is required as an input parameter, arising the need for considering such a validation mechanism of the generated clusterings to select the most reliable stratification for each metric. Furthermore, when a metric is used in two or more different datasets or set of primary studies on the same topic, the most reliable stratification for such a metric might be obtained for different number k of groups, which can be interpreted as a finding of additional heterogeneity due to the instances and trial design of the datasets.

Therefore, our approach helps researchers in getting information about the reliability of the metrics and the characteristics of the datasets that they want to analyze. This information should be relevant for the selection of metrics and meta-analysis studies.

2. Methods

In this section, we first describe our analytic framework that can serve as a decision support tool in the evaluation of quantitative metrics. A general overview of the methodology implemented in it can be seen in Figure 1.

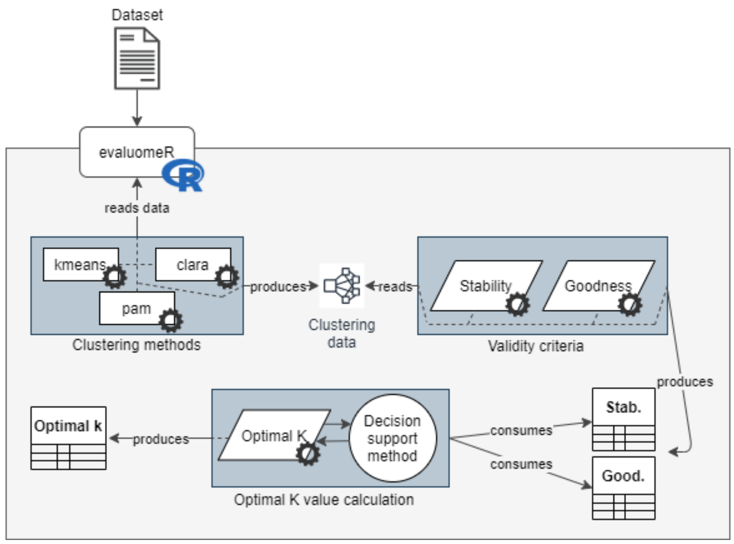

Figure 1: The *evaluomeR* overall architecture. Clustering-based data analysis is applied, and then validity criteria are calculated, so that the Optimal k module computes the optimal setting for the metric based on both criteria.

Clustering techniques such as *k-means* are used for unsupervised classification in order to perform class discovery, cluster analysis or unsupervised pattern recognition [19]. These clustering techniques consider data tuples as objects, which are then arranged into groups, or clusters, according to a distance matrix. However, the outputs of the unsupervised methods depend on the clustering algorithms used. In addition to *k-means*, our implemented method offers to users other clustering methods as Partitioning Around Medoids (PAM) or Clustering LARge Applications (CLARA) for their analysis.

2.1. Stability

Our method can evaluate the effect of small alterations on the data according to the stability analysis by means of bootstrap resamplings and the similarity between categories reported by the Jaccard coefficient [20], which is used as an external validation criterion when the gold standard is available.

This coefficient is also used to obtain the stability index by assessing the similarity between each category of the clustering generated on a metric and the most similar cluster in each bootstrapped clustering [16]. The stability values fall in the interval $[0, 1]$, and can be interpreted in terms of statistical stability degrees [21] as shown in Table 1:

Range	Category
$[0, 0.60)$	Unstable
$[0.60, 0.75]$	Doubtful
$(0.75, 0.85]$	Stable
$(0.85, 1]$	Highly stable

Table 1: Stability classification.

2.2. Goodness

This analysis supplies an internal validation measurement of the clustering based on how closely related q the instances in a category are, and how well-separated a category is from the rest of categories. We use the Silhouette width [17] as goodness index of the clusters, since it enables to compute and compare the quality of the clusters generated on a metric. More precisely, the Silhouette width estimates the similarity between a given instance and the rest of instances in the same cluster and the dissimilarity with the instances in the nearest neighboring cluster. The global goodness is the average Silhouette width value obtained on all the instances. These goodness values are in the range [-1, 1] and are interpreted as shown in Table 2 [22]:

Range	Clustering Structure
$[-1, 0.25)$	There is no substantial clustering structure
$[0.25, 0.50]$	The clustering structure is weak and could be artificial
$(0.50, 0.70]$	There is a reasonable clustering structure
$(0.70, 1]$	Strong clustering structure has been found

Table 2: Structure classification.

2.2.1. Optimal setting

In this section, we propose a method that allows to select automatically the optimal k value for a metric in a given dataset. It is based on the analysis of *evaluomeR* regarding stability and goodness of the clusters for a range of values of k, more concretely, on finding the optimal k setting based on the value of k_s, which provides the highest stability and the value of k_g, which provides the highest goodness. Note that each metric is analyzed independently:

- If $k_s = k_g$, then that value is the optimal number of clusters.
- If $k_s \neq k_g$, then additional criteria are needed. In this work, we propose the following criteria:
  - If both k_s and k_g provide at least stable classifications or both provide non stable classifications, the optimal number of clusters is the one with the largest Silhouette width, i.e., $k = k_g$.
  - If k_s provides at least stable and reasonable classifications and k_g does not provide stable classifications, then $k = k_s$.
  - If k_s provides at least stable classifications but less than reasonable, and k_g does not provide stable classifications, then if k_g provides an at least reasonable Silhouette width, then $k = k_g$. Otherwise, $k = k_s$.

For a set of metrics m_i, this criterion obtains the optimal number of clusters k_i for each metric m_i. Then, the metrics can be ranked by the stability and goodness obtained for their optimal number of clusters, thus enabling to make decisions about which one is the most suitable for evaluating the dataset depending on the data analysis requirements.

3. Results

In this section we present the main results of this work. First, our software tool *evaluome* will be described (see Section 3.1). Then, three use cases of its application will be presented (see Section 3.2).

3.1. evaluomeR

In this section we describe *evaluomeR*, and the functionality offered to different types of users. First, we describe the functionality included in the Bioconductor package *evaluomeR*. This R package permits to apply the *evaluomeR* methods in R environment in combination with other data analysis packages. Second, we describe the web portal, which permits the online execution of the methods and that is intended for non-programmers.

3.1.1. The evaluomeR package

The package *evaluomeR* provides R functions that implement the methods aforementioned, see Figure 1. The package *evaluomeR* v1.6.2 is available in Bioconductor 3.12 [23] and depends on the following packages: *fpc* [21], *cluster* [24], *corrplot* [25], *Rdpack* [26], *SummarizedExperiment* [27] and *MultiAssayExperiment* [28]. It requires R version 3.6 or higher to run. Other dependencies such as Bioconductor or CRAN R packages are automatically downloaded via Bioconductor install manager. The package has MIT license.

A summary of the functionality is provided next:

- '*stability*' and '*stabilityRange*': The package calculates the stability for a set of metrics for a single value for k or for range of values, and specifying the number of bootstrap replicates. By default, the functions calculate the stability indices with 100 bootstrap replicates and also generate stability plots.

- '*quality*' and '*qualityRange*': The package calculates the goodness of the clusters for a single value for k or for a range of values. By default, the functions calculate the goodness and also generate the plots of the Silhouette widths for the metrics.

- '*getOptimalKValue*': The functionality of the optimal setting is mentioned in Section 2.2.1. It takes into account the results of the stability of the metrics as well as the goodness of the clusters to compute the criterion of which is the best suitable k value. Additionally, this method reports the best k value considering only the stability or the goodness data independently.

- Additional plots: '*plotMetricsBoxplot*', '*plotMetricsCluster*', '*plotMetricsClusterComparison*', '*plotMetricsMinMax*' and '*plotMetricsViolin*'. The package generates four additional plots using the input data, so enabling a global analysis of the metrics: violin plots, boxplots, clustering of the set of metrics, and the min/max/sd of each metric.

3.1.2. The web portal

The *evaluomeR* portal [29] is a Shiny [30] application which permits general users to apply our method by proceeding as follows (see Figure 2):

- Input data: Upload a CSV file or select one of the examples provided by us. The names of the metrics must be provided in the first row of the file, which plays the role of header. Each column in the CSV file represents a metric and each row represents an instance in the dataset. The measurements of each metric are provided for each instance in the dataset.

- Output configuration: The user may select one of the four *evaluomeR* methods: *Stability*, *Quality*, *Correlations* and *Optimal K*. Every method, upon execution, shows output data tables and plots. Each one provides a download button where users can fetch the resulting data shown on the tables in CSV format. Moreover, plots are interactive and also downloadable.

- Execution configuration: The minimum and maximum number of clusters (k), which must be in the range [2,15] are set by the user. The user can also set the number of bootstrap replicates and the seed.

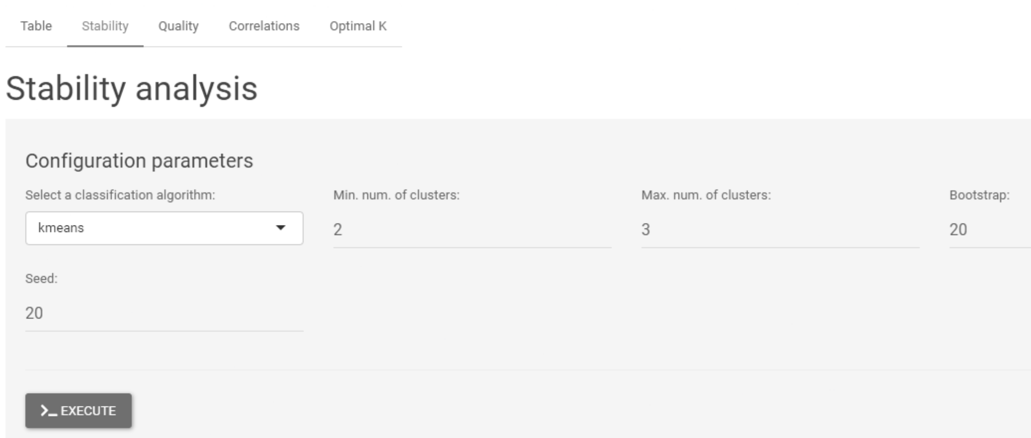

Figure 2: Screen snapshot of the *evaluomeR* portal.

3.2. Use cases

We illustrate the application of *evaluomeR* to support decisions in three use cases: (1) analysis of the behavior of the impact factor metric; (2) analysis of the behavior of nineteen metrics in ontology repositories, and (3) analysis of the behavior of effect sizes of primary studies. The source data of the first two use cases and the results of the three use cases are available at GitHub[1]. The source data of the third use case were extracted from the R package *metafor* [31].

3.2.1. Use case 1: bibliometric study

In recent years, the impact factor has been the most relevant bibliometric indicator for the quality of research journals. The impact factor is a metric whose value for a given journal depends on the number of papers published and the number of citations received by the papers published in the journal in a period of time. The impact factor is calculated by Clarivate Analytics and nearly every journal publishes it on its web page. Clarivate Analytics classifies each journal in a series of categories in the Journal Citations Report (JCR) and then, journals are ranked in such categories by quartiles. In some countries, the assessment of the scientific quality of the work of researchers is mostly determined by the ranking of the journal in which they publish. Those assessment schemes use sometimes tertiles and sometimes quartiles. In the last years, there have been criticisms to the use of the impact factor to evaluate the quality of research. Recently, it has been abandoned by Dutch universities for supporting promotion and hiring decisions [32]. Consequently, the behavior of the impact factor metric deserves to be studied, to determine which is the optimal number of clusters suggested by the category data. It should be noted that the optimal number of clusters may vary for different categories.

In this use case, we studied the series of impact factor data in the period 2016-20 for three JCR categories: "Computer Science, Artificial Intelligence" (CSAI), "Computer Science, Information Systems" (CSIS) and "Operations Research & Management Science" (ORMS). We analyzed the behavior of the metric per year and per category (Figures 3 and 4). In this case we had fifteen series of data, which were independently processed using *evaluomeR*.

Computer Science, Artificial Intelligence. Figure 3 shows the results of the application of *evaluomeR* to the metric *impact factor* for the category "Computer Science, Artificial Intelligence" (CSAI) for the years 2016, 2017, 2018, 2019 and 2020 in the k range [2,15]. Figure 3 (A) shows the stability of the metric across years, and Figure 3 (B) shows the goodness of the clusters generated by such metric.

[1] `https://github.com/neobernad/evaluomeR/tree/master/usecases`

In 2016, all the stability scores were in the range [0.60,1], meaning that the clusterings had at least reasonable structure. A highly stable clustering was obtained for $k = 2$ (0.921), which would mean that the journals in the category could be grouped in two categories. The stability for $k = 4$, that is, classification based on quartiles was 0.701, thus being a doubtful classification. However, the stability for $k = 5$ was higher and stable, 0.805. Figure 3 (B) shows the goodness of the clusters generated for k in the range [2,15]. Most of the Silhouette widths were in the range [0.50,1], meaning that they were not unstable. The unstable exceptions occurred when $k = 8$ (0.498) and $k = 9$ (0.491). The goodness for $k = 4$ was 0.562, thus having a reasonable structure. Again, the result for $k = 5$ was higher, 0.572. The best option for the 2016 data is to use two categories for classifying the journals. Since two could be considered a very reduced number of categories, we could state that the classification of the journals in five categories is more reliable than using quartiles. In the next studies, we did not take into account the results for $k = 2$.

In the 2017 case, the stability for $k = 4$ (0.948) was higher than for $k = 5$ (0.86), and $k = 3$ (0.961) was closer to the stability of $k = 4$. In terms of goodness, the result for $k = 3$ (0.617) was better than for $k = 4$ (0.607) or $k = 5$ (0.552). Hence, using tertiles would be the best option for the 2017 data. Regarding 2018, $k = 4$ provided the clustering with highest stability (0.797), however $k = 3$ (0.784) was also close to this high score. The largest width of the Silhouette was reached for $k = 3$ (0.612), therefore tertiles are again a suitable option.

The value $k = 4$ (0.924) achieved the highest stability for 2019, $k = 3$ (0.888) being the closest one. However, regarding the goodness, $k = 3$ (0.634) provided a higher Silhouette width than $k = 4$ (0.600), thus tertiles are suggested. Finally, 2020 data presented a similar behavior, where stability of $k = 4$ (0.899) outperformed $k = 3$ (0.770), but in terms of goodness $k = 3$ (0.683) produced a better result than $k = 4$ (0.585), therefore tertiles are again the suggested option.

Computer Science, Information Systems. Figure 3 shows the results of the study for the category "Computer Science, Information Systems" (CSIS). Figure 3 (A) shows the stability of the clusters generated for k in the range [2,15] for the 2016-2020 data.

For 2016, stable clusters were obtained for k between 3 and 6, with values ranging from a minimum of 0.769 ($k = 5$) and a maximum of 0.925 ($k = 3$).

The results of the goodness of the clusterings are shown in Figure 3 (B). The best result was for $k = 3$ (0.604), which means a reasonable clustering structure. Consequently, tertiles seem to be the best option.

In the case of 2017, we can see lower stability values with high stable clusters for k in $\{3, 6\}$. As for 2016, the largest Silhouette width was obtained for $k = 3$ (0.604), that is, reasonable structure. For higher values of k, the structure of the clustering was reasonable (< 0.70). Thus, septiles ($k = 7$) are the best option for 2017 with stability (0.766) and goodness (0.554), whereas $k = 3$ results in a lower stability (0.762) but a higher goodness (0.615). Regarding 2018, we obtained only one stable cluster for $k = 6$ (0.768), whilst the scores of $k = 3$ (0.697) and $k = 4$ (0.668) were significantly lower. The values of the Silhouette widths showed values suitable to a reasonable clustering structure, being $k = 3$ (0.609) the largest Silhouette, and $k = 6$ providing a goodness of 0.560. The usage of sextiles provided the best results in terms of reliability. In summary, we observed a similar behavior of the metric for the three years included in the study, and the optimal k is 3.

In 2019 the most stable classification was obtained with $k = 6$ (0.805), being $k = 4$ (0.799) the second most stable one. The goodness score for $k = 6$ (0.573) presented a reasonable clustering structure as well as for $k = 4$ (0.562), thus sextiles are the suggested option. On the other hand, for 2020 data, septiles would be the optimal partition as the value for stability in $k = 7$ (0.815) provided a highly stable classification and, additionally, the Silhouette width score for $k = 7$ (0.574) produced a reasonable clustering.

Operations Research & Management Science. For the 2016 data (see Figure 3), $k = 3$ provided the highest clustering stability (0.880), whereas the rest of the clusters provided a doubtful clustering structure. Regarding the goodness of the clusters, the best result was also obtained for $k = 6$, since the Silhouette width was 0.594, and $k = 3$ was close (0.583). The structure of the clusters was not strong for any k. Consequently, a classification based on tertiles seemed the best option. For the 2017 data, high stable clusters were only obtained for $k = 3$ (0.959). The structure of the clusters was reasonable for $k = 3$ (0.5923), this score being the second highest value, as $k = 11$ results in a Silhouette of 0.5927. Given these results, a classification based on tertiles seemed appropriate. For the 2018 data, the most stable cluster was obtained for $k = 3$ (0.845). The structure of the clusters was reasonable $k = 3$ (0.580). Given these results, a classification based on tertiles seemed the best decision. In summary, it seems that a classification based on tertiles

Table 3: Summary of the results of the impact factor use case. CSAI stands for 'Computer Science, Artificial Intelligence', CSIS for 'Computer Science, Information Systems' and ORMS for 'Operations Research & Management Science'

Category/Year	**2016**	**2017**	**2018**	**2019**	**2020**
CSAI	3	3	3	3	3
CSIS	3	7	6	6	7
ORMS	3	3	3	3	4

provided the most reliable clusters.

In the case of 2019, we obtain a highly stable clustering for $k = 3$ (0.981). The stability scores for the rest of the partitions are stable. The highest goodness value was obtained for $k = 3$ (0.605) and $k = 6$, hence a partition based on tertiles is recommended. For 2020 data, we also detected a high stability for $k = 3$ (0.922) although $k = 4$ (0.871) was nearby. Thus, the Silhouette width score determined the optimal k value. Concretely, $k = 4$ (0.606) was the reported one since it provided a higher value than $k = 3$ (0.560).

Table 3 summarizes the results for the three studies described in the previous subsections. For year and JCR category, each cell in the table includes the optimal k by applying the decision criterion described in Section 2.2.1. We can see that the optimal k was the same for the "Computer Science, Artificial Intelligence" category. Furthermore, the impact factor shown the same stochastic behavior for the "Operations Research & Management Science" category from 2016 to 2019. However, the impact factor was a different stochastic behavior for the three categories in 2020.

3.2.2. Use case 2: structural ontology metrics

Ontologies have gained popularity in the biological domain because of their four main properties. Ontologies provide (1) standard identifiers for classes and relations that represent the phenomena within a domain, (2) a vocabulary for a domain, (3) metadata providing the intended meaning of the classes and relations, (4) and machine-readable axioms and definitions that enable computational access to some aspects of the meaning of classes and relations [33]. There exist several repositories hosting biological ontologies, some of the most relevant being the OBO Foundry [34], AgroPortal [35], OntoBee [36], the Ontology Lookup Service [37], AberOWL [38], or NCBO BioPortal [39].

The use of metrics is common to describe properties of ontologies. Ontology metrics are used for measuring facets such as cohesion, the existence of multiple inheritance, or the richness of the ontology in terms of properties or comments for humans. Analyzing the general properties of the repositories of biological ontologies requires to combine the results by the metrics in the repositories under study. This can also be achieved by creating datasets that include the ontologies of those repositories. Despite the fact that some ontologies are included in more than one repository, some repositories are specific of particular subdomains. For example, AgroPortal is for the agriculture domain and the OBO Foundry is general for biology and biomedicine. Consequently, ontologies of different repositories might have different properties, which could imply different stochastic behavior of the metrics. This is why in this case study we analyzed the behavior of the 19 ontology structural metrics (see Table 4) included in the OQuaRE ontology quality framework [40] in two corpora of ontologies: AgroPortal and the OBO Foundry. 78 AgroPortal ontologies and 119 OBO Foundry ones constituted the datasets for this study. Both repositories have more ontologies but some ones failed to be retrieved by our automatic process.

In the next subsections, we describe first the behavior of the 19 metrics on the AgroPortal dataset, then on the OBO Foundry one and, finally, on the aggregated dataset. Our main aim in this use case was to identify which metrics are more appropriate for generalizing the findings on the particular repositories. In this use case, we used values of k in the range [2,6] for simplicity. Although it is shown in the figures, we did not take into account the results for $k = 2$ as an optimal value in the analysis for avoiding elementary dichotomous classifications. Given the number of metrics, we do not perform a detailed study of each metric, but justify the selections done of the optimal k value for each metric.

AgroPortal. Figure 5 shows the results of the study of the behavior of the 19 metrics on the AgroPortal dataset (AGRO) in terms of stability (A) and goodness (B) of the clusters. Next, we justify the optimal k for those metrics with different optimal value for stability and goodness:

- CROnto: $k_s = 6$ and $k_g = 3$. Both k values produce non-stable classifications. We select 3 as optimal since it provides higher Silhouette width, i.e., the clustering is more consistent.

- LCOMOnto: $k_s = 5$ and $k_g = 3$. Both k values provide stable classifications, thus we select 3 since it provides higher Silhouette width.
- NACOnto: $k_s = 3$ and $k_g = 6$. Both k values produce stable classifications, and 6 achieves higher Silhouette width.
- NOCOnto and TMOnto2: $k_s = 4$ and $k_g = 3$. Both k values produce stable classifications, we select 3 since it provides higher Silhouette width in both metrics.
- POnto: $k_s = 5$ and $k_g = 4$. Both k values produce stable classifications, and 4 achieves higher Silhouette width.
- PROnto and RROnto: $k_s = 3$ and $k_g = 4$. Both k values generate stable classifications, but 4 provides higher Silhouette width in both metrics.
- WMCOnto2: $k_s = 6$ and $k_g = 4$. Both k values generate strong Silhouette width, 6 produces a stable classification but 4 does not, then we use 6 as the optimal setting.

OBO Foundry. Figure 5 shows the results of the study of the behavior of the 19 metrics on the OBO Foundry dataset (OBO) in terms of stability and goodness of the clusters. Next, we justify the optimal k for those metrics with different optimal value for stability and goodness:

- CBOOnto, CBOOnto2 and NOMOnto: $k_s = 6$ and $k_g = 3$. Both k values provide stable classifications. We select 3 since it provides higher Silhouette width in these metrics.
- DITOnto: $k_s = 3$ and $k_g = 5$. Both k values generate reasonable Silhouette width, 3 produces a stable classification but 5 does not, then 3 is selected.
- NACOnto, RFCOnto and WMCOnto2: $k_s = 4$ and $k_g = 3$. Both k values produce stable classifications. We select 3 since it provides higher Silhouette width in these metrics.
- POnto: $k_s = 3$ and $k_g = 4$. Both k values generate reasonable Silhouette width, 3 produces a stable classification but 4 does not, then 3 is selected as the optimal setting.

Aggregated dataset. We repeat the same procedure on the aggregated dataset, which consists of both AgroPortal and OBO Foundry content. This study is also shown in Figure 5 as AGRO+OBO. Next, we justify the optimal k for those metrics with different optimal value for stability and goodness:

- AROnto: $k_s = 4$ and $k_g = 5$. Both k values provide stable classifications. We select 5 since it provides higher Silhouette width.
- CBOOnto and CBOOnto2: $k_s = 6$ and $k_g = 5$. Both k values produce non-stable classifications. We select 5 since it provides higher Silhouette width in both metrics.
- CROnto: $k_s = 6$ and $k_g = 3$. Both k values generate non-stable classifications, and 3 provides higher Silhouette width.
- DITOnto: $k_s = 3$ and $k_g = 5$. Both k values produce non-stable classifications, and 5 achieves higher Silhouette width.
- INROnto: $k_s = 6$ and $k_g = 4$. Both k values generate at least reasonable Silhouette width, 6 produces stable classification but 4 does not. Thus, we select 6 as the optimal setting.
- LCOMOnto: $k_s = 3$ and $k_g = 4$. Both k provide stable classifications, and 4 achieves higher Silhouette width.
- NACOnto: $k_s = 4$ and $k_g = 3$. Both k produce stable classifications. We select 3 since it provides higher Silhouette width.
- PROnto and RROnto: $k_s = 3$ and $k_g = 6$. the optimal k for stability is 3 and the one for goodness is 6. Both k generate stable classifications, and 6 provides higher Silhouette width in both metrics.
- WMCOnto: $k_s = 6$ and $k_g = 3$. Both k values produce stable classification, and we select 3 since it provides higher Silhouette width.

Table 5 summarizes the optimal value of k for each metric in the three datasets. There we can see that the metrics ANOnto, CROnto, NOCOnto, NOMOnto, RFCOnto, TMOnto2, and WMCOnto have the same stochastic behavior in the three datasets. CBOnto, CBOnto2, DITOnto and INROnto have the same stochastic behavior in the two individual datasets but different in the aggregated one. The metrics LCOMOnto, NACOnto, POnto, TMOnto and WMCOnto2 have the same stochastic behavior in the aggregated dataset and in one of the individual datasets. Finally, AROnto, PROnto and RROnto exhibit a different stochastic behavior in each dataset.

3.2.3. Use case 3: effect sizes of primary studies

As previously mentioned, meta-analysis is a statistical methodology for integrating the research results reported in a pool of published empirical studies on a particular topic. These combinations usually involve studies with differences in their design and conduct which can lead to heterogeneous outcomes [45]. This is why studying the presence of this variability in outcome measures emerges as a recurring issue in meta-analysis.

In this use case, we focused our efforts on demonstrating the value of our software tool provided and its usefulness for assisting in exploring and examining the sources of heterogeneity. Indeed, we used our automated process for clustering the studies combined in a meta-analysis to assess whether the effect sizes vary across the latent classes reported. By assuming that each study belongs to one of such classes, the iterative classification method implemented in [2] is based on the maximization of the within-class compactness and between-class separability of the studies. Along with the validation cluster criteria described previously, the best option of clustering reported by *evaluomeR* can help in identifying such underlying classes of studies leading to find features of the studies which enable to yield a more precise explanation of the exhibited heterogeneity in such outcome measures. This latent factor can be handled as a potential moderator of the overall results which is said to be an effect moderator. In addition, different effect size metrics are available (e.g. the standardized mean difference, the odds ratio, the correlation coefficient and so on) depending on the kind of study and data used in the primary studies (e.g. mean and standard deviation in two groups, binary outcomes or correlation). Therefore, to that end, we applied our automated process to three meta-analysis datasets from the R package *metafor* [31] to evaluate the moderating effect of the latent factor on different effect size metrics. Furthermore, we will examine these potential sources of within- and between-study heterogeneity reported by *evaluomeR* using the functions provided in the R package *metafor*.

Correlational data. To begin with, we recalled *dat.molloy2014* from *metafor* combining 16 primary studies used by Molloy et al. [46] for analyzing the correlation between the patient's levels of conscientiousness and medication adherence. This dataset consists of observed correlations, sample sizes of the studies, continuous and categorical variables such as mean age and methodological quality, which may be examined as moderators. By assuming that the studies were drawn from different populations, we conducted a meta-analysis under the random-effects model and the restricted maximum-likelihood (REML) estimator on the metric of Fisher's r-to-z transformed correlation coefficient. Converted back to Pearson's correlation, the point estimate expresses the average correlation which was equal to 0.150 (95% CI of 0.088 to 0.212, $p < 0.0001$) reflecting a significant modest relationship. The total amount of the residual heterogeneity τ^2 was 0.0081 ($SE = 0.006$), I^2 was 61.73% and the Q-test was 38.160 ($df = 15$, $p = 0.0009$). Moreover, there was no potential outlier in the studies combined in this meta-analysis [47]. Additionally, we performed a moderator analysis for methodological quality defined by the author on a scale from 1 (lower quality) to 4 (higher quality). The results provided evidence that methodological quality had a significant moderating effect ($Q(3) = 25.648$, $p < 0.0001$). Nevertheless, the estimated residual heterogeneity τ^2 only dropped to 0.0073 ($SE = 0.006$) with respect to the previous meta-analysis revealing that this moderator itself explains 9.93% of the total amount of the residual heterogeneity. In addition, the Q-test was 26.879 ($df = 13$, $p = 0.0129$) and $I^2 = 53.72\%$, which indicates that other moderators are influencing the correlation between conscientiousness and medication adherence. A customized forest plot generated from the results of this moderator analysis is presented in Figure 6A including the heterogeneity statistics within and between classes of effect size.

For our purpose, we conducted a moderator analysis for estimating whether the observed correlation can be explained by the classification reported by *evaluomeR*. To identify the underlying classes of studies, we first ran our automated process with the k value varying from 2 to 6. According to the validation criteria, the output revealed stable classifications both for $k = 2$ and $k = 4$, the second option being the best one since it provided higher Silhouette width score. This resulting latent factor was added in a mixed-effects model as a potential moderator supplying the output used for creating the forest plot represented in Figure 6B. The results reflected evidence that this optimal classification had a significant moderating effect ($Q(4) = 90.921$, $p < 0.0001$). The Q-test was no significant (1.470, $df = 12$, $p = 0.9999$) and $I^2 = 0.00\%$, suggesting that nearly 100% of the heterogeneity can be explained by including this latent factor in the model. For each latent class of effect size, the forest plot depicts the within-class heterogeneity statistics, which reported no evidence of heterogeneity. Furthermore, there was no relationship between conscientiousness and medication adherence (0.016, 95% CI of -0.037 to 0.068) in the latent class 1, whereas significant modest increases in the average correlation were found in the class 2 (0.257, 95% CI 0.140 to 0.374), in the class 3 (0.162, 95% CI of 0.113 to 0.212), and the class 4 (0.357, 95% CI of 0.231 to 0.482).

Mean differences. A second example showing the usefulness and effectiveness of our software tool to provide information about the heterogeneity of the datasets was carried out employing *dat.bangertdrowns2004*, taken from a meta-analysis on the outcome measures derived from 48 studies about the effectiveness of school-based writing-to-learn interventions on academic achievement [48]. Firstly, the random-effects model with the standardized mean difference included in *dat.bangertdowns2004* as effect size metric was used throughout. The point estimate was equal to 0.222 (95% CI of 0.132 to 0.312, $p < 0.0001$) which pointed out a higher mean level of academic achievement in the intervention group. The total amount of the residual heterogeneity τ^2 was 0.0499 ($SE = 0.020$), I^2 was 58.37% and the Q-test was 107.106 ($df = 47$, $p < 0.0001$). All the results reported from the meta-analysis were graphically displayed as a forest plot (Figure 7A). This dataset also contains variables which can be explored as moderators of effect size. Among them, Grade is a categorical moderator indicating the grade in which the intervention was carried out, with four levels: elementary (1), middle (2), high school (3) and college (4). A moderator analysis was carried out for Grade as moderator of effect size. The results provided evidence that Grade had a significant moderating effect ($Q(4) = 28.536$, $p < 0.0001$), but the Q-test was also significant (102.004, $df = 44$, $p < 0.0001$) and $I^2 = 59.15\%$ suggesting that other moderators influence the effectiveness of interventions on academic achievement. The point estimates and a 95% CI as well as the rest of results are presented in Figure 7A.

To identify underlying effect size patterns of studies, we selected an interval for the value of k varying from 2 to 6 to run our automated procedure on the outcome measures. The higher stability and goodness values matched the same k value equal to 2, i.e., two underlying classes of studies were identified by *evaluomeR*. From the latent factor detected, we performed a moderator analysis for testing the significance. The forest plot displayed in Figure 7B shows the output of the moderator analysis for this factor, which revealed a significant moderating effect ($Q(2) = 108.724$, $p < 0.0001$). The Q-test was not significant (36.204, df = 46, p = 0.8493) and $I^2 = 0.00\%$, suggesting that nearly 100% of the heterogeneity might be explained by including this latent factor in the model. In within-class analyses, there was no evidence of heterogeneity. Moreover, there was no difference in the mean levels between the two groups (0.048, 95% CI of -0.011 to 0.108) in the class 1 whereas a significant higher mean level in the intervention group was revealed (0.604, 95% CI of 0.489 to 0.719) in the class 2.

Binary data. Finally, the dataset named *dat.li2007* was employed to illustrate the usability of our computer tool to stratify the effect size when heterogeneity is found. This review consists of 22 randomized clinical trials to examine the effectiveness of intravenous magnesium versus placebo in the prevention of death following acute myocardial infarction [49]. We conducted the meta-analysis for log odds ratios. The random-effects model summary result of -0.546 (95% CI of -0.841 to -0.251) suggested that magnesium might significantly reduce mortality. Moreover, there was evidence of heterogeneity since the Q-test was 57.716 ($df = 21$, $p < 0.0001$) with $I^2 = 82.23\%$ and the total amount of the residual heterogeneity τ^2 was 0.1766 ($SE = 0.123$). The meta-analysis output is displayed in Figure 8A.

In order to pool the effect owing to the exhibited heterogeneity, we executed our automated process for the k value ranging from 2 to 6 on the logarithm of the odds ratios to cluster the trials into well-separated and compact underlying classes. According to the output, the higher stability and goodness were achieved classifying the trials in the 2 latent classes disclosed by *evaluomeR*. The moderator analysis for this latent factor provided a significant moderating effect ($Q(2) = 34.068$, $p < 0.0001$). In addition, there was no evidence of heterogeneity as the Q-test indicated (22.232, $p = 0.3281$), being $I^2 = 45.72\%$ and $\tau^2 = 0.0317$ ($SE = 0.033$), suggesting that nearly 82.06% of the heterogeneity might be accounted for this factor. In within-class analyses, presence of heterogeneity was not significant in the class 2. Nevertheless, there was evidence of heterogeneity in the class 1, although it was reduced. Actually, this class 1 includes both types of primary studies, large and small studies, which shows variability in the clinical trials and possible discussion in the meta-analysis literature (for more detail, among others see Li et al.[49] and Mawdsley et al. [50]). Anyway, no difference on mortality was found in the magnesium group with respect to placebo (-0.117, 95% CI of -0.310 to 0.076) in the first class of trials, whereas the second one reflected a significant decrease in mortality (-1.173, 95% CI of -1.575 to -

0.770). A customized forest plot from this moderator analysis was generated (see Figure 8B).

4. Discussion

Decision support systems need to use, analyze and classify different types of datasets. Most datasets have variables that correspond to types of quantitative measurements, and they are the metrics that describe a particular scenario. Decision-making is based on those metrics. The decision support models learned using those metrics are applied to other datasets, but without validating that the stochastic behavior of the metrics is homogeneous across datasets. Analyzing the stochastic behavior of quantitative metrics in different datasets is therefore important, but there is currently a lack of software tools able to support in such a process. In this paper we have presented a software tool to help researchers to understand the stochastic behavior of metrics by identifying latent classes which account for the variability in such outcome measures. The package *evaluomeR* provides two different ways for accessing its functionality, each access being tailored for a particular user.

The users of *evaluomeR* should be aware that the method requires the dataset to contain at least k different outcome measures of a metric to build k classes. In addition, a feasible range of k values may be used to select the most reliable stratification of such a metric. The reliability of the clustering generated from a metric is determined by both unsupervised classification validation criteria, stability and goodness of the clusters. Due to the lack of the gold standard, the bootstrap resampling technique is applied to assess the stability of the latent classes built with respect to each bootstrapped clustering. We have chosen a number of replicates $bs = 100$ in our use cases since [51] suggested that bs in the range 50 to 200 usually makes a good standard error estimator, and $bs = 100$ usually gives quite satisfactory results. Nevertheless, the number of bootstrap replicates can be defined by the user in *evaluomeR*. Besides, the Silhouette width is also used to measure the cohesion and separation of instances of such underlying classes.

In this work, we have illustrated the use of the tool in three use cases: impact factors, ontology structural metrics and effect sizes of primary studies. Regarding the first use case, it should be noted that a thorough analysis of the JCR is out of the scope of the present paper, since we have focused on describing the usefulness of *evaluomeR*. Nevertheless, the results found for the three categories studied reveal that analyzing all the JCR categories would be of interest for those researchers whose scientific activity is mainly evaluated by the quartiles of the journals that publish their work.

The second use case is related to the interest of our research group for analyzing ontology metrics, which made us realize of the potential benefits of *evaluomeR* for researchers. This use case is richer in terms of number of metrics, thus permitting a more detailed discussion of the results. According to the optimal value of k for each metric in the three datasets summarized in Table 5, the metrics ANOnto, CROnto, NOCOnto, NOMOnto, RFCOnto, TMOnto2, and WMCOnto exhibit the same behavior in the three datasets. Thereby, the heterogeneity was stratified by the same number of latent classes. However, this does not happen with the rest of metrics, which could be interpreted as less reliable metrics on those datasets.

The third use case has been devoted to showing the usefulness and effectiveness of the supplied computer tool to stratify the heterogeneity in effect size estimates of primary studies. On three different meta-analysis datasets, this software has provided a categorical moderator formed of the underlying classes of studies discovered by the automated process. The moderator analyses for each latent factor performed to pool the overall effect sizes that explain within- and between-study heterogeneity have reported significant moderator effects and no evidence of heterogeneity.

We are currently working on implementing functions for suggesting the optimal value of k such as the one presented in Section 2.2.1, and to include a preprocessing step that would suggest an upper limit for k by analyzing the size of the dataset and the distribution of values.

5. Conclusions

Clustering-based data analysis plays an important role as a decision support tool in the evaluation of the stochastic behavior and reliability of quantitative metrics on datasets, by improving the search process of the optimal classification. The use of statistical properties such as stability and goodness of classifications allows for a useful analysis of the behavior of quantitative metrics, which can be used for supporting decisions about which metrics to apply on biomedical datasets. *evaluomeR* is a software tool that provides an easy, flexible and automated way for analyzing such a behavior.

Acknowledgments

None.

Ethical approval

Ethics approval was not required for this study,

Funding

This research is part of the grants TIN2017-85949-C2-1-R and PID2020-113723RB-C22 funded by MCIN/AEI/10.13039/501100011033/.

Competing interests

The authors have no conflicts to disclose.

References

[1] S. Imbeaud, E. Graudens, V. Boulanger, X. Barlet, P. Zaborski, E. Eveno, O. Mueller, A. Schroeder, C. Auffray, Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces, Nucleic Acids Research 33 (6) (2005) e56–e56.

[2] M. Franco, J. M. Vivo, M. Quesada-Martínez, A. Duque-Ramos, J. T. Fernández-Breis, Evaluation of ontology structural metrics based on public repository data, Briefings in Bioinformatics 21 (2) (2020) 473–485.

[3] M. Singh, T. Singh, S. Soni, Pre-operative assessment of ablation margins for variable blood perfusion metrics in a magnetic resonance imaging based complex breast tumour anatomy: Simulation paradigms in thermal therapies, Computer Methods and Programs in Biomedicine 198 (2021) 105781.

[4] R. García-de León-Chocano, C. Sáez, V. Muñoz-Soler, A. Oliver-Roig, R. García-de León-González, J. M. García-Gómez, Robust estimation of infant feeding indicators by data quality assessment of longitudinal electronic health records from birth up to 18 months of life, Computer Methods and Programs in Biomedicine 207 (2021) 106147.

[5] X. Luo, L. Yang, H. Cai, R. Tang, Y. Chen, W. Li, Multi-classification of arrhythmias using a hcrnet on imbalanced ecg datasets, Computer Methods and Programs in Biomedicine 208 (2021) 106258.

[6] M. D. Wilkinson, S.-A. Sansone, E. Schultes, P. Doorn, L. O. B. da Silva Santos, M. Dumontier, A design framework and exemplar metrics for fairness, Scientific data 5 (1) (2018) 1–4.

[7] S. Moccia, E. De Momi, S. El Hadji, L. S. Mattos, Blood vessel segmentation algorithms—review of methods, datasets and evaluation metrics, Computer methods and programs in biomedicine 158 (2018) 71–91.

[8] J. Chen, H. You, K. Li, A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images, Computer methods and programs in biomedicine 185 (2020) 105329.

[9] J.-M. Vivo, M. Franco, D. Vicari, Rethinking an roc partial area index for evaluating the classification performance at a high specificity range, Advances in Data Analysis and Classification 12 (3) (2018) 683–704.

[10] M. Franco, J.-M. Vivo, Evaluating the performances of biomarkers over a restricted domain of high sensitivity, Mathematics 9 (21) (2021) 2826.

[11] L. Souza-Pereira, N. Pombo, S. Ouhbi, V. Felizardo, N. Garcia, Clinical decision support systems for chronic diseases: A systematic literature review, Computer Methods and Programs in Biomedicine 195 (2020) 105565.

[12] J. Gurevitch, J. Koricheva, S. Nakagawa, G. Stewart, Meta-analysis and the science of research synthesis, Nature 555 (7695) (2018) 175.

[13] M. M. Islam, H.-C. Yang, T. N. Poly, W.-S. Jian, Y.-C. J. Li, Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis, Computer Methods and Programs in Biomedicine 191 (2020) 105320.

[14] D. E. Nkhoma, C. J. Soko, P. Bowrin, Y. B. Manga, D. Greenfield, M. Househ, Y.-C. Li, U. Iqbal, Digital interventions self-management education for type 1 and 2 diabetes: A systematic review and meta-analysis, Computer Methods and Programs in Biomedicine 210 (2021) 106370.

[15] F. Siddi, A. Amedume, A. Boaro, A. Shah, A. M. Abunimer, P. A. Bain, J. Cellini, Q. R. Regestein, T. R. Smith, R. A. Mekary, Mobile health

and neurocognitive domains evaluation through smartphones: A meta-analysis, Computer Methods and Programs in Biomedicine 212 (2021) 106484.

[16] C. Hennig, Cluster-wise assessment of cluster stability, Computational Statistics & Data Analysis 52 (2007) 258–271.

[17] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65.

[18] A. R. M. Forkan, I. Khalil, H. Kumarage, Patient clustering using dynamic partitioning on correlated and uncertain biomedical data, Computer methods and programs in biomedicine 190 (2020) 105483.

[19] M. Franco, J.-M. Vivo, Cluster analysis of microarray data, in: Microarray Bioinformatics, Springer, 2019, pp. 153–183.

[20] P. Jaccard, Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines, Bull Soc Vaudoise Sci Nat 37 (1901) 241–272.

[21] C. Hennig, fpc: Flexible Procedures for Clustering, r package version 2.2-3 (2019).
URL `https://CRAN.R-project.org/package=fpc`

[22] L. Kaufman, P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 1990.

[23] J. A. Bernabé-Díaz, M. Franco-Nicolás, J. M. Vivo-Molina, M. Quesada-Martínez, A. Duque-Ramos, J. T. Fernández-breis, Bioconductor evaluomer package, `https://doi.org/doi:10.18129/B9.bioc.evaluomeR`, accessed on 2021-08-10 (August 2021).

[24] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, K. Hornik, cluster: Cluster Analysis Basics and Extensions, r package version 2.1.2 — For new features, see the 'Changelog' file (in the package source) (2021).
URL `https://CRAN.R-project.org/package=cluster`

[25] T. Wei, V. Simko, R package 'corrplot': Visualization of a Correlation Matrix, (Version 0.92) (2021).
URL `https://github.com/taiyun/corrplot`

[26] G. N. Boshnakov, Rdpack: Update and manipulate rd documentation objects, r package version 2.1.3 (2021). doi:10.5281/zenodo.3925612.

[27] M. Morgan, V. Obenchain, J. Hester, H. Pagès, SummarizedExperiment: SummarizedExperiment container, (Version 1.24.0) (2021). URL https://bioconductor.org/packages/SummarizedExperiment

[28] M. Ramos, L. Schiffer, A. Re, R. Azhar, A. Basunia, C. Rodriguez, T. Chan, P. Chapman, S. R. Davis, D. Gomez-Cabrero, et al., Software for the integration of multiomics experiments in bioconductor, Cancer research 77 (21) (2017) e39–e42.

[29] J. A. Bernabé-Díaz, M. Franco-Nicolás, J. M. Vivo-Molina, M. Quesada-Martínez, A. Duque-Ramos, J. T. Fernández-breis, Webpage evaluomer shiny, https://semantics.inf.um.es/shiny/evaluomeR-shiny/, accessed on 2021-05-16 (May 2021).

[30] W. Chang, J. Cheng, J. Allaire, Y. Xie, J. McPherson, shiny: Web Application Framework for R, r package version 1.4.0 (2019). URL https://CRAN.R-project.org/package=shiny

[31] W. Viechtbauer, Conducting meta-analyses in R with the metafor package, Journal of Statistical Software 36 (3) (2010) 1–48.

[32] C. Woolston, et al., Impact factor abandoned by dutch university in hiring and promotion decisions, Nature 595 (7867) (2021) 462–462.

[33] R. Hoehndorf, P. N. Schofield, G. V. Gkoutos, The role of ontologies in biological and biomedical research: A functional perspective, Briefings in Bioinformatics 16 (6) (2015) 1069–1080.

[34] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, et al., The obo foundry: coordinated evolution of ontologies to support biomedical data integration, Nature Biotechnology 25 (11) (2007) 1251.

[35] C. Jonquet, A. Toulet, E. Arnaud, S. Aubin, E. D. Yeumo, V. Emonet, J. Graybeal, M.-A. Laporte, M. A. Musen, V. Pesce, et al., Agroportal: A vocabulary and ontology repository for agronomy, Computers and Electronics in Agriculture 144 (2018) 126–143.

[36] E. Ong, Z. Xiang, B. Zhao, Y. Liu, Y. Lin, J. Zheng, C. Mungall, M. Courtot, A. Ruttenberg, Y. He, Ontobee: a linked ontology data server to support ontology term dereferencing, linkage, query and integration, Nucleic Acids Research 45 (D1) (2016) D347–D352.

[37] R. Côté, F. Reisinger, L. Martens, H. Barsnes, J. A. Vizcaino, H. Hermjakob, The ontology lookup service: bigger and better, Nucleic Acids Research 38 (suppl_2) (2010) W155–W160.

[38] R. Hoehndorf, L. Slater, P. N. Schofield, G. V. Gkoutos, Aber-owl: A framework for ontology-based data access in biology, BMC Bioinformatics 16 (1) (2015) 26.

[39] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, M. A. Musen, Bioportal: Enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications, Nucleic Acids Research 39 (suppl_2) (2011) W541–W545.

[40] A. Duque-Ramos, J. T. Fernández-Breis, R. Stevens, N. Aussenac-Gilles, Oquare: A square-based approach for evaluating the quality of ontologies, Journal of Research and Practice in Information Technology 43 (2) (2011) 159–176.

[41] S. R. Chidamber, C. F. Kemerer, A metrics suite for object oriented design, IEEE Transactions on software engineering 20 (6) (1994) 476–493.

[42] W. Li, Another metric suite for object-oriented programming, Journal of Systems and Software 44 (2) (1998) 155–162.

[43] M. D. Wilkinson, S.-A. Sansone, E. Schultes, P. Doorn, L. O. B. da Silva Santos, M. Dumontier, A design framework and exemplar metrics for fairness, Scientific Data 5.

[44] S. Tartir, I. B. Arpinar, Ontology evaluation and ranking using OntoQA, in: ICSC '07: Proceedings of the International Conference on Semantic Computing, IEEE Computer Society, Washington, DC, USA, 2007, pp. 185–192. doi:http://dx.doi.org/10.1109/ICSC.2007.65.

[45] D. Langan, J. P. Higgins, D. Jackson, J. Bowden, A. Angeliki, V. Evangelos, W. Viechtbauer, M. Simmonds, A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses, Research Synthesis Methods 10 (1) (2019) 83–98.

[46] G. J. Molloy, R. E. O'Carroll, E. Ferguson, Conscientiousness and medication adherence: A meta-analysis, Annals of Behavioral Medicine 47 (1) (2014) 92–101.

[47] D. S. Quintana, From pre-registration to publication: a non-technical primer for conducting a meta-analysis to synthesize correlational data, Frontiers in Psychology 6 (2015) 83–98.

[48] R. L. Bangert-Drowns, M. M. Hurley, B. Wilkinson, The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis, Review of Educational Research 74 (1) (2004) 29–58.

[49] J. Li, Q. Zhang, M. Zhang, M. Egger, Intravenous magnesium for acute myocardial infarction, Cochrane Database of Systematic Reviews (2).

[50] D. Mawdsley, J. Higgins, A. J. Sutton, K. Abrams, Accounting for heterogeneity in meta-analysis using a multiplicative model–an empirical study, Research Synthesis Methods 8 (1) (2017) 43–52.

[51] R. J. Tibshirani, B. Efron, An introduction to the bootstrap, Monographs on statistics and applied probability 57 (1993) 1–436.

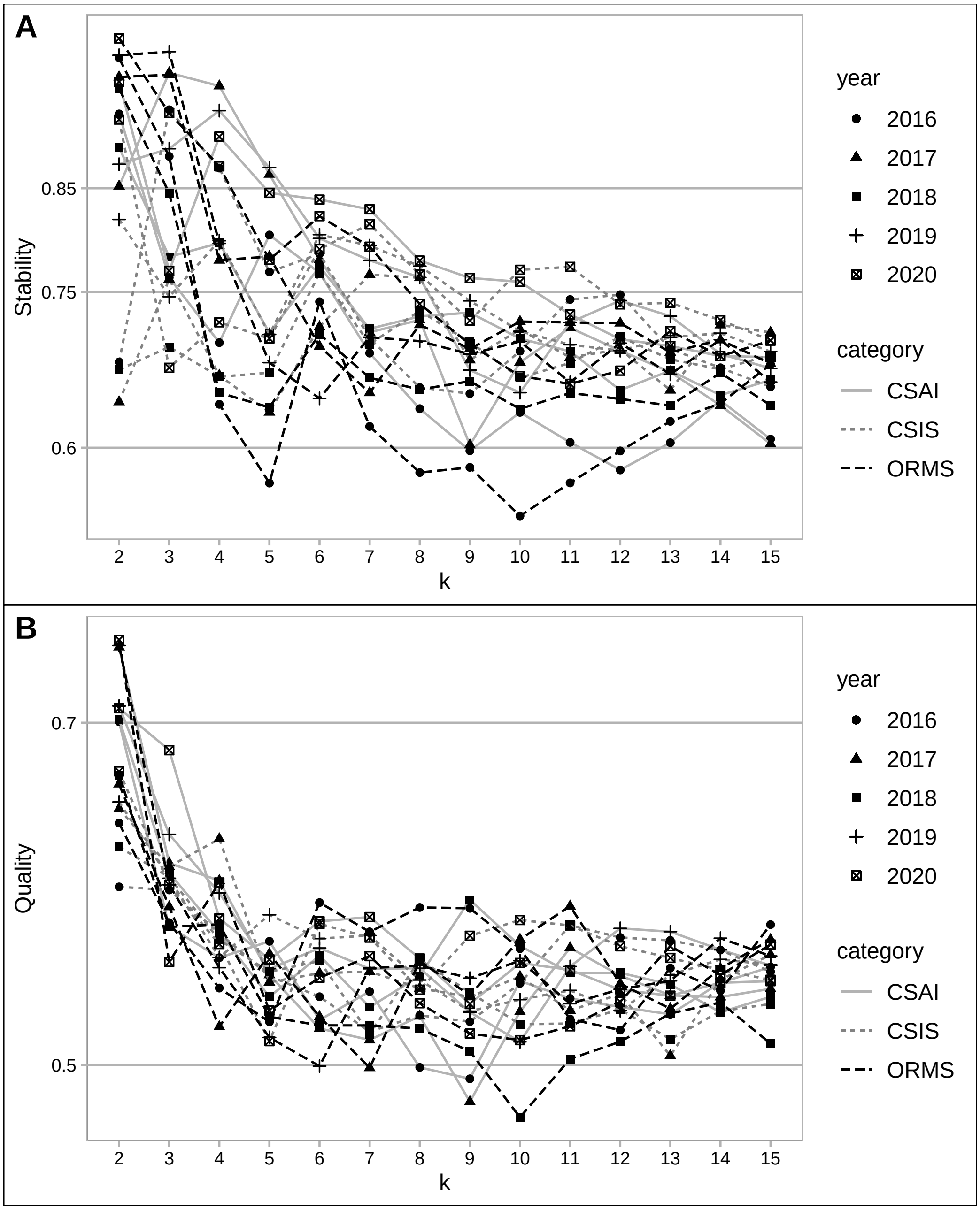

Figure 3: The stability (A) and goodness (B) of the classification of the impact factor for the JCR category "Computer Science, Artificial Intelligence" (CSAI), "Computer Science, Information Systems" (CSIS) and "Operations Research & Management Science" (ORMS) in the period 2016-2020.

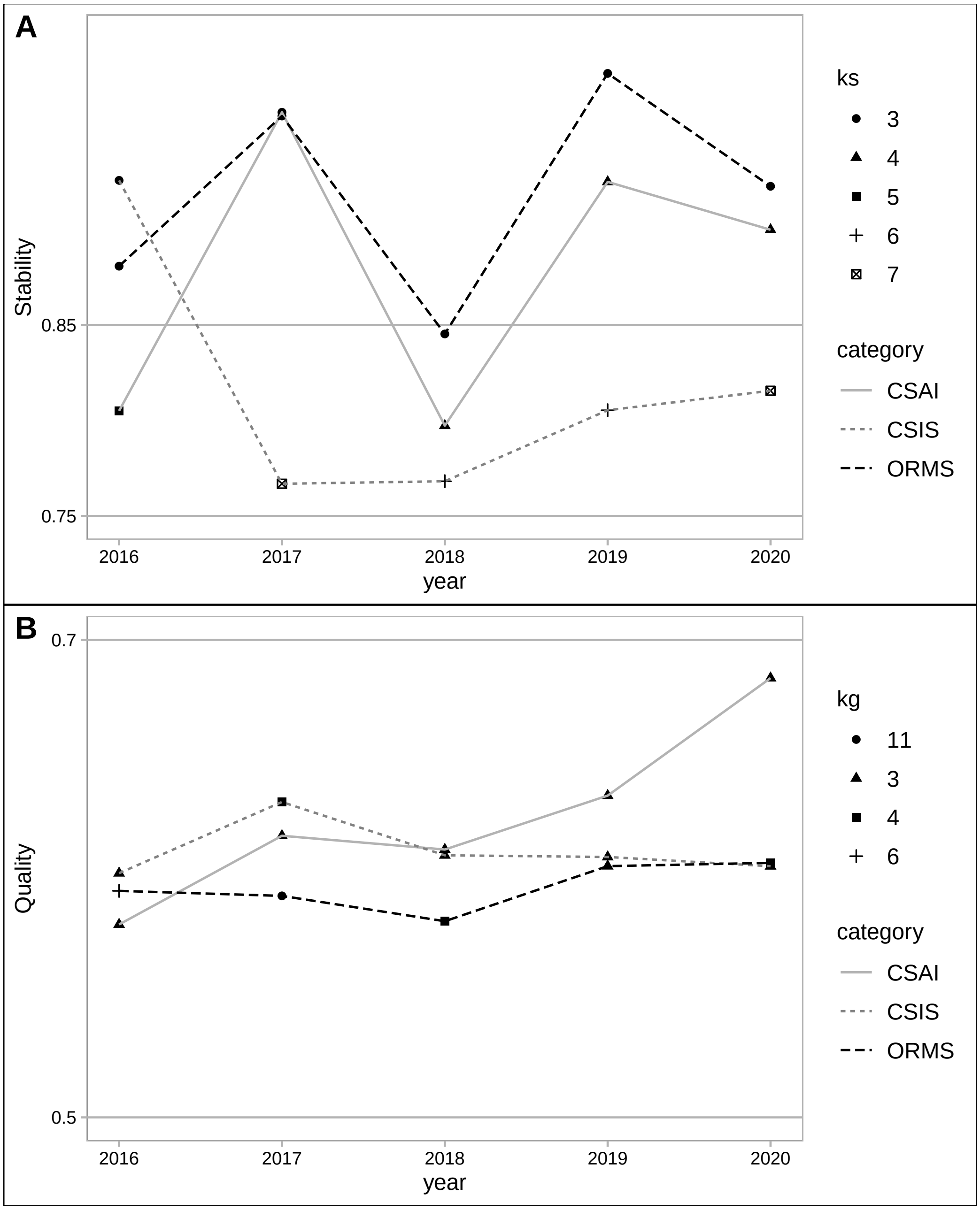

Figure 4: (A) Stability scores for k_s and (B) goodness scores for k_g per year, corresponding to the classification of the impact factor in the three JCR categories "Computer Science, Artificial Intelligence" (CSAI), "Computer Science, Information Systems" (CSIS) and "Operations Research & Management Science" (ORMS).

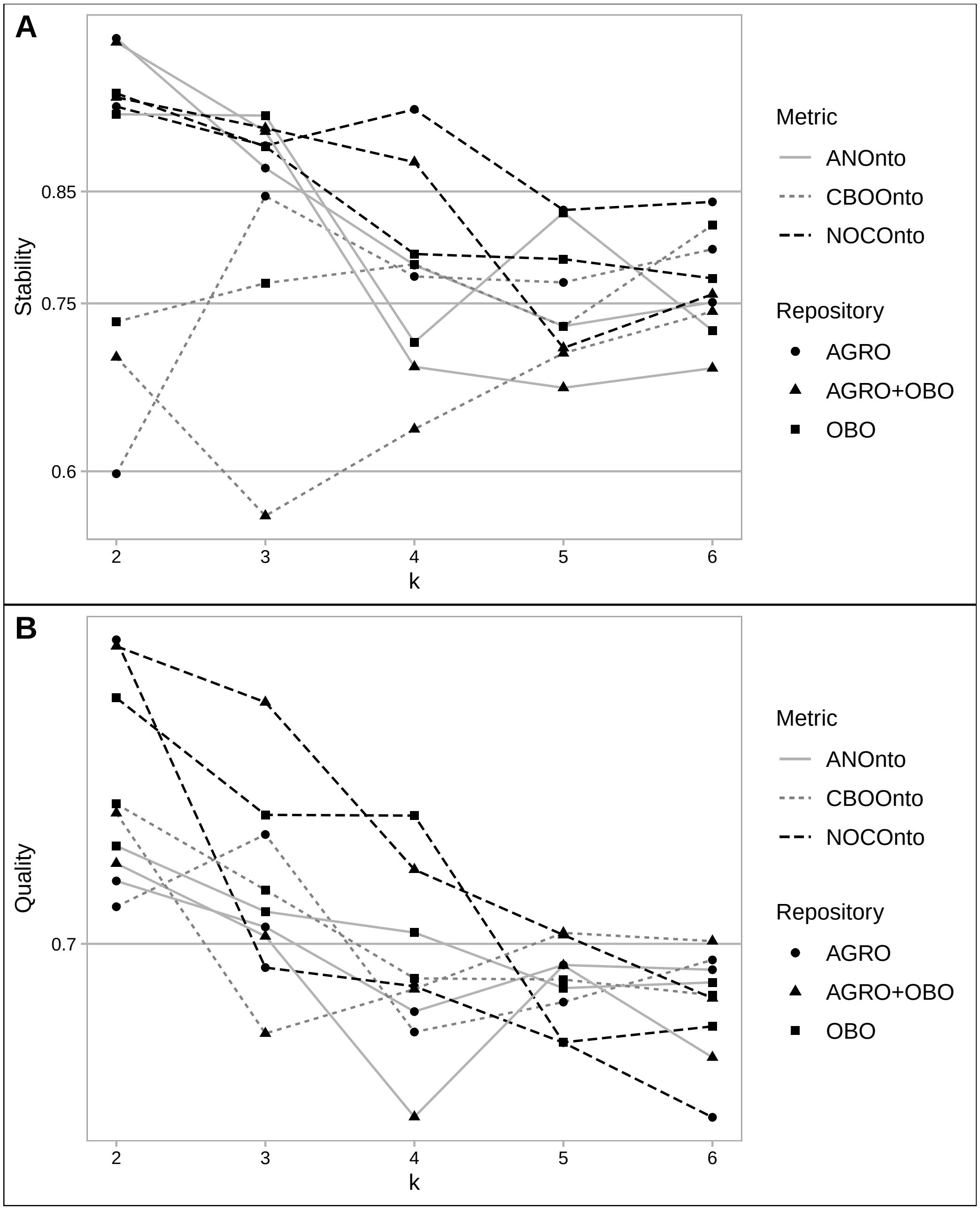

Figure 5: The stability (A) and goodness (B) of the classifications of the ontology metrics ANOnto, CBOOnto and NOCOnto for AgroPortal (AGRO), OBO Foundry (OBO) and the aggregated set of both (AGRO+OBO) datasets.

Table 4: Definition of the 19 metrics evaluated: column 1 shows the acronym of the metric, column 2 describes the ontology facet measured by the metric, column 3 describes how the metric is calculated, and column 4 includes the references in which the metrics have been proposed or adapted to ontologies.

Metric name	**Facet**	**Description**
CBOnto[40, 41]	Coupling	Number of direct ancestors of classes divided by the number of classes minus subclasses of thing
DITOnto[40, 41]	Depth of the hierarchy	Length of the longest path from thing to a leaf classes
NOCOnto[40, 41]	Descendants	Number of the direct subclasses divided by the number of classes minus the number of leaf classes
RFCOnto[40, 41]	Properties usage	Number of usages of object and data properties and superclasses divided by the number of classes
WMCOnto[40, 41]	Complexity	Mean length of the paths from thing to a leaf classes
NOMOnto[42]	Properties	Mean number of object and data property usages per class
NACOnto[42]	Ancestors of leaf classes	Mean number of superclasses per leaf classes
LCOMOnto[43]	Cohesion	Mean length of all paths from leaf classes to thing
ANOnto[44]	Annotations	Mean number of annotations properties per classes
CROnto[44]	Individuals	Mean number of individuals per classes
AROnto[44]	Attribute richness	Number of restrictions of the ontology per classes
INROnto[44]	Descendants	Mean number of subclasses per classes
PROnto[44]	Property richness	Number of subclass of relationships divided by the number of subclass of relationships and properties
RROnto[44]	Properties usage	Number of usages of object and data properties and super classes divided by the number of classes
TMOnto[40]	Multiple inheritance	Mean number of classes with more than one ancestor
POnto[40]	Ancestors	Mean number of direct ancestors per class
CBOnto2[40]	Coupling	Mean number of direct ancestors per classes
TMOnto2[40]	Multiple inheritance	Mean number of direct ancestors of classes with more than 1 direct ancestor
WMCOnto2[40]	Complexity	Mean number of paths from thing to a leaf classes

Table 5: Optimal value of k for each metric in each dataset

	AgroPortal	OBO Foundry	AgroPortal + OBO Foundry
ANOnto	3	3	3
AROnto	3	4	5
CBOOnto	3	3	5
CBOOnto2	3	3	5
CROnto	3	3	3
DITOnto	3	3	5
INROnto	3	3	6
LCOMOnto	3	4	4
NACOnto	6	3	3
NOCOnto	3	3	3
NOMOnto	3	3	3
POnto	4	3	3
PROnto	4	3	6
RFCOnto	3	3	3
RROnto	4	3	6
TMOnto	6	3	3
TMOnto2	3	3	3
WMCOnto	3	3	3
WMCOnto2	6	3	3

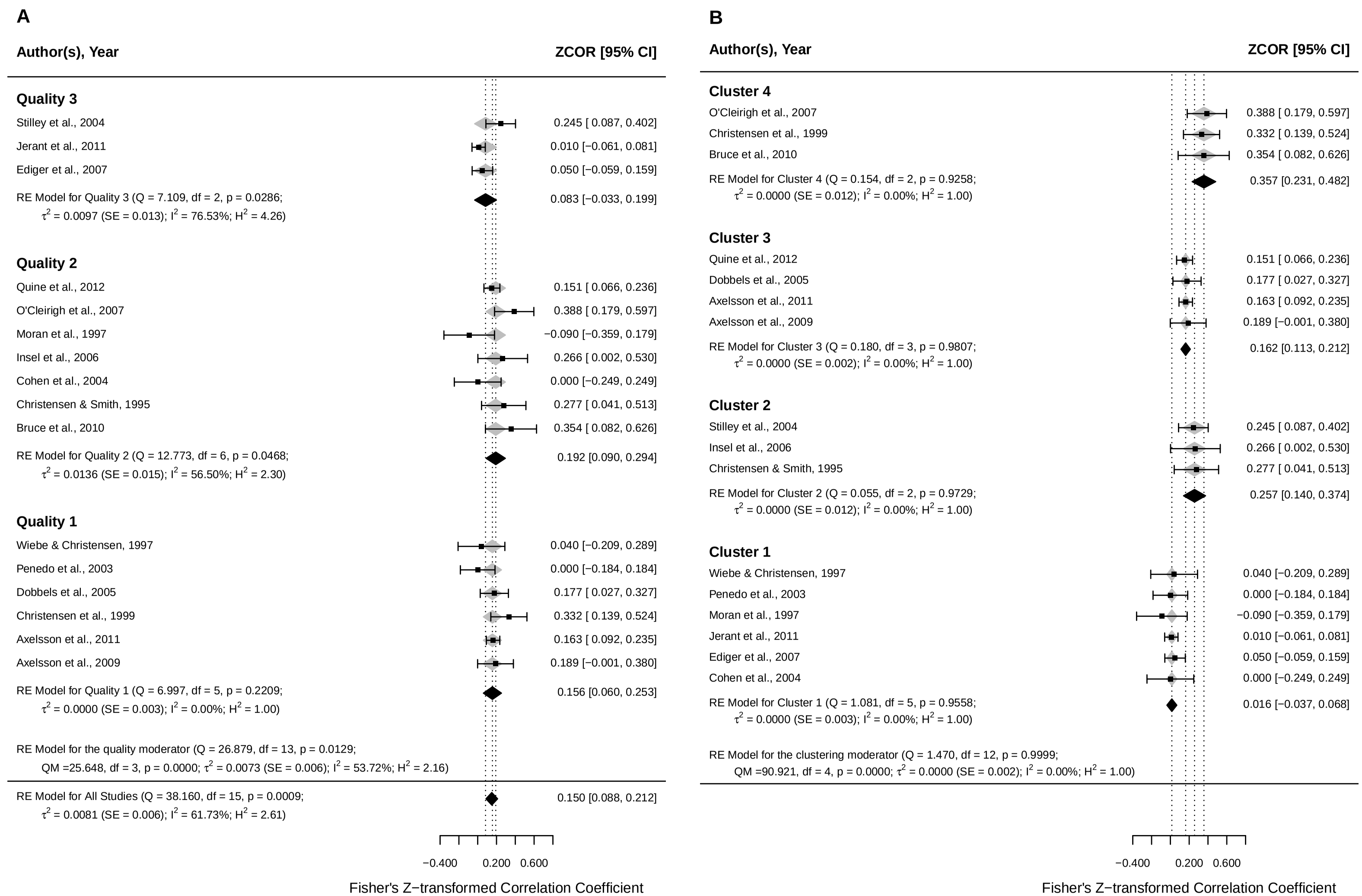

Figure 6: Forest plot for Fisher's transformed correlation coefficient from dat.molloy2014 dataset of the R package metafor.

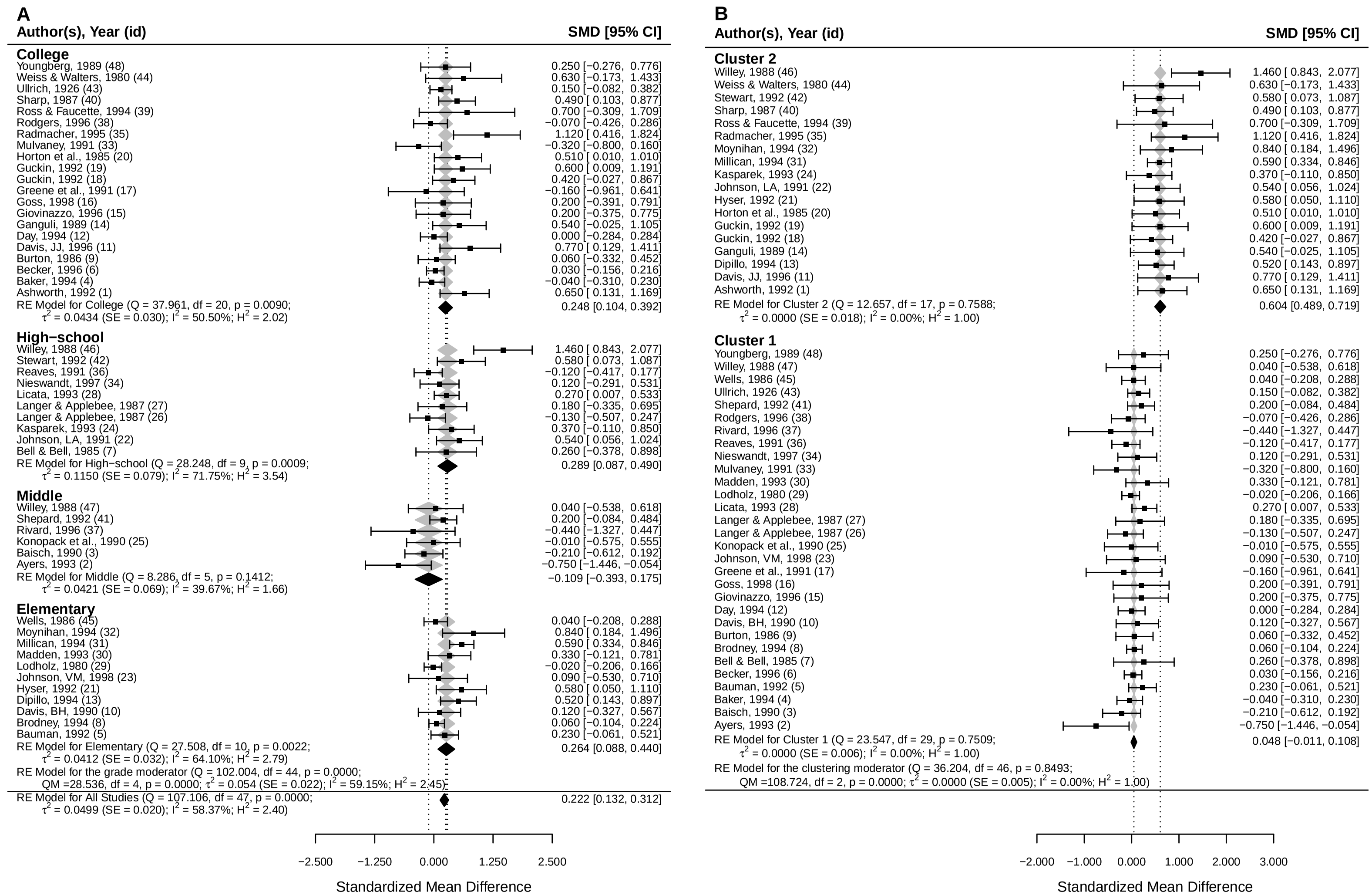

Figure 7: Forest plot for the standardized mean difference from dat.bangertdrowns2004 dataset of the R package metafor.

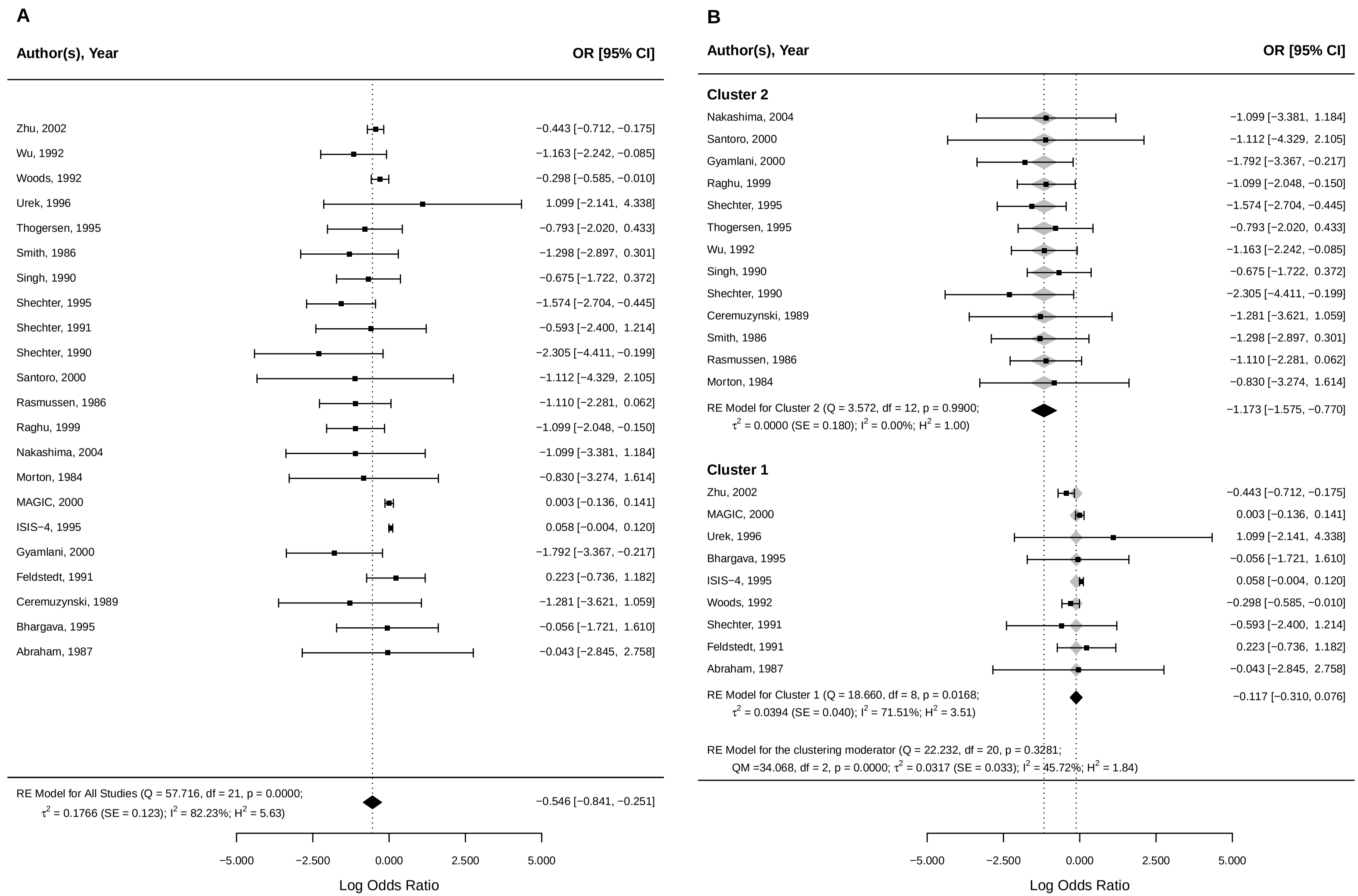

Figure 8: Forest plot for the log odds ratio from dat.li2007 dataset of the R package metafor.