Identifying benchmark units for research management and evaluation

Qi Wang1 · Tobias Jeppsson2

Received: 31 October 2021 / Accepted: 19 May 2022 / Published online: 19 June 2022 © The Author(s) 2022

Abstract
While normalized bibliometric indicators are expected to resolve the subject-field differences between organizations in research evaluations, the identification of reference organizations working on similar research topics is still of importance. Research organizations, policymakers and research funders tend to use benchmark units as points of comparison for a certain research unit in order to understand and monitor its development and performance. In addition, benchmark organizations can also be used to pinpoint potential collaboration partners or competitors. Therefore, methods for identifying benchmark research units are of practical significance. Even so, few studies have further explored this problem. This study aims to propose a bibliometric approach for the identification of benchmark units. We define an appropriate benchmark as a well-connected research environment, in which researchers investigate similar topics and publish a similar number of publications compared to a given research organization during the same period. Four essential attributes for the evaluation of benchmarks are research topics, output, connectedness, and scientific impact. We apply this strategy to two research organizations in Sweden and examine the effectiveness of the proposed method. Identified benchmark units are evaluated by examining the research similarity and the robustness of various measures of connectivity.

Keywords Benchmark identification · Bibliometrics · Scientometrics · Research evaluation · Research topic identification

Introduction

In the practice of research management and evaluation, the identification of benchmark units is of importance. Research organization managers, policymakers and research funding providers tend to use benchmark units as points of comparison for a given research body in order to
understand and monitor its development and performance. They can further promote researchers to establish collaborations with colleagues at such organizations to learn and improve from benchmark units with excellent performance. Relatively often, comparison to benchmark organizations is also one of the cornerstones for further policy initiatives. For this reason, as bibliometric researchers, we often receive commissions in relation to the identification of benchmark units from both university administrators and research funder managers.

Few publications in the fields of scientometric studies and research evaluation have focused on this problem. To our knowledge, only three studies have explored the identification of benchmark units. One is an early study by Noyons et al., (1999, in which they introduce a strategy to identify benchmarks for a micro-electronic research center. Later, Carayol et al. (2012) propose a method for selecting peer universities and departments, and a study by Andersen et al. (2017) develop a bibliometric-based approach for choosing proper benchmarks. The explanations for the lack of studies on this topic might be threefold. One reason might be that normalized bibliometric indicators are frequently used to deal with differences in research fields and publication output between organizations, and this solution may be seen as sufficient. However, size still matters especially when evaluating the research performance of organizations that change in size over time (Andersen et al., 2017; Katz, 2000). Frenken et al. (2017, as cited in Rousseau, 2020) also consider that benchmarking of universities can be misleading, and indicate that “benchmarking is most meaningful between universities of a similar size” Field-normalized citation indicators that work well for large organizations and at a certain granularity of field classification may also be misleading for smaller organizations with a more narrow research focus (Ruiz-Castillo & Waltman, 2015; Zitt et al., 2005), where comparisons against relevant benchmarks may be more suitable. Another reason for this situation might be the inferior data quality of research addresses. The identification of benchmark units tends to be used for detecting comparable groups for university departments or research organizations at the equivalent level. As widely recognized, however, research address data at that level are tremendously confusing (e.g. Glänzel, 1996; Melin & Persson, 1996), which paralyzes publication retrieval for such research organizations. Therefore in practice, expert opinions have usually been adopted for the identification of benchmark units, which implies the understanding of benchmark depends on the expert’s perspective. Yet, experts may point out those outstanding organizations in their research areas, but it remains uncertain if those are comparable, and the experts may be unfamiliar with the research organizations that are located outside their particular regions. Furthermore, the lack of access to large-scale bibliographic databases may also hamper explorations on this topic.

Under such circumstances, the present work attempts to contribute to this important topic in the field of research evaluation and policy through a precise definition of benchmark research units and a well-connected framework for operationalization. To do so, we begin with an overview of the benchmark concept. It is then followed by the definition of a benchmark research unit and a set of criteria for the identification. Next, two research institutions are used as examples to demonstrate the present approach and evaluate its effectiveness. Finally, several different strategies are applied for the evaluation and validation of our results.

Overview of the benchmark concept

To provide an overview of how the term “benchmark” is used in general, we reviewed related studies. According to Ettorchi-Tardy et al. (2012), the benchmark method was first implemented by the Xerox Company where they compared the features and quality of their
productions with related competitors in order to lower the production costs (Fedor et al., 1996). At the same time, numerous companies, such as Ford, General Motors and AT&T (Pryor, 1989), also applied such methods for continuous quality improvement in production and management practices (Bi, 2017).

Since then, benchmark methods have become conceptualized (e.g. Jac Fitz-enz, 1993; Kumar et al., 2006; Moriatry, 2011). Here we list a few definitions. Camp (1989) consider benchmarks as the best industry practices and by the implementation of these practices, counterparts would achieve exceptional performance (Anand & Kodali, 2008); Lucerti et al. (1995) define benchmark method as an approach used for evaluating and improving the company performance in comparison with the companies with the best performance; Jansen et al. (2010) consider benchmark method as “systematic comparison of the performance of (parts of) organizations and their similar services, processes and routines, on the basis of predetermined indicators, with the goal that organizations can improve their performance by learning from one another” (pp. 214). As Amaral and Sousa (2009) summarized, despite subtle differences among these definitions, the core procedures and central purposes of applying benchmark methods remain rather similar, which is to improve the performance of organizations through comparing with and learning from the best.

In management practices, the use of benchmark methods has spread over companies in various industries “to improve their understanding of the most successful practices in their field” (González & Álvarez, 2001, pp. 518). Moreover, public administration has also adopted benchmark methods to enhance their services. Sankey and Padró (2016) present the findings of utilizing benchmarks that the Australasian Council on Open, Distance and E-learning created to assist higher education institutions to promote their e-learning activities. In their study, benchmarks are considered a quality improvement measure where universities are encouraged to self-assess against benchmarks and consequently improve their activities to meet expectations. Besides this, benchmark methods have been applied to promote e-government services (Jansen et al., 2010; Kromidha, 2012; Petrović et al., 2012), and also applied in New Zealand’s transportation system to improve the performance of transport and achieve sustainability and environmental targets (Henning et al., 2011).

We note that the term “benchmark” has also been used in many other scientific fields in recent years, for instance, researchers use benchmark methods in medical science to improve healthcare service quality (Ettorchi-Tardy et al., 2012; Tuzkaya et al., 2019). In the field of engineering and computer science especially machine learning and artificial intelligence, benchmark methods have been seen as essential for unbiased comparisons of the performance between relevant algorithms (Hoffmann et al., 2019; Hothorn et al., 2005; Kistowski et al., 2015).

Furthermore, studies exploring principles and methods to build and select benchmarks can be found in various research areas (e.g. Carpinetti & De Melo, 2002; Dattakumar & Jagadeesh, 2003; Huppler, 2009). For instance, Bast and Korzen (2017) build a benchmark for evaluating the quality of tools for text extraction from PDF; Lou and Yuen (2019) propose a method to construct benchmark test suites for examining the performance of evolutionary algorithms; Hougaard and Tvende (2002) and Hougaard and Keiding (2004) propose models of benchmark selections in operational research.
Definition and attributes of a benchmark research unit

In this section, we turn to the specific use of benchmarks in research evaluation and policy. As mentioned above, few previous studies in this field have investigated the identification of benchmark research organizations. Noyons et al. (1999) select benchmark institutes for a micro-electronics research center in order to analyze its position in the world. The identified benchmarks in their study present similar research profiles and publish a similar number of publications with the evaluative center. Carayol and colleagues (2012) consider that the quantity of scientific production and their impact are the two essential attributes for a benchmark unit. In addition to these two attributes, Andersen and colleagues stress the importance of research topics in the selection of benchmarks (2017). More specifically, they hold that “the topicality or subject profile” of a benchmark unit should be approximately similar to the treatment.

In practice, identifying benchmark research units is done for various reasons. Benchmark units can be used to compare, track and monitor the research performance of a given organization, while others might be interested in discovering similar research environments and further establishing collaboration relationships. Despite the difference in starting points, the premise of conducting such analyses lies in the existence of research similarities between a treatment and its benchmark units. We, therefore, consider similar research topics as the most important attribute for a benchmark research unit.

Second, a benchmark unit is expected to have a similar scale as the treatment. This is because the size of research organizations could greatly affect our assessment when tracking their academic performance over time. Scales of research organizations can be measured from various aspects, such as the number of researchers, research grants, and scientific output, etc. Yet, if the number of researchers is used to assess scales of organizations, other issues need to be considered, for instance, the structure of academic age, the combination of research and teaching time, and so forth. Besides this, it is less possible and feasible to acquire the detailed information of researchers and research funding for research organizations. Given these limitations, scientific output is more appropriate to measure the scales of organizations.

Furthermore, researchers at a benchmark unit are expected to be well connected, in other words, they should already have established collaboration relations to some extent. Of course, we acknowledge that researchers may not collaborate with their colleagues for many reasons. From the perspective of research management, a research unit with a certain degree of internal collaboration is more meaningful for further comparison and investigation.

The last attribute of benchmark units is research impact, which implies that a benchmark unit is expected to have a comparable or better research impact compared to the treatment. This is because the identification of excellent units could be helpful for the researchers at the treatment to learn from, collaborate with, and even surpass. The attribute of collaboration relations and research impact may not be necessary for certain analyses, however, it should be stressed that the present approach is quite flexible which allows users to adjust parameter values according to their specific perspective of analyses.

Based on these considerations, we conceive of a benchmark unit as a well-connected research environment where researchers work on the same topics, publish a similar number of papers, and present a comparable or better research impact in relation to the
treatment group during the same time period. The four essential attributes of a research organization to consider when identifying benchmark units are therefore, (i) research topics, (ii) scale, (iii) connectedness, and (iv) research impact.

**Methodology**

In this section, we elaborate on how to operationalize our definition of benchmark units, and we go through the four attributes in turn below.

**Research topics**

Delineating the research topics of an evaluative organization is the most important procedure for the identification of benchmarks. In this study, research topics were defined as publication clusters and constructed with the use of the clustering method of Waltman and Van Eck (2012, 2013) based on direct citation links. The reasons for choosing this strategy were summarized in Wang (2018). Our employed clustering system assigned around 36 million publications (namely, articles and reviews) into 5053 clusters.¹ These publications were published from 1980 to 2019 and were covered by the Web of Science (WoS) database. It has a similar scale compared to the one used in Leiden Ranking (2020), which consists of around 4000 clusters.²

---

¹ This study is based on data from the in-house bibliometric database (BIBMET) of the library at KTH.
² More detailed information on the fields in Leiden Ranking is available at https://www.leidenranking.com/information/fields.
Research topics of a treatment unit are, then, described by the clusters that include its publications. By mapping the distributions of the 15 research centers in Sweden, we conclude that the distributions of publications over clusters are usually highly skewed, as shown in Fig. 1, despite that most of the centers are not interdisciplinary oriented. It is of little significance to delineate the research profile of a treatment group using the clusters that only have an extremely small number of its publications. In addition to this, it should be noted that the size of clusters varies greatly in our clustering system. Over the 38 years, the largest cluster consists of around 56,000 publications, whereas the smallest has only 500. Therefore, it would be more reasonable to consider the relative number of a treatment group’s publications in each cluster when defining its dominant research topics.

To be specific, let \( p_{i,ts,te}^k \) denote the number of publications of evaluative unit \( i \) in cluster \( k \) from the year \( ts \) to \( te \), then its total number of publications can be expressed as \( P_{i,ts,te} = \sum p_{i,ts,te}^k \). Let \( k^i_{ts,te} \) denote the total number of publications in cluster \( k \) from the year \( ts \) to \( te \), and \( d_{i,ts,te}^k \) denote the share of publications for \( i \) in cluster \( k \), then we have

\[
d_{i,ts,te}^k = \frac{p_{i,ts,te}^k}{\sum p_{i,ts,te}^k}
\]

To pinpoint dominant clusters for unit \( i \), we intend to select the clusters with a high \( a \). Let \( r \) denote the rank of \( a \) in decreasing order, and it hence satisfies \( d_{i,ts,te}^{k(r)} \geq d_{i,ts,te}^{k(r-1)} \). Let \( d_{i,ts,te} \) denote the share of publications for unit \( i \) in the first \( n \) clusters, then we have

\[
d_{i,ts,te} = \frac{p_{i,ts,te}^{k(1)} + p_{i,ts,te}^{k(2)} + \cdots + p_{i,ts,te}^{k(r)} + p_{i,ts,te}^{k(n)}}{\sum p_{i,ts,te}^k} = \frac{\sum p_{i,ts,te}^{k(r)}}{\sum p_{i,ts,te}^k}
\]

To identify the dominant clusters, we require \( d_{i,ts,te} \geq d_{\text{min}} \) where \( d_{\text{min}} \) is a parameter that determines a minimum share of publications for unit \( i \) in the first \( n \) clusters. Moreover, we also need to avoid selecting the clusters with an extremely small number of publications for evaluative unit \( i \) and for treatment organization \( j \), which requires \( p_{i,ts,te}^{k(r)} \geq p_{\text{min}} \) and \( p_{j,ts,te}^{k(r)} \geq p_{\text{min}} \). The parameter \( p_{\text{min}} \) is used to define the minimum number of publications. By setting the parameter values, \( n \) clusters can be selected to delineate the research profile for unit \( i \).

**Scales**

As discussed above, the number of publications is used to assess the scale of research organizations. After the selection of the dominant clusters, publications between year \( ts \) and \( te \) in the \( n \) clusters can be aggregated into research organizations based on the affiliations of authors. Let \( P_{j,ts,te} \) denote the publications of organization \( j \) in the \( n \) clusters, which can be expressed as \( P_{j,ts,te} = \sum p_{j,ts,te}^{k(r)} \). As discussed, for a research unit to be considered as a benchmark, it should have a similar number of publications with the treatment, that is \( P_{j,ts,te} \in (P_{i,ts,te} - \Delta p, P_{i,ts,te} + \Delta p) \). \( \Delta p \) refers to the differences in the number of publications between evaluative unit \( i \) and treatment unit \( j \).

**Connectedness**

Researchers at a benchmark unit are expected to be connected to some extent. In this work, a weighted measure of the clustering coefficient (Opsahl & Panzarasa, 2009) has been applied to examine the extent to which researchers working at a benchmark unit are well connected. According to Opsahl and Panzarasa (2009), the clustering coefficient evaluates the total value of closed triplets in a weighted network in comparison with the total value.
of triplets. The higher the clustering coefficient, the more connected a network is. They have proposed several measures for calculating the weighted triplet value, including arithmetic, mean, geometric mean, maximum and minimum, and further indicated the choice of the weighted measure should be based on the research question at hand.

In our network, nodes represent researchers and links refer to the number of their collaborated publications. We assume that as long as researchers A and B, A and C have established collaboration relations respectively, B and C are then likely to collaborate. We, therefore, choose the minimum measure. Let \( h_{j,bs,te} \) denote the clustering coefficient of a benchmark unit, then we require that \( h_{j,bs,te} \geq h_{\text{min}} \), in which the parameter \( h_{\text{min}} \) is used to set the minimum clustering coefficient for treatment \( j \).

**Research impact**

A benchmark unit should present at least a comparable scientific impact in relation to the evaluative unit. This work uses PP(top 10%) to measure impact. PP(top 10%) is the proportion of the publications of an organization that is compared with the other publications with the same document type, in the same scientific fields and same year belong to the top 10% most frequently cited (Waltman & Schreiber, 2013). PP(top 10%) has been frequently used in research evaluation and university rankings. As introduced in the Leiden Ranking, fields of science tend to be defined by aggregations of related journals such as the WoS subject category system and the Scopus All Science Journal Classification. However, this approach is problematic to assign multidisciplinary journals to a specific field. Therefore, following the strategy applied in the Leiden Ranking, this analysis use citation-based clusters, as explained above, to define fields of science. The individual publication level classification system can assign each publication into one field, resolving the problem caused by multidisciplinary journals. Let \( c_{i,bs,te} \) denote the PP(top 10%) of unit \( i \). For an organization \( j \) to be considered as a benchmark, it should have at least a similar impact compared to the treatment, that is \( c_{j,bs,te} \geq c_{\text{min}} \). The parameter \( c_{\text{min}} \) defines the minimum PP(top 10%) for treatment \( j \).

In summary, a benchmark research unit should satisfy each of the following criteria:

1. it should present similar research interests with the treatment;
2. it should have a certain amount of research output published in the main research topics of the treatment;
3. it should have comparable research output;
4. its researchers should be well connected; and
5. it should exert at least a comparable scientific impact.

**Data**

Two research organizations are served as examples to demonstrate the proposed methods. The first is the Hero-m,\(^3\) at the Royal Institute of Technology. This center aims “to develop tools and competence for fast, intelligent, sustainable and cost-efficient product development for Swedish industry. Continuous scientific breakthroughs are exploited to enable

\(^3\) The current name of Hero-m is Hero-m 2 Innovation (Hero-m 2i) and that is a continuation of Hero-m.
design of materials from atomistic scales to finished products”. Based on the information available on their webpage, we collected 255 publications between 2007 and 2019. PP(top 10%) and connectivity of Hero-m are 19% and 0.39 respectively.

Another example is the Science for Life Laboratory (SciLifeLab), which uses as a national research infrastructure “for the advancement of molecular biosciences in Sweden”. It was established in 2010 and aims to provide access to advanced and latest facilities for life science researchers across major universities in Sweden, and hence research carried out at SciLifeLab have a broad range, “from understanding, diagnosing and treating diseases to microbial bioenergy production and environmental monitoring”. We collected 5284 WoS-covered publications for the SciLifeLab between 2007 and 2019, and its PP(top 10%) and connectivity are 22% and 0.37.

The two research institutes differ greatly in terms of scales and scopes of research activities. SciLifeLab covers rather broad research topics despite under the field of life science, while Hero-m shows a rather concentrated research interest on microstructure and mechanical properties of materials such as steels, cemented carbides, and advanced electronic materials. The use of these two cases enables us to further examine the performance of the present method. Detailed information regarding the research topics of Hero-m and SciLifeLab can be found in supplementary materials (Supplementary1 Research topics of Hero-m and SciLifeLab).

### Results

To examine the effectiveness and sensitiveness of this approach, two sets of parameter values were employed, which were summarized in Table 1. Specifically, we expect the first \( n \) clusters, ranked by the share of a treatment’s publications over clusters, to have about 80% of its total publications (\( d_{\text{min}} = 80P_{i,ts,te} \)). Furthermore, these clusters would be considered dominant only if they include more than five of the treatment’s publications (\( p_{\text{min}} = 5 \)). We further require the total publications of a potential benchmark in those dominant clusters to be within the range between 0.7 and 1.3 times the treatment’s total publications, which means \( \Delta p = 0.3 \). Also, the connectivity and PP(top 10%) of a benchmark

| Table 1 | Two sets of parameter values |
|---------|-------------------------------|
| Parameter sets for Hero-m | Parameter sets for SciLifeLab |
| Set—1 | Set—2 | Set—1 | Set—2 |
| \( d_{\text{min}} \) | 212 (80% \( P_{i,ts,te} \)) | 159 (60% \( P_{i,ts,te} \)) | 4227 (80% \( P_{i,ts,te} \)) | 3170 (60% \( P_{i,ts,te} \)) |
| \( p_{\text{min}} \) | 5 | 10 | 5 | 10 |
| \( \Delta p \) | 53 (30% \( P_{i,ts,te} \)) | 132.5 (50% \( P_{i,ts,te} \)) | 3785 (30% \( P_{i,ts,te} \)) | 2642 (50% \( P_{i,ts,te} \)) |
| \( c_{\text{min}} \) | 0.39 (\( c_{i,ts,te} \)) | 0.3 (80% \( c_{i,ts,te} \)) | 0.37 (\( c_{i,ts,te} \)) | 0.3 (80% \( c_{i,ts,te} \)) |
| \( h_{\text{min}} \) | 19% (\( h_{i,ts,te} \)) | 15% (80% \( h_{i,ts,te} \)) | 22% (\( h_{i,ts,te} \)) | 17.6% (80% \( h_{i,ts,te} \)) |

---

4 More detailed information regarding the Hero-m center is available at [https://www.kth.se/hero-m-2i/about-hero-m-2i](https://www.kth.se/hero-m-2i/about-hero-m-2i).

5 More information regarding SciLifeLab is available at [https://www.scilifelab.se/about-us](https://www.scilifelab.se/about-us).

6 Research at SciLifeLab: [https://www.scilifelab.se/research/?filter=all](https://www.scilifelab.se/research/?filter=all).
should be equal to or larger than the treatment \((c_{\text{min}} = c_{i,t,s,e})\) and \((h_{\text{min}} = h_{i,t,s,e})\). For the second set, the treatment’s publications in the dominant clusters should account for 60% of its total output and each dominant should include no fewer than 10 of the treatment’s publications. Accordingly, potential benchmarks are the research organizations that have publications in these dominant clusters. More specifically, a benchmark should satisfy the criteria that its total publications are between 0.5 and 1.5 times the treatment’s publications, and its connectivity and scientific impact are equal to or larger than 80% of the treatment. In short, the second set of parameters is

\[
d_{\text{min}} = 0.80P_{i,t,s,e}, \quad p_{\text{min}} = 5, \quad \Delta p = 0.3, \quad c_{\text{min}} = 0.80c_{i,t,s,e} \quad \text{and} \quad h_{\text{min}} = 0.80h_{i,t,s,e}.
\]

For Hero-m, the two sets of parameter values yielded 2 and 18 benchmark units respectively, and all identified from the first set of parameters were also founded by the second set. Table 2 presents the name and country of the identified organizations and also indicates that they were identified according to which set of parameters. The number of publications, connectivity, and PP(top10%) of each benchmark can be found in this table as well.

For SciLifeLab, no research organizations were identified with the use of the first set of parameters, this is mainly because of the criterion of research impact. The PP(top10%) of SciLifeLab is quite high at 22%, and unfortunately no potential benchmark units display an equivalent research impact using our methodology. As we relax parameter values, 102 benchmark units were identified for the SciLifeLab. We consider the number of identified organizations reasonable, the first set of parameter values is comparatively rigorous since it requires benchmark units to have quite similar characteristics with the treatment in each of

| Identified benchmark unit | Country | Set | No. of publications | Connectivity | PP(top10%) |
|---------------------------|---------|-----|---------------------|--------------|------------|
| University of Oxford     | UK      | Set—1 | 219                  | 0.44         | 28%        |
| University of Tennessee  | US      | Set—1 | 299                  | 0.46         | 28%        |
| University of Tennessee  | US      | Set—2 | 214                  | 0.62         | 29%        |
| Los Alamos National Laboratory | US | Set—2 | 343                  | 0.38         | 27%        |
| University of Oxford     | UK      | Set—2 | 143                  | 0.40         | 25%        |
| University of Michigan   | US      | Set—2 | 227                  | 0.49         | 25%        |
| Pennsylvania State University | US      | Set—2 | 153                  | 0.37         | 24%        |
| Delft University of Technology | NL | Set—2 | 249                  | 0.35         | 23%        |
| Pacific Northwest National Laboratory | US | Set—2 | 168                  | 0.50         | 20%        |
| Deakin University        | AU      | Set—2 | 141                  | 0.30         | 19%        |
| University of Paris-Saclay | FR | Set—2 | 137                  | 0.59         | 19%        |
| RWTH Aachen University   | GE      | Set—2 | 194                  | 0.38         | 18%        |
| The University of Texas at El Paso | US | Set—2 | 179                  | 0.48         | 18%        |
| Pohang University of Science and Technology | KR | Set—2 | 370                  | 0.35         | 18%        |
| Uppsala University      | SE      | Set—2 | 223                  | 0.56         | 17%        |
| Argonne National Laboratory | US | Set—2 | 179                  | 0.62         | 16%        |
| Northwestern Polytechnical University | CN | Set—2 | 210                  | 0.60         | 16%        |
| University of California, Berkeley | US | Set—2 | 141                  | 0.38         | 16%        |
| Shanghai Jiao Tong University | CN | Set—2 | 332                  | 0.53         | 16%        |
| Karlsruhe Institute of Technology | GE | Set—2 | 289                  | 0.55         | 16%        |
the four attributes. Due to space constraints, we visualized the identified benchmark units for the SciLifeLab with the program VOSviewer (Van Eck & Waltman, 2010). A network of collaboration relations was constructed based on the publications of the SciLifeLab and its identified benchmark units, as shown in Fig. 2. Specifically, nodes represent identified research organizations, with sizes proportional to the number of publications and colors proportional to the connectivity values, and links indicate the strength of collaboration. As shown, most identified organizations have already established collaboration relations between themselves. But it is also acceptable that some relatively isolated organizations were shown on the map such as Tsinghua University and the University of Barcelona, which we assume is mainly because they are the only ones identified from their respective countries.

It should be noted that we label the identified benchmark units using the name of their affiliated universities or institutions at an equivalent level. This is because address data of research organizations in WoS at the department level is relatively messy. Departments are also likely to experience transformations of organization structure, which leads to changes in research addresses. Therefore, to avoid such issues, the name of affiliated universities were used instead. Second, note that an identified unit could be a mixture of several departments at a university or could be a subgroup of a department. It should be stressed that researchers at a benchmark unit are not necessary to be physically located in the same working place, instead, they are required to present a well-connected collaboration network. A further discussion of this point can be found in the final section.

**Validation**

In this section, we further test the effectiveness of the proposed method. Fundamentally, the validation depends on determining whether the benchmark units and the treatment have a relatively large overlap in research interests. Accordingly, it is important to examine if the
identified benchmark units are consistent with our treatment in terms of research topics. Next, we also need to evaluate the robustness of using the minimum measure for assessing the connectivity of benchmark units. The reliability of using the number of publications and PP(top10%) will not be evaluated here, since they are well-developed measures to assess scientific output and research impact, and are frequently used in various types of research evaluations.

First, the Latent Dirichlet Allocation (LDA) was employed to generate underlying research topics from the publications of each identified benchmark. LDA is a natural language processing method in accordance with the assumption of “bag of words”. For LDA, each document can be characterized by a mixture of latent topics and each topic can be further represented by a set of terms (Blei et al., 2003). This work adopted an R package for fitting topic models (Hornik and Grün, 2011). In our work, abstracts of publications were used to generate topics. Some common terms were considered as stop words, such as “theory”, “data”, and “method” and so forth. Since Hero-m is not a multidisciplinary research center with broad and diverse research topics, we decided to generate simply two topics for this center and each identified benchmark. Furthermore, for each topic, we listed the first 20 terms according to their probabilities of the inferred topics. The result of LDA for both treatment and benchmarks are presented in Table 3 and supplementary materials (Supplementary2 Topics of benchmarks) respectively. Comparing the results in the two tables, the topics of Hero-m and its benchmarks are rather similar, focusing on properties, structure, and calculation of materials. But we have noticed that some benchmarks have one of the topics regarding nuclear, ray, radiation, and irradiation, which seems not to be a major research interest at Hero-m, at least from the Hero-m publications we collected. For instance, one of the research focuses at Pacific Northwest National Laboratory 7 is relevant to nuclear materials, which “refers to the metals uranium, plutonium, and thorium, in any form, according to the IAEA”.8 However, Hero-m has indeed a small number of publications on this topic, such as Li and Korzhavyi’s study in 2015 (Interactions of point defects with stacking faults in oxygen-free phosphorous-containing copper) and Xu and colleagues in 2020 (Nuclear and magnetic small-angle neutron scattering in self-organizing nanostructured Fe1–xCrx alloys). For this reason, some organizations with a focus on nuclear materials, such as Pacific Northwest National Laboratory, have been identified.

For SciLifeLab, we examined its similarity, in terms of research topics, with those identified benchmark units that they have not established any collaborations with, for instance the University of Barcelona. This selection was done since it is reasonable to assume that SciLifeLab and its collaborative organizations share research interests, in part determined by their co-published articles, so it provides a poor basis for evaluation. To evaluate the extent of subject overlap, we constructed a term map of keywords that were collected from the publications of both SciLifeLab the University of Barcelona (Fig. 3). The yellow nodes indicate that a keyword was used in the publications from both organizations, purple otherwise. As shown, yellow nodes scatter in the entire map, suggesting the overlap of research topics for the two organizations. Some differences in research interests for the two organizations were also observed. For instance, SciLifeLab presents fewer publications that have used the keywords “Chagas disease” (i.e. a tropical parasitic disease caused

---

7 More detailed information on Pacific Northwest National Laboratory are available at https://www.pnnl.gov/materials-science and https://en.wikipedia.org/wiki/Pacific_Northwest_National_Laboratory.

8 More information regarding nuclear materials is available at https://en.wikipedia.org/wiki/Nuclear_material.
| Topic-1                                                                 | Topic-2                                                                 |
|------------------------------------------------------------------------|------------------------------------------------------------------------|
| Hero-m Phase, alloys, experimental, properties, thermodynamic, calculations, magnetic, materials, simulations, interactions, temperature, calculated, material, energy, parameters, diffusion, principles, modeling, composition, order | Formation, grain, steels, carbon, high, surface, martensite, size, energy, density, ferrite, structure, growth, due, low, degrees, alloy, temperatures, atoms, micro-structure |
Despite that, we consider that the treatment and benchmark units having slightly different research profiles, is practical and acceptable. An identical distribution of publications over dominant research topics between a treatment and its benchmark units is not required in this study.

As discussed in the methods section, there are various measures to calculate connectivity in a weighted network. Taking the context of this analysis into account, we have chosen

![Fig. 3 Co-occurrence term map between SciLifeLab and the University of Barcelona](image)

|        | Arithmetic mean | Geometric mean | Minimum | Maximum |
|--------|-----------------|----------------|---------|---------|
| Arithmetic mean | 1               |                |         |         |
| Geometric mean   | 0.99            | 1              |         |         |
| Minimum          | 0.96            | 0.98           | 1       |         |
| Maximum          | 0.99            | 0.97           | 0.91    | 1       |

by Trypanosoma cruzi\(^9\) and “solid phase synthesis” (i.e. a type of chemical synthesis\(^10\)). Despite that, we consider that the treatment and benchmark units having slightly different research profiles, is practical and acceptable. An identical distribution of publications over dominant research topics between a treatment and its benchmark units is not required in this study.

As discussed in the methods section, there are various measures to calculate connectivity in a weighted network. Taking the context of this analysis into account, we have chosen

\(^9\) More detailed information regarding Chagas disease is available at [https://en.wikipedia.org/wiki/Chagas_disease](https://en.wikipedia.org/wiki/Chagas_disease).

\(^10\) More detailed information regarding solid phase synthesis is available at [https://en.wikipedia.org/wiki/Solid-phase_synthesis](https://en.wikipedia.org/wiki/Solid-phase_synthesis).
the minimum measure. However, it is necessary to examine if the results are robust when another measure was used. We have therefore applied the four different measures to calculate connectivity for the benchmark units of Hero-m (i.e. as shown in Table 2) identified with the use of both sets of parameter values. The results can be found in supplementary materials (Supplementary3 Measures of connectivity). We used the Pearson correlation coefficient to examine their consistency, which results are summarized in Table 4. As can be seen, connectivity values obtained with different measures present a high correlation with each other.

Discussion and conclusions

This study has defined benchmark research units and presented a model for how appropriate benchmarks can be discovered. In accordance with the definition, a framework for the identification was proposed. Using two research institutions as examples, we demonstrated the presented method and reported the identified benchmark units. Finally, various analyses were applied to assess the effectiveness of the proposed method.

Despite being frequently used in research management and evaluation, the term “benchmark” still lacks a precise conceptual framework. We hope that our elaboration on the definition of benchmark research units, as well as their most important attributes can help to rectify this situation. We, as bibliometric researchers, have noted that the expectation of benchmark units differs greatly among stakeholders from our practical work. Therefore, the construction of a precise definition can improve our understanding of the different aspects of benchmark research units, as well as clarify potential misunderstandings. Another merit is the flexibility of the proposed method, which allows users to adjust parameter values according to the purpose of their analyses. If, for instance, research impact is not the main concern in a certain analysis, one can easily set the parameter $h_{\text{min}}$ as zero; on the other hand, one can increase the parameter value $c_{\text{min}}$ if relatively integrated research environment are required.

One may argue that a treatment and identified benchmark units may show inconsistent distributions of publications over their dominant research topics. For instance, the treatment may have a similar number of publications in three dominant clusters; however, the publications of an identified benchmark unit can belong to merely one of these three clusters. We acknowledge the existence of such a scenario. Nevertheless, further restrictions on the distribution of publications seem unnecessary, since this study was carried out on a fine-grained science classification system and hence the three dominant clusters can be quite similar to each other. Even so, other ways to compare the full distribution of publications over research topics could improve the identification of benchmarks, and is an area of future research.

Besides the validation tests mentioned in the previous section, we have also examined the performance of using the WoS subject categories as research topics, instead of publication clusters. However, 69% of Hero-M’s publications belong to the category, Material Science Multidisciplinary, which is a rather broad research topic and cannot precisely describe the research focus of the center. Thus we believe a finer classification system must be used for this type of study. However, using a classification that is too granular can instead run the risk of creating self-referential categories, meaning that most publications in individual clusters are from one individual research group. By analyzing the share of organizations in each cluster, we found that, only in 21 clusters in our classification system, a single
organization accounts for more than 40% of the total publications. In other words, the classification system used here does not seem to present a self-referential problem.

Another challenge of this approach can be its reliability in terms of identifying benchmark units for multi- or inter-disciplinary research organizations. We consider that one may suffer from such problems if a coarse system, for instance, the WoS journal classification system, was applied to delineate research topics of organizations. For a fine-grained system constructed at the individual publication level, publications focusing on interdisciplinary research questions can be aggregated to form a cluster. We, therefore, suggest proceeding with the identification of benchmark units on fine-grained, publication-level classification systems.

**Limitations**

A potential limitation of this study is that an identified benchmark unit might consist of a group of connected researchers from several departments at the same university and might also be a subgroup of a university department. Considering the data quality issue, we avoid using the information of research addresses at the department level, which raises the difficulties in verifying if researchers at a potential benchmark are physically working together. Even if the data quality problem can be resolved to some extent by data cleaning or better algorithms to delineate research groups, other issues like organization transformations and researcher mobility are still quite thorny. However, the method requires the benchmark units to have a similar level of connectivity as the evaluative unit, which should be sufficient for most purposes, and we argue that co-location is not a mandatory requirement for benchmark units (or evaluative units), as long as they collaborate. To further understand this issue, we have examined the publications’ research address for each identified benchmark unit of Hero-m, through extensive data cleaning. In this case, we found that researchers at an identified benchmark unit tend to be a subgroup at the same department-level research organization, which summary can be found in Table 5. The reason for this might be the small size of Hero-m. But this may also suggest researchers working on the same topics are more likely to be assigned to the same sub-organization at a university. Therefore we believe a well-connected research environment is sufficiently meaningful for an appropriate benchmark unit.

Second, we acknowledge that our approach has a trade-off between precision and recall. Unfortunately, there is no well-defined pool of suitable benchmarks for each evaluative unit, and hence the magnitude of the trade-off is difficult to estimate. Nevertheless, we consider that precision is more important than recall in terms of the present research question. In other words, identifying appropriate comparable research units seems to be more meaningful in practical applications. It should be stressed that our approach is quite flexible, which allows users to adjust parameter values to detect benchmarks according to their specific purposes.

Finally, we also acknowledge that our approach may be hard to implement in practice, since it requires direct access to a bibliometric database and the ability to process the data. However, the current trend towards open bibliometric datasets (as part of the open science movement), such as OpenAlex, may make the application of these bibliometric methods more accessible in the near future.

---

11 More detailed information regarding OpenAlex is available at https://openalex.org.
Appendix

See Table 5.

### Table 5 Research address of identified benchmark units for Hero-m

| Identified benchmark unit | Research address |
|--------------------------|------------------|
| University of Oxford     | Dept Mat, University of Oxford |
| University of Tennessee  | Dept Mat Sci & Engn, Knoxville, University of Tennessee |
| Los Alamos National Laboratory | Los Alamos National Laboratory, US |
| University of Michigan   | Dept Nucl Engn & Radiol Sci, University of Michigan |
| Pennsylvania State University | Dept Mat Sci & Engn, Pennsylvania State University |
| Delft University of Technology | Dept Mat Sci & Engn, Delft University of Technology |
| Fac Aerosp Engn, Delft University of Technology | |
| Pacific Northwest National Laboratory | Pacific NW Natl Lab, Richland, WA 99,352 USA |
| Deakin University        | Ctr Mat & Fibre Innovat, Deakin University |
| University of Paris-Saclay | Lab Anal Microstruct Mat, University of Paris-Saclay |
| DEN Serv Rech Met Appl, University of Paris-Saclay | |
| RWTH Aachen University   | Dept Ferrous Met, Univ Aachen, Rhein Westfal TH Aachen |
| The University of Texas at El Paso | Dept Mat & Mat Engn, The University of Texas at El Paso |
| Ctr Struct & Funct Mat Res & Innovat, The University of Texas at El Paso | |
| Pohang University of Science and Technology | Grad Inst Ferrous Technol, Pohang University of Science and Technology |
| Uppsala University      | Dept Phys & Astron, Uppsala University |
| Argstrom Tribomat Grp, Uppsala University | |
| Argonne National Laboratory | Div Mat Sci, Argonne National Laboratory |
| Northwestern Polytechnical University | Natl Key Lab thermostruct Composite Mat, Northwestern Polytechnical University |
| Sch Mat Sci & Engn, Northwestern Polytechnical University | |
| University of California, Berkeley | Dept Nucl Engn, Berkeley, University of California, Berkeley |
| Lawrence Berkeley Lab, Div Mat Sci, University of California, Berkeley | |
| Shanghai Jiao Tong University | Natl Die & Mold CAD Engn Res Ctr, Shanghai Jiao Tong University |
| Karlsruhe Institute of Technology | Inst Appl Mat, Karlsruhe Institute of Technology |

**Supplementary Information** The online version contains supplementary material available at [https://doi.org/10.1007/s11192-022-04413-7](https://doi.org/10.1007/s11192-022-04413-7).

**Acknowledgements** The paper is an extended version of the ISSI2021 conference paper, Wang and Jeppsson (2021). A bibliometric strategy for identifying benchmark research units. 18th International Conference on Scientometrics & Informetrics. Leuven, Belgium. We would like to thank Lennart Stenberg for his helpful comments and suggestions. We are grateful to VINNOVA for their support of this study. We also thank reviewers for their comments on this paper.
Funding  Open access funding provided by Royal Institute of Technology.

Open Access  This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Amaral, P., & Sousa, R. (2009). Barriers to internal benchmarking initiatives: An empirical investigation. *Benchmarking: An International Journal, 16*(4), 523–542.

Anand, G., & Kodali, R. (2008). Benchmarking the benchmarking models. *Benchmarking: An International Journal, 15*(3), 257–291.

Andersen, J. P., Didegah, F., & Schneider, J. W. (2017). The necessity of comparing like with like in evaluative scientometrics: A first attempt to produce and test a generic approach to identifying relevant benchmark units. In STI conference science and technology indicators conference.

Bast, H., & Korzen, C. (2017). A benchmark and evaluation for text extraction from pdf. In 2017 ACM/IEEE joint conference on digital libraries (JCDL).

Bi, H. H. (2017). Multi-criterion and multi-period performance benchmarking of products and services: Discovering hidden performance gaps. *Benchmarking: An International Journal, 24*(4), 934–972.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993–1022.

Camp, R. C. (1989). *Benchmarking: The search for industry best practices that lead to superior performance*. Quality Press.

Carayol, N., Filliatreau, G., & Lahatte, A. (2012). Reference classes: a tool for benchmarking universities’ research. *Scientometrics, 93*(2), 351–371.

Carpinetti, L. C., & De Melo, A. M. (2002). What to benchmark? *Benchmarking: An International Journal, 9*(3), 244–255.

Dattakumar, R., & Jagadeesh, R. (2003). A review of literature on benchmarking. *Benchmarking: An International Journal, 10*(3), 176–209.

Ettorchi-Tardy, A., Levif, M., & Michel, P. (2012). Benchmarking: A method for continuous quality improvement in health. *Healthcare Policy, 7*(4), e101.

Fedor, D. B., Parsons, C. K., & Shalley, C. E. (1996). Organizational comparison processes: Investigating the adoption and impact of benchmarking-related activities. *Journal of Quality Management, 1*(2), 161–192.

Frenken, K., Heimeriks, G. J., & Hoekman, J. (2017). What drives university research performance? An analysis using the CWTS Leiden Ranking data. *Journal of Informetrics, 11*(3), 859–872.

Glänzel, W. (1996). The need for standards in bibliometric research and technology. *Scientometrics, 35*(2), 167–176.

González, E., & Álvarez, A. (2001). From efficiency measurement to efficiency improvement: The choice of a relevant benchmark. *European Journal of Operational Research, 133*(3), 512–520.

Henning, T. F., Muruvan, S., Feng, W. A., & Dunn, R. C. (2011). The development of a benchmarking tool for monitoring progress towards sustainable transportation in New Zealand. *Transport Policy, 18*(2), 480–488.

Hoffmann, F., Bertram, T., Mikut, R., Reischl, M., & Nelles, O. (2019). Benchmarking in classification and regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(5), e1318.

Hornik, K., & Grün, B. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1–30.

Hothorn, T., Leisch, F., Zeileis, A., & Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics, 14*(3), 675–699.

Hougaard, J. L., & Keiding, H. (2004). Continuous benchmark selections. *Operations Research Letters, 32*(1), 94–98.

Hougaard, J. L., & Tvede, M. (2002). Benchmark selection: An axiomatic approach. *European Journal of Operational Research, 137*(1), 218–228.
Huppler, K. (2009). The art of building a good benchmark. In Technology conference on performance evaluation and benchmarking. Springer.

Fitz-Enz, J. (1993). Benchmarking staff performance: How staff departments can enhance their value to the customer. Pfeiffer.

Jansen, J., de Vries, S., & van Schaik, P. (2010). The contextual benchmark method: Benchmarking e-government services. Government Information Quarterly, 27(3), 213–219.

Katz, J. S. (2000). Scale-independent indicators and research evaluation. Science and Public Policy, 27(1), 23–36.

V. Kistowski, J., Arnold, J. A., Huppler, K., Lange, K. D., Henning, J. L., & Cao, P. (2015). How to build a benchmark. In Proceedings of the 6th ACM/SPEC international conference on performance engineering.

Kromidha, E. (2012). Strategic e-government development and the role of benchmarking. Government Information Quarterly, 29(4), 573–581.

Kumar, A., Antony, J., & Dhakar, T. S. (2006). Integrating quality function deployment and benchmarking to achieve greater profitability. Benchmarking: An International Journal, 13(3), 290–310.

Li, Y., & Korzhavyi, P. A. (2015). Interactions of point defects with stacking faults in oxygen-free phosphorus-containing copper. Journal of Nuclear Materials, 462, 160–164.

Lou, Y., & Yuan, S. Y. (2019). On constructing alternative benchmark suite for evolutionary algorithms. Swarm and Evolutionary Computation, 44, 287–292.

Lucertini, M., Nicolò, F., & Telmon, D. (1995). Integration of benchmarking and benchmarking of integration. International Journal of Production Economics, 38(1), 59–71.

Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. Scientometrics, 36(3), 363–377.

Moriarty, J. P. (2011). A theory of benchmarking. Benchmarking: An International Journal, 18(4), 588–612.

Opsahl, T., & Panzarasa, P. (2009). Clustering in weighted networks. Social networks, 31(2), 155–163.

Noyons, E. C., Moed, H. F., & Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. Journal of the American Society for Information Science, 50(2), 115–131.

Petrović, M., Bojković, N., Anić, I., & Petrović, D. (2012). Benchmarking the digital divide using a multi-level outranking framework: Evidence from EBRD countries of operation. Government Information Quarterly, 29(4), 597–607.

Pryor, L. S. (1989). Benchmarking: A self-improvement strategy. The Journal of Business Strategy, 10(6), 28.

Rousseau, R. (2020). Benchmarkings and rankings. In R. Ball (Ed.), Handbook bibliometrics (pp. 299–309). De Gruyter Saur.

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. Journal of Informetrics, 9(1), 102–117.

Sankey, M., & Padró, F. F. (2016). ACODE Benchmarks for technology enhanced learning (TEL): Findings from a 24 university benchmarking exercise regarding the benchmarks’ fitness for purpose. International Journal of Quality and Service Sciences, 8(3), 345–362.

Tuzkaya, G., Sennaroglu, B., Kalender, Z. T., & Mutlu, M. (2019). Hospital service quality evaluation with IVIF-PROMETHEE and a case study. Socio-Economic Planning Sciences, 68, 100705.

Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. Journal of the American Society for Information Science and Technology, 64(2), 372–379.

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. Journal of the American Society for Information Science and Technology, 63(12), 2378–2392.

Waltman, L., & Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. The European Physical Journal B, 86(11), 471.

Wang, Q. (2018). A bibliometric model for identifying emerging research topics. Journal of the Association for Information Science and Technology, 69(2), 290–304.

Wang, Q., & Jeppsson, T. (2021). A bibliometric strategy for identifying benchmark research units. In 18th International Conference on Scientometrics & Informetrics (pp. 1229–1234).

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523–538.

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. Scientometrics, 63(2), 373–401.