Relevance and Quantification of Diversity in Bibliometric and Scientometric Applications

NICOLAS ROBINSON-GARCIA
EC3 Research Group, University of Granada

WOLFGANG GLÄNZEL
ECOOM and Dept. MSI, KU Leuven, Belgium

european
summer school
for scientometrics

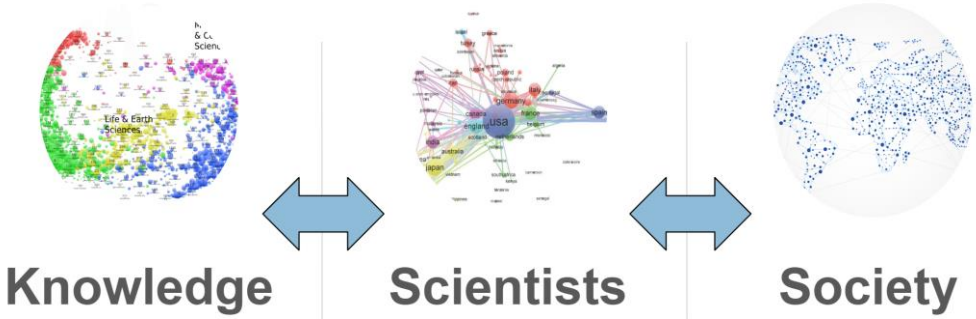
www.scientometrics-school.eu

https://doi.org/10.5281/zenodo.8354264

esss european summer school for scientometrics

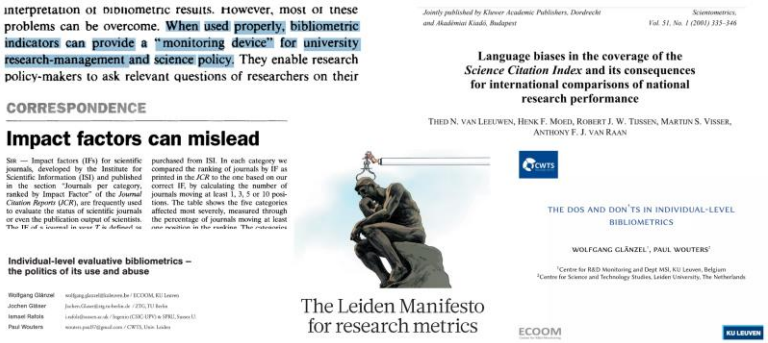
Part I: Relevance of Diversity in Bibliometric and Scientometric Applications

Diversity in scientometrics – Why does it matter?



The generation of knowledge as the result of co-existing **social, cognitive and cultural processes and actors**.

esss european summer school for scientometrics



Why does it matter?

- A far-fetched and deterministic normative view of the scientific reward system has led on many occasions to misuse and abuse of metrics.
- End with an approximation in which:

> *"cause" is a meaningful word, superlatives can be used, dichotomous thinking is realistic, with a resultant "zero-sum" mentality, and the "make a hypothesis -find a correlation" method makes sense.*

Moravcsik, 1987, p. 75

esss european summer school for scientometrics

Why does it matter?

- Providing context for interpreting citing and publishing differences
- Monitoring inclusion policies in research practices
- **Fostering a successful, sustainable, balanced and socially responsible scientific ecosystem**

esss european summer school for scientometrics

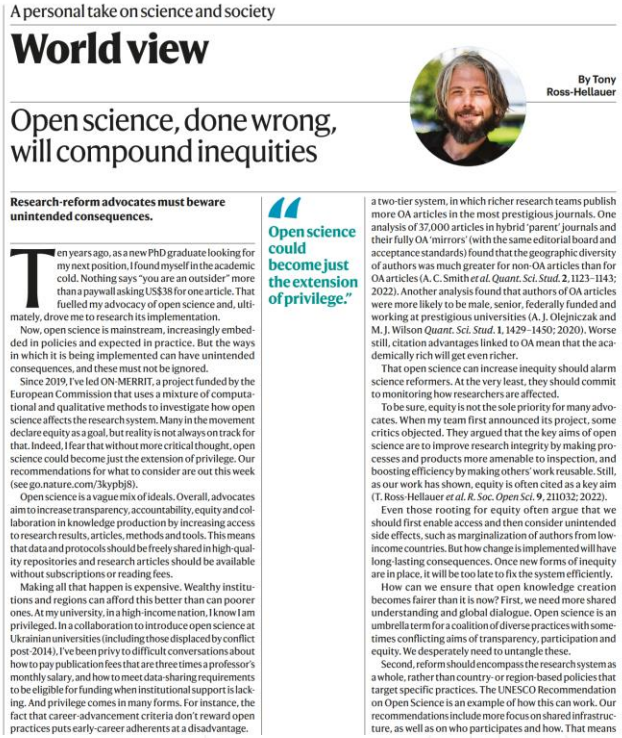
Part II: Quantification of Diversity in Bibliometric and Scientometric Applications

Diversity has many facets and manifestations

- Not all of those may be relevant to or measurable in bibliometrics and research assessment.
- Let's first revise some key concepts before looking into specific measurements of diversity.

How to proceed?

- **Object** or **actor** under study
  - E.g., people, outputs, institutions
- **Dimension**
  - E.g., cultural, racial, religious, gender, age, disability
- **Manifestation**
  - E.g., co-authorship, committees, publication types, language
- **Level** of manifestation
  - Relates to the durability and intensity of the manifestation



A personal take on science and society

World view

By Tony Ross-Hellauer

Open science, done wrong, will compound inequities

Research-reform advocates must beware unintended consequences.

Ten years ago, as a new PhD graduate looking for my next position, I found myself in the academic cold. Nothing says "you are an outsider" more than a paywall asking US$38 for one article. That fuelled my advocacy of open science and, ultimately, drove me to research its implementation.

Now, open science is mainstream, increasingly embedded in policies and expected in practice. But the ways in which it is being implemented can have unintended consequences, and these must not be ignored.

Since 2019, I've led ON-MERRIT, a project funded by the European Commission that uses a mixture of computational and qualitative methods to investigate how open science affects the research system. Many in the movement declare equity as a goal, but reality is not always on track for that. Indeed, I fear that without more critical thought, open science could become just the extension of privilege. Our recommendations for what to consider are out this week (see go.nature.com/3kypbj8).

Open science is a vague mix of ideals. Overall, advocates aim to increase transparency, accountability, equity and collaboration in knowledge production by increasing access to research results, articles, methods and tools. This means that data and protocols should be freely shared in high-quality repositories and research articles should be available without subscriptions or reading fees.

Making all that happen is expensive. Wealthy institutions and regions can afford this better than can poorer ones. At my university, in a high-income nation, I know I am privileged. In a collaboration to introduce open science at Ukrainian universities (including those displaced by conflict post-2014), I've been privy to difficult conversations about how to pay publication fees that are three times a professor's monthly salary, and how to meet data-sharing requirements to be eligible for funding when institutional support is lacking. And privilege comes in many forms. For instance, the fact that career-advancement criteria don't reward open practices puts early-career adherents at a disadvantage.

> "Open science could become just the extension of privilege."

a two-tier system, in which richer research teams publish more OA articles in the most prestigious journals. One analysis of 37,000 articles in hybrid 'parent' journals and their fully OA 'mirrors' (with the same editorial board and acceptance standards) found that the geographic diversity of authors was much greater for non-OA articles than for OA articles (A. C. Smith *et al. Quant. Sci. Stud.* **2**, 1123–1143; 2022). Another analysis found that authors of OA articles were more likely to be male, senior, federally funded and working at prestigious universities (A. J. Olejniczak and M. J. Wilson *Quant. Sci. Stud.* **1**, 1429–1450; 2020). Worse still, citation advantages linked to OA mean that the academically rich will get even richer.

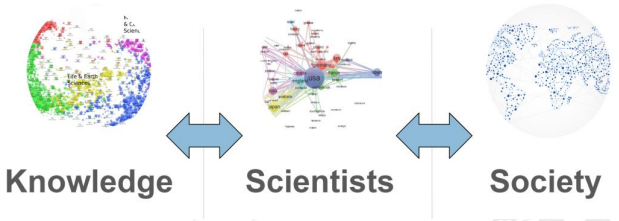
That open science can increase inequity should alarm science reformers. At the very least, they should commit to monitoring how researchers are affected.

To be sure, equity is not the sole priority for many advocates. When my team first announced its project, some critics objected. They argued that the key aims of open science are to improve research integrity by making processes and products more amenable to inspection, and boosting efficiency by making others' work reusable. Still, as our work has shown, equity is often cited as a key aim (T. Ross-Hellauer *et al. R. Soc. Open Sci.* **9**, 211032; 2022).

Even those rooting for equity often argue that we should first enable access and then consider unintended side effects, such as marginalization of authors from low-income countries. But how change is implemented will have long-lasting consequences. Once new forms of inequity are in place, it will be too late to fix the system efficiently.

How can we ensure that open knowledge creation becomes fairer than it is now? First, we need more shared understanding and global dialogue. Open science is an umbrella term for a coalition of diverse practices with sometimes conflicting aims of transparency, participation and equity. We desperately need to untangle these.

Second, reform should encompass the research system as a whole, rather than country- or region-based policies that target specific practices. The UNESCO Recommendation on Open Science is an example of how this can work. Our recommendations include more focus on shared infrastructure, as well as on who participates and how. That means



Actor/object

- **Knowledge**
  - Outputs – Language, document types
  - Subject – Topics, disciplines, words
- **Scientists**
  - Personal traits
  - Trajectory / Experience
- **Society**
  - Outreach – Social media, citations
  - Impact – Social, economic, scientific

esss european summer school for scientometrics

Dimensions

- Subject diversity (multi-/inter-disciplinarity/
- Regional diversity (e.g. in the context of internationalisation)
- Cultural diversity
  - Gender
  - Ethnicity
  - Language
  - Religion/confession and ideology/paradigms
  - Age
  - Sexual orientation
  - Disabilities

Manifestation (examples)

- *General terms* (examples)
  - Action: Collaboration in general and particular terms
  - Consolidation: Constitution of entities
  - Organisation: Diversity in processes
- *Particular terms* (examples)
  - co-authorship of publications
  - diversity in editorial boards, committees
  - peer reviewing, expert opinion

Level of manifestation

Usually four types or levels a distinguished

- *internal*, *external*, *organisational* and *global*

Most relevant manifestations in bibliometrics are found in scholarly communication (academic publications at various levels, such as individual documents, aggregated to journals, subjects, academic staff: doctoral students, teaching staff; research teams, departments, institutions)

From the viewpoint of producers: constitution of editorial boards, peer-reviewing processes, bibliodiversity, etc.

esss european summer school for scientometrics

Quantification

➢ *Quantification is a precondition for measurement.*

It requires the availability of relevant data. In the case of bibliometrics, this may include bibliographic databases, publication repositories, the Web.

- *First rule*: Only valid and reliable (available) information can be quantified.
- *Second rule*: Only information that can be made countable in an unambiguous manner may be quantified.
- *Third rule*: Not all attempts of quantification may result in meaningful data.

Quantification

Local vs. global approach

- **Local level** (meso/micro)
  - Quantification is often feasible (cf. teams, depts., institutes)
  - Large-scale analysis via bottom-up approach possible
- **Global level** (macro)
  - Top-down approach with breakdown to local units is problematic and requires tools with high degree of uncertainty and ambiguity (see examples below)

esss european summer school for scientometrics

Quantification (The case of gender identification)

Assumptions

1. *Author names provide evidence of gender*
2. *Gender is binary*

Operationalization

- Mix between universal and country-based approaches for gender identification
- Use of lists to identify gender (which have male overrepresentation, e.g., Wikidata)
- Geographical bias on the identification of gender

esss european summer school for scientometrics

Measurement

- *Not everything that can be quantified can also be measured.*
  - Prerequisites of measurement are commensurability of scales, applicability on a large scale, possibility of standardisation and normalisation (for benchmarking).
  - Beyond that, there are legal restrictions as well: Competing interests, confidentiality and possible conflicts, e.g., privacy issues.
    Actors may not wish to reveal their origin, sexual orientation, confession, disabilities, etc.

Examples from bibliometrics (1)

1. Subject and regional diversity

1.1 Subject variety as manifestation of interdisciplinarity

- Interdisciplinarity, which integrates knowledge from several disciplines, can be measures at many different levels of aggregation and using various granularities.
  - Quantification from the cognitive perspective is possible through citation-based or textual document analysis. − Can be achieved using bibliographic databases.
  - The organizational perspective provides limited opportunities via collaboration analysis. − Requires information from supplementary sources.

esss european summer school for scientometrics

Examples from bibliometrics (1)

1. Subject and regional diversity

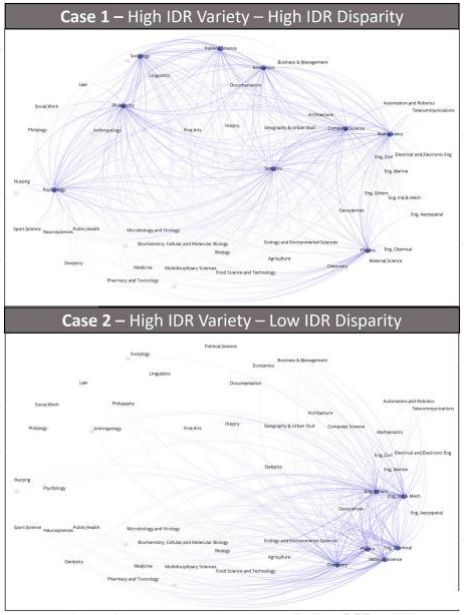
1.2 (Team) constitution as manifestation of internationalisation and multi-disciplinarity

Both subject profiles (multi-disciplinarity) and internationalisation is quantifiable at the local level.

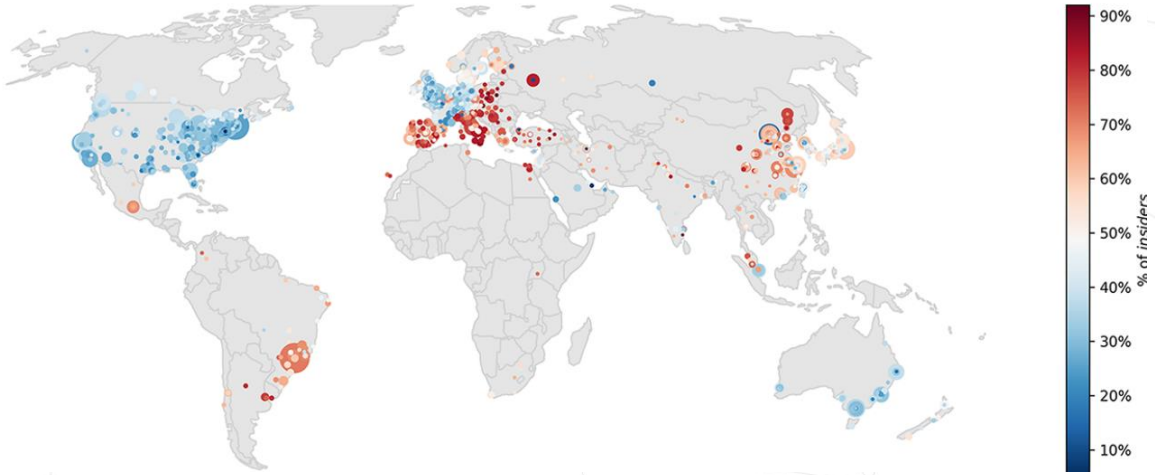
- Internally: through personal information of academic staff
- Externally (subject diversity): through projects, patents and publication output

Examples from bibliometrics (1)

1. Subject and regional diversity



Disciplines to which team members belong (D'Este & Robinson-Garcia, 2023)



Share of insiders (started working in the same institution) with 6 years of trajectory by university (Machacek et al., 2022)

Examples from bibliometrics (1)

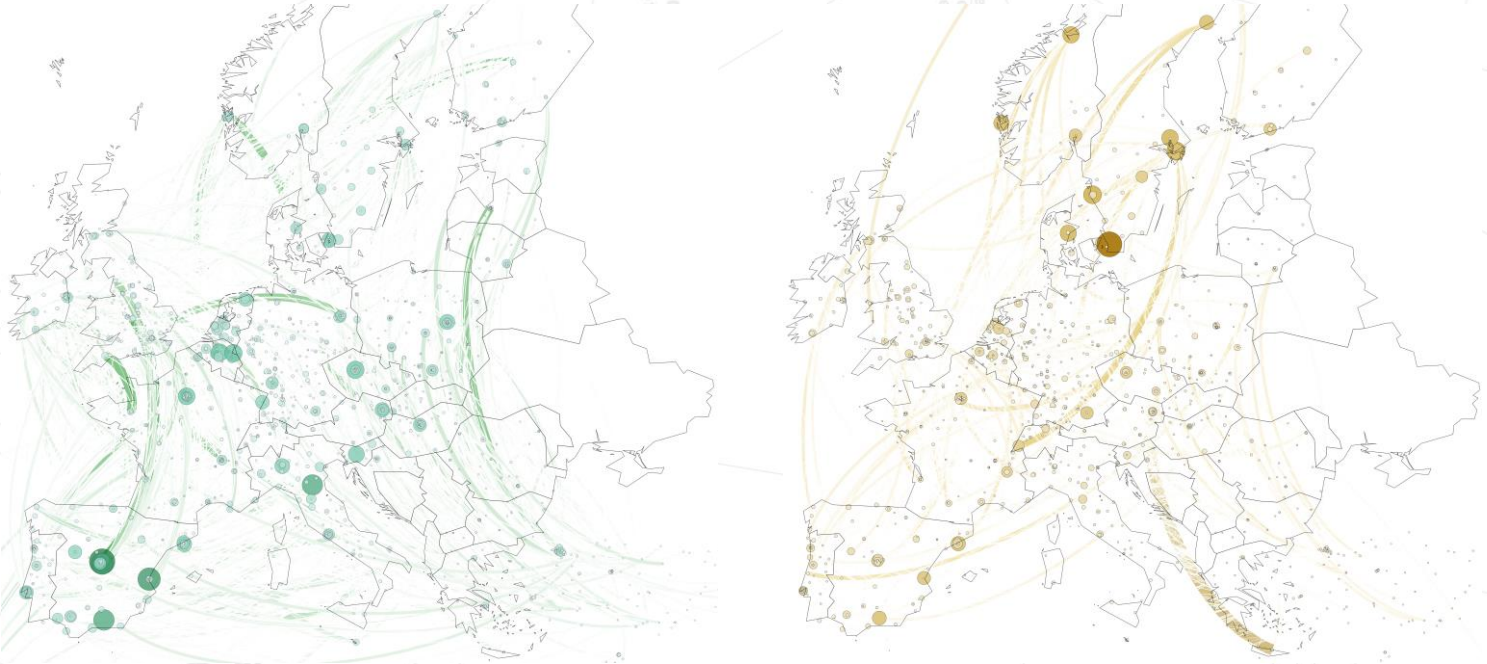
1. Subject and regional diversity

1.3 Mobility in the context of Internationalisation

- In practice, the distinction between internationality and international collaboration is rather difficult as the motivation for both may differ.
  - Academic mobility may serve as proxy for regional diversity, if properly separated from mere collaboration.
  - Data sources exist, even on the large scale (e.g., the Erasmus project, European Tertiary Education Register (ETER), supplemented by the Global Research Identifier Database (GRID) – see example on the next slide)

Examples from bibliometrics (1)

Example: International mobility of students



Students' mobility networks in Europe in the 2013-2014 academic year (Chi et al. (ISSI2020); Left: Social sciences, Right: Mathematics and Computing

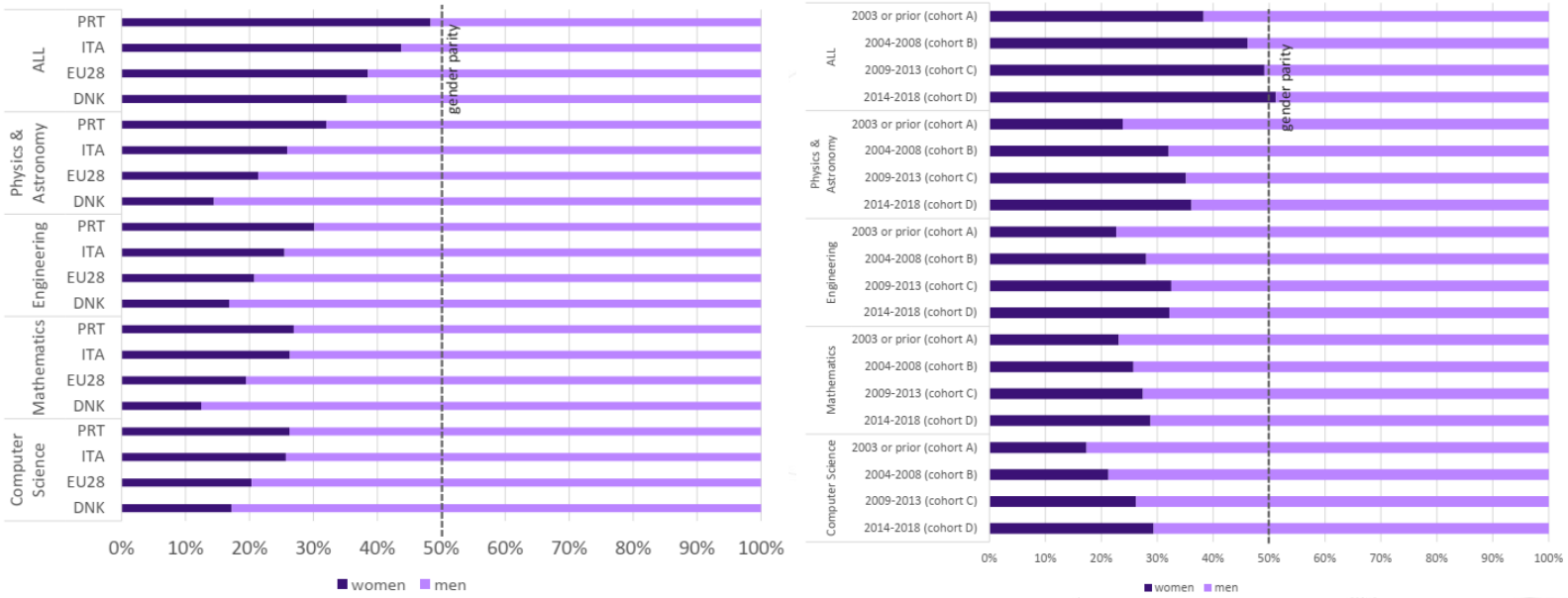
Examples from bibliometrics (2)

2. Gender/Ethnicity

- Quantification feasible at the local level using information from personal. Assignment of individuals may be unique.
- Several (large-scale) macro studies use information on doctoral students and teaching staff as well on authors of scientific publications.
  - This often requires supplementary sources and determination tools providing ambiguous information (e.g., genderize.io; namsor.app),
  - But, in turn, it allows comparative trend studies.
  - Multiple assignment if often made binary by “participation of”.

Examples from bibliometrics (2)

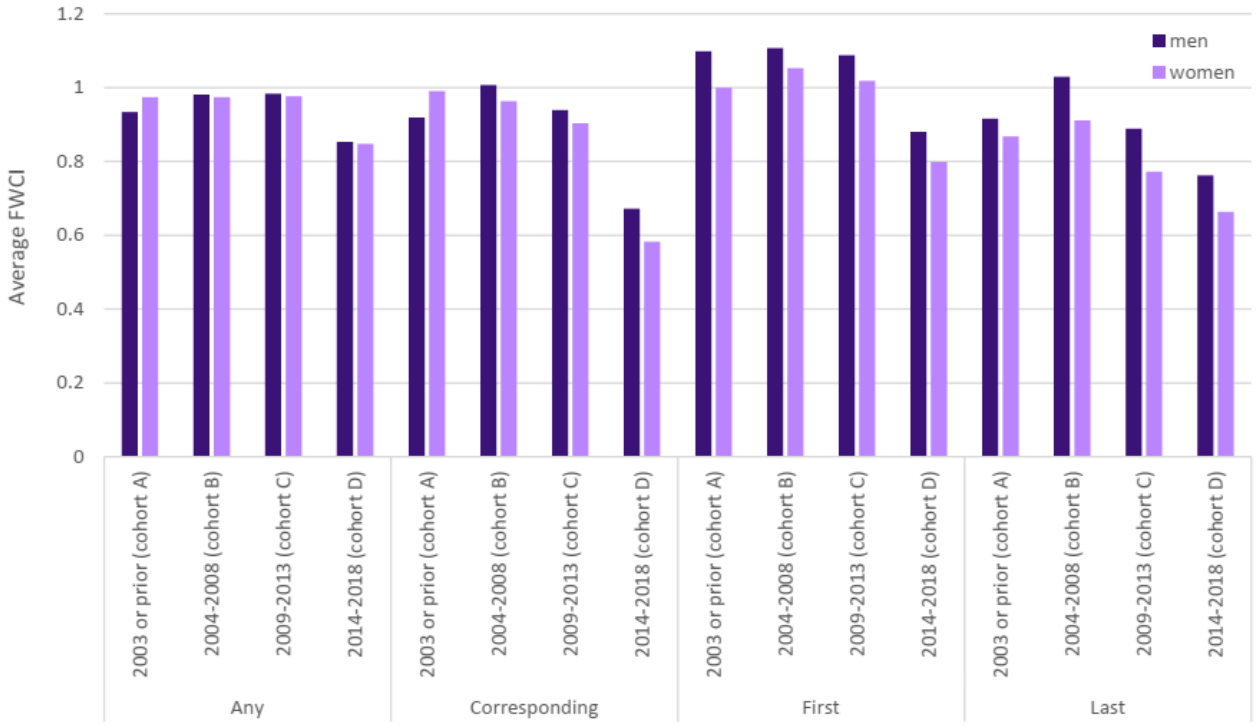
Example: Gender representation in Portugal's research



Share of women and men researchers in selected fields within the physical sciences in Portugal and comparators (2014–2018) – Left: all, Right: decomposed by seniority. *Source:* Elsevier (2021), Gender in the Portugal Research Arena: A Case Study in European Leadership. Data based on *Scopus and NamSor*

Examples from bibliometrics (2)

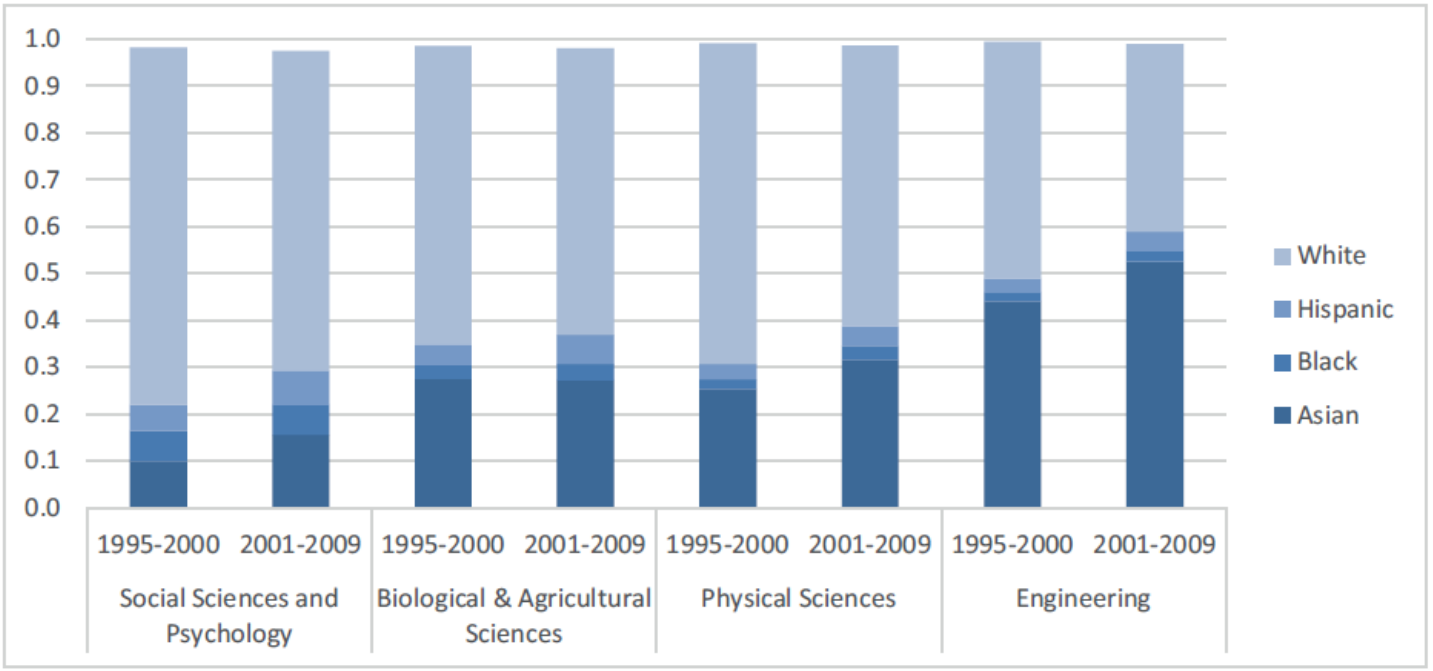
Example: Gender differences in citations (Portugal)



Mean field-weighted citation impact by gender, seniority and authorship position according to Elsevier's gender report on research in Portugal

Examples from bibliometrics (2)

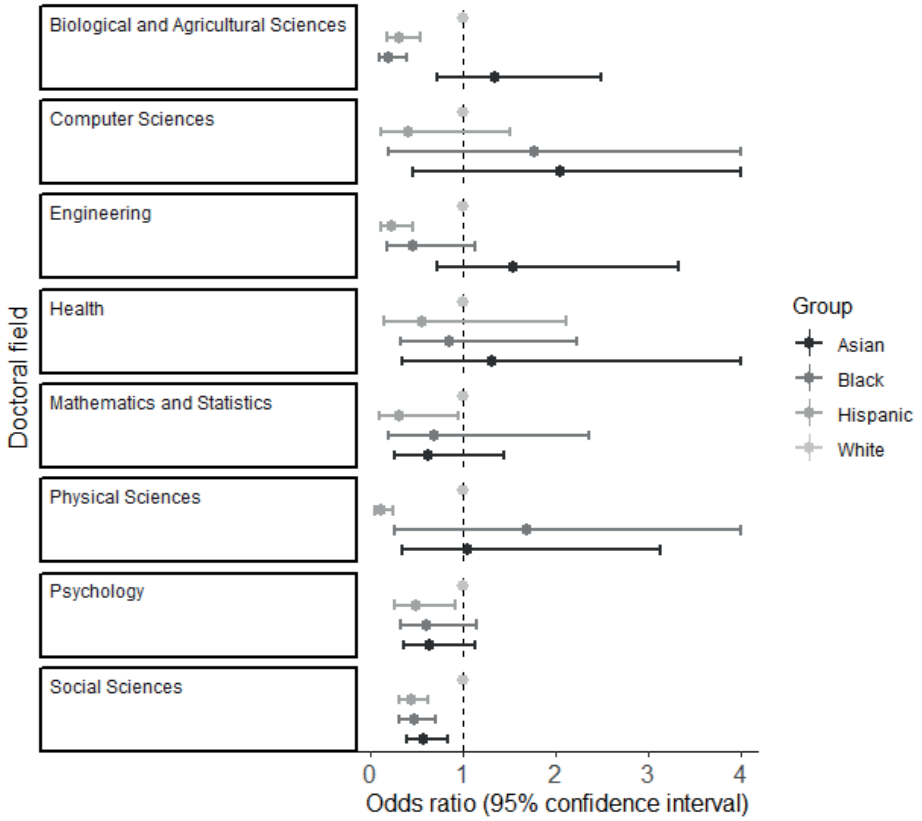
Example: Race/ethnicity of U.S. doctorate recipients



Share of doctorate recipients by selected doctoral field, race/ethnicity and cohort Group. *Source*: Chang et al. (ISSI, 2019). Data based on NSF, NCSES, SDR 2013

Examples from bibliometrics (2)

Example: Race/ethnicity of U.S. doctorate recipients

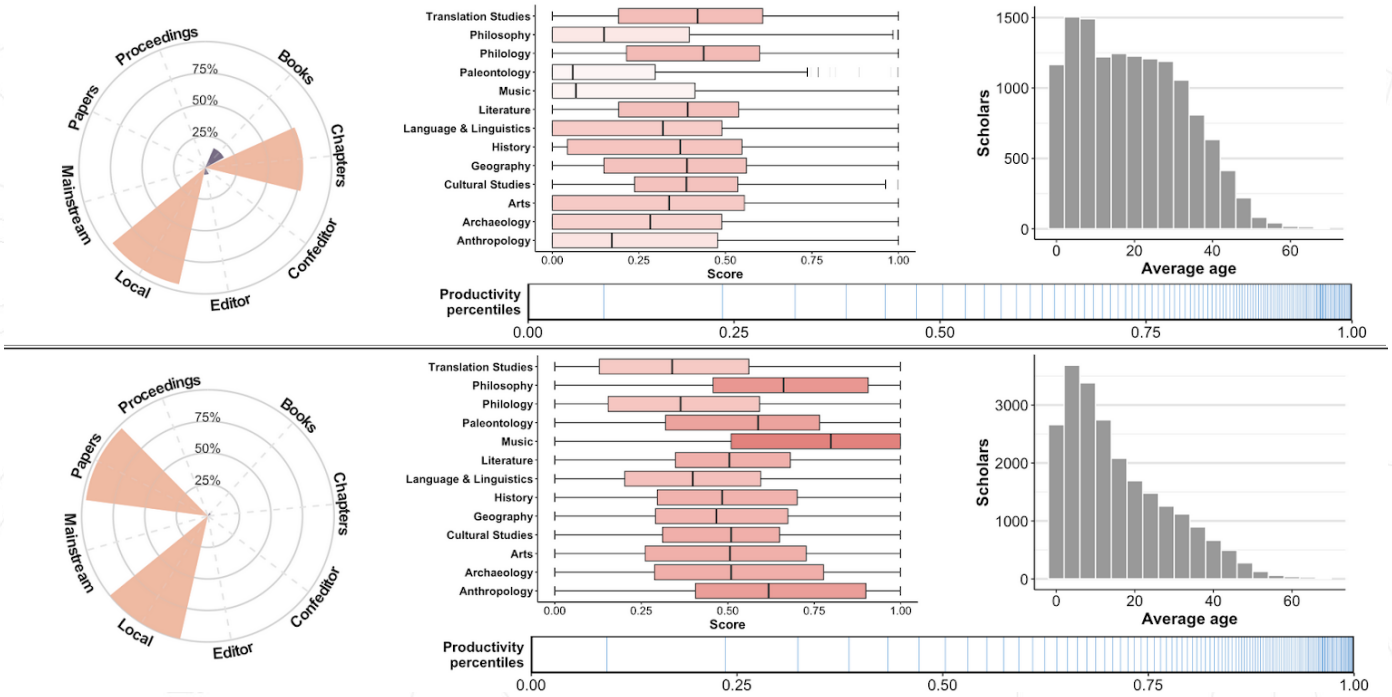


Estimated odds ratio of publishing of doctorate recipients employed in academia by race/ethnicity
Source: Chang et al. (ISSI, 2019). Data based on NSF, NCSES, Survey of Doctoral Recipients 2013

Examples from bibliometrics (3)

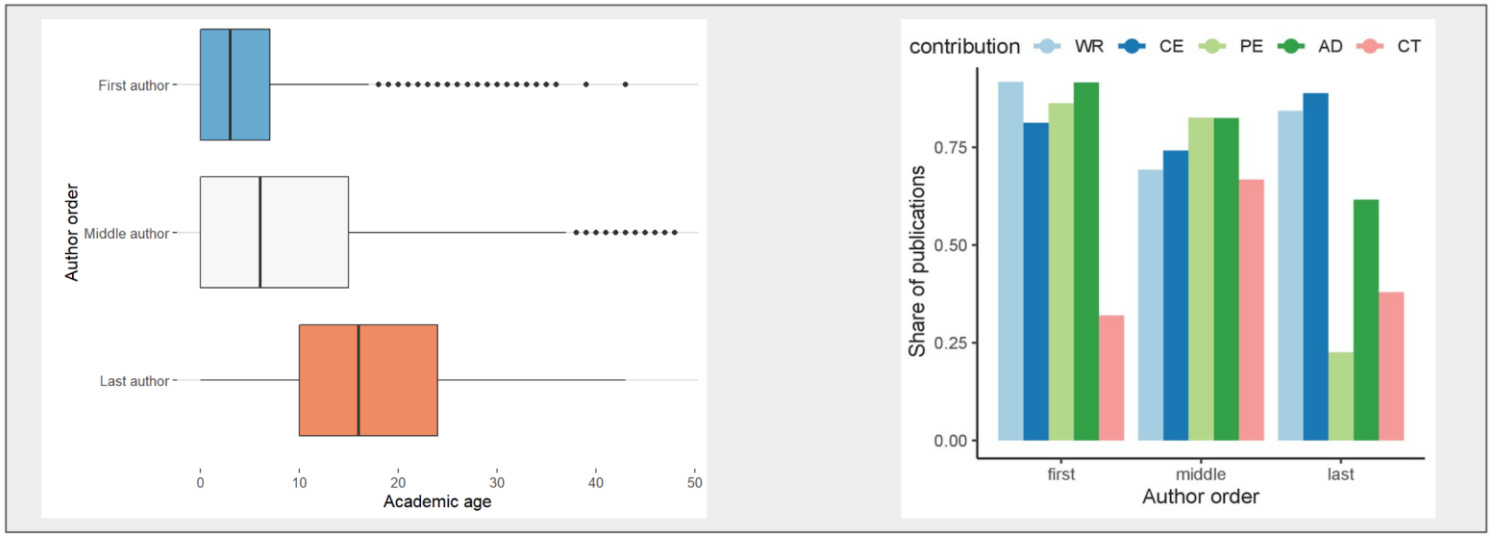
3. Bibliodiversity

Source: Author profiles from the Spanish SSH database Dialnet



Examples from bibliometrics (4)

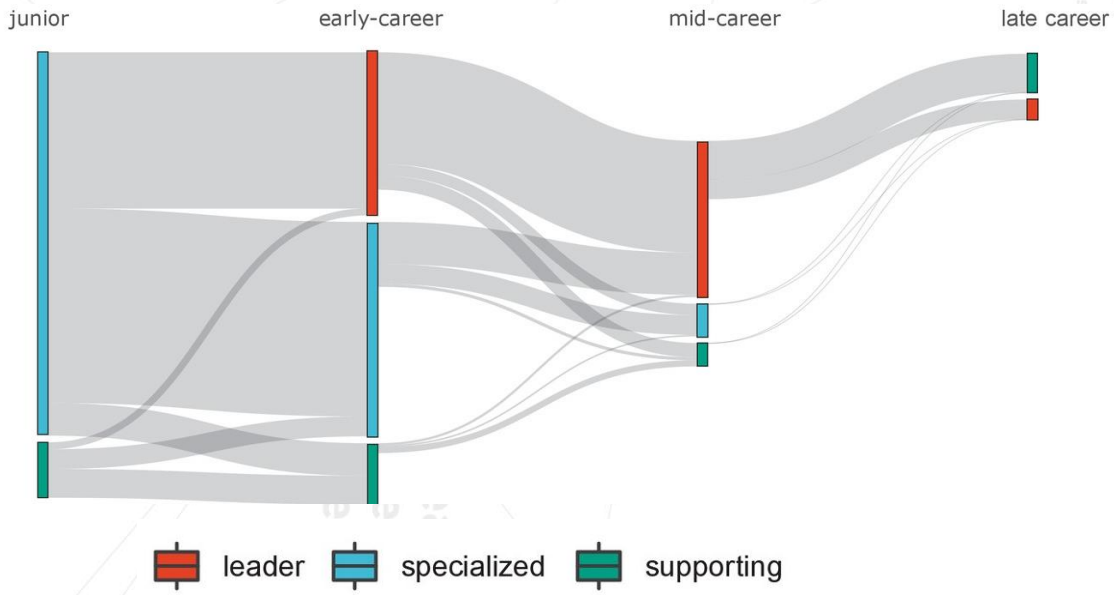
4. Age and contributions



Left: Academic age of researchers based on their author order. Right: Relation between author order and type of contribution (Robinson-Garcia et al., 2020)

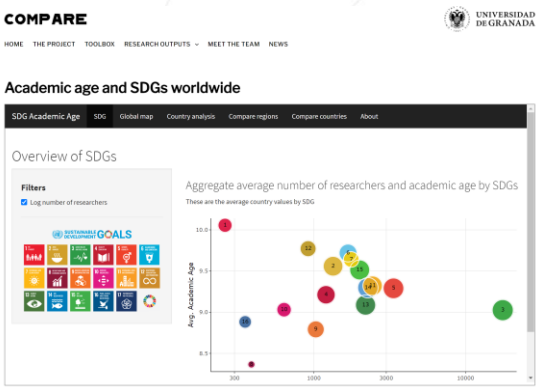
Examples from bibliometrics (4)

4. Age and contributions



SOME REMARKS

- Author order only used in predictive model but not archetypes
- Different generations of researchers included
- Researchers are *forced* into an archetype



Examples from bibliometrics (5)

5. Age and topics

Click here!

How old are on average researchers by SDGs?

Are there differences by countries and/or regions?

Other questions:

How does the age of the scientific workforce relate to citation impact?

Are there dependencies between countries?

Examples from bibliometrics (6)

6. Language and impact

Institution	Language	Share	Avg. Cit
University of Granada	English	88.8%	2.7
University of Granada	Spanish	10.8%	0.2
University of Granada	French	0.2%	0.1
University of Granada	Others	0.2%	0.0
Sapienza University Rome	English	98.11%	2.7
Sapienza University Rome	Italian	1.64%	0.1
Sapienza University Rome	Spanish	0.12%	0.1
Sapienza University Rome	Others	0.10%	0.1

Source: Web of Science
Time period: 2022

Examples from bibliometrics (6)

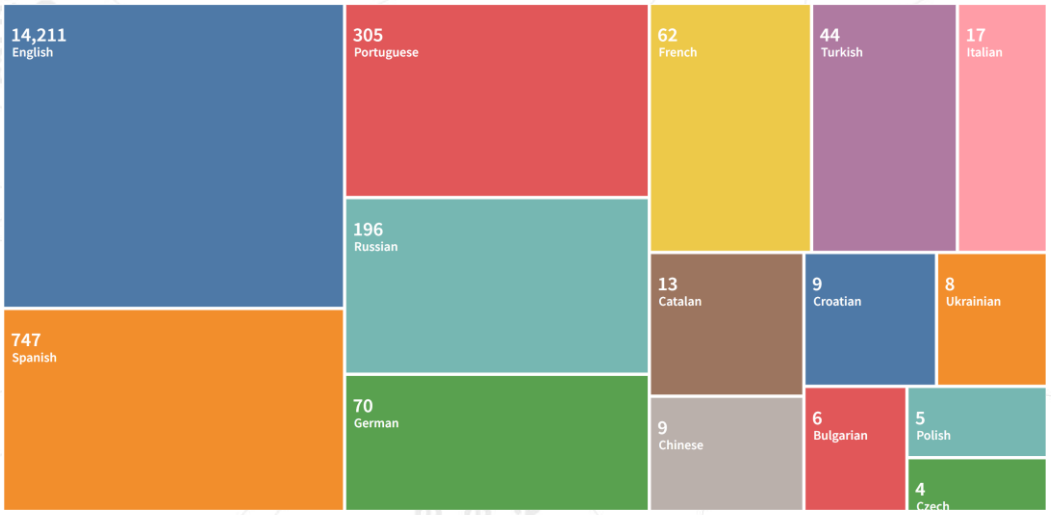
6. Language and impact

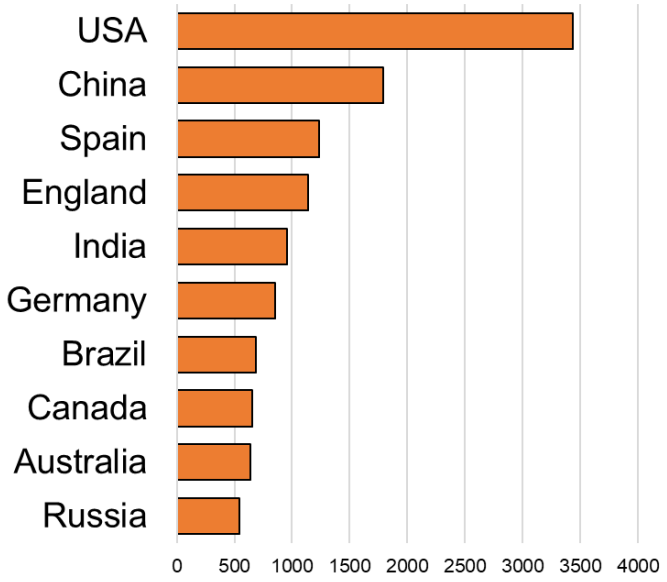
Here it is important to consider the following when interpreting these figures:

- Subject profile
- Journal coverage and differences by language and country
- Size of the target audience. *E.g., there are more Spanish speaking countries than Italian*

Diversity in the mirror of scientometric research

Publications within the micro topic of Bibliometrics between 2018 and 2022. Source: Web of Science





Left: Composition of papers by language Right: Top 10 most productive countries

Some conclusions

Scientometrics allows the analysis of effects of diversity in terms of academic productivity, collaboration and various impacts, including citation and broader impact.

Despite these opportunities, there are a number of caveats to avoid most serious pitfalls as pointed to below.

- The effects of diversity are influenced by several factors that are often not independent from each other. Subject field, Age, Seniority are among the most important ones.
- The communities and entities to which these factors can be attributed have their own specific peculiarities.

Some conclusions

- Ignoring this may cause strong biases, even distortions of the results.
  - The example on slide 29 may be biases i.a. by geographical regions, the subject field, language, and the coverage of the data source and the superposition of these and other effects.
  - The separation of motivation factors for mobility (slide 19) on the large scale, notably for academic staff is difficult.
- The quality of underlying the data is crucial for building and applying responsible metrics (cf. lecture by S. Gauch on “New Assessment Systems”).
  - Kozlowski et al. (PLoS ONE, 2022) showed that the validity of name-based inference varies by race/ethnicity and may result in biases.

esss european summer school for scientometrics

Some conclusions

- Identification, e.g., of gender, and ethnicity of actors using automated techniques on the large scale is subject to limitations.
  - Assignment to cohorts also remains an often haphazard decision. Determination of author contribution on the basis of position in the co-authorship list is an error-prone approach.
  - Publication record as used in the context of social stratification of authors in research collaboration is as questionable proxy for seniority or academic position in the context of diversity studies.
- Scientometric diversity studies may be useful in depicting specific scholarly communication patters and monitoring trends by quantitative methods, but the use of “diversity indicators” in research assessment exercises or any evaluative context would require *responsible metrics*.

Epilogue: "Unimpeded bibliometric standards"

István Örkény (1912–1979): More One-Minute Stories. Corvina, Budapest, 2007. Selected and translated by Judith Sollosy.

Unimpeded production standards

"Hello? Machine shop?"
"Skultéti here."
"How much, Skultéti?"
"Thirty-three, Comrade."
"What's thirty-three, Skultéti?"
"What's thirty-three, Comrade?"
"Yes, what's thirty-three, Skultéti."
"Why? Wasn't thirty-three the right answer, Comrade?"
"The right answer to what, Skultéti."
"To your question, Comrade."
"Never mind, Skultéti, just resume where you left off."

(*Heavy industry folklore*, 1978)