Andres, E. et al.

Conference Paper — Manuscript Version (Preprint)

Who studies whom? An Analysis of Geo-Contextualized Sustainable Development Goal Research

Suggested Citation: Andres, E. et al. (2023) : Who studies whom? An Analysis of Geo-Contextualized Sustainable Development Goal Research, 27th International Conference on Science, Technology and Innovation Indicators (STI 2023), International Conference on Science, Technology and Innovation Indicators, Leiden,
https://dapp.orvium.io/deposits/6442ef3679d562b66027191b/view

This Version is available at:
http://hdl.handle.net/11108/612

Who studies whom? An Analysis of Geo-Contextualized Sustainable Development Goal Research

E. Andres[1], E. Calò[2], R. Maria del Rio-Chanona[3,4], Marcia R. Ferreira[3,6], Maria Henkel[5], J. Számely[1], L. Yang[3]

first.author@address.com
ORCID
[1]Central European University, Vienna, Austria

second.author@address.com
ORCID
[2]IMT School for Advanced Studies Lucca, Lucca, Italy

third.author@address.com; fourth.author@address.com
https://orcid.org/0000-0001-5337-4637; https://orcid.org/0000-0002-0189-7919
[3]Complexity Science Hub, Vienna, Austria

[4]Center for International Development, Harvard Kennedy School, United States

fifth.author@address.com
https://orcid.org/0000-0002-5000-104X
[5]ZBW Leibniz Information Centre for Economics, Kiel, Germany

[6]Faculty of Informatics, Vienna University of Technology, Vienna, Austria

Abstract

This study examines the representation of countries in research related to the United Nations Sustainable Development Goals (SDGs), with a focus on potential biases due to factors such as geography, language, and access to resources. We start with a dataset of 6.7 million SDG-related publications, then extract the country of focus (i.e. the country that the publication's research focuses on) and country of origin (i.e. authors' institutions of affiliation). The resulting subsample of almost 50,000 publications is used to study country frequencies and construct a geographical research network. The results indicate that there are significant imbalances in research attention and funding for the SDGs, with wealthier countries and those with greater research resources being overrepresented. This study highlights the importance of greater global cooperation to ensure that research on the SDGs accurately reflects the needs and priorities of all countries.

1. Introduction

Since 2015, all 193 United Nations member states have adopted the 2030 Agenda for Sustainable Development (United Nations, n.d.). The 17 Sustainable Development Goals (SDGs) are at its core. They are interconnected and aim at improving social, economic, and environmental sustainability by and for "all countries – developed and developing – in a global partnership" (United Nations General Assembly, 2015, n.p.). Achieving them requires a global collaborative effort – especially in research, which is an important driver of evidence-based decision-making for, as well as implementation and monitoring of the SDGs, as emphasized by the 2030 Agenda (United Nations General Assembly, 2015).

Consequently, the SDGs have gained significant research attention in recent years. Much of the research in this context has a specific geographical focus, with most studies examining specific continents, countries, or other specific regions. This regional focus can lead to representation inequities in research between countries.

Differences in the representation of countries, languages, and cultures can be observed in published research in general. In 2003, Kofi Annan, secretary-general of the United Nations at the time, emphasized the need for a more balanced distribution of research efforts and resources to bridge the gap between developed and developing countries, ensuring that the benefits of scientific advancements reach all of humanity. (Annan, 2003). Salager-Meyer (2008) called attention to the increasing North/South disparities and the importance of multilingualism for the success of developing countries in the international research field. More recently, Matthews et al. (2020) surveyed 9000 researchers from eight countries about international research collaboration. Responses indicated that, while research is increasingly conducted internationally, researchers from developing countries still face various barriers, including a lack of resources, bureaucratic paperwork or support from local institutions as well as national, racial, and ethnic biases. Vieira et al. (2022) analysed publications from periods between 1990–1999, 2000–2009 and 2010–2018 regarding international research collaborations and found a decreased impact from geographical and cultural factors, while the impact from socioeconomic, political, and intellectual differences increased.

At the same time, few investigations have been conducted regarding discrepancies in the representation of different countries in activities relating to the UN's Sustainable Development Goals. Blicharska et al. (2021) examined the involvement of countries from the global North and South in SDG partnerships, concluding that "partners from low-income countries (...) were involved in far fewer partnerships" (p. 8) and fearing perpetuation of the North–South divide.

Together, these and other factors may lead to biases in research and funding that may not accurately reflect the needs and priorities of the SDGs.

To examine the likelihood of possible over- or under-representation of the countries under study, we create a global network analysis of geo-contextualized SDG research. Our aim in this ongoing study is to uncover and raise awareness for the inequalities concerning SDG research in the data.

2. Methods

Our analysis is based on a dataset of roughly 6.7 million publications that fall within the scope of the SDGs (United Nations General Assembly, 2015). The publication data was collected from the Dimensions database by Digital Science[1] for the period 1980 to 2022. This dataset records publication metadata such as title, abstract, author-affiliation linkages, geo-localised affiliations, citations, and whether the publication is assigned to one (or more) of the 17 SDGs.

[1] https://www.dimensions.ai

We extracted, for as many publications as possible, the *country of focus*, i.e. the country that the publication's research focuses on, and the *country of origin* of the study, i.e. the home countries of authors' institutions of affiliations. We allow for a publication to have multiple countries of focus and countries of origin. For example, when the publication studies a region comprising several countries or when there are authors with affiliations based on different countries. We do this as follows.

2.1 Identifying countries of origin and focus

To identify the geographic focus of SDG-classified publications, we utilize Name Entity Recognition (NER). Made possible by recent advances in Deep Learning, today's state-of-the-art NER allows automated and accurate categorization of Named Entities (NEs), such as public figures, cities or countries, through pre-trained neural network models (Li et al., 2022). To analyse the available texts, titles, and abstracts of each publication, we use the spaCy Natural Language Processing (NLP) library[2]. Using its pre-trained transformer model for the English language[3], we extracted locations (LOC), geopolitical entities (GPE), languages (LANGUAGE), nationalities, and religious or political groups (NORP). The used model achieves high precision NER on general English texts[4]. In our tests with selected publications, it performed similarly well on titles and abstracts of these publications, with reduced precision due to linguistic peculiarities of scientific texts (e.g., formulas, scientific terminology).

To identify the country of focus, we used the identified NEs that corresponded to a country, a region, or a city. The related country of those locations is the country of focus (e.g., 'New York' and 'United States'). We discarded all NEs that referred to non-geographic entities (e.g., names of proteins) and natural geographical entities such as lakes, mountains, and rivers. For this first analysis, we also discarded continents, if we could not infer the country from the context. To identify the countries of origin, we use the metadata provided by Dimensions. For this part, we discarded publications that did not have the affiliation of *all* authors. This led to a substantial reduction in our dataset (see section 3. Results).

2.2 Building the SDGs geographical research network

Using the identified countries of origin and countries of focus for each publication related to at least one SDG, we build a weighted and directed network. In this network, an edge connects the country of origin (determined by the authors' affiliation) with the country of focus (determined by NER). The weight is the total number of authors from the origin country studying the country of focus. We note that one publication can contribute to the weight of different edges. This is possible since we allow publications to have multiple origin countries (when there are authors of different countries of affiliations) and multiple focus countries (when more than one country is mentioned in the abstract).

2.3 Analysis methodology

We first took a descriptive approach to analyse the data. We measured the number of times a country appears in publications as an origin country and how many times as a country of focus

[2] https://spacy.io/

[3] https://spacy.io/models/en#en_core_web_trf

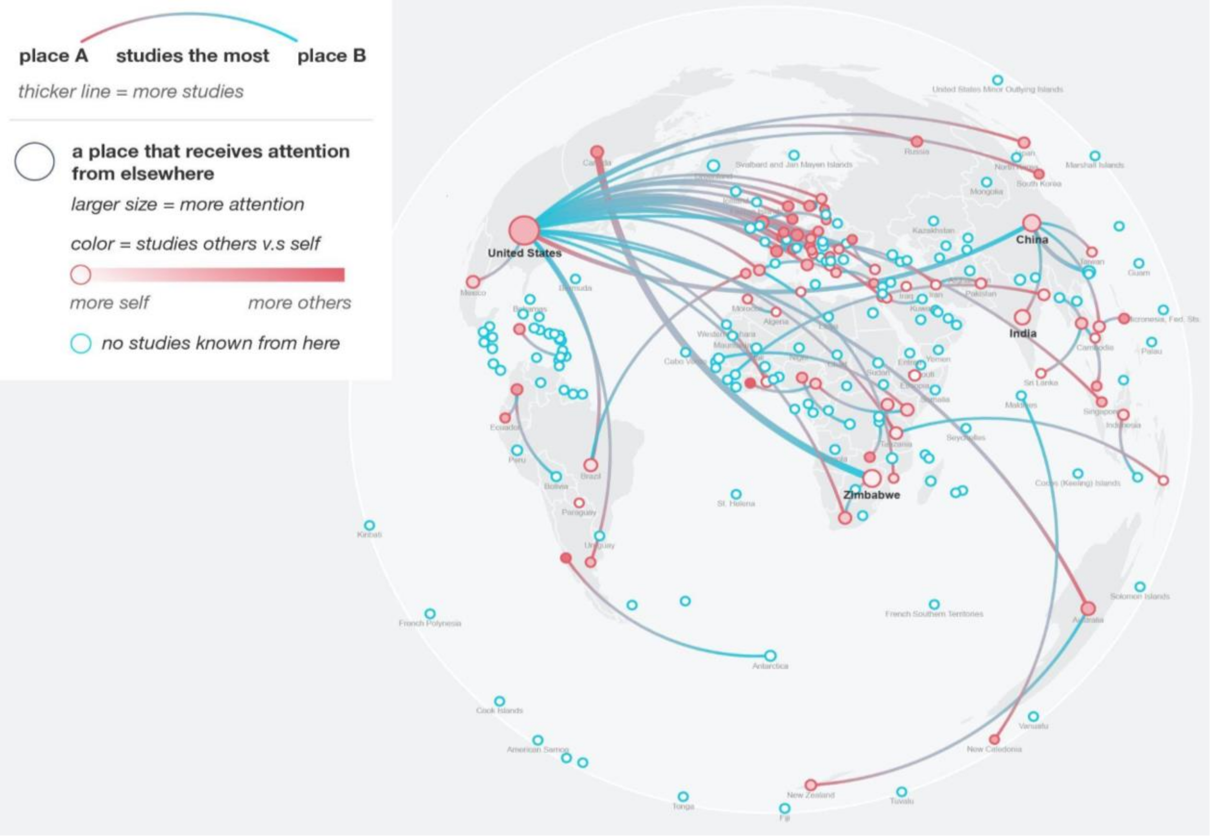
[4] https://spacy.io/models/en#en_core_web_trf-accuracy

for each SDG. We then used network analysis tools to investigate the community structure of the geographical research network using the Louvain algorithm (Blondel et al., 2008), which is the community detection algorithm that resulted in the highest modularity out of the state-of-the-art community detection algorithms we tested. Furthermore, we calculated the eigenvector centrality of the nodes (Newman, 2018). Finally, we use country metadata to understand our results better. In particular, we used size and economic indicators, such as the GDP per capita and population of the countries and their relationship to the number of publications where a country appears as the focus country.

3. Results

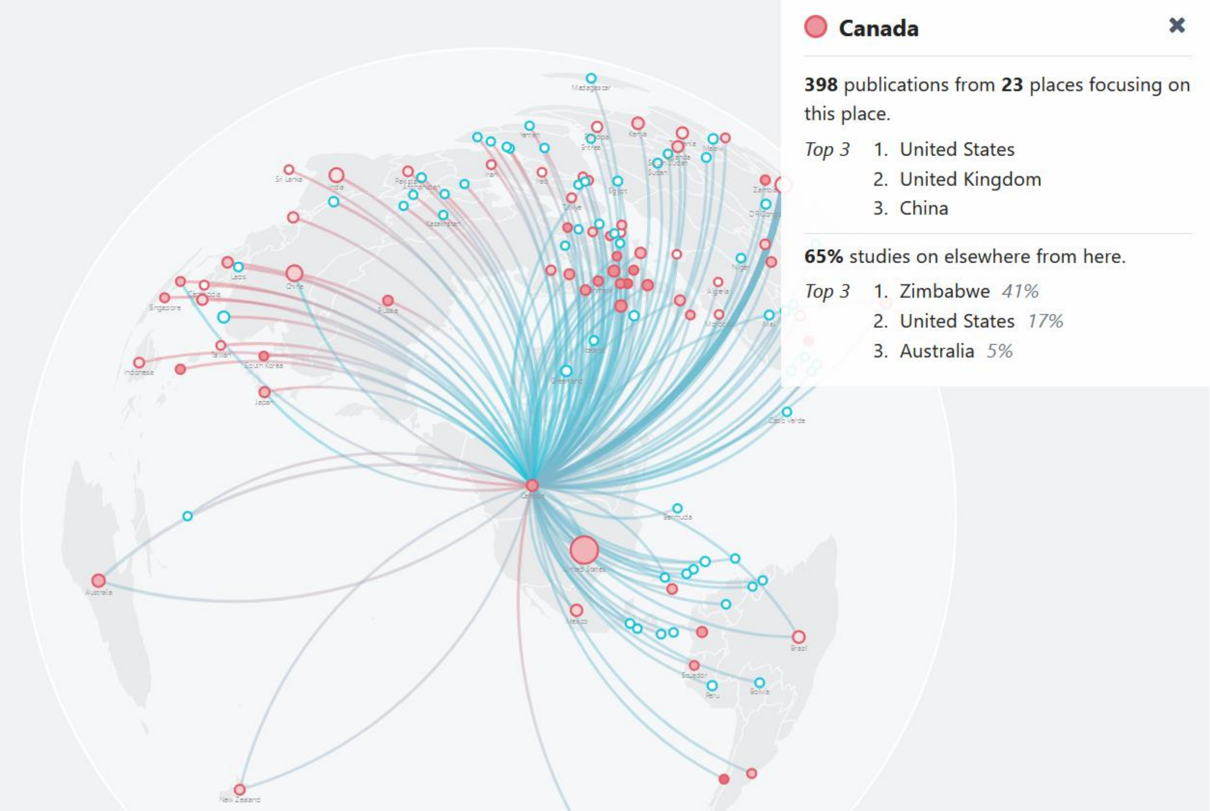
Out of the 6.7 million publications categorized under at least one SDG, we were able to extract Named Entities that were countries for 1.8 million publications. Out of these publications, roughly 49,000 had geo-locatable affiliations for *all* the authors. The results from this article are based upon this dataset of 49,000 publications.

Figure 1: Geographical SDG Research Network. Nodes with blue source links represent countries that study other countries or themselves, nodes with red targeted links represent countries of focus. Find the interactive version of the network here: https://vis.csh.ac.at/who-studies-whom/



We find that for 61% of the publications, there was at least one author with an affiliation of the country of focus. This result indicates that researchers are relatively likely to study their country of affiliation. Figure 1 shows the geographical SDG research network – a weighted directed network of country of origin and country of focus. Nodes with blue source links represent countries that study other countries or themselves, nodes with red targeted links represent countries of focus. The network can be explored online via https://vis.csh.ac.at/who-studies-whom/. Clicking on a country will show all countries of origin and countries of focus connected with that particular place by the publications in our dataset (see Figure 2).

Figure 2: Geographical SDG Research Network. Detail view for Canada, showing Top countries of origin and Top countries of focus.



Canada

398 publications from **23** places focusing on this place.

Top 3
1. United States
2. United Kingdom
3. China

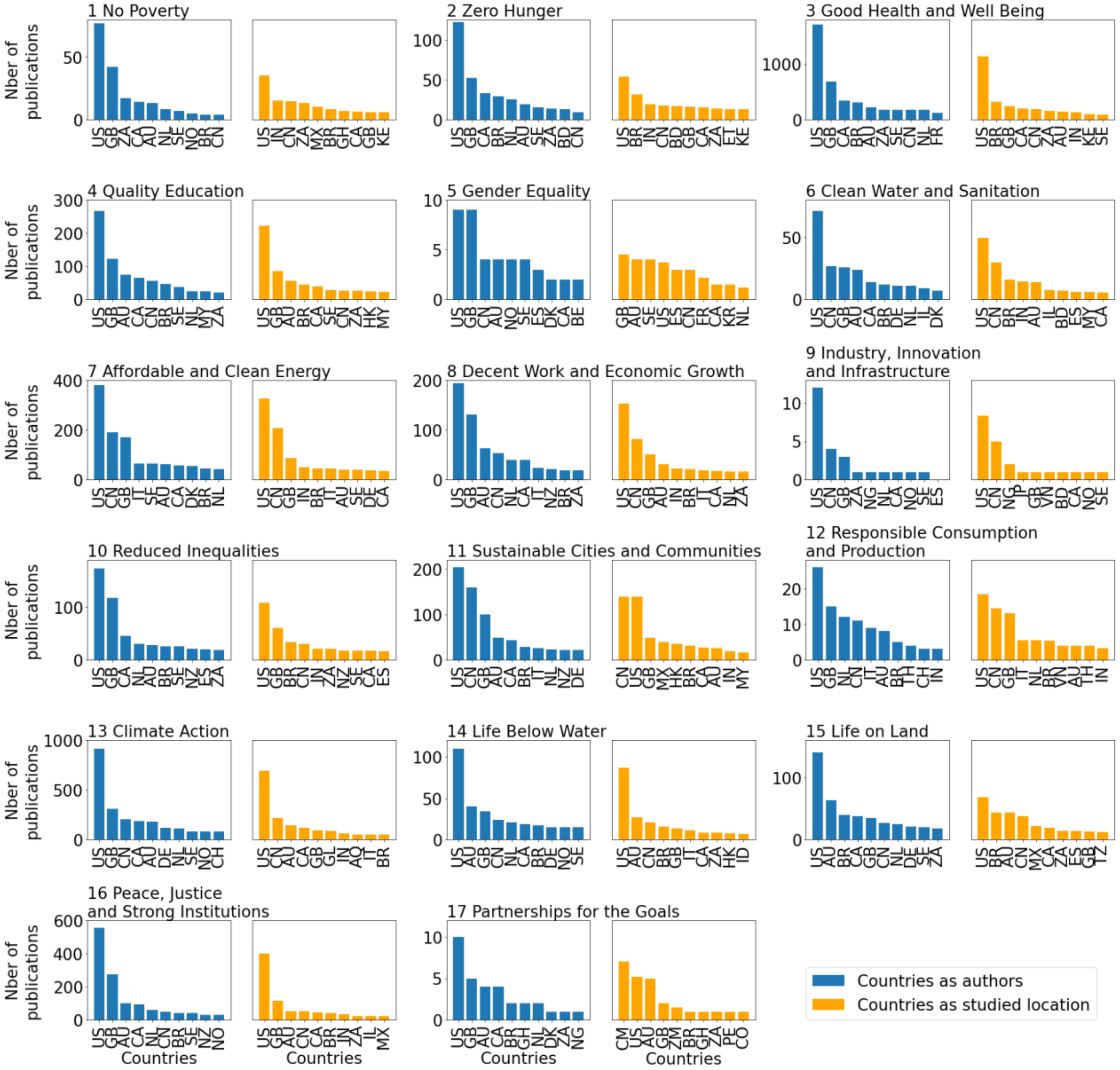
65% studies on elsewhere from here.

Top 3
1. Zimbabwe *41%*
2. United States *17%*
3. Australia *5%*

Analysing the geographical SDG research network, we can find language effects (for example, Portugal studies Brazil the most) and also geographical effects. The community detection analysis of the SDGs geographical research networks shows communities closely overlapping with continents, which means that authors are more likely to study geographically nearby countries, and countries are more likely to be studied by other countries nearby. Another example is the strong research link between Canada and Zimbabwe (see Figure 2). This can be explained with their shared history as members of the British Commonwealth and their development cooperations (Government of Canada, 2022).

Our results also show that there are non-geographical factors that may explain the frequency with which a country is studied. Figure 3 shows the Top 10 countries publishing (country of origin) the most articles about each SDG (blue) and the most studied countries (country of focus; orange) by SDG. While China, India, and Brazil (emerging economies) are among the top countries of focus in 'zero hunger' and 'no poverty', the United States are still the most studied country. They attract a significant focus and are the most frequent, most studied country across SDGs. A significant portion of the research originates from Europe (see Figure 1).

Figure 3: Top 10 of the countries which produce the more articles about an SDG (blue) and the most studied countries (orange) by SDG.



We find that high population, but also GDP per capita explain, to a significant extent, the centrality of a country in the SDGs geographical research network, as shown in Figures 4 and 5. Countries with higher GDP per capita also tend to have a high eigenvector centrality in the network. This result suggests that wealthier countries are more likely to both produce research and be studied in the context of the SDGs.

Figure 4: The population of every country as a function of the eigenvector centrality in the network. The two quantities are positively correlated, with a Pearson coefficient of 0.543, p-value <0.05.

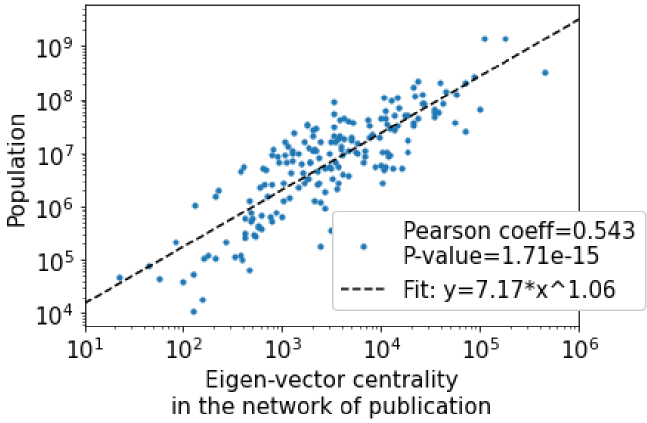
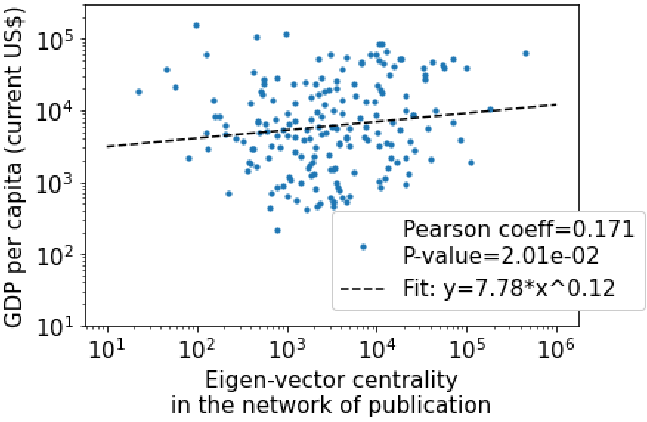


Figure 5: The Gross Domestic Product per capita of every country as a function of the eigenvector centrality in the network. The two quantities are positively correlated, with a Pearson coefficient of 0.171, p-value <0.05.



4. Discussion

The Sustainable Development Goals (SDGs) are inherently interconnected, and as such, an equitable distribution of resources and research efforts is crucial for fostering a comprehensive understanding of intricate global developments. This approach not only facilitates the creation of efficient solutions, but also promotes a more inclusive and collaborative environment in addressing the multifaceted challenges faced by countries worldwide, as intended in the 2030 Agenda (United Nations General Assembly, 2015).

As part of our ongoing study, initial findings indicate that research under-prioritises developing countries and is instead predominantly focusing on resource-rich nations like the United States or those in close geographical proximity. This study emphasizes the importance of adopting a more discerning approach to research priorities and funding distribution, advocating for a focus on vulnerable countries in order to address the disparity in knowledge production between the global North and South (Castro Torres & Alburez-Gutierrez, 2022; Blicharska et al. 2021).

4.1. Limitations

Our study in progress has several limitations. First, although we identified continent names and bodies of water during the Named Entity Recognition (NER) process, we excluded them from the analysis as it focused on the country level. Second, out of 6.7 million articles, we identified 2.8 million that mention a location in their title or abstract. However, articles not mentioning locations in their abstract may still focus on certain regions. This may especially be the case for publications from the global North. Castro Torres & Alburez-Gutierrez (2022, p. 1) note that "articles studying the global North are systematically less likely to mention the name of the country they study in their title compared to articles on the global South". Furthermore, we discarded publications where the metadata did not include the affiliation of all authors, resulting in a substantial reduction in our dataset (see Results section). Demonyms have not been resolved yet, which may impact the analysis further. Lastly, the used SpaCy transformer model is not specifically trained for scientific texts, which may affect the precision of the NER process. However, this limitation seems to influence precision rather than recall, and we post-processed the entities to mitigate this issue.

4.2. Outlook

We are still in the process of analysing the data and improving the network. The next steps are to mitigate aspects mentioned in the limitations. Further in-depth analysis of the data is needed, including using a larger dataset that incorporates publications with missing author affiliations and resolving demonyms. Furthermore, we will explore which topics are studied the most by SDGs and where, by analysing the semantic network of topics.

All nations, irrespective of their development status, should reap the benefits of research. Raising awareness is essential in promoting more inclusive research practices and fostering diverse representation within the academic and research communities. By encouraging data-sharing and collaboration across borders, we can facilitate a more equitable distribution of knowledge and resources, further strengthening our collective ability to tackle complex global challenges. In doing so, we pave the way for a future where research and innovation are driven by a truly global and diverse community, ultimately contributing to the successful realisation of the SDGs.

References

Annan, K. (2003). A Challenge to the World's Scientists. Science, 299(5612), 1485–1485. https://doi.org/10.1126/science.299.5612.1485

Blicharska, M., Teutschbein, C., & Smithers, R. J. (2021). SDG partnerships may perpetuate the global North–South divide. Scientific Reports, 11(1), 22092. https://doi.org/10.1038/s41598-021-01534-6

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Castro Torres, A. F., & Alburez-Gutierrez, D. (2022). North and South: Naming practices and the hidden dimension of global disparities in knowledge production. Proceedings of the National Academy of Sciences, 119(10), e2119373119. https://doi.org/10.1073/pnas.2119373119

Government of Canada (2022). Canada and Zimbabwe. Retrieved April 21, 2023, from https://www.international.gc.ca/country-pays/zimbabwe/index.aspx?lang=eng

Li, J., Sun, A., Han, J., & Li, C. (2022). A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1), 50–70. https://doi.org/10.1109/TKDE.2020.2981314

Matthews, K. R. W., Yang, E., Lewis, S. W., Vaidyanathan, B. R., & Gorman, M. (2020). International scientific collaborative activities and barriers to them in eight societies. Accountability in Research, 27(8), 477–495. https://doi.org/10.1080/08989621.2020.1774373

Newman, M. E. J. (2018). Networks (Second edition). Oxford University Press.

Salager-Meyer, F. (2008). Scientific publishing in developing countries: Challenges for the future. Journal of English for Academic Purposes, 7(2), 121–132. https://doi.org/10.1016/j.jeap.2008.03.009

United Nations. (n.d.). THE 17 GOALS | Sustainable Development. Retrieved April 21, 2023, from https://sdgs.un.org/goals

United Nations General Assembly. (2015). A/RES/70/1 - Transforming our world: the 2030 Agenda for Sustainable Development. https://undocs.org/en/A/RES/70/1

Vieira, E. S., Cerdeira, J., & Teixeira, A. A. C. (2022). Which distance dimensions matter in international research collaboration? A cross-country analysis by scientific domain. Journal of Informetrics, 16(2), 101259. https://doi.org/10.1016/j.joi.2022.101259

Open science practices

This is a work in progress. The code and data from our analysis is openly available via our GitHub repository (https://github.com/RMariaDelRioChanona/SDGs_sustainability) and is continuously expanded and updated.

The used dataset from Dimensions is private, and hence we cannot make it openly available. However, aggregated statistics can be obtained through our interactive visualization (https://vis.csh.ac.at/who-studies-whom/).

Acknowledgments

The authors thank the Complexity Science Hub Vienna for their feedback during previous presentations and funding for the CEU-CSH workshop and MH's research visit. MDRC acknowledges funding from the James S. McDonnell Foundation.

Author contributions

The research project was devised by MDRC and MF. MF collected the data. MH performed the Named Entity Recognition. JS and EA performed the network and regression analysis with inputs from other authors. LY constructed the online visualisation tool. All authors discussed and analysed the results and contributed to the writing of the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Funding information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.