Exploring the evolution of GIS research using bibliographic data

Berry R[*1], Hafferty C[†2], Orford S[‡3] and Clarke L[§4]

[1]Countryside and Community Research Institute (CCRI), University of Gloucestershire, UK
[2]Environmental Change Institute, Oxford University, UK
[3]School of Geography and Planning, Cardiff University, UK
[4]School of Natural, Social and Sports Sciences, University of Gloucestershire, UK

March 31, 2023

Summary

This paper provides new insight into the evolution of geographical information science and systems (GIS) research via a computational analysis (in R) of 120,000 bibliographic records (from 1970 to 2022) downloaded from Scopus. We conduct an exploratory analysis of the data, then attempt to discover the thematic/topical structure of the GIS literature using the Structural Topic Model (STM) framework. We show how topics in GIS have evolved and discuss how our findings contribute to the understanding of the evolution and trajectory of GIS research. We conclude by highlighting the limitations of the approach and explaining our future research plans.

KEYWORDS: GIS, bibliographic analysis, topic modelling, Structural Topic Model (STM), R

1. Introduction

The availability of rich bibliographic datasets and the emergence of novel computational techniques for analysing large text corpora has created opportunities for generating new insight into academic research areas (Asmussen and Møller, 2019). Bibliographic data can be used to chart the development of a field of study, to examine, for example, changes in its magnitude, geographical extent, and thematic structure. It can unlock hidden knowledge and relations in a subject domain and help to identify the evolution of 'hot' topics and how these relate to real-world problems; charting how knowledge, skills and technology move from being nascent to mainstream, particularly in relation to uptake in different regions of the world and the impact on society in terms of solving problems (e.g., climate change) / causing problems (e.g., geoprivacy). While "manual" exploration of research literature is necessary and important, it is usually not a practical way of analysing the large number of records that can be returned from a broad bibliographic database query.

1.1. Bibliographic analysis and GIS

There have been a small number of empirical studies focussed on providing new insight into the GIS research literature using bibliographic data. Of these studies, most have been concerned with the scientometric analysis of the literature, with an emphasis on evaluating research performance (i.e., impact and output) and collaboration (e.g., Huang, 2022; de Melo and Queiroz, 2019; Biljecki, 2016; Liu et al. 2016; Tian et al. 2008). To date, there has been no comprehensive large-scale empirical analysis of the GIS literature focussed on its thematic structure, though some small-scale studies have

[*] rberry@glos.ac.uk
[†] caitlin.hafferty@ouce.ox.ac.uk
[‡] orfords@cardiff.ac.uk
[§] lclarke@glos.ac.uk

paid attention to the analysis of topics. These include recent studies by Wu et al. (2023) who analysed 9,400 publications between 1991-2020, and Wei et al. (2015) who analysed a database of 3,290 publications between 2003-2012. Through our research, we estimate that the GIS research literature now comprises c.120,000 peer-reviewed papers from over 8,000 journals. In this paper we aim to analyse this large database and provide the most comprehensive analysis of bibliographic data relating to global research output in GIS, from its early development to the present day (up to and including the year 2022). Additionally, we attempt to differentiate between fundamental GIScience topics and domain application topics in our analysis – i.e., separating the "GISci" from the "GIS" (Longley et al., 2015).

1.2. Aims and objectives

The aim of the research is to analyse bibliographic data relating to GIS research, using exploratory bibliographic analysis and topic modelling, to further our understanding of how the field of GIS has evolved globally in terms of its:

1. Size
2. Geographical distribution
3. Thematic/topical structure

2. Methodology

2.1. Data

Bibliographic data was acquired from Scopus (https://www.scopus.com). The search query used to retrieve articles from the GIS research literature can be accessed here.[**] The query searches *Title*, *Abstract*, and *Keywords* fields and selects published research from peer-reviewed journal articles only. A total of 139,491 bibliographic records (from 1970-2022) were retrieved using this query, which were downloaded in BibTex Format.

2.2. Software

The data processing, analysis, and visualisation is conducted in the R programming language. The full R code and resources will be made available via the project's GitHub site - a development version of the code can be accessed here[††]. References for the R packages used are listed here[‡‡].

2.3. Data processing and exploratory bibliographic analysis

The BibTex files downloaded from Scopus were combined into a single data frame in R and the data was cleaned, tidied, and transformed for analysis (e.g., removing non-GIS research papers and extracting the country of origin from the first author address string). The final cleaned database comprised 119,185 records. An exploratory analysis of the data was then conducted, where statistics and visualisations related to the research objectives (Section 1.3) were produced.

2.4. Topic modelling

Topic modelling (TM) is a form of computational content analysis. Topic models are generative statistical machine learning algorithms which are used to discover latent topics/semantic structures in a body of text with an efficiency that cannot be matched by manual analysis – though the approach still

[**] https://github.com/robertberryuk/GIS-Topics-Dev/blob/main/ScopusSearch_Berry_etal_2023.txt
[††] https://github.com/robertberryuk/GIS-Topics-Dev/blob/main/GIS-Topics-Berry-GISRUK23-DevCode.R
[‡‡] https://github.com/robertberryuk/GIS-Topics-Dev/blob/main/References.txt

requires significant human input. The Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003) and its derivatives have been applied in other fields of study to help understand the topical structure of academic research areas (e.g. Droste et al., 2018). Recent innovations include the Structural Topic Model (STM) which advances TM by allowing topic prevalence to covary with document-level metadata (Roberts et al., 2013). The STM was used to perform the modelling, implemented in R using the *stm* package (Roberts et al., 2019), which provides a range of functions for handling each stage of the STM modelling process, from data pre-processing to model estimation, interpretation, and visualisation (Figure 1).

One of the main challenges with TM is determining the optimum number of topics (*k*) and interpreting the topics generated by the unsupervised approach of STM. The conventional approach to "finding *k*" is to generate a series of models with differing values of *k*, and to examine the outputs to determine a good fit (in relation to the research question). For our research we generated a series of models with values of *k* ranging from 10 to 150 and used the helper functions in the *stm* package to identify a suitable value range for *k*. Currently, we are interpreting the latent topics generated from models ranging from 80 to 140 topics and experimenting with plotting the prevalence of the topics over time.

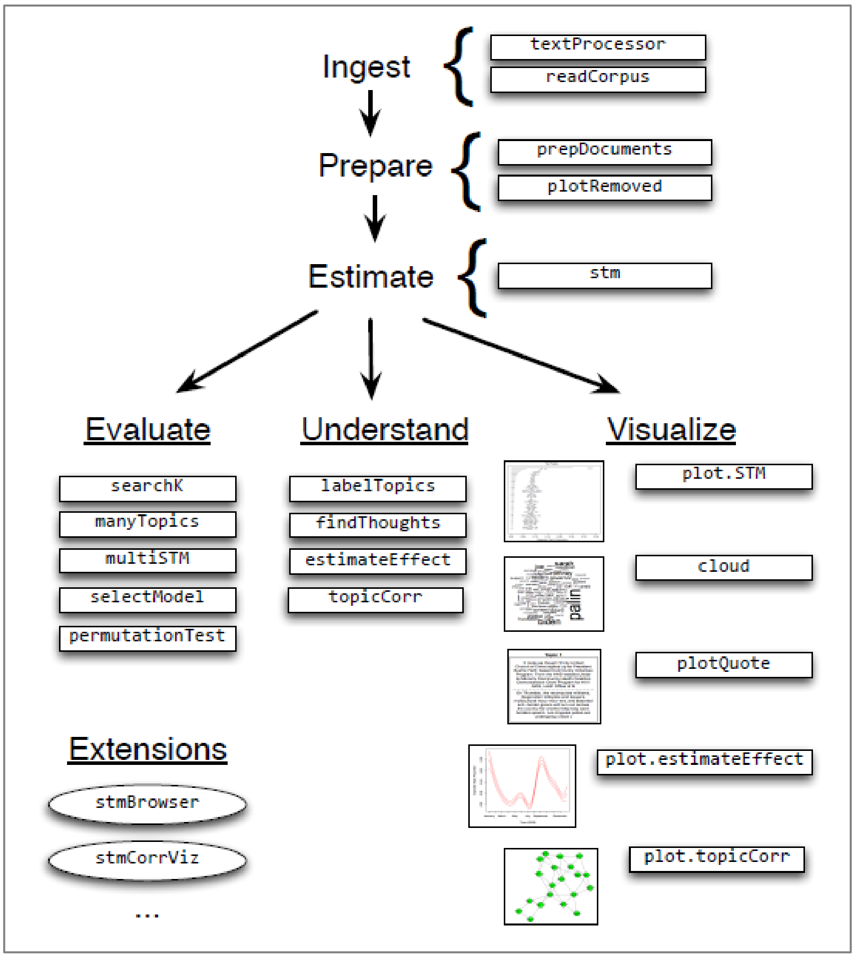


Figure 1 Overview of the *stm* package functions (from Roberts et al. 2019)

3. Preliminary Results

This paper reports some early headline results from our analysis. Firstly, the field of GIS research has seen a huge growth in the output of publications, from seven papers in 1970 to almost 11,000 in 2022. Secondly, in terms of the geographical distribution of research outputs, the key findings are: a) GIS research has become more geographically widespread, with only four countries producing publications in 1980, compared with 134 countries in 2022; and b) the center of gravity of productivity (measured by volume of research output) has shifted towards Asia – in 1991 the United States dominated the GIS research landscape but it has been overtaken by China in recent years, with India's share of global output also growing (Figure 2).

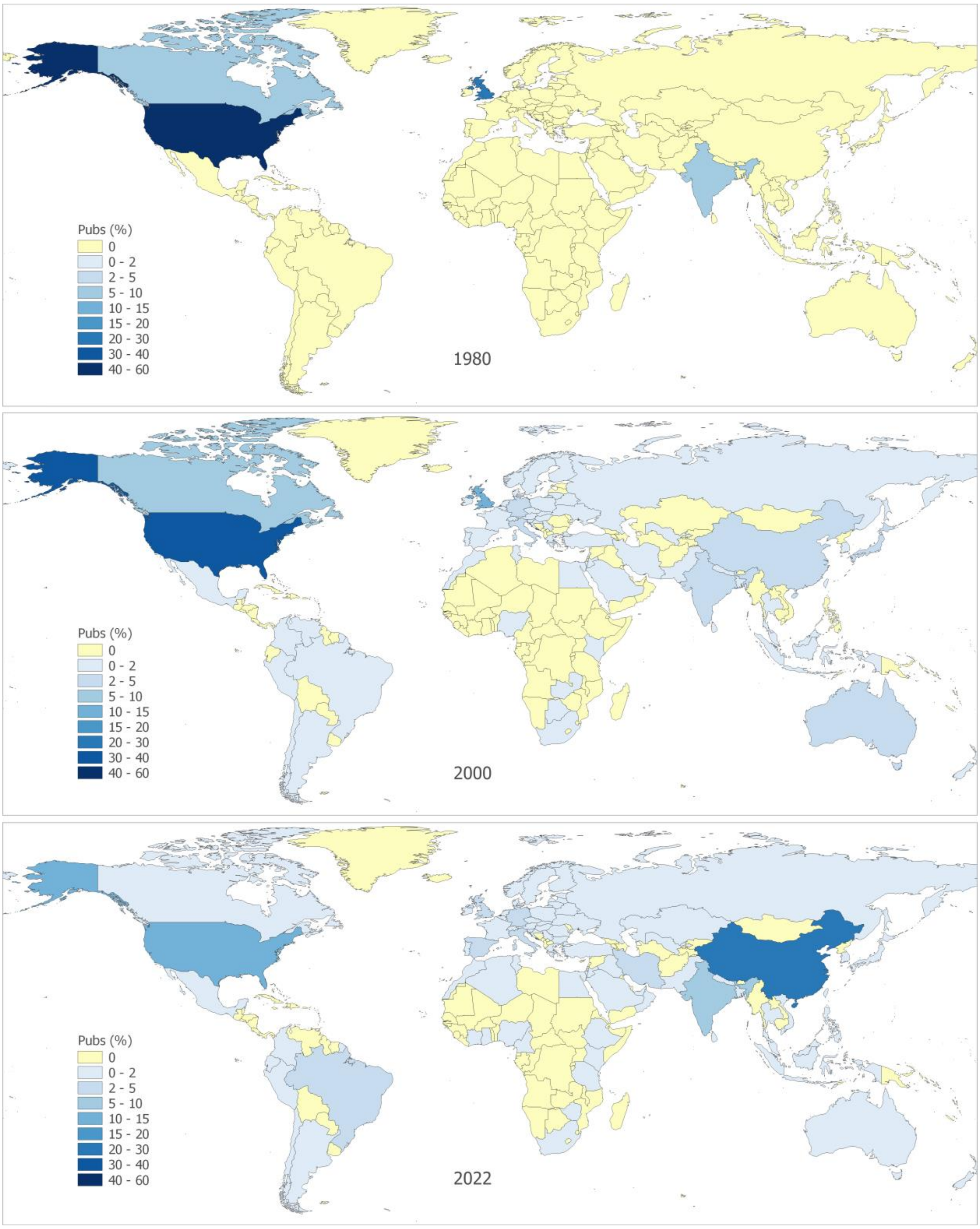
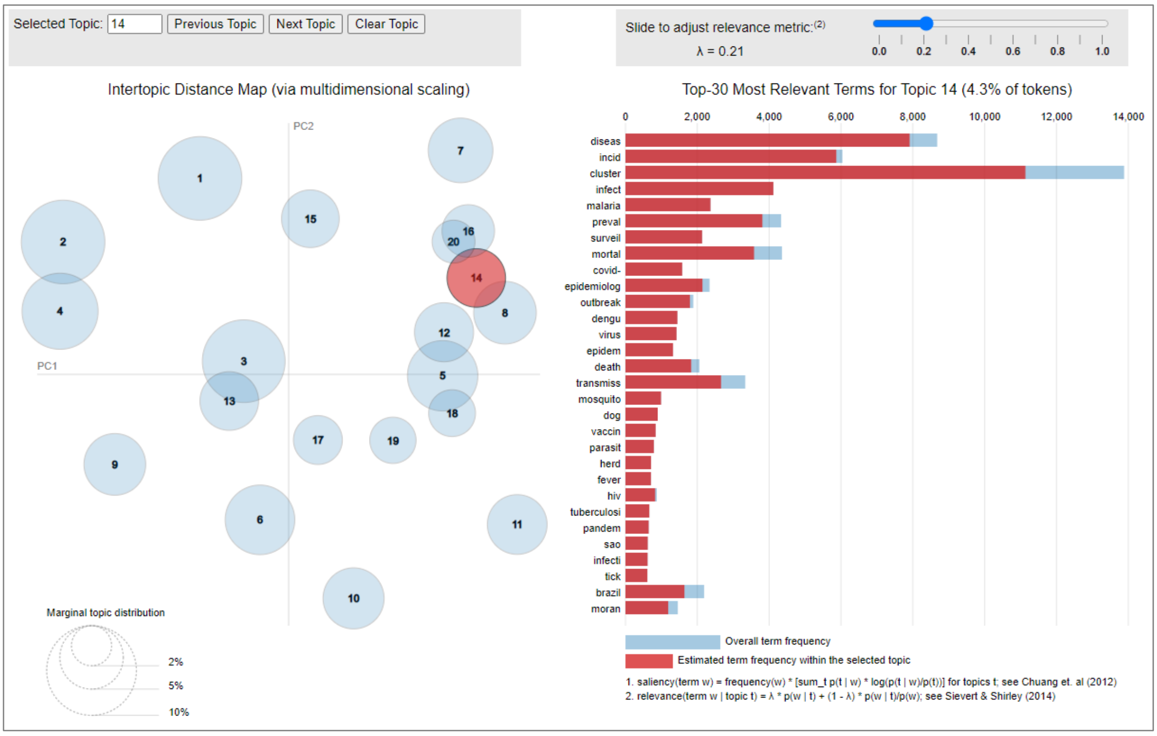


Figure 2 Publication output by country (as a percentage of total annual GIS publications) for the years 1980 (top), 2000 (centre) and 2022 (bottom)

There are numerous outputs from the topic modelling process which are currently being analysed to: a) determine an optimum value for *k*; b) interpret (with the assistance of ChatGPT) and assign names to the generated anonymous topics (e.g., Figure 3); and c) visualise the prevalence of topics over time (Figure 4). Notable findings include strong growth trends in human geography and health (Figure 4).



Selected Topic: 14

Slide to adjust relevance metric:(2)

λ = 0.21

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3 Interactive browser output produced by the *LDAvis* R package for an STM where *k* = 20. The left-hand panel shows relative topic size and intertopic distance (closer = more similar). Topic 14 is highlighted – note the associated terms in the right-hand panel suggesting this topic relates to spatial epidemiology.

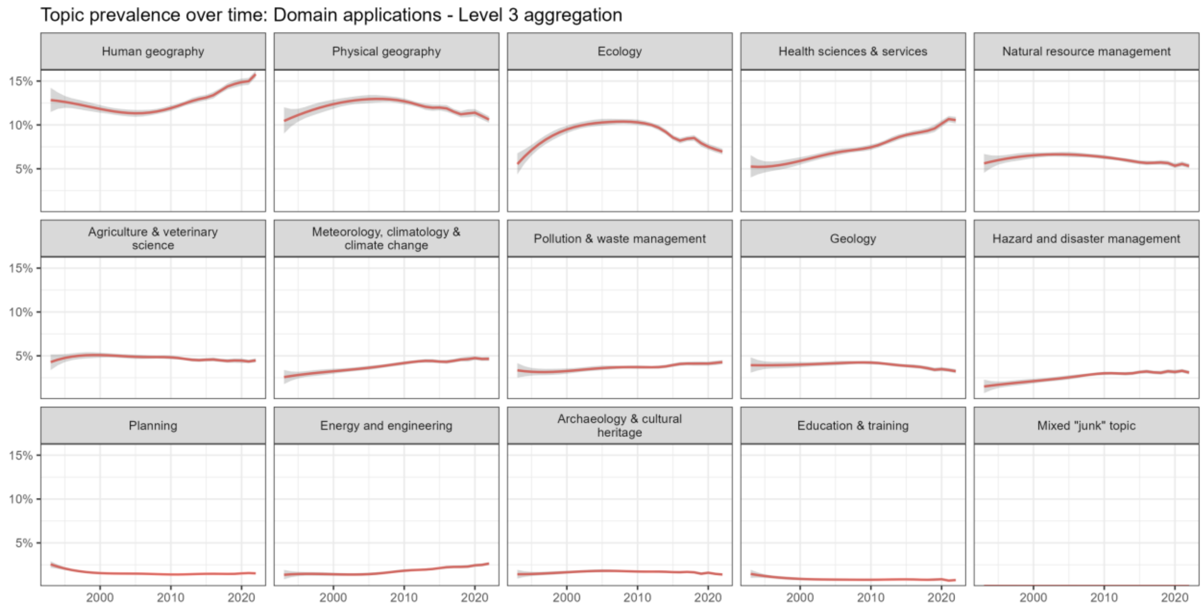


Figure 4 Facet plots showing the prevalence of aggregated "domain application" topic groups over time, for an STM where *k* = 135. Values in the y axis represent the topic proportion in relation to the entire GIS literature. Time (in years) is shown on the x axis.

4. Conclusions and further work

This paper is the first step towards building a comprehensive computational analysis of the GIS literature. It provides new insight into understanding the evolution of GIS research and its current direction. The Scopus data used in this project offers a rich database of bibliographic information, with noteworthy results emerging from preliminary exploratory and topic modelling analysis.

One limitation of this research is that Scopus is one of many academic bibliographic databases and it also privileges the English language. Another constraint is that we were only able to fit our topic models to the abstract of each article, and not to the full paper itself (which would be impractical without significant additional resources).

The next stage of the research is to build on our work by interpreting the STM outputs to extract a set of 'final' named topics/themes, and to plot the prevalence of these topics over time. We also intend to continue mining the Scopus data to answer more complex questions about the evolution and thematic structure of GIS research, for example – how has the inter- and multi-disciplinary nature of GIS changed over time?

5. Acknowledgements

This research is funded by the Countryside and Community Research Institute (CCRI), University of Gloucestershire (https://www.ccri.ac.uk/).

References

Asmussen, C.B. and Møller, C. (2019) 'Smart literature review: a practical topic modelling approach to exploratory literature review', *Journal of Big Data*, 6(1), p. 93. doi:10.1186/s40537-019-0255-7.

Biljecki, F., 2016. 'A scientometric analysis of selected GIScience journals'. *Int. J. Geogr. Inf. Sci.* 30, 1302–1335. https://doi.org/10.1080/13658816.2015.1130831.

Blei, D., Ng, A. and Jordan, M. (2003) 'Latent dirichlet allocation', *Journal of Machine Learning Research*, 2, pp. 993–1022.

de Melo, A.V.F. de, Queiroz, A.P. de, (2019) 'Bibliometric mapping of papers on geographical information systems (2007-2016)'. *Bol. Ciênc. Geodésicas* 25, e2019015. https://doi.org/10.1590/s1982-21702019000300015.

Droste, N., D'Amato, D. and Goddard, J.J. (2018) 'Where communities intermingle, diversity grows – The evolution of topics in ecosystem service research', *PLOS ONE*. Edited by W. Glanzel, 13(9), p. e0204749. doi:10.1371/journal.pone.0204749.

Huang, W., (2022) 'What Were GIScience Scholars Interested in During the Past Decades?' *J. Geovisualization Spat. Anal.* 6, 7. https://doi.org/10.1007/s41651-021-00098-3.

Liu, F. *et al.* (2016) 'Global research trends of geographical information system from 1961 to 2010: a bibliometric analysis', *Scientometrics*, 106(2), pp. 751–768. doi:10.1007/s11192-015-1789-x.

Longley, P., Goodchild, M.F., Maguire, D., Rhind, D. (2015) 'Geographic information science & systems', Fourth edition. ed. Wiley, Hoboken, NJ.

Roberts, M.E. *et al.* (2013) 'The Structural Topic Model and Applied Social Science', Neural *Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*. Neural Information Processing Society. Nevada, USA, 5-8 December 2013.

Roberts, M.E., Stewart, B.M., Tingley, D., 2019. stm: 'An R Package for Structural Topic Models' *J. Statisitcal Software*. 91. https://doi.org/10.18637/jss.v091.i02

Tian, Y., Wen, C. and Hong, S. (2008) 'Global scientific production on GIS research by bibliometric analysis from 1997 to 2006', *Journal of Informetrics*, 2(1), pp. 65–74. doi:10.1016/j.joi.2007.10.001.

Wei, F., Grubesic, T.H. and Bishop, B.W. (2015) 'Exploring the GIS Knowledge Domain Using CiteSpace', *The Professional Geographer*, 67(3), pp. 374–384. doi:10.1080/00330124.2014.983588.

Wu, X., Dong, W., Wu, L., Liu, Y. (2023) 'Research themes of geographical information science during 1991–2020: a retrospective bibliometric analysis' *International Journal Geographic Information Science*. 37, 243–275. https://doi.org/10.1080/13658816.2022.2119476

Biographies

Robert Berry is a Senior Research Fellow at the Countryside and Community Research Institute (CCRI), University of Gloucestershire. His work involves the development of reproducible geodata science applications across a range of environmental and rural research areas.

Caitlin Hafferty is a Postdoctoral Researcher in Environmental Social Science at the Environmental Change Institute, University of Oxford. Her research broadly explores the governance and equity dimensions of nature recovery and nature-based solutions. She is particularly interested in effective and meaningful public and stakeholder engagement in environmental decision-making processes.

Scott Orford is a Professor in GIS and spatial analysis at Cardiff University School of Geography and Planning and WISERD. His research focusses on the spatial and statistical modelling of social and economic processes.

Lucy Clarke is a Senior Lecturer in Physical Geography in the School of Natural, Social and Sports Sciences, University of Gloucestershire. She specialises in environmental hazards and landscape change using image analysis of aerial photography and satellite imagery, as well as spatial analysis using GIS.