Scientometrics
https://doi.org/10.1007/s11192-022-04408-4

A systematic method for identifying references to academic research in grey literature

Matthew S. Bickley[1] · **Kayvan Kousha**[1] · **Michael Thelwall**[1]

Received: 31 October 2021 / Accepted: 12 May 2022

Abstract
Grey literature encompasses documents not published in academic journals or books. Some grey literature has substantial societal importance, such as medical guidelines, government analyses and pressure group reports. Academic research cited in such documents may therefore have had indirect societal impact, such as in policy making, clinical practice or legislation. Identifying citations to academic research from grey literature may therefore help assess its societal impacts. This is difficult, however, due to the variety of document and referencing formats used in grey literature, even from a single organisation. In response, this study introduces and tests a semi-automatic method to match academic journal articles with unstandardised grey literature cited references. For this, the metadata (lead author last name, title, year) of 2.45 million UK Russell Group university outputs was matched against a 100-document sample of UK government grey literature to assess the accuracy of 21 matching heuristics. The optimal method (lead author last name and title in either order, maximum of 200 characters apart) is sufficiently accurate and scalable to make the task of matching research outputs to grey literature references feasible. The method was then applied to 3347 government publications, showing approximately 23% of UK government grey literature in this study contained at least one reference to UK Russell Group university output, and of this grey literature, an average of 3.79 references were present per document. The applied method also shows that economics and environmental science academic research is most cited between 2010 and 2018.

Keywords Grey literature · Impact assessment · Citation analysis · UK government · Academic references · Bland–Altman analysis

✉ Matthew S. Bickley
M.Bickley@wlv.ac.uk

Kayvan Kousha
K.Kousha@wlv.ac.uk

Michael Thelwall
M.Thelwall@wlv.ac.uk

[1] Statistical Cybermetrics Research Group (SCRG), University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK

Published online: 23 June 2022

Introduction

Grey literature is a term which describes text-based documents not published in a standard academic format (e.g., books or journal articles, IGLWG, 1995). These documents are usually produced by organisations that do not focus on publishing (Schöpfel, 2010, p. 11). Grey literature includes, but is not limited to, unpublished research, governmental reports, policies, some conference proceedings, theses and dissertations (GreyNet International, 2019; UNE, 2019). The use of standardised reference lists in grey literature has not been widespread (Benzies et al., 2006), complicating their automated extraction and assessment.

Academic research impact has been mainly based upon counting citations from formal academic publications (e.g., journal articles) with traditional citation indexes such as Scopus or Web of Science. Nevertheless, other types of non-standard publications may be needed for monitoring the wider benefits of academic research. For instance, in the context of the UK Research Excellence Framework, impacts on "economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia" must be demonstrated (REF, 2019, p. 68). Cited references in grey literature can form part of this evidence by showing specific non-academic uses of research. For instance, World Health Organization (2020) guidance on water, sanitation, hygiene and waste management for COVID-19 has cited several scientific journal articles for healthcare policy making. The historical roots of research can also be found by analysing cited references (Marx & Bornmann, 2016), which can potentially show differences in impact across various literature.

Grey literature is both cited by academic research (Bickley et al., 2019, 2020) and cites it. Altmetric (n.d.) counts references to academic research from grey literature, Overton.io specialises in this, and some previous studies have analysed documents citing grey literature in narrow contexts dating back varying years (Alberani et al., 1990; Cordes, 2004; Pelzer & Wiese, 2003; Woods et al., 2020). Amongst these studies, it has been found that technical reports were highly used in the United States, the United Kingdom and Canada (Alberani et al., 1990), and that citation analysis can measure impact from an organisation's grey literature output (Cordes, 2004), both showing that grey literature has impact in its use by academic research, but it is unclear if this relationship is two-way. In fields such as nursing, it has been shown that there is a high incidence of grey literature citations (Woods et al., 2020), however some studies have shown that related disciplines such as medicine and biology have a lower-than-average incidence of these types of references (Pelzer & Wiese, 2003). This could be due to the publication dates of these studies being somewhat different and the ever-changing format and use of grey literature across this time but could also indicate the need for clarification and further study into the impact of grey literature in academia.

Nevertheless, no practical method has been developed to systematically identify academic research citations in grey literature across reference formats. In response, this article demonstrates a method to semi-automatically match a list of academic documents with the references of grey literature, wherever and however they occur. This is designed to help institutions or researchers find evidence for the societal impact of their traditional publications.

Research questions

The goal of this study is to introduce and assess a method to reliably match academic documents with grey literature references. UK government publications are used as the test case because of the importance of government documents, their ad-hoc format and referencing in

the UK, and the large number that are freely available online. The primary goal is to answer the technical aspect of this study: whether a method of detecting references in grey literature is feasible, and if so, which method proposed is optimal. The secondary goal is to apply the optimal method discovered to a large set of UK government documents to analyse links between grey literature and academic output in various ways. Thus, the following research questions drive this study:

1. Can references in grey literature in multiple formats be automatically matched with academic publications with high accuracy?
2. What proportion of UK government grey literature cited UK Russell Group university outputs?
3. Are there changes in time in the proportions of UK Russell Group university outputs cited by UK government grey literature?
4. Which disciplines associated with UK Russell Group university outputs are most cited by UK government grey literature?

Methodology

The new data extraction method uses free *Publish or Perish* and *Webometric Analyst* software to gather and analyse data. It was tested on a sample of grey literature. Twenty-one heuristics to extract citations were compared using Bland–Altman analyses and plots (Bland & Altman, 1986). The UK government website was chosen to test the method because it contains high value documents with varied reference formats. It has previously been studied to investigate the impact of these documents (Bickley et al., 2019, 2020), but not the references in them.

Step 1: Google Scholar search (via Publish or Perish)

Publish or Perish (Harzing, 2010) was used to identify digitised UK government reports indexed by Google Scholar during 2010–2018, which are assumed to be most likely to cite academic research (e.g., in contrast to documents without many references, such as committee minutes). Other programs such as Dimensions (Hook et al., 2018) could also have been used. To generate effective searches, the '*site:*' command was used together with the '*filetype:*' command to limit the results to UK Government website documents in PDF format: https://www.com/site:gov.ukfiletype:pdf.

The query was also amended and submitted to search for Microsoft Word documents (.doc and.docx), which are also present in the repository. Publish or Perish allows searching by year which was used to allow for discrepancies between years to be isolated so subject area differences can be analysed in a future study. From this, each query term above was searched 9 times, one for each of 2010–2018, leading to a total of 27 different search queries (9 years, 3 filetypes). This method found the URLs of about 65% (4280 of 6591) search results indexed in Google Scholar (4206 PDF files, 74 Word documents; Table 1).

Step 2: Downloading UK government reports

Webometric Analyst (*Services—> Download binary or text URLs*) was used to download each full text document found by Publish or Perish. About 80% of the documents were

Table 1 Documents found using Google Scholar, extracted using Publish or Perish, downloaded using Webometric Analyst and converted using PowerShell, with totals and success ratios for each step compared to the previous step (number of documents found using Publish or Perish out of number of documents shown in Google Scholar, for example), split by year/filetype

Year	Filetype	Google Scholar	Publish or Perish	Webometric Analyst	Converted
2010	pdf	801	281	242	236
	doc	11	11	10	10
	docx	1	1	1	1
2011	pdf	1070	500	406	396
	doc	11	11	11	11
	docx	0	0	0	0
2012	pdf	1030	565	462	451
	doc	14	14	11	11
	docx	2	2	1	1
2013	pdf	905	586	469	464
	doc	12	12	7	7
	docx	1	1	0	0
2014	pdf	911	612	486	479
	doc	6	6	0	0
	docx	2	2	0	0
2015	pdf	633	536	435	432
	doc	5	5	1	1
	docx	0	0	0	0
2016	pdf	496	461	361	358
	doc	1	1	1	1
	docx	4	4	2	2
2017	pdf	372	370	301	297
	doc	1	1	1	1
	docx	1	1	0	0
2018	pdf	299	295	216	214
	doc	1	1	1	1
	docx	1	1	0	0
Total		6591	4280	3425	3374
Success ratio		N/A	64.9%	80.0%	98.5%

successfully downloaded by Webometric Analyst (3435 of 4280; Table 1). The documents missed were likely due to some unstandardised PDF settings or document protection causing the automatic download of the file to fail.

Step 3: Converting PDF files to text files

The Xpdf command line tool (Glyph & Cog, 2019) through Webometric Analyst (*Text—>Convert PDF files in folder to text*) was used to automatically convert the PDF and Word documents to text, of which over 98% (3374 of 3425; Table 1) were successfully converted. Manual conversion of a PDF or Word document is possible, but the methods

presented here are intended to form a basis for potentially larger scale studies, where manual conversion is infeasible.

Step 4: Identification and extraction of references

The above steps produced 3374 plain text UK government grey literature documents. Two random samples of 50 documents were selected for initial testing. One sample had reference lists and the other did not, produced using manual checking.

The metadata and references for the 50 documents with references were manually extracted from the original PDF/Word documents to bypass automatic document processing errors (e.g., incorrect line breaks). The following metadata was also manually extracted: lead author last name, title and publication year. Reference titles with fewer than 5 words were removed because short titles may contribute to inaccurate results, as titles like, "Introduction to Philosophy" are inherently more generic and could easily provide a false positive when looking for a match. Moreover, a maximum of 10 words from reference titles were used to avoid matching problems such as line wrapping and line breaks. Table A1 (Online Resource, Appendix A1; https://doi.org/10.6084/m9.figshare.16895485) shows the number of references in each document used, which only includes those with titles of 5 or more words (i.e., each document may contain more references that are not be included in the matching process).

The metadata extracted as above from the documents formed a set of potential references that a computer program could try to match with the same (or another) set of documents. This would test the ability of the program to recognise the pre-selected reference list. This would be an unrealistically small set of references to match, however. Scopus references were therefore used to expand the set, as follows. Scopus Advanced Search was used to download the metadata (authors, title and year) for all UK Russell Group university outputs across all years (until 2020 inclusive) and publication types, giving 2.45 million unique records, which was deemed to be a feasible dataset to collect whilst being a relatively large and diverse collection of records. The list of universities in the UK Russell Group is shown in Table 2. Initially, search terms using university names were used such as:

AFFIL ("University of Birmingham")

However, it was realised that some search terms may cause false positives (e.g., "University of York" was also showing some results for "New York University"), so another method was needed. It was found that Scopus also assigns institutions unique identifiers (Table 2), so these were used in combination to limit the result to only UK Russell Group universities. The search term was finalised as:

AF-ID (60019702) OR AF-ID (60020650) OR AF-ID (60031101) OR AF-ID (60023998) OR AF-ID (60022175) OR AF-ID (60027272) OR AF-ID (60026479) OR AF-ID (60001490) OR AF-ID (60015150) OR AF-ID (60011520) OR AF-ID (60012070) OR AF-ID (60020661) OR AF-ID (60003059) OR AF-ID (60003771) OR AF-ID (60006222) OR AF-ID (60015138) OR AF-ID (60026851) OR AF-ID (60022109) OR AF-ID (60029738) OR AF-ID (60001881) OR AF-ID (60025225) OR AF-ID (60022148) OR AF-ID (60022020) OR AF-ID (60016418).

Table 2 List of UK Russell Group universities with Scopus affiliation ID and number of outputs across all years (until 2020 inclusive)

Russell Group university	Scopus affiliation ID	No. outputs across all years (until 2020 inclusive)
University of Birmingham	60019702	115330
University of Bristol	60020650	114537
University of Cambridge	60031101	240901
Cardiff University	60023998	79284
Durham University	60022175	54768
University of Edinburgh	60027272	131427
University of Exeter	60026479	45217
University of Glasgow	60001490	99475
Imperial College London	60015150	200955
King's College London	60011520	155865
University of Leeds	60012070	101136
University of Liverpool	60020661	94622
London School of Economics and Political Science	60003059	28702
University of Manchester	60003771	177340
Newcastle University	60006222	78622
University of Nottingham	60015138	95706
University of Oxford	60026851	246908
Queen Mary University of London	60022109	56841
Queen's University Belfast	60029738	62842
University of Sheffield	60001881	103947
University of Southampton	60025225	99553
University College London	60022148	269011
University of Warwick	60022020	64156
University of York	60016418	50542
Total[a]		2767687
Total (excluding duplicates)		2458111

[a]Includes duplicates as some outputs are affiliated with multiple Russell Group universities

From this point, data collection was done manually as a maximum of 20,000 Scopus records can be extracted at one time, and as the search term provided more than this number, splitting of the results (and re-combining after download) was required. Using the filters available on Scopus, the data was able to be refined into groups over less than 20,000 whilst not introducing duplicate results and allowing for the subject areas to be recorded at the same time as an individual Scopus record to does explicitly show which Scopus-defined subject area it is categorised as. Table 2 also shows the number of outputs by each institution.

Step 5: Matching search terms using Webometric Analyst

Webometric Analyst (*Text—> Find two/three strings close together in a set of files*) was used to match the references in the expanded list with the original documents. A bespoke

routine was added to allow a batch of text files to be imported and then a matching file be added to search each document in the batch for matching terms. Two versions were created (one to match 2 terms, another to match 3), and each version allowed for adjustable options; to allow for the pair (or triple) of terms to appear in any order or with a specific one first, and for the maximum distance in characters between the terms found to be classed as a match.

The different methods changed which indicators to include. Four combinations of matching terms were chosen: (1) lead author last name, title and year, (2) lead author last name and title, (3) lead author last name and year, and (4) title and year. Because many reference formats start with author names (Pears & Shields, 2019), each option was repeated with the author forced to appear before the other terms (when included). Options of 50-, 100- and 200-character maximum distances between terms were chosen by inspection of references seen in many documents.

A total of 21 different methods were created using combinations of these options and used in the results, and each is numbered in Table A1 (Online Resource, Appendix A1; https://doi.org/10.6084/m9.figshare.16895485) and Table 3 for identification purposes. Table 3 also presents a brief description of each of the methods; which pair or triple of indicators are used, whether the lead author last name must be first (not applicable to the method only involving title and year), and the maximum distance (in characters) between indicators pairs (including pairwise matching in the triple involving lead author last name, title and year, but ignoring if the first and third indicator in the order they appear is more than the maximum distance).

Results 1: Development of an optimal method to identify references to academic research in grey literature (Research Question 1)

Table A1 (Online Resource, Appendix A1; https://doi.org/10.6084/m9.figshare.16895485) shows the overall results for all methods investigated. For 18 of the 21 methods (1–12 and 19–21), 49 of the 50 documents without references are always correctly predicted to have no references. In Table A1, these are combined into one row to avoid repetition of many rows of zeroes. The remaining methods (13–18) are relatively large overpredictions for these 49 documents, so although cited reference counts for each document is different, the fact that they are much larger than zero when zero is predicted by methods 1–12 and 19–21, means the exact value is unimportant. The remaining document without references is shown on its own row in Table A1 for clarity as all methods predict at least one cited reference when the true value is zero. For the 50 documents with known reference counts of at least one, each document is shown on a separate row in Table A1 to illustrate the difference between the number of cited references known to exist in the document and each of the 21 methods' predicted measure.

Once calculated, Bland–Altman analysis (Bland & Altman, 1986) was used to compare the methods to the known number of manually counted references. Bland–Altman analysis is used traditionally in clinical contexts to compare two methods for estimation across multiple observations and is a simple way to estimate an agreement interval (limits of agreement) containing 95% of the differences of one method compared to the other (Giavarina, 2015). However, it should be noted that these limits of agreement should be defined in advance for what the 95% central limit should be in terms of the variable being measured—it could be that the method proposed is still unacceptable if the limits are too large.

Table 3 All methods, with descriptions, showing bias (mean of known and method-reported number of references), standard deviation and rank of each (smaller absolute bias and standard deviation is better)

Method number	Description of method			Bias	Standard deviation	Bias rank	Standard deviation rank
	Combination of indicators used ('Author' means 'Lead author last name')	Author last name first?	Maximum distance between indicator pairs				
1	Author, title, year	Yes	50	− 7.35	19.01	15	15
2	Author, title, year	Yes	100	− 3.82	15.04	11	13
3	Author, title, year	Yes	200	− 1.36	6.04	8	8
4	Author, title, year	Not necessarily	50	− 7.18	18.66	14	14
5	Author, title, year	Not necessarily	100	− 3.52	14.18	10	12
6	Author, title, year	Not necessarily	200	− 1.13	5.61	7	7
7	Author, title	Yes	50	− 5.27	13.27	13	11
8	Author, title	Yes	100	− 0.87	2.67	6	5
9	Author, title	Yes	200	− 0.28	1.78	2	2
10	Author, title	Not necessarily	50	− 5.19	13.27	12	10
11	Author, title	Not necessarily	100	− 0.80	2.64	=4	4
12	Author, title	Not necessarily	200	− 0.21	1.76	1	1
13	Author, year	Yes	50	747.04	809.24	16	16
14	Author, year	Yes	100	1132.91	1209.32	18	18
15	Author, year	Yes	200	1636.76	1720.39	20	20
16	Author, year	Not necessarily	50	798.48	835.59	17	17
17	Author, year	Not necessarily	100	1213.97	1250.06	19	19
18	Author, year	Not necessarily	200	1759.42	1783.49	21	21
19	Title, year	N/A	50	− 1.96	8.85	9	9
20	Title, year	N/A	100	− 0.80	3.04	=4	6
21	Title, year	N/A	200	− 0.44	1.78	3	3

For this study, as 21 varying methods are being proposed via different combinations of indicator-pairs/triples and distances, Bland–Altman analysis can be used to show which of the 21 methods is best by finding the smallest (absolute) bias along with limits of agreement compared to the known reference count in each document. For the best method, it is also necessary to state if the limits of agreement are acceptably small as it is possible that the bias for a method is close to zero but if the limits of agreement are too large, the method is deemed unreliable.

In general, if a measurement is to be taken of a specific variable, but this measurement is difficult or costly, it may be of interest to know the level of another variable which can be taken easily or relatively inexpensively. Bland–Altman analysis is appropriate here as large-scale manual counting of cited references in grey literature would be time consuming if not expensive. A paired Student's t-test for the total number of references found by each method would be insufficient because it only compares group means rather than differences observation-by-observation, leading to ignoring false matches in this context.

Table 3 shows the biases and standard deviation calculated using Bland–Altman analysis for each method and ranks them within both statistics (smaller absolute bias and standard deviation is better). The most accurate method in terms of the smallest absolute bias (-0.21) and standard deviation (1.76) uses the lead author last name and title (author not necessarily first), with a maximum of 200 characters between the two (method 12). The limits of agreement are calculated similarly to a confidence interval, meaning that 95% of the predictions would be within 3.4496 cited references ($3.4496 = 1.96 \times 1.76$, or approximately 3 rounded to the nearest integer number of references) of the true value. It is deemed that this is acceptable for a 'best-method' proposed, but further research into the distance parameter may yield improvements, reducing this uncertainty.

The same indicators used in the best method above, but with the author last name required to be first (method 9) was the second-best method for both statistics (bias $= -0.28$, standard deviation $= 1.78$). All other methods have comparatively worse bias and/or standard deviation (e.g., method 21, using title and year with distance of 200, is ranked third in both statistics but the bias is more than double the best method).

Negative bias implies that the method underpredicts the number of references in each document, whereas positive bias indicates overprediction. For the 15 methods with biases and standard deviations much closer to zero (methods 1–12 and 19–21), they show underpredictions, so these methods are missing matches (false negatives) rather than including (m)any false positives. The remaining 6 methods (13–18, using author and year) are wildly overpredicting the number of references (also large standard deviations), likely due to the generic nature of a single author last name and a 4-digit integer, both of which could commonly be contained within other strings of text (e.g., "Ng 2018" is a common last name and potential year combination, but would show as a match in the string "predicting 2018 as a more fruitful year"). A space in front of the lead author last name could remove this type of false positive, but if the author last name were preceded by other characters, text markup language (such a new line delimiters), or by nothing at all (if the author last name were the first text in the document), then there would be many false negatives.

Bland–Altman plots of the methods can illustrate the difference between the known and predicted number of references against mean of the two. A 'good' prediction method (Fig. 1) shows a scatter plot which has similar spread across all parts of the x-axis (mean of known and method-reported number of references), ideally with points being close to zero on the y-axis (difference between known and method-reported number of references), meaning narrow limits of agreement. It follows that this method would then have low bias and standard deviation (limits of agreement), meaning the method presented is a good predictor of the true measure.

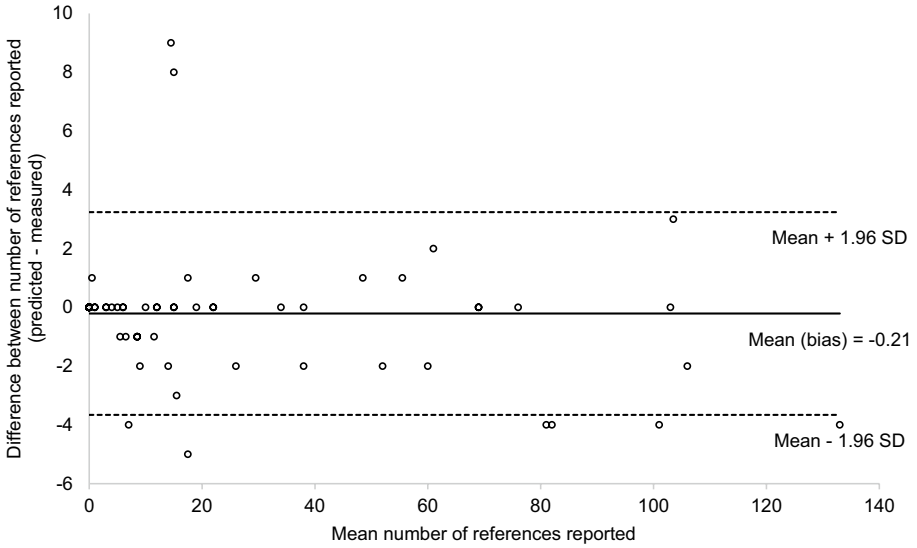

Fig. 1 Bland–Altman plot of agreement between method 12 (lead author last name and title, author not necessarily first, distance 200) and known reference count of 100-document sample, the best method proposed

Unsuitable methods would show either a higher absolute bias (Fig. 2), or worse showing 'fanning-out' along the axis from left to right (Fig. 3). Here, as all measures are positive integers and the worst of the prediction methods presented wildly overpredict rather than underpredict, 'one-sided fanning-out' is shown in Fig. 3, but leads to the same conclusion; larger bias and limits of agreement are visible.

Figures 1, 2, and 3 show Bland–Altman plots from the best ranked method (12), the worst ranked of those with relatively close-to-zero bias/standard deviation (method 1) and the worst ranked overall (method 18). Bland–Altman plots for the other methods are shown in Figs. B1–B18 (Online Resource, Appendices B1–B18; https://doi.org/10.6084/m9.figshare.16895485).

Results 2: Application of the optimal method to UK Russell Group University outputs (Research Questions 2–4)

The documents which were excluded from the sample used in testing for the best or 'optimal' method above were gathered to use in a demonstration of this method in practice. In this way, all 3374 were chosen, and all 2.45 million Scopus metadata records were used to test the feasibility of this method on a larger dataset. During original data collection of the Scopus metadata, results were collected by subject area so that disciplinary differences could be analysed. Due to the method of data collection from the grey literature repository used not allowing for subject area to be known, it is assumed that the subject area of the references they may contain may be indicative of the grey literature topic, if not useful for assessing the documents' impact on specific academic areas. The year of publication of the Scopus reference was also known, so results can also be split by year, shown below.

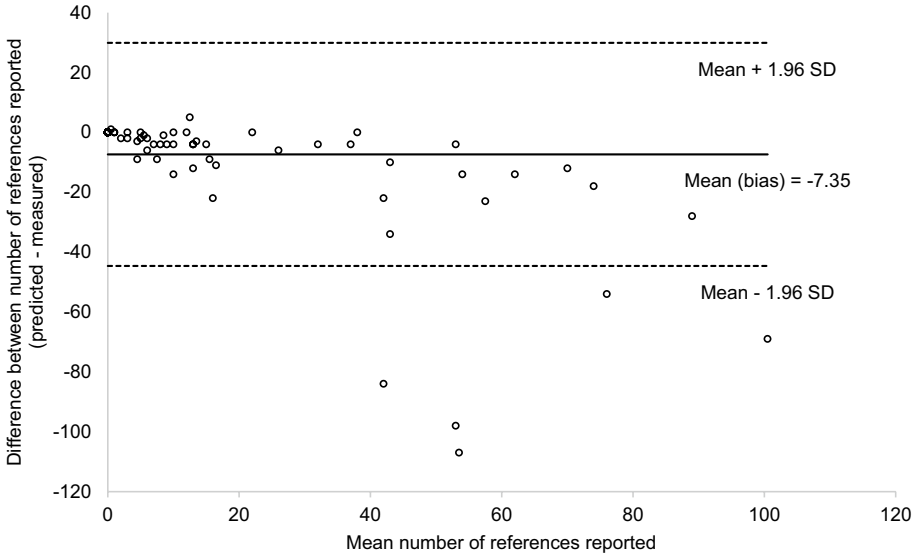

Fig. 2 Bland–Altman plot of agreement between method 1 (lead author last name, title and year, author first, distance 50) and known reference count of 100-document sample, showing larger bias/limits of agreement (bias ± 1.96 SD) to Fig. 1

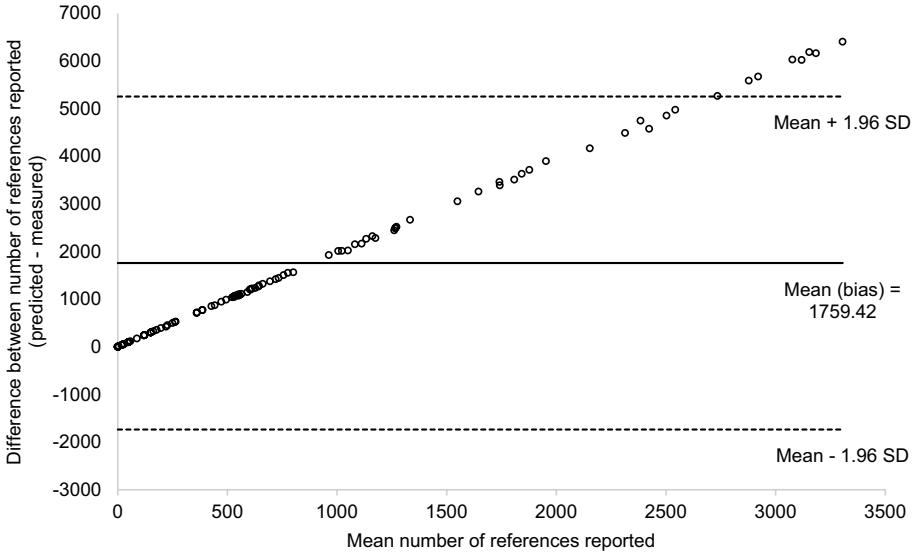

Fig. 3 Bland–Altman plot of agreement between method 18 (lead author last name and year, author not necessarily first, distance 200) and known reference count of 100-document sample, demonstrating a poor method with 'one-sided fanning-out'

The method—although not the main use for it—can also be used to look at two other insights. Firstly, an estimate for the number of references present in a grey literature article can be approximated. As this is a large dataset, the time taken to run can also be predicted

using regression analysis, after acquiring the information from several likely or possible independent variables. This regression equation can then be used in future studies or as part of a program that this methodology may be implemented into. Both phenomena are also analysed in this paper.

Number of references per grey literature document

In addition to flagging when a match has been made, the raw output from the method described in the first part of the results section also records the details of the match. It shows which document the match was made in (as a batch of documents can be checked in the same instance of the method) which is recorded as the title of the document, and which indicator pair was found in the match, showing both strings searched for (here, the lead author last name and article title). The grey literature documents are named with sequential numbers (generated by Webometric Analyst) in the download phase using the list from Publish or Perish, which can then be matched back to the original list. Using a combination of these, it can be determined how many matches per document have been found, indicating the number of references to Russell Group universities per document. However, as some Russell Group outputs have multiple subject areas, the true number of Russell Group outputs referenced is less than would be shown from the raw method outputs. As the indicators pairs are known, duplicate reference mentions can be removed, and the number of unique matches calculated. This would indicate the true number of Russell Group outputs mentioned in each grey literature document, and from this, the average (arithmetic mean) number of Russell Group outputs mentioned in the grey literature analysed overall and per year can be calculated (for both all documents and all documents with at least one reference).

Overall, Table 4 shows that on average there are approximately 3.79 references to Russell Group output from those that have at least one Russell Group reference present. If looking at the entire population of UK government grey literature, whether at least one known reference is present or not, this drops to approximately 0.87 references per document. This gives a lower bound for the number of references per document, but not an upper bound as other references could be present to non-Russell Group studies. Extrapolating this number based upon the prevalence of Russell Group university output within Scopus or in academia could be treated as an upper bound due to the probability of the UK government referencing the UK Russell Group likely being higher over any completely random academic paper but would also be more of a guess than estimate as further investigations into the source of articles referenced in this grey literature must be undertaken.

Proportion of grey literature documents with at least one reference, split by year

The overall proportion of grey document literature documents that contain at least one reference to Russell Group output is approximately 23% (777 of 3374; Table 4). As the grey literature documents were collected separately by year via Publish or Perish, it is possible to compare the proportion containing at least one reference across each of the nine grey literature document years. The absolute values here are much higher than the previous section, but this is the proportion of grey literature documents with at least one reference, not the proportion of Russell Group output mentioned in the grey literature—hence the much higher proportions (at least 17% for all years). There are no clear and obvious statistically significant differences between consecutive years, and although there appears to be a slight

Table 4 Average number of references per grey literature document across 9 years (2010–2018) and overall, and average number of references per grey literature document that contains at least one known reference across the same 9 years (2010–2018) and overall

Grey literature document year	No. of grey literature documents	Arithmetic mean no. of references per document	No. of grey literature documents (with at least one reference)	Arithmetic mean no. of references per document (with at least one reference)
2010	247	0.66	57	2.86
2011	407	0.54	73	2.99
2012	463	0.71	89	3.71
2013	471	0.98	113	4.07
2014	479	0.86	97	4.27
2015	433	0.96	111	3.73
2016	361	0.88	89	3.58
2017	298	1.13	86	3.92
2018	215	1.35	62	4.68
All	3374	0.87	777	3.79

trend of increasing proportion as the years go on, this is not a solid conclusion as there are at least 3 years (2011, 2014 and 2016) where the proportion decreases from the previous year. It is important not to cherry pick a pattern where one may not exist so the overall inference from this result is deemed to be inconclusive (Fig. 4).

Proportion of Russell Group output referenced, split by subject area

Using the optimal method described in the first part of the results section, Fig. 5 shows that disciplinary differences do exist in terms of the references made in grey literature documents, hence the non-academic impact of grey literature may be assessed using this automated method. Across all years, the top 2 subject areas referenced by the grey literature documents are featured statistically significantly higher than any other subject, and between the two, the difference is also statistically significant. The most referenced subject area is Economics, Econometrics and Finance with the proportion of Russell Group output referenced of 0.00642. Scopus groups these 3 together, so it cannot be discerned if any of these are different to the others in the group, but their grouping makes sense as their use and focus is arguably interchangeable in most situations.

The second most referenced topic is Environmental Science (proportion 0.00444), and although significantly lower than Economics, Econometrics and Finance, it is significantly higher than the third ranked topic, Social Sciences (proportion 0.00305). Although these number may seem very low (less than 1% for all proportions and confidence interval limits), it should be remembered that this is a relatively small set of grey literature compared to the relatively large set of Russell Group outputs. The numbers are relative, so could be normalized for easier reading, however the proportion value itself is still interesting for estimation of impact.

Below the top two, there are less obvious differences between consecutive positions, however it is worth noting that there appears to be 2 distinct groups—from rank 3 (Social Sciences) to rank 18 (Biochemistry, Genetics and Molecular Biology) which has a lower confidence limit of 0.000353, and then rank 19 (Mathematics) through rank 28

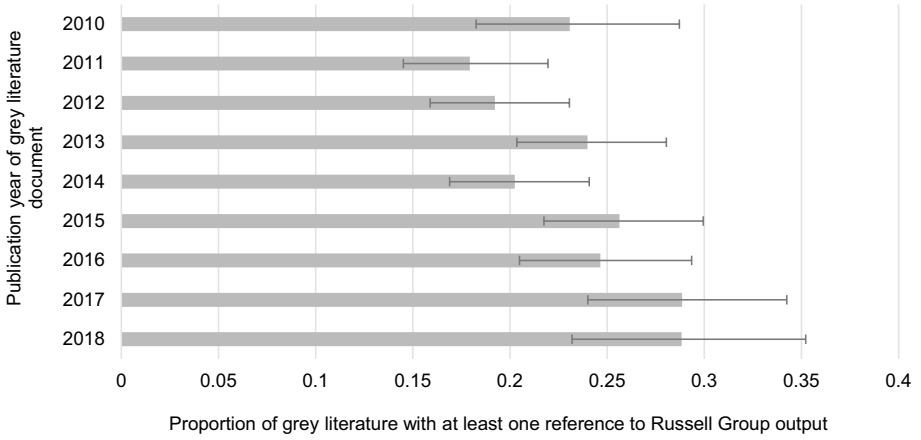

Fig. 4 Proportion of UK government grey literature documents (2010–2018) with at least one reference to Russell Group output (all years) across 9 years, split by publication year of UK government grey literature document

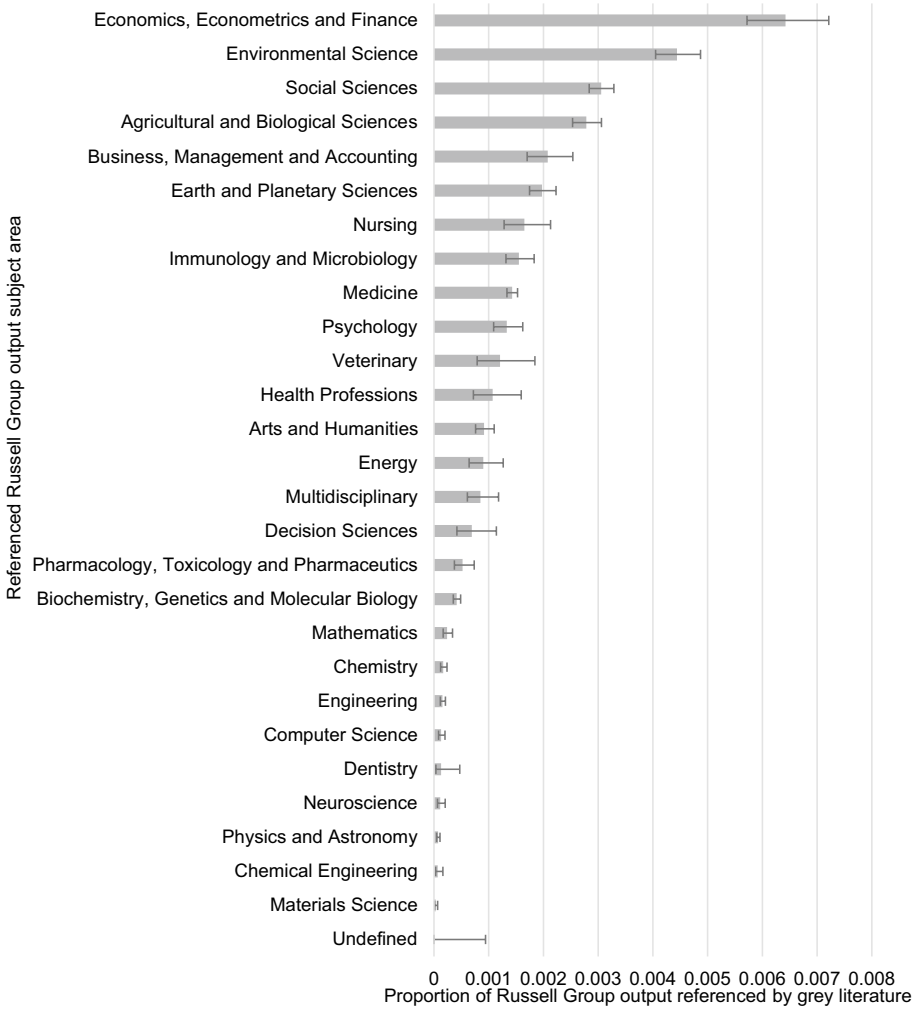

Fig. 5 Proportion of Russell Group output (all years) referenced by UK government grey literature documents (2010–2018) across 28 Scopus-defined subject areas, sorted from highest to lowest with 95% confidence intervals shown

(Undefined), where Mathematics has an upper confidence limit of 0.000339, which does not overlap with Biochemistry, Genetics and Molecular Biology. It is worth noting that the lowest ranked topic (Undefined) has no mentions in any of the grey literature searched, and with it also being the subject area of lowest references on Scopus in this dataset, the wider confidence interval could be misleading. The reason for this is likely due to the subject area being 'Undefined'—the research contained within is likely obscure or niche for a wider audience and contains relatively few records compared to other disciplines.

In terms of a year-by-year comparison shown in Figs. B19–B27 (Online Resource, Appendices B19–B27; https://doi.org/10.6084/m9.figshare.16895485), the overall order is similar for most subject areas, with Economics, Econometrics and Finance ranked first for all but 2 years (third in 2013, second in 2018), Environmental Science ranked in the

top four for every year, and two other subject areas (Social Sciences, and Agricultural and Biological Sciences) ranked in the top 10 for all 8 years. Economics, Econometrics and Finance is significantly higher than the second ranked subject area for 2011, 2015 and 2016, and close to significance in 2014. Undefined is consistently ranked last (or joint last), obviously as no references pointed to the topic in any year. Dentistry was referenced only in two years (2016 and 2017), but was ranked fifteenth for both years, meaning its overall rank was higher than it may have otherwise been.

Regression analysis of time taken for method

To help with possible use of this method, a 'time to completion' may be useful information before a research group implements it. To estimate this, linear regression was undertaken on the variables recorded and mentioned in the methodology in this chapter. Regression was performed in SPSS v26.0.0.0 and was first undertaken with all variables present. Any variables SPSS automatically excluded were removed in the first instance, and then repeated regression analyses were performed, removing one variable at a time, that being the variable with the least significance, until only variables remained that were all significant ($\alpha = 0.05$).

In the process, and after seeing the best models, it could be seen that the number of citations metadata pairs in the method was a significant factor, but the coefficient in the model was shown to be zero by SPSS. As it was significant ($p < 0.001$), the coefficient must be non-zero, so it was assumed that SPSS was having issues showing the coefficient to the correct number of significant figures or decimal places. This assumption was proven to be the case by adjusting the variable to divide the number of indicator pairs by different powers of 10 and re-running the regression analysis. It was found that dividing this number by 10,000 provided enough adjustment to show the coefficient to a sensible number of significant figures. The reason for this issue is unknown and proved to be even more confusing when looking at some of the coefficients from models partway through the regression models as some coefficients were given in scientific notation (e.g., $2.089\text{E}-6 = 2.089 \times 10^{-6}$).

Multiple models are presented to allow for a balance of best model through to easiest to calculate based on the difficulty to access the variables involved. The rank of best model (order 1, 2) is given by the highest *R* Square/Adjusted *R* Square value, shown from the SPSS output for each model. The rank of easiest to calculate (order 2, 1) is given by the difficulty of finding the information for the variables involved. Model 2 is included as the *R* Square value is very close to that of model 1, and the variables for the equation from model 2 will all be readily available before any part of the method is undertaken (i.e., before grey literature document conversion from PDF to plaintext), specifically knowing the total size of all the PDF documents involved in megabytes. The equation from model 1 is the hardest to calculate as the same information as model 2 must be known, in addition to two pieces of information about the RAM used in the computer and the conversion to plaintext must also have been completed (as total size of the PDFs/Word documents in plaintext form is required).

As can be seen by the equations, only 2 variables are related to the data itself; how many citation-indicator pairs (lead author last name and article title) to be checked, and how large the collection of grey literature is in megabytes (either as PDFs/Word documents or as plaintext files). The other variable(s) involved relate to the performance of the computer

being used to run the method, either purely the CPU clock speed in gigahertz, or additionally the RAM size and speed in megahertz. If all these variables are known, these regression equations can be used to predict the time to completion, useful to know if the impact of the grey literature documents needed to be calculated in a specific timeframe.

The final regression models are shown in Table 5 and the corresponding equations are shown below. Raw SPSS outputs for these models with extra information are shown in Figs. B28–B29 (Online Resource, Appendices B28–B29; https://doi.org/10.6084/m9.figshare.16895485).

Equation presented by model 1 (best model with highest *R* Square):

$$\begin{aligned}\text{Time taken (hours)} = {} & 43.132 + 0.0004282 \times \text{No. of citation indicator pairs from metadata} \\ & - 43.371 \times \text{CPU clock speed (GHz)} - 3.130 \\ & \times \text{RAM size (GB)} + 0.012 \times \text{RAM speed (MHz)} + 1.954 \\ & \times \text{Grey literature plaintexts total size (MB)}\end{aligned}$$

Equation presented by model 2 (easiest to calculate model with acceptable *R* Square):

$$\begin{aligned}\text{Time taken (hours)} = {} & 66.797 + 0.0004282 \\ & \times \text{No. of citation indicator pairs from metadata} \\ & - 45.534 \times \text{CPU clock speed (GHz)} + 0.107 \\ & \times \text{Grey literature PDF/Word documents total size (MB)}\end{aligned}$$

Both models and equations presented are to be tested in a future study looking at a different database of grey literature within a specific subject area to see the robustness of the equations in a related but more specific and 'research group-like' context.

Table 5 *R* Square values and coefficients with *p* value significances for predictor variables used in regression analysis (performed in SPSS), showing two separate models for time taken to find Scopus citation metadata indicator pairs on a database of 3,374 UK government grey literature documents

Regression analysis	Model 1 (highest *R* Square)		Model 2 (easiest to calculate)	
R square	0.913		0.899	
Predictor variable	Unstandardised B coefficient value	*p* value	Unstandardised B coefficient value	*p* value
Constant	43.132	0.005	66.797	< 0.001
No. of citation indicator pairs from metadata	0.0004282	< 0.001	0.0004282	< 0.001
CPU clock speed (GHz)	− 43.371	< 0.001	− 45.534	< 0.001
RAM size (GB)[a]	− 3.130	0.007	–	–
RAM speed (MHz)[a]	0.012	0.007	–	–
Grey literature plaintexts total size (MB)[a]	1.954	< 0.001	–	–
Grey literature PDF/Word documents total size (MB)[b]	–	–	0.107	< 0.001

[a]Predictor variable only used in Model 1

[b]Predictor variable only used in Model 2

Discussion and limitations

The optimal method found here does not necessarily represent the new 'gold-standard' to measure impact of grey literature upon standard academic output but is an accurate and reliable approach for use in further research. In this way, the combination of options above that lead to the smallest difference (bias) and variation across all records is the aim, allowing the conclusion that the optimal method proposed has both high recall and precision. The method was deemed to having an acceptably low standard deviation/limits of agreement, but further research into the methodologies here may present a way at reducing this variability while keeping the bias small.

To illustrate the difficulties in identifying references in grey literature, after conducting this research, a relatively low proportion (23%) of UK government grey literature documents contain references to UK Russell Group academic work, and where they do, they do not necessarily appear at the end of the document, in a titled reference section or even in a specific style. Some examples were found where a citation may have been present in the main body of the text but was not explicitly referenced at any point.

There are many reasons why the optimal method presented (method 12) may have performed best in this situation. Firstly, the distances between indicators chosen were arbitrary. A pattern in the result shows that of the more feasible approaches (methods 1–12 and 19–21), the larger the distance between matching terms, the smaller the absolute value of the bias and the standard deviation, hence more accurate and precise. It is plausible that a larger distance may obtain a more accurate method, although this was not tested in this research. Now that the combination of lead author last name and title has been deemed to be the best indicator pair (with lead author last name not necessarily first), future work could fix these and focus on finding a more optimal maximum distance between the author and title of paper. This could be looked at in the same dataset presented here, looking at a full scale of distance (up to the maximum length of the grey literature document being tested) and analysed using a similar or different statistical method to see where a crossover from underprediction to overprediction occurs, considering false positives and negatives if possible.

Secondly, although just stated that the pair of lead author last name and title are the best combination, it is important to consider why this may be. Thinking about the characteristics of the three options for indicators, the more generic one is the year of the academic document. As stated previously in this work, years being a predictable 4-digit number would be a good indicator to use, but likely are too generic as they may well appear in other parts of the study such as table of figures. The most specific of these is the title of the document, which can vary in length but are usually many words long. Here, the titles were truncated to no more than 10 words, but this is still the most specific of the three indicators considered. This may have the opposite problem of using year; they are so specific that a punctuation discrepancy or line wrap halfway through the citation could cause the match to be missed. The lead author last name sits in the middle of these two—generally more specific than a year but less so than a title, even with double-barrelled or similar names.

As this method requires at least two of the three indicators to be used, it is proposed that lead author last name along with title rather than year was more accurate as some references may state the year far away from the author last name in the main body of research (for example, stating the author at the start of a quote and the year after the quote). This leaves the main 'matching' section to likely be in the reference list, where authors are generally listed first, and then depending on the referencing style used, the year or title could

come next. Future work in this area could consider trying to detect the exact style of referencing used in grey literature. In the UK government grey literature documents tested here, it is currently unclear what style, if any, is generally used.

It is possible that due to the short nature of some lead author last names and years, they could be contained within other strings, leading to false positives. The optimal method does not make use of the year of a reference but does include lead author last names. An extended method could include looking for a preceding line break (in the case of most reference lists where each reference starts with the lead author last name) and/or a following comma (a common format in references, Pears & Shield, 2019). However, this brings extra complexity to the method and may cause missed matches if the grey literature uses a non-standard format.

The results are limited to using a single case study (UK government grey literature publications matched with Russell Group university academic output). It is possible that other grey literature repositories may have more standardised ways of reporting references within documents or recording these within metadata (e.g., Scopus), but this method has been designed to be applicable to any text documents, regardless of definition or knowledge of content.

Although the optimal method presented in this study has been limited to a sample of documents taken from a single grey literature repository, the inclusion of tools such as Publish or Perish, and Webometric Analyst allow for much larger scale studies to use the methodology presented here. This process has been shown to work on a larger dataset in the second part of this study, with promising results assessing impact of an entire repository and allowing for subject areas differences to be seen.

Once applied to a larger dataset, the prediction of the number of references per grey literature document was shown to be assessable, however is at best a lower bound for the true number of references present in the document. An upper bound could be calculated if the prevalence of Russell Group references in either UK government grey literature (for this study) or more general grey literature was known, although that may only be predicted from a novel method such as those presented by this study.

Conclusions

In answer to the research questions, it is possible to match references in grey literature to a known list of academic publications. This was possible as a large list of metadata was collected from Scopus—the hardest part of this method due to the manual collection. If these tests were to be repeated by other users, it may be useful to find an automated way of collecting data like this, and to make sure the origin of potential matches is suitable (i.e., US-based journals/books if undertaken on US-based grey literature repositories or collections).

The exact number of citations here may be trusted as the lead author last name, title and year of the known matches were deliberately included in the matching process to check if the method is reliable for both false positives and negatives. In addition, 21 different combinations of options were tested extensively to determine a best approach. The ultimate method proposed (lead author last name and title in either order, a maximum of 200 characters apart) has low underestimation of reference count and can be treated as effective. However, when performing on a dataset with potentially any number of references, any method is likely to present many missed matches due to the unknown nature of potential citations—all academic output from all recorded years would have to be included in the list

of matching terms to remove this problem. It may be more appropriate to use the method proposed here to assess impact in terms of proportion between different subject areas rather than pure numbers or if the scope of academic references used is small and known.

Answering the second research question, applying the optimal method established here on a wider scale has been shown to work as intended, allowing analyses to be undertaken that involved comparing differences in grey literature impact within different subject areas. Here, the results were able to show clear disciplinary differences across the top few subjects, with differences between them becoming less evident as subjects receive less citations.

Although this method allows us to see the proportion per year of grey literature that references academic research, the trend shown here is slight across the period 2010–2018, and any variations may be due to factors not considered as part of this study. An estimate for the proportion of UK government grey literature that references UK Russell Group academic research was also available by this method, showing approximately 23% of UK government grey literature references UK Russell Group outputs.

This method also allows for future expansion and robustness-testing by means of regression analysis for time predictions when running the method. Characteristics of references in specific grey literature documents may also be assessed from this method, which are undertaken in a future study.

To aid with impact assessment of this type of content, it is also suggested that researchers potentially wishing to measure impact of their grey literature should endeavour to create DOIs where appropriate so persistent identifiers already used to assess impact by other mediums (Altmetric, n.d.) can be extended into the grey literature realm.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-022-04408-4.

Acknowledgements This paper is a substantially extended version of the conference paper presented virtually and published in the conference proceedings at ISSI2021, Belgium (Bickley, Kousha & Thelwall, 2021). This version of the paper continues with the methodology described at ISSI2021 and applies it to a larger case study, assessing the feasibility of the method in a wider framework.

Declarations

Conflict of interest Two of the three authors of this paper (Kousha and Thelwall) currently sit on the distinguished reviewers board for *Scientometrics*.

References

Alberani, V., Pietrangeli, P. D. C., & Mazza, A. M. (1990). The use of grey literature in health sciences: A preliminary survey. *Bulletin of the Medical Library Association, 78*(4), 358.

Altmetric. (n.d.). *Altmetric for Institutions*. Retrieved April 26, 2019 from https://www.altmetric.com/audience/institutions/

Benzies, K. M., Premji, S., Hayden, K. A., & Serrett, K. (2006). State-of-the-evidence reviews: Advantages and challenges of including grey literature. *Worldviews on Evidence-Based Nursing, 3*(2), 55–61. https://doi.org/10.1111/j.1741-6787.2006.00051.x

Bickley, M. S., Kousha, K., & Thelwall, M. (2019). Can the impact of grey literature be assessed? An investigation of UK government publications cited by articles and books. In G. Catalano, C. Daraio, M. Gregori, H. F. Moed, & G. Ruocco (Eds.), *17th International Conference on Scientometrics &*

Informetrics, ISSI2019: Proceedings, Volume II (pp. 1801–1812). International Society for Scientometrics and Informetrics/Edizione Efesto.

Bickley, M. S., Kousha, K., & Thelwall, M. (2020). Can the impact of grey literature be assessed? An investigation of UK government publications cited by articles and books. *Scientometrics, 125*(2), 1425–1444. https://doi.org/10.1007/s11192-020-03628-w

Bickley, M. S., Kousha, K., & Thelwall, M. (2021). A systematic method for identifying references to academic research in grey literature. In W. Glänzel, S. Heeffer, P. Chi, & R. Rousseau (Eds.), *18th International Conference on Scientometrics & Informetrics, ISSI2021: Proceedings* (pp. 121–132). International Society for Scientometrics and Informetrics.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307–310. https://doi.org/10.1016/s0140-6736(86)90837-8

Cordes, R. (2004). Is grey literature ever used? Using citation analysis to measure the impact of GESAMP, an international marine scientific advisory body. *Canadian Journal of Information and Library Science, 28*(1), 49–70.

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica, 25*(2), 141–151.

Glyph & Cog, LLC. (2019). *Download Xpdf and XpdfReader*. Retrieved March 24, 2019 from https://www.xpdfreader.com/download.html

GreyNet International. (2019). *Document types in grey literature*. Retrieved January 4, 2019 from http://www.greynet.org/greysourceindex/documenttypes.html

Harzing, A. W. K. (2010). *The publish or perish book*. Tarma Software Research.

Hook, D. W., Porter, S. J., & Herzog, C. (2018). Dimensions: Building context for search and evaluation. *Frontiers in Research Metrics and Analytics., 3*, 23. https://doi.org/10.3389/frma.2018.00023

Interagency Gray Literature Working Group. (IGLWG, 1995). *Gray information functional plan (GIFP)*. Retrieved January 7, 2019 from https://apps.dtic.mil/dtic/tr/fulltext/u2/b300928.pdf

Marx, W., & Bornmann, L. (2016). Change of perspective: Bibliometrics from the point of view of cited references—a literature overview on approaches to the evaluation of cited references in bibliometrics. *Scientometrics, 109*(2), 1397–1415. https://doi.org/10.1007/s11192-016-2111-2

Pears, R., & Shields, G. J. (2019). *Cite them right: The essential referencing guide*. Macmillan International Higher Education.

Pelzer, N. L., & Wiese, W. H. (2003). Bibliometric study of grey literature in core veterinary medical journals. *Journal of the Medical Library Association, 91*(4), 434.

Research Excellence Framework. (2019). *Guidance on submissions, 68*. Retrieved March 26, 2020 from https://www.ref.ac.uk/media/1092/ref-2019_01-guidance-on-submissions.pdf

Schöpfel, J. (2010). Towards a Prague definition of grey literature. In: *Twelfth international conference on grey literature: Transparency in grey literature*. Grey Tech Approaches to High Tech Issues. (pp. 11–26).

University of New England. (UNE, 2019). *Grey literature*. Retrieved January 4, 2019 from https://www.une.edu.au/library/support/eskills-plus/research-skills/grey-literature

Woods, S., Phillips, K., & Dudash, A. (2020). Grey literature citations in top nursing journals: A bibliometric study. *Journal of the Medical Library Association, 108*(2), 262. https://doi.org/10.5195/jmla.2020.760

World Health Organization. (WHO, 2020). *Water, sanitation, hygiene and waste management for COVID-19*. Retrieved March 25, 2020 from http://www.dwi.gov.uk/2020-03-03%20WASH-IPC_EN.pdf