



CONTRIBUTIONS

Variation in Results of Three Biology-Focused Search Engines: A Case Study Using North American Tree Species

Pete Bettinger¹ , Krista Merry¹ , Steve McNulty², Jacek Siry¹, and Michael Gavazzi²

¹Warnell School of Forestry and Natural Resources, University of Georgia, 180 E. Green St, Athens, Georgia 30602 USA

²Southern Research Station, U.S. Department of Agriculture, 3041 E. Cornwallis Rd, Research Triangle Park, North Carolina 27709 USA

Introduction

A publication is a permanent, accessible, isolated, and definitive piece of work; it should not be difficult to determine who wrote it, when they wrote it, and where it was published (Cole and Eales 1917). Publications that communicate research findings to others include those works found in peer-reviewed journals and conference proceedings, along with books, reports, and other published materials. A search and synthesis of the published literature is a first step in the development of a research paper or proposal and is important for setting the stage and for defining the gaps in science that need addressing. Given advances in scientific and computing technologies and an increase in both the number of scientists and scientific outlets (e.g., journals) over the last few decades, the volume of published peer-reviewed journal articles in science and engineering has grown by about 4% per year over the last decade to about 2.6 million articles annually in 2018; 38% of these publications were produced or developed by researchers from China and the United States (White 2019). It has been suggested that the number of published works on a particular topic can be used as an indicator of importance or awareness assigned to it by the scientific community (Cohen et al. 2008, Schöffel et al. 2016).

Today, several search engines can be used to identify and compile bibliographic information on topics of interest to scientists and the general public. Search engines assess the bibliographic metadata (e.g., keywords and topics) associated with publications against a search query and report matches. However,

Bettinger, P., K. Merry, S. McNulty, J. Siry, and M. Gavazzi. 2020. Variation in Results of Three Biology-Focused Search Engines: A Case Study Using North American Tree Species. *Bull Ecol Soc Am* 102(1):e01808. <https://doi.org/10.1002/bes2.1808>

the volume of bibliographic records produced by different search engines can differ considerably while using the same search query. Therefore, we designed this study to examine the results provided by three major search engines (Web of Science, AGRICOLA, and CAB Abstracts) and their ability to identify the available scholarly literature for biology-focused studies that form a bibliographic database. In this case study, we examined the findings of a search of articles published between 1900 and 2020 concerning 696 species of plants with tree-like growth characteristics that may be found in North America. The case study is inspired by recent activity aimed at the potential development of a third edition of *Silvics of North America* (SNA). The second SNA (Burns and Honkala 1990) included life-history accounts of about 200 of the most commercially and ecologically important tree species of the United States, Canada, Mexico, and the Caribbean Islands. This compendium updated the first SNA published 35 years earlier (Fowells 1965) and complements other similar works such as the 14-volume description of trees in North America developed over a century ago by Sargent (1890, 1902). However, since the second edition was published over 30 years ago, climate change, invasive species, and improved ecosystem distribution model and projection capabilities may have rendered the most current version of the SNA obsolete. A future revision of this important reference resource is currently being considered by government agencies in North America. One consideration in an SNA revision would be the potential inclusion of additional trees species. Therefore, knowledge is needed to inform discussions and guide the selection of tree species for a possible third installment and update of this resource. However, there is a shortage of literature, as reflected through published works, on measuring the importance of tree species (and thus which additional species, if any, to include in a revised SNA). In our literature review on this subject, we did not find any papers that cover the broad scope of publications on tree species of North America.

Three major search engines (Web of Science, AGRICOLA, and CAB Abstracts) are commonly used to comb the available scholarly literature for biology-related studies that form a bibliographic database. We used these three search engines and SNA as a case study for comparative publication-based species importance, and we provide information regarding which species might be included in an updated SNA. Therefore, the goal of this research is to assess the variation in scholarly publication results of North American tree species through publication counts, and the development of a metric to rank the importance of tree species that could be included in a revised SNA. Further, we discuss the usefulness of such an approach for describing the societal importance of tree species. Although the importance of a topic can be measured in many ways (Hood and Wilson 2001), we used publication count as one proxy for societal interest and prioritization in the research community (Elsevier Analytical Services 2019). We suggest that the synthesis provided here can provide useful information to researchers regarding the importance of North American tree species and that this information could also be used to guide the selection of trees to include in a revised SNA. This work involves bibliometrics: a field with the goal of creating knowledge from the scholarly output, using bibliographic information such as publication counts to describe the evolution and trends in topical research, or the productivity of researchers or research groups. Bibliometrics also seeks to understand where one person or organization might lie within scientific society (Martínez-Gómez 2015). Bibliometrics and citation analysis began in the field of law at least 200 years ago (Shapiro 1992) and generally involves the quantitative treatment and analysis of literature. Therefore, the basis for any bibliometric study is a database of publications. Bibliometrics is often viewed as a subset of scientometrics, which focuses on the measurement and analysis of all aspects of science and scholarship (Hood and Wilson 2001). The use of bibliometrics (publication counts) as a source of information for policy formulation discussions is well established, and outcomes are used as an indication of research output (Pardey and Christian 2002). However, the

disadvantages of using these methods include a bias toward popular fields and the possibility of incompleteness (Martínez-Gómez 2015).

Methods

We use an indication of publication count (Wildgaard et al. 2014) as the method for analyzing the variation in scholarly publication results and the relative importance of many North American tree species (those native or introduced to the United States, Canada, Mexico, and the Caribbean Islands). We consider tree species in this research as perennial plants that often have a single stem that may grow to 4 or 5 m in height, as suggested in the US Department of Agriculture (2020) PLANTS database. Publication production and trends have long described the broad array of scientific fields using citation counts (Cole and Eales 1917, White 2019). The popularity of the subject matter has been described using database searches for the frequency of occurrence of keywords (Hood and Wilson 2001). Bias and interest toward one subject or another, from disciplines as diverse as bees, liver cancer, wastewater irrigation, and operations research methods, have been suggested (Bettinger and Chung 2004, Maassen 2016, Sumner et al. 2018, Klingelhöfer et al. 2018). While many analyses focus on specific journals within specific fields of study, our focus is on individual works or publications (similar to other studies such as Waltman and van Eck 2012 and Wang et al. 2013) from any indexed publication outlet.

In this study, we use the number of published works, through queries of bibliographic databases, to infer the importance of certain tree species. While perhaps the simplest bibliographic parameter (Choudhri et al. 2015), publication count is still an informative measure. For example, other metrics have been used to analyze literature concerning e-cigarettes (Briganti et al. 2019) and evidence-based medicine, such as published clinical trials, case reports, and other works (Hung et al. 2015). While a publication count, or volume of output, may provide an estimate of activity and productivity in the literature of a given field, it does not provide differentiation among publications according to their value (Choudhri et al. 2015). A publication count is simply the number of unique publications that a search query locates using a set of search criteria. A publication count differs from a citation count in that a publication count refers to a person's or subject's entire body of work. In contrast, a citation count refers to a specific piece of work. A citation count (the number of other publications that cite a particular work) associated with an individual work can infer the importance of that work (Albert et al. 1991). However, citation counts also have weaknesses, including the impact of self-citation and other manipulations that affect the metric (Choudhri et al. 2015).

Search queries were made using three bibliographic databases: Web of Science, AGRICOLA, and CAB Abstracts. The Web of Science contains over 74.8 million records from over 21,000 peer-reviewed journals and tens of thousands of books, reports, and conference proceedings from a variety of scientific disciplines that extend back as far as 1900 (Clarivate 2020). In our investigation, we used the Clarivate interface to search the Web of Science Core Collection through a basic search of keywords (genus and species and variants of these for each). AGRICOLA (AGRICultural OnLine Access) contains 5.2 million records representing citations for publications (journal articles and abstracts) and other bibliographic records (monographs, serials, etc.) of agriculture and allied disciplines that extend back to the 15th century (US Department of Agriculture 2013). For peer-reviewed journals, AGRICOLA maintains a database of records for 1,715 journals from 127 publishers. The databases within AGRICOLA are updated daily. In our investigation, we used a University of Georgia interface for AGRICOLA, search-

ing all publication types using a Boolean/phrase search of keywords (genus and species and variants of these for each). CAB Abstracts contains over 9.7 million records representing abstracts, journal articles, conference proceedings, books, and other published works across applied life sciences fields. The database contains works that extend back as far as 1973 and include publications from 120 countries. The Centre for Agriculture and Bioscience International (2019) suggested that around 80% of the records in the database cannot be found anywhere else. We again used a University of Georgia interface for CAB Abstracts, searching all publication types using a Boolean/phrase search of keywords (genus and species and variants of these for each). The use of different databases can yield results that largely overlap but have also been shown to differ considerably based on the source material (Choudhri et al. 2015).

Search queries were conducted in English, using contemporary taxonomy (genus and species identifiers and their surrogates), as suggested through SNA (Burns and Honkala 1990) and the U.S. Department of Agriculture PLANTS Database. The initial set of tree species were derived from SNA and supplemented with species that were identified as having a tree-like growth quality from the U.S. Department of Agriculture PLANTS Database. Duplications of genus and species identifiers were resolved by carefully examining each genus/species combination. Some tree species were combined into one instance of a species when they used the same genus/species combination. Other tree species had to be carefully evaluated to avoid double-counting search results. For example, *Pinus monophylla* (western white pine) is also known as a variant of another species, *Pinus edulis* var. *fallax*. Therefore, the query for western white pine contained a criterion using the Boolean operator OR ("*Pinus monophylla*" or "*Pinus edulis* var. *fallax*"). However, when *Pinus edulis* (two-needle pine) was assessed, it required a criterion using the Boolean operator NOT ("*Pinus edulis*" not "*Pinus edulis* var. *fallax*"). Common names were considered for inclusion in the queries using the Boolean operator OR (e.g., OR "*common name*"). However, several tree species shared common names and the query results could not be disentangled. Therefore, common names were not included in search queries. The term "North America" was not used as a keyword (as topic or in "all fields") in these queries as it acted to severely limit the results. In fact, in limited tests using this term, no results were returned when queries included AND "North America."

Ultimately, 696 tree species were assessed for the number of publications containing their genus/species appearance in all fields within each of the three bibliographic databases. Two timeframes were considered: 1900 to 2020, and 1985 to 2020. The former represents a realistic assessment of the publications mentioning each species during a contemporary period of science. The latter represents the general period since the last SNA (Burns and Honkala 1990) concluded. In communications with authors who contributed to the previous SNA, Burns and Honkala noted that while the publication date is 1990, the publication process began around 1980. Most material was ready for publication by mid-decade (1985).

As in other studies (e.g., Martínez-Gómez 2015), different types of papers were considered for this analysis, including peer-reviewed (or edited) journal articles, monographs, conference proceedings, book chapters, graduate and undergraduate theses, and peer-reviewed (or written) research reports. No attempt was made, from the thousands of returns generated through these searches, to exclude items based on their relative contribution to the state of knowledge of the life history of North American tree species. While this limits human error in judgment of the comparable quality of published works, it also can introduce a false sense of importance on a tree species when based only on publication count. However, there should be no reason to suspect bias across the inflated importance values. If all publication counts are equally inflated, then no rules of normality have been violated.

As opposed to an evaluative study, this study is purely descriptive and quantitative, relying on counts of publications to assess relative effort applied to the study of tree species. In this regard, the publication counts might be viewed as a proxy for species importance. Although not as complex as the relevance estimator suggested by von Korff et al. (2017) and the scoring scheme proposed by Mørk et al. (2014) for keywords found in an abstract, a bibliographic analysis can be used to develop an indicator of importance. After publication counts were noted, tree species were ranked using each of the three bibliographic databases and the two timeframes (1900–2020, and 1985–2020). Coniferous and deciduous species and others (e.g., Mohave yucca [*Yucca mohavensis*]) were ranked separately. An average rank was then computed using the six categories (three bibliographic databases, two timeframes) and was used for reporting purposes here. Tree species were also placed into quartiles based on their average rank.

Results

Without accounting for duplicate publications, or those that were contained in query results for more than one tree species, we located 111,204 publications through Web of Science, 145,080 publications through AGRICOLA, and 645,457 publications through CAB Abstracts. The publication dates of these spanned all years (1900–2020). Interestingly, from 1985 forward, beginning at about the time that the development of the last SNA (Burns and Honkala 1990) was concluding, 95.1%, 74.8%, and 53.4% of these papers were published, according to the Web of Science, AGRICOLA, and CAB Abstracts, respectively. These percentages indicate that there is temporal and extent search bias between the three search engines, with the Web of Science queries having the highest percentage of results in their database from the last three or four decades than the other two publication search processes. These findings suggest that the choice of search engine can impact the amount and recentness of the comparative literature for use in the development, analysis, and discussion of biological research.

If the three publication search processes were considered to be equal in quality, the average ranking of tree species (by search query returns) might be used as a measure of interest in tree species. A complete list of the rankings can be found in the Dryad Digital Repository (Bettinger 2020). The top 20 coniferous and deciduous species (Tables 1 and 2) include several commercially important species, not only for North America, but also elsewhere. *Eucalyptus globulus* and *Pinus radiata*, for example, are more commercially crucial in South America and Oceania. The natural geographic range of *Eucalyptus globulus* is also outside of North America, and it is considered an introduced or exotic tree species on this continent. These and other species, such as *Pseudotsuga menziesii* and *Pinus patula*, are also planted in areas outside of their native range. Therefore, it might be presumed that a healthy percentage of the query returns for some species are focused on their economic, ecological, and social importance outside of North America. Some of the findings from research outside of North America may be transferrable for understanding a species' life history and socioeconomic significance, but to what extent is uncertain. The results for coniferous, deciduous, and other tree species (Figs. 1, 2) show a negative exponential amount of publications in Web of Science that level off after the median-ranked tree species. Although not displayed here, the AGRICOLA and CAB Abstracts results had a similar trend. Between 82% and 85% of the query returns for the entire set of coniferous tree species, depending on the bibliographic database used, were related to the top 25% of the tree species (32 species). For deciduous and other species, this was more dramatic, as between 90% and 95% of the query returns for the entire set were related to the top 25% of the tree species (142 species).

Interestingly, the highest amount of query returns for a tree species from a search of one bibliographic database did not necessarily yield the highest amount of query returns from another bibliographic database. For example, the top-ranked coniferous species (*Pinus radiata*) had nearly 2,000 more query returns using the AGRICOLA bibliographic database and twice (over 80,000) as much as the next species (*Pinus sylvestris*) using the CAB Abstracts bibliographic database. Yet *Pinus sylvestris* had 2.5 times more query returns (11,387 vs. 4,497) when the Web of Science bibliographic database was used. For deciduous and other tree species, *Populus balsamifera* and *Populus fremontii* both had higher amounts of query returns (31,920 and 30,521) when the CAB Abstracts bibliographic database was used than the overall top-ranked species, *Eucalyptus globulus* (26,739 query returns using CAB Abstracts). However, both *Populus* species had less than 10% of the *Eucalyptus globulus* query returns when the Web of Science bibliographic database was used. Several deciduous tree species had zero query returns when searched in the bibliographic databases: 62 species when using Web of Science, 68 when using AGRICOLA, and 26 when using CAB Abstracts. All coniferous tree species had at least one query return from searches in the three bibliographic databases.

Table 1. Top-ranked coniferous tree species, through returns of queried publications in Web of Science, AGRICOLA, and CAB Abstracts (27 January–10 February 2020).

Rank	Genus and species	Common name
1	<i>Pinus radiata</i>	Monterey pine
2	<i>Pinus sylvestris</i>	Scots pine
3	<i>Pseudotsuga menziesii</i>	Douglas-fir
4	<i>Pinus taeda</i>	Loblolly pine
5	<i>Pinus contorta</i>	Lodgepole pine
6	<i>Pinus ponderosa</i>	Ponderosa pine
7	<i>Abies balsamea</i>	Balsam fir
8	<i>Picea glauca</i>	White spruce
9	<i>Picea mariana</i>	Black spruce
10	<i>Pinus nigra</i>	Austrian pine
11	<i>Picea sitchensis</i>	Sitka spruce
12	<i>Pinus banksiana</i>	Jack pine
13	<i>Pinus elliottii</i>	Slash pine
14	<i>Pinus strobus</i>	Eastern white pine
15	<i>Juniperus communis</i>	Common juniper
16	<i>Tsuga heterophylla</i>	Western hemlock
17	<i>Pinus resinosa</i>	Red pine
18	<i>Juniperus virginiana</i>	Eastern redcedar
19	<i>Pinus palustris</i>	Longleaf pine
20	<i>Thuja plicata</i>	Western redcedar

Notes. Ranks are based on an average, using two periods of time (1900–2020, 1985–2020) and three bibliographic databases (Web of Science, AGRICOLA, and CAB Abstracts). Genus and species denote the main genus and species. The query may have included synonyms and other versions of the genus and species. The common name denotes one of the foremost common names for this tree species. A tree species may potentially be referenced by several other common names.

Table 2. Top-ranked deciduous and other tree species, through returns of queried publications in Web of Science, AGRICOLA, and CAB Abstracts (27 January –10 February 2020).

Rank	Genus and species	Common name
1	<i>Eucalyptus globulus</i>	Tasmanian bluegum
2	<i>Acer saccharum</i>	Sugar maple
3	<i>Populus tremuloides</i>	Quaking aspen
4	<i>Robinia pseudoacacia</i>	Black locust
5	<i>Quercus rubra</i>	Northern red oak
6	<i>Eucalyptus grandis</i>	Grand eucalyptus
7	<i>Populus deltoides</i>	Eastern cottonwood
8	<i>Juglans nigra</i>	Black walnut
9	<i>Populus balsamifera</i>	Balsam poplar
10	<i>Acer rubrum</i>	Red maple
11	<i>Prunus serotina</i>	Black cherry
12	<i>Carya illinoensis</i>	Pecan
13	<i>Populus trichocarpa</i>	Black cottonwood
14	<i>Alnus glutinosa</i>	European alder
15	<i>Populus fremontii</i>	Fremont cottonwood
16	<i>Liriodendron tulipifera</i>	Yellow-poplar
17	<i>Acer negundo</i>	Boxelder
18	<i>Artemisia tridentata</i>	Big sagebrush
19	<i>Quercus alba</i>	White oak
20	<i>Betula papyrifera</i>	Paper birch

Notes. Ranks are based on an average, using two periods of time (1900–2020, 1985–2020) and three bibliographic databases (Web of Science, AGRICOLA, and CAB Abstracts). Genus and species denote the main genus and species. The query may have included synonyms and other versions of the genus and species. The common name denotes one of the foremost common names for this tree species. A tree species may potentially be referenced by several other common names.

The ten lowest ranked coniferous species included in the 1990 version of SNA (Burns and Honkala 1990) all had fewer than 50 query returns when we used the Web of Science bibliographic database (Table 3). *Juniperus silicicola*, *Picea breweriana*, and *Pinus glabra* all had 11 or fewer query returns when we used the Web of Science bibliographic database. Similar results were obtained when we used the AGRICOLA bibliographic database, and each had about four times as many query results when we used the CAB Abstracts bibliographic database. The ten lowest ranked coniferous species included in the SNA were also ranked in the bottom half of the coniferous species rankings (Table 3). In contrast, all ten of the lowest ranked deciduous and other species (Table 4) had nine or fewer query returns when we used the Web of Science bibliographic database. Similar results were obtained when we used the AGRICOLA bibliographic database. As many as 60 query results were obtained for one of these species when we used the CAB Abstracts bibliographic database. However, three of these species (i.e., *Carya myristiciformis*, *Ulmus serotina*, and *Castanopsis chrysophylla*) only yielded one query return when we used the Web of Science bibliographic database. One species (*Carya myristiciformis*) that was included in the SNA yielded no results when we used the AGRICOLA bibliographic database.

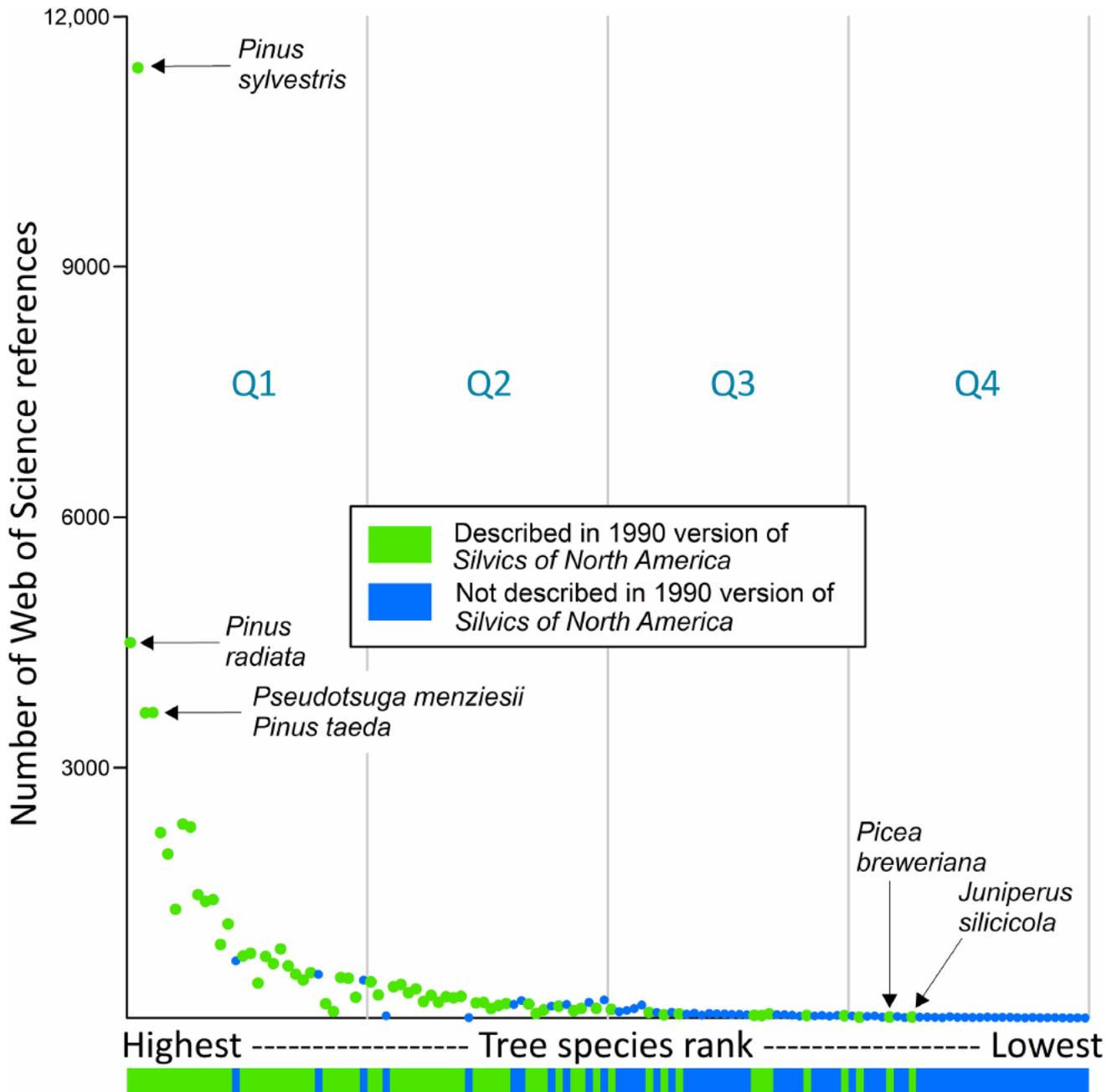


Fig. 1. Frequency distribution of coniferous tree species returns of queried publications in Web of Science (27 January –10 February 2020), over the time range 1900–2020.

The highest ranked coniferous species that were not included in the 1990 version of SNA (Table 5) had some notable characteristics. These included several species (juniper and cypress) associated with interior xeric landscapes, and three pine species (*Pinus caribaea*, *Pinus patula*, and *Pinus oocarpa*) more prevalent in the tropical or subtropical southern part of North America. The highest ranked deciduous and other species that were omitted from the 1990 version of SNA include three species (red mangrove [*Rhizophora mangle*], black mangrove [*Avicennia germinans*], and false tamarind [*Lysiloma latisiliquum*]) associated with wet tropical or subtropical landscapes, several

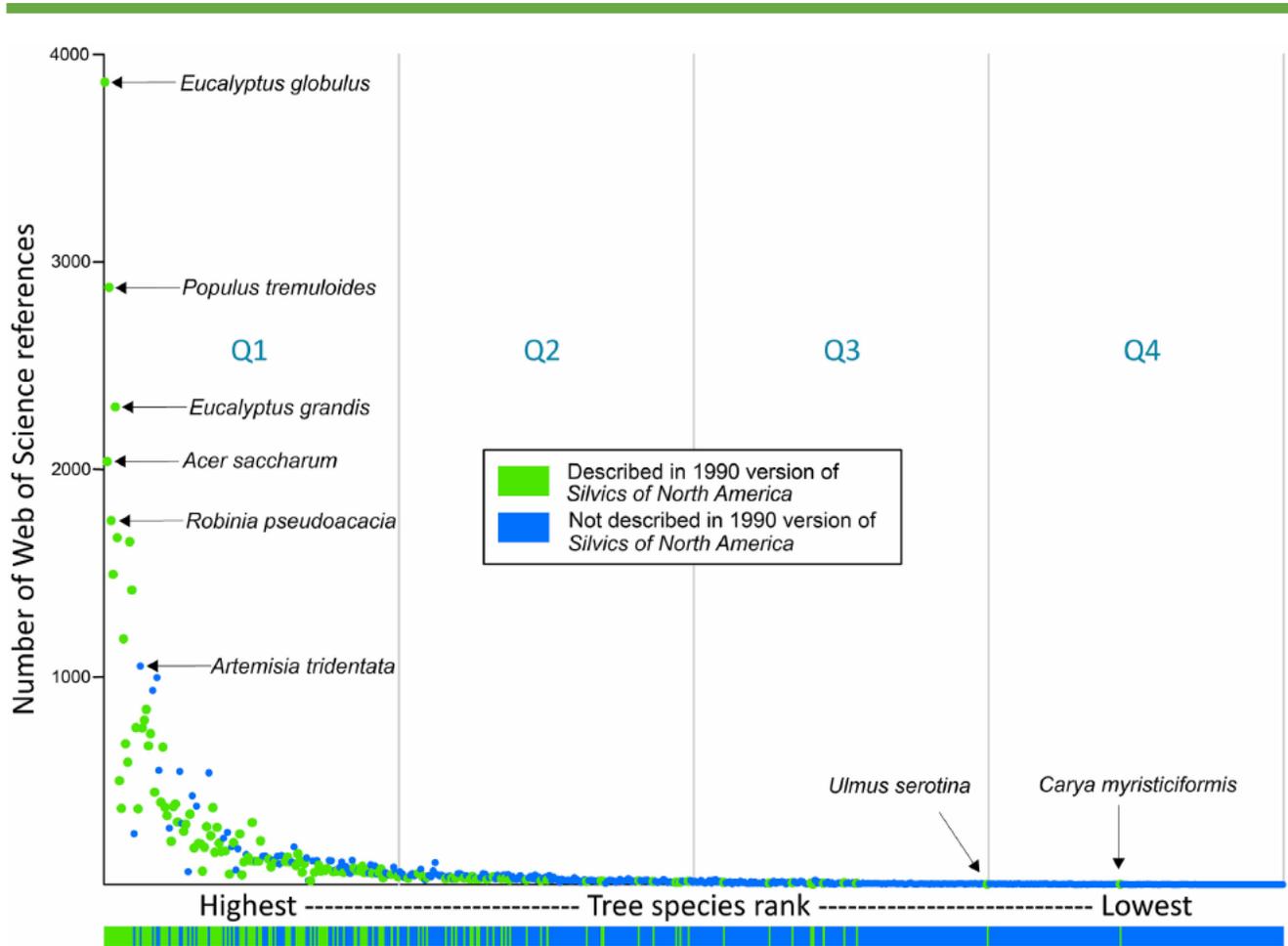


Fig. 2. Frequency distribution of deciduous and other tree species returns of queried publications in Web of Science (27 January –10 February 2020), over the time range 1900–2020.

species (Fremont cottonwood [*Populus fremontii*], big sagebrush [*Artemisia tridentata*], mesquite [*Prosopis juliflora*], and honey mesquite [*Prosopis glandulosa*]) again associated with interior xeric landscapes, one decimated species (American chestnut [*Castanea dentata*]) that continues to survive as a small tree in the Appalachian Mountains, and an evergreen monocot (*Yucca mohavensis*) (Table 6).

One of the more interesting findings involved Washoe pine (*Pinus washoensis*), whose status as a potential variant of *Pinus ponderosa* remains an open question (Willyard et al. 2017), but is recognized in our study as a tree species from one of the databases we used to form our species list. We received five query returns when we used the Web of Science bibliographic database (“*Pinus washoensis*” OR “*Pinus ponderosa* var. *washoensis*” OR “*Pinus ponderosa* ssp. *washoensis*”). However, similar queries used in the AGRICOLA and CAB Abstracts returned about 92% to 95%, respectively, of the returns for *Pinus ponderosa* (2,500+ in each case) even when the ponderosa pine query included NOT “*Pinus ponderosa* var. *washoensis*” and NOT “*Pinus ponderosa* ssp. *washoensis*.” These inflated and most likely duplicate query returns for Washoe pine distorted its rank among the coniferous tree species.

Table 3. Ten lowest ranked coniferous tree species (of 128 analyzed) that were included in the 1990 version of *Silvics of North America*, through returns of queried publications in Web of Science, AGRICOLA, and CAB Abstracts (27 January –10 February 2020).

Rank	Genus and species	Common name
105	<i>Juniperus silicicola</i>	Southern redcedar
102	<i>Picea breweriana</i>	Brewer spruce
98	<i>Pinus glabra</i>	Spruce pine
96	<i>Torreya taxifolia</i>	Florida nutmeg
91	<i>Larix lyallii</i>	Subalpine larch
86	<i>Pinus pungens</i>	Table Mountain pine
85	<i>Pinus sabiniana</i>	California foothill pine
84	<i>Pinus serotina</i>	Pond pine
74	<i>Taxodium ascendens</i>	Pond cypress
72	<i>Pinus clausa</i>	Sand pine

Notes. Ranks are based on an average, using two periods of time (1900–2020, 1985–2020) and three bibliographic databases (Web of Science, AGRICOLA, and CAB Abstracts). Genus and species denote the main genus and species. The query may have included synonyms and other versions of the genus and species. The common name denotes one of the foremost common names for this tree species. A tree species may potentially be referenced by several other common names.

Table 4. Ten lowest ranked deciduous and other tree species (of 568 analyzed) that were included in the 1990 version of *Silvics of North America*, through returns of queried publications in Web of Science, AGRICOLA, and CAB Abstracts (27 January –10 February 2020).

Rank	Genus and species	Common name
490	<i>Carya myristiciformis</i>	Nutmeg hickory
435	<i>Ulmus serotina</i>	September elm
363	<i>Populus heterophylla</i>	Swamp cottonwood
357	<i>Nyssa ogeche</i>	Ogeechee tupelo
347	<i>Ulmus thomasii</i>	Rock elm
342	<i>Castanopsis chrysophylla</i>	Giant chinquapin
341	<i>Fraxinus profunda</i>	Pumpkin ash
332	<i>Acer barbatum</i>	Southern sugar maple
321	<i>Didymopanax morototoni</i>	Matchwood
299	<i>Magnolia fraseri</i>	Mountain magnolia

Notes. Ranks are based on an average, using two periods of time (1900–2020, 1985–2020) and three bibliographic databases (Web of Science, AGRICOLA, and CAB Abstracts). Genus and species denote the main genus and species. The query may have included synonyms and other versions of the genus and species. The common name denotes one of the foremost common names for this tree species. A tree species may potentially be referenced by several other common names.

Discussion

As others have noted, bibliographic analyses can add a quantitative aspect to a somewhat imperfect qualitative assessment process (Choudhri et al. 2015). While the availability and accessibility of data

support the method we chose may be presupposed, the selection of other methods, and the performance (in our case publication count) can only act as a proxy for impact (Wildgaard et al. 2014). Impact, or importance, of tree species from this analysis is attributable across North America. However, some of the lower-ranked trees may have a significant local impact, and there may be a relationship between species location and research facilities. For example, some of the lowest ranked species naturally occur in interior xeric areas; these areas are generally sparsely inhabited, contain few universities, and therefore few local researchers. By comparison, another species may also have a small native range but located in a more mesic, higher populated areas that contain more universities. This would facilitate a high likelihood of study compared to the xeric tree, and essentially amounts to a form of sampling bias.

Further, some species that were highly ranked and not included in the 1990 SNA may benefit from uses not realized before 1985. Examples include *Taxus brevifolia* (taxol), mangrove (preventing coastal erosion), and mesquite (grilling). These species highlight the fact that the SNA, like all science, was directed by needs and interests at the time it was produced. Therefore, consideration of the impact of emerging tree species, or trending species uses and locational biases, should be incorporated into a discussion of the trees and be included in a potential third installment and update of this resource.

While the analysis we conducted is not perfect, it does provide a relative indication of the amount of interest applied to each tree species since 1900 and since 1985. The amount of literature available for many tree species is quite extensive. By comparison, in the collection of conifers described in the previous SNA, the number of references attributed to coniferous species accounts ranged from 10 to 195 (average = 48.9, standard deviation = 29.7). In fact, the species with the lowest number of references cited (10) in the previous SNA (*Torreya taxifolia*) only included three references that were specific to that species. Therefore, the volume of published works since 1985 should inform important updates of

Table 5. Top ten ranked coniferous tree species (of 128 analyzed) that were not included in the 1990 version of *Silvics of North America*, through returns of queried publications in Web of Science, AGRICOLA, and CAB Abstracts (27 January –10 February 2020).

Rank	Genus and species	Common name
15	<i>Juniperus communis</i>	Common juniper
26	<i>Pinus caribaea</i>	Caribbean pine
32	<i>Pinus patula</i>	Mexican weeping pine
35	<i>Pinus attenuata</i>	Knobcone pine
46	<i>Pinus washoensis</i>	Washoe pine
52	<i>Juniperus osteosperma</i>	Utah juniper
53	<i>Cupressus arizonica</i>	Arizona cypress
57	<i>Pinus oocarpa</i>	Ocote pine
59	<i>Juniperus monosperma</i>	Oneseed juniper
61	<i>Juniperus ashei</i>	Ashe's juniper

Notes. Ranks are based on an average, using two periods of time (1900–2020, 1985–2020) and three bibliographic databases (Web of Science, AGRICOLA, and CAB Abstracts). Genus and species denote the main genus and species. The query may have included synonyms and other versions of the genus and species. The common name denotes one of the foremost common names for this tree species. A tree species may potentially be referenced by several other common names.

Table 6. Top ten ranked deciduous and other tree species (of 568 analyzed) that were not included in the 1990 version of *Silvics of North America*, through returns of queried publications in Web of Science, AGRICOLA, and CAB Abstracts (27 January –10 February 2020).

Rank	Genus and species	Common name
15	<i>Populus fremontii</i>	Fremont cottonwood
18	<i>Artemisia tridentata</i>	Big sagebrush
24	<i>Prosopis juliflora</i>	Mesquite
26	<i>Rhizophora mangle</i>	Mangrove
27	<i>Prosopis glandulosa</i>	Honey mesquite
32	<i>Asimina triloba</i>	Pawpaw
37	<i>Avicennia germinans</i>	Black mangrove
38	<i>Castanea dentata</i>	American chestnut
41	<i>Lysiloma latisiliquum</i>	False tamarind
42	<i>Yucca mohavensis</i> [†]	Mohave yucca

Notes. Ranks are based on an average, using two periods of time (1900–2020, 1985–2020) and three bibliographic databases (Web of Science, AGRICOLA, and CAB Abstracts). Genus and species denote the main genus and species. The query may have included synonyms and other versions of the genus and species. The common name denotes one of the foremost common names for this tree species. A tree species may potentially be referenced by several other common names.

[†] Plants with a tree-like growth habit that were not gymnosperms were placed in this group.

tree species life histories. Interestingly, in our analysis there are also several instances (e.g., 47 coniferous species) where the number of references cited in the previous SNA was greater than the number of query returns from the period 1900–1985 using Web of Science. By comparison, this only occurred with 15 coniferous species when using AGRICOLA, and four coniferous species when using CAB Abstracts. However, these types of comparisons are not clean enough to comment on the completeness of the bibliographic databases. Species life histories from the previous SNA included references to personal communications, books, and other broad resources (e.g., *Insects and Diseases of Trees in the South* [U.S. Forest Service 1972]) that likely would not have been located during a query search of the three bibliographic databases for a specific tree species.

While useful in providing an overview of the importance of a subject, bibliometric analyses such as those conducted here do not adequately describe issues related to scientific impact and research quality (Milfont et al. 2019). An examination of the scientific value of publications would further help inform the discussion (Cole and Eales 1917). A tree species with several hundred small ephemeral papers would seem more important using our methods than a tree species with only a handful of publications of great importance. Counts of publications, which assume equal value among the entire set (Olson and Bae 2019), might correctly infer popularity yet incorrectly conclude importance concerning economic, environmental, or social value. Therefore, one drawback to the approach conducted here is that the significance of each paper was not assessed, which could be addressed by investigating the relationships among the references contained within each paper.

Publication counts can be used as an indicator of research output. Still, they can be a subject of concern since an analysis of publication counts ignores the degree of knowledge provided through research

articles, the quality and size of accompanying data sets, and incentives that may have been employed to promote academic publication (White 2019). For the citation of individual works, some have also shown that papers with no citations were no different in importance than those cited a few times. However, highly cited papers were deemed significantly more important (Albert et al. 1991). As with other studies that involved a significant amount of literature located during keyword queries, a more in-depth content analysis was not conducted, nor was any effort expended to find unpublished studies, which may bias the importance of a subject (Zhao et al. 2018). Had the queries returned fewer results, we may have been able to conduct a complete assessment of the relevancy of each to each of the tree species. Still, as it stands, our analysis is limited in this respect, similar to other analyses (e.g., Thornley et al. 2011).

One of the purposes of this investigation was to assess the amount of information available for updating the SNA. The result of our keyword searches resulted in a large volume of query returns. Some studies determined the importance of individual papers based on their relationship to other literature (e.g., Nakatoh et al. 2016). Perhaps a sophisticated search scheme (query) using constructs such as citation graphs could be developed to ignore less essential papers and only report literature that is deemed relevant to the subject. On a positive note, the approach we took and the results we found indicate the likelihood of finding new knowledge on the subject of individual tree species. Regardless of whether this knowledge is more or less valuable or impactful, the magnitude of scientific publications may be important when deciding which new species should be included in new editions of works such as SNA. We also made the methodological decision of using a set of specific databases for data mining, which primarily include journal articles and reports, perhaps at the expense of counts of other publication types. In this respect, the rankings we developed are observations of importance based on the bibliographic databases chosen and may be subject to change if other sources of information are used.

Most of the publications returned through queries of tree species were from 1985 forward. This may be related to the availability of digital versions of published works, and the pace of digitizing older published works. One analysis of older agricultural experiment station reports that involved digitizing the content, conducting quality control assessments, and making these available in an online bibliographic service, suggested that each publication required over one hour of time to convert from print to digital format (McGeachin 2018). Therefore, organizations that own or manage older collections certainly must consider the time and cost of making these available via the Internet. Consequently, it may be reasonable to assume that access to some of these collections may be restricted to the printed versions. Nearly 95% of the Web of Science returns were from 1985 forward, which is interesting as others have suggested that the Web of Science may be of high value for locating research published before 1996 (Choudhri et al. 2015). Given our findings, AGRICOLA and CAB Abstracts may be of higher importance in finding older research publications, even though CAB Abstracts databases contain works that extend only as far back as 1973. Finally, the queries related to each tree species were made in a manner where the genus and species appeared as a topic of the paper. Given the thousands of publications identified during these queries, it was impossible within the time frame of the work to differentiate those publications that were intimately associated with studies involving the ecology and management of a specific tree species, and those publications that simply briefly mentioned certain tree species in discussions of the gaps or status quo of scientific knowledge.

Conclusions

This study provides a keyword (genus and species) search of 696 North American tree-like plant species and provides the first quantitative analysis of the importance that each species has acquired through published works. Findings suggest that the choice of search engine can impact the amount and recentness of the comparative literature for use in the development, analysis, and discussion of biological research. From our case study, the results indicate that there is temporal and extent search bias between the three search engines we employed, with the Web of Science queries having the highest percentage of results in their database from the last four decades than the other two. The results also suggest that AGRICOLA and CAB Abstracts may be of higher value in finding older (before 1985) research publications compared to the Web of Science. While the impetus for each published work, whether economic, ecological, or social, is unclear, the case study results indicate the relative importance of each tree species, and therefore may inform discussions focused on which tree species to include in broad reference materials that focus on trees of North America. Further, the reasons why some lower-ranked species were included in the 1990 version of SNA and why some higher-ranked species were not are unclear. A broader discussion of the socioeconomic and ecological criteria for inclusion in this valuable resource, therefore, seems necessary. And finally, the choice of search engine can impact the amount and recentness of the comparative literature for use in the development, analysis, and discussion of biological research, as the variation in scholarly results through search queries suggests.

Acknowledgments

We would like to thank the Southern Research Station, U.S. Forest Service for funding this research (Agreement number: 19-CS-11330110-068).

Literature Cited

- Albert, M. B., D. Avery, F. Narin, and P. McAllister. 1991. Direct validation of citation counts as indicators of industrially important patents. *Research Policy* 20:251–259.
- Bettinger, P. 2020. Tree species rankings based on publication rate (North America). Dryad, Dataset, <https://datadryad.org/stash/share/DOyKIIKhzaIVGLcTsHfKRwVERo9EIbepDXZTerN8LqY>
- Bettinger, P., and W. Chung. 2004. The key literature of, and trends in, forest-level management planning in North America, 1950–2001. *International Forestry Review* 6:40–50.
- Briganti, M., C. D. Delnevo, L. Brown, S. E. Hastings, and M. B. Steinberg. 2019. Bibliometric analysis of electronic cigarette publications: 2003–2018. *International Journal of Environmental Research and Public Health* 16:Article 320.
- Burns, R. M., and B. H. Honkala. (technical coordinators). 1990. *Silvics of North America*. USDA Forest Service Agriculture Handbook 654.
- Centre for Agriculture and Bioscience International. 2019. CAB Abstracts. Centre for Agriculture and Bioscience International. <https://www.cabi.org/publishing-products/cab-abstracts/> [accessed 16 April 2020].
- Choudhri, A. F., A. Siddiqui, N. R. Khan, and H. L. Cohen. 2015. Understanding bibliometric parameters and analysis. *Radiographics* 35:736–746.
- Clarivate. 2020. Web of science core collection. Clarivate, Philadelphia, Pennsylvania, USA. <https://clarivate.com/webofsciencegroup/solutions/web-of-science-core-collection/> [accessed 16 April 2020].

-
- Cohen, C., J. A. R. Vianna, L. R. Battistella, and E. Massad. 2008. Time variation of some selected topics in bioethical publications. *Journal of Medical Ethics* 34:81–84.
- Cole, F. J., and N. B. Eales. 1917. The history of comparative anatomy: part I. -A statistical analysis of the literature. *Science Progress* 11, 578–596.
- Elsevier Analytical Services. 2019. Alzheimer’s disease research insights: impacts, trends, opportunities [online]. Available from https://www.elsevier.com/___data/assets/pdf_file/0020/910910/Alzheimers-research-insights_report_RL_AS_WEB.pdf [accessed 8 July 2020]
- Fowells, H. A. (comp.). 1965. *Silvics of forest trees of the United States*. USDA Forest Service Agricultural Handbook No. 271.
- Hood, W. W., and C. S. Wilson. 2001. The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics* 52:291–314.
- Hung, B. T., et al. 2015. Research trends in evidence-based medicine: a joinpoint regression analysis of more than 50 years of publication data. *PLoS One* 10:e0121054.
- Klingelhöfer, D., Y. Zhu, M. Braun, M. H. K. Bendels, D. Brügmann, and D. A. Groneberg. 2018. Aflatoxin – Publication analysis of a global health threat. *Food Control* 89:280–290. <https://doi.org/10.1016/j.foodcont.2018.02.017>
- von Korff, M., T. Fink, and T. Sander. 2017. A new relevance estimator for the compilation and visualization of disease patterns and potential drug targets. *Pacific Symposium on Biocomputing*, World Scientific Publishing Company, Singapore. Volume 22. Pages 312–323.
- Maassen, S. 2016. Bibliometric analysis of research on wastewater irrigation during 1991–2014. *Irrigation and Drainage*. 65:644–653.
- Martínez-Gómez, A. 2015. Bibliometrics as a tool to map uncharted territory: a study on non-professional interpreting. *Perspectives: Studies in Translatology* 23:205–222.
- McGeachin, R. 2018. Time costs to digitize print agricultural serial publications. *Journal of Agricultural & Food Information* 19:116–120.
- Milfont, T. L., K. Amirbagheri, E. Hermanns, and J. M. Merigó. 2019. Celebrating half a century of *Environment and Behavior*: a bibliometric review. *Environment and Behavior* 51:469–501.
- Mørk, S., S. Pletscher-Frankild, A. P. Caro, J. Gorodkin, and L. J. Jensen. 2014. Protein-driven inference of miRNA-disease associations. *Bioinformatics* 30:392–397.
- Nakatoh, T., H. Nakanishi, T. Minami, K. Baba, and S. Hirokawa. 2016. A visual citation search engine. *In Human interface and the management of information: information, design and interaction*. Edited S. Yamamoto. *Lecture Notes in Computer Science*, Springer International Publishing, Cham, Switzerland. LNCS 9734:168–178.
- Olson, N., and J. Bae. 2019. Biosensors—publication trends and knowledge domain visualization. *Sensors* 19:Article 2615.
- Pardey, P. G., and J. E. Christian. 2002. The production and diffusion of policy knowledge: a bibliometric evaluation of the International Food Policy Research Institute. *International Food Policy Research Institute. Impact Assessment Discussion Paper No. 14*.
- Sargent, C. S. 1890. *The silva of North America. A description of the trees which grow naturally in North America exclusive of Mexico. Volume I. Magnoliaceae–Ilicineae*. Houghton, Mifflin and Company, Boston, Massachusetts, USA.
- Sargent, C. S. 1902. *The silva of North America. A description of the trees which grow naturally in North America exclusive of Mexico. Supplement Volume XIV. Caricaceae–Coniferae*. Houghton, Mifflin and Company, Boston, Massachusetts, USA.

- Schöffel, N., F. Domnitz, D. Brüggmann, D. Klingelhöfer, M. H. K. Bendels, and D. A. Groneberg. 2016. Eine kritische betrachtung der globalen forschungsaktivität zum thema harnblasenkrebs. *Der Urologe* 55:1470–1480.
- Shapiro, F. R. 1992. Origins of bibliometrics, citation indexing, and citation analysis: the neglected legal literature. *Journal of the American Society for Information Science*. 43:337–339.
- Sumner, S., G. Law, and A. Cini. 2018. Why we love bees and hate wasps. *Ecological Entomology* 43:836–846.
- Thornley, C. V., A. C. Johnson, A. F. Smeaton, and H. Lee. 2011. The scholarly impact of TRECVID (2003–2009). *Journal of the American Society for Information Science and Technology* 62:613–627.
- U.S. Department of Agriculture. 2013. About AGRICOLA [online]. Available from <https://agricola.nal.usda.gov/help/aboutagricola.html> [accessed 16 April 2020].
- U.S. Forest Service. 1972. Insects and diseases of trees in the South. U.S. Department of Agriculture, Forest Service, Southeastern Area State and Private Forestry, Forest Pest Management Group, Atlanta, Georgia, USA.
- Waltman, L., and N. J. van Eck. 2012. A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology* 63:2378–2392.
- Wang, H., M. Liu, S. Hong, and Y. Zhuang. 2013. A historical review and bibliometric analysis of GPS research from 1991–2010. *Scientometrics* 95:35–44.
- White, K. 2019. Publications output: U.S. trends and international comparisons. National Center for Science and Engineering Statistics, National Science Foundation, National Science Board Science and Engineering Indicators 2020, NSB-2020-6.
- Wildgaard, L., J. W. Schneider, and B. Larsen. 2014. A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics* 101:125–158.
- Willyard, A., et al. 2017. *Pinus ponderosa*: A checkered past obscured four species. *American Journal of Botany*. 104:161–181.
- Zhao, J., X. Liu, W. Zhang, Y. Xing, S. W. Cho, and Y. Hao. 2018. Evidence-based nursing outputs and hot spot analysis of the last 5 years in mainland China: results of a bibliometric analysis. *International Journal of Nursing Practice* 24:e12628.