ECOGRAPHY

Review and synthesis

Text-analysis reveals taxonomic and geographic disparities in animal pollination literature

Joseph W. Millard, Robin Freeman and Tim Newbold

J. W. Millard (https://orcid.org/0000-0002-3025-3565) ✉ (joseph.millard.17@ucl.ac.uk) and T. Newbold (https://orcid.org/0000-0001-7361-0051), Dept of Genetics, Evolution and Environment, Univ. College London, London, UK. – R. Freeman and JWM, Inst. of Zoology, Zoological Society of London, London, UK.

Ecography
42: 1–16, 2019
doi: 10.1111/ecog.04532

Subject Editor: Bo Dalsgaard
Editor-in-Chief: Miguel Araújo
Accepted 17 September 2019

Ecological systematic reviews and meta-analyses have significantly increased our understanding of global biodiversity decline. However, for some ecological groups, incomplete and biased datasets have hindered our ability to construct robust, predictive models. One such group consists of the animal pollinators. Approximately 88% of wild plant species are thought to be pollinated by animals, with an estimated annual value of $230–410 billion dollars. Here we apply text-analysis to quantify the taxonomic and geographical distribution of the animal pollinator literature, both temporally and spatially. We show that the publication of pollinator literature increased rapidly in the 1980s and 1990s. Taxonomically, we show that the distribution of pollinator literature is concentrated in the honey bees (*Apis*) and bumble bees (*Bombus*), and geographically in North America and Europe. At least 25% of pollination-related abstracts mention a species of honey bee and at least 20% a species of bumble bee, and approximately 46% of abstracts are focussed on either North America (32%) or Europe (14%). Although these results indicate strong taxonomic and geographic biases in the pollinator literature, a large number of studies outside North America and Europe do exist. We then discuss how text-analysis could be used to shorten the literature search for ecological systematic reviews and meta-analyses, and to address more applied questions related to pollinator biodiversity, such as the identification of likely interacting plant–pollinator pairs and the number of pollinating species.

Keywords: animal pollination, ecological systematic review, global biodiversity, named-entity recognition, pollination ecology, text-analysis

Introduction

The number of publications and journals in the academic sciences is vast and continuing to increase (Ferreira et al. 2015). The field of ecology and biodiversity is no exception. Between 1990 and 2014, the total number of ecological research articles increased more than tenfold, from fewer than 10 000 in 1989 to at least 125 000 in 2014 (Nunez-Mir et al. 2016). In conjunction with this increase, digitisation of the literature, indexing tools (such as Scopus, Web of Science and Google Scholar), and the research structures of systematic review and meta-analysis have all become standard

NORDIC SOCIETY OIKOS

www.ecography.org

practice (Lortie 2014, Gurevitch et al. 2018). Understanding of global biodiversity decline in particular has reaped the benefits of these changes (Loh et al. 2005, Butchart et al. 2010, Pereira et al. 2010, Tittensor et al. 2014, Newbold et al. 2015). However, for some important ecological and taxonomic groups, incomplete and biased datasets have hindered our ability to construct robust, predictive models (De Palma et al. 2016, Bartomeus et al. 2018). One such ecological group consists of the animal pollinators, animals that act as a vector for the transfer of pollen from the male to the female reproductive parts of a flowering plant, causing fertilisation and the production of a fruit and seed (Proctor et al. 1996). Animal pollination is highly important, especially in tropical humid and warm environments where approximately 95% of flowering plant species are animal pollinated (Rech et al. 2016). Globally, approximately 88% of wild plant species are thought to be pollinated by animals (Ollerton et al. 2011), providing an ecosystem service valued at $230–410 billion dollars per annum (Lautenbach et al. 2012). Although disputed by some on the basis of taxonomic and geographic biases in the data used (Ghazoul 2005, 2015), many papers have suggested that pollinators are declining in the face of several environmental pressures (Biesmeijer et al. 2006, Steffan-Dewenter and Westphal 2007, Potts et al. 2010b, Winfree et al. 2011, Goulson et al. 2015, Woodcock et al. 2016).

IPBES (2017) summarised the anthropogenic threats to pollinators as change in land cover, chemical application (pesticides, fertilisers, herbicides and fungicides), disease, pollinator management, the introduction of invasive species and climate change. Through the interacting effect of these threats, populations of wild invertebrate pollinators have declined (Ollerton 2017 for a summary of the evidence for pollinator decline), although we know little about the status of wild pollinators outside North America and Europe (IPBES 2017). In Britain and the Netherlands, wild bee species richness has declined over ~50–70% of the total land area (Biesmeijer et al. 2006). Bumble bee declines are some of the best studied, with at least 3 bumble bee species having gone extinct in the UK, and at least 4 species across 11 European countries (Kosior et al. 2007, Goulson et al. 2008). Regional colony losses in European (1985–2005) and USA (1947–2005) honey bees have also been well documented (Stokstad 2007, Potts et al. 2010a), at 25% and 60% respectively, despite a global increase in managed colonies (Aizen and Harder 2009).

Predictive models are important in understanding pollinator biodiversity change, but are a challenge to implement robustly. This difficulty is in part driven by the geographical and taxonomic distribution of available pollinator biodiversity data (De Palma et al. 2016). Pollinators are represented across a variety of taxonomic groups, including bats, birds and multiple insect taxa, but many of the key syntheses of pollinator decline have been restricted to the bees of North America and Europe (Winfree et al. 2011, Ghazoul 2015, Goulson et al. 2015, De Palma et al. 2016), but see Regan et al. (2015) for a global study on the status of mammal and bird pollinators. Although widely accepted, the degree of this bias and the extent to which it might influence biodiversity models is uncertain (Ghazoul 2005, 2015, De Palma et al. 2016, Ollerton 2017). Some studies have made progress towards quantifying the geographical or taxonomic distribution of the animal pollination literature (Archer et al. 2014, Ollerton 2017), but the way in which taxonomy interacts with spatial distribution globally has not to our knowledge been the subject of a thorough review. This lack of research is in part a symptom of article indexing tools, which despite their contribution, still have significant limitations, a problem not confined to the animal pollination literature (Westgate et al. 2015, 2018, Westgate and Lindenmayer 2016). Indexing search tools such as Scopus do have functions to account for differences in spelling (fuzzy-matching), and variable suffixes for the same family of words (stemming), but searching for geographical and taxonomic names and identifying overall text topic, is only possible through discrete search terms and phrases. As a result, returning literature fully representative of a particular theme, geographical region, or taxonomic group is difficult to accomplish, given the semantic ambiguity of search terms across academic fields (Westgate and Lindenmayer 2016, Roll et al. 2017). In the context of the animal pollination literature, better tools for extracting pollinator information could be used as the basis for more taxonomically and geographically representative meta-analyses and systematic reviews, in turn increasing the robustness of synthetic analyses.

We highlight here two text-analysis tools (taxonomic and geographical entity extraction), which could be used to search more efficiently for literature on animal pollination. We then demonstrate one application of these tools in a systematic review: the quantification of the taxonomic and geographical distribution of the animal pollination literature. This analysis builds on the reviews of Archer et al. (2014) and Ollerton (2017), introducing new text-analysis methods, examining temporal changes in pollinator publications and investigating the interaction between the taxonomic and geographical distributions of pollinator studies. Finally, we discuss how these tools fit within the debate around the robustness of pollinator meta-analyses and systematic reviews, proposing that biases could be mitigated by making searches of the literature more efficient. We summarise by emphasising two different although related points: firstly, the geographical and taxonomic distribution of the animal pollination literature is indeed highly concentrated in North America and Europe in the honey bees and bumble bees, although many studies do exist for other species and geographic regions; and secondly, the development of text-analysis tools shows significant promise in optimizing the search for information on animal pollination, both through capturing data on underrepresented regions and taxa, and through speeding up the search process.

CLIFF-CLAVIN and Taxize: a brief overview of geographic and taxonomic entity extraction tools in ecology

Text-analysis could help to mitigate the problem of biased and incomplete pollinator response data. Also often called text-mining, text-analysis refers to the automated extraction of information from large volumes of text (Cohen and Hunter 2008), most notably across multiple documents (Griffiths and Steyvers 2004, Grimmer and Stewart 2013, Westgate et al. 2015). Given the very large numbers of published papers containing potentially useful information, such technologies are invaluable in automatically drawing together results across lots of studies (Grimmer and Stewart 2013), thereby reducing the duration of the 'synthesis gap', or in other words the lag between the practice of science and the synthesis of evidence (Westgate et al. 2018). Text-analysis tools can be used to optimise the systematic review and meta-analysis literature search path. For example, topic categorisation algorithms can be used to allocate articles automatically to particular fields of study, enabling the curator to discard articles of low relevance (O'Mara-Eves et al. 2015, Westgate 2018). Particularly in the context of pollinator data, the application of such tools could increase recall of the relevant literature and decrease the effort required, in turn reducing data biases in systematic reviews and meta-analyses. Here we discuss two tools that could be particularly useful in the context of pollination ecology: geoparsing and taxonomic entity extraction.

Geoparsing allows place names in text to be identified, resolved and assigned geographical coordinates (Leidner and Lieberman 2011). Geoparsing can therefore be broken into two steps: firstly, the identification of geographical mentions (known as toponyms); and secondly, the resolution of mentions as the most likely physical coordinates (D'Ignazio et al. 2014). The first step is a key obstacle; the problem being semantic (D'Ignazio et al. 2014). Identical words can be used to describe both place and non-place information, interpretable only in the context the term is written (Leidner and Lieberman 2011). For example, the words 'Rio' and 'Alexandria' could be used to describe both a geographic location and the name of a person. High performance machine learning algorithms will therefore attempt to resolve locations through contextual information (Leidner and Lieberman 2011, Gritta et al. 2018). CLIFF-CLAVIN is one such tool (D'Ignazio et al. 2014, Gritta et al. 2018). An open-source geoparser, CLIFF-CLAVIN was developed for extracting geographical information from news articles (D'Ignazio et al. 2014). CLIFF-CLAVIN also has an implementation of focus, meaning it attempts to resolve the primary country location of a given piece of text, even when the country is not mentioned (D'Ignazio et al. 2014). CLIFF-CLAVIN estimates focus on the basis of the most frequently mentioned country, and in the absence of country mentions, the frequency of specific locations within countries (D'Ignazio et al. 2014). CLIFF-CLAVIN will attempt to find geographical locations from the local to continental level. For example, CLIFF-CLAVIN is able to find correctly the records 'Krakatoa', 'Sumatra', 'Indonesia' and 'Asia'. Although still in the early stages of development, and requiring significant improvements in accuracy and speed (Gritta et al. 2018), geoparsers have previously been used to identify the main geographical location of news reports (Imani et al. 2017), to geotag museum specimens (Beaman and Conn 2003), and to digitise historical maps (Chiang 2017).

Taxonomic entity extraction refers to the identification of taxonomic names (in theory of any taxonomic rank) from blocks of text (Sarkar 2007). Such algorithms tend to use taxonomic dictionary string matches, rule-based inference and machine learning, either independently or in combination (Akella et al. 2012). Dictionary match algorithms search for each word (unigram) and pair of words (bigram) in a taxonomic database such as NameBank (Leary et al. 2007), returning a record if the strings match. Similarly, rule-based inference searches for regular expressions indicating a form often associated with a species record, such as bigram capitalisation and abbreviation. Machine learning approaches identify text likely to represent a taxonomic record, inferring from both context and string structure (Akella et al. 2012). The R package 'taxize' has implementations for two of these algorithm types in the function scrapenames: dictionary string match (Taxonfinder) and machine-learning (Neti Neti) (CRAN 2018). Scrapenames will search for strings resembling taxonomic records at any taxonomic rank, including abbreviated records, hybrids and higher taxa. For example, scrapenames is able to correctly find the records '*A. manicatum*', 'Apidae' and '*Viburnum macrocephalum* f. Keteleeri'. Many authors have emphasised the value of extracting taxonomic information (Sarkar 2007, Guralnick and Hill 2009, Parr et al. 2012, Thessen et al. 2012), and others its associated methodological difficulties (Correia et al. 2018), but to our knowledge few studies have explored potential applications.

In the following section we demonstrate how CLIFF-CLAVIN and taxize can be used in combination to quantify the taxonomic and geographical distribution of the animal pollinator literature. These patterns reveal disparities in pollinator literature, reinforcing the problem of biases in the context of pollinator biodiversity modelling. Finally, in exploring future directions, we discuss how these tools may be used to decrease data biases in biodiversity meta-analyses and systematic reviews. We discuss how text-analysis could make the literature search process more efficient, reducing the duration of time required for review preparation, increasing the recall of relevant literature, and enabling the prioritisation of underrepresented taxa and regions.

Quantifying the taxonomic and geographic distribution of the animal pollination literature

We scraped the pollination literature for mentions of animal species and location data to investigate the taxonomic, geographical and temporal distribution of studies on animal pollination. We considered any primary research article

published in English returned through a search for the term 'pollinat*' in Scopus, that mentioned an animal species in the abstract. Animal species scraping and geographical entity extraction were accomplished through a methodology built on the 'taxize' R package and the geoparser CLIFF-CLAVIN. Our rationale for applying this semi-automatic approach, rather than manually checking all abstracts, was that identifying all Latin binomials and geographic locations would not be feasible given the volume of text. We first describe the methodology applied (Supplementary material Appendix 1 for additional validation), before discussing change over time, the taxonomic breakdown of the animal pollinator literature, overall geographical distribution of information, and finally geographical distribution for individual taxonomic groups.

Taxonomic extraction

We queried Scopus using the stemmed term 'pollinat*' (28/03/18–29/03/18), before subsetting for primary research articles in English (Supplementary material Appendix 1 Fig. A1). Duplicated records were filtered out by removing duplicated titles. Any records without titles were also removed. We then retained any papers with abstracts that mentioned a taxonomic name, applying in conjunction both the Neti Neti and Taxonfinder algorithms implemented in the package 'taxize' (Supplementary material Appendix 1 Fig. A1). Taxonfinder represents a dictionary match algorithm, searching against multiple dictionaries for potential taxonomic records. The Neti Neti algorithm applies a machine learning approach to extract strings deemed likely to be taxonomic records.

We then carried out a series of data-cleaning steps to identify Latin binomial animal species within our initial scrape. We chose to use the Catalogue of Life (COL) in validating species records as animal species, due to its greater coverage (84% of all described species: Roskov et al. 2017). Animal species were validated by performing character string matches against the Latin binomials of a Metazoan subset of the COL (Supplementary material Appendix 1 Fig. A1). Matches with the COL were carried out at two levels (1 and 2). Level 1 represented direct string matches, and level 2 any matches with an abbreviated record. Within level 1, we distinguished a series of sub-levels, reflecting the approach used to resolve a record as an animal species: level 1a represents any direct match with the original string, level 1b any direct match following the removal of punctuation, and level 1c any direct match following the removal of punctuation and the string 'spp'. An abbreviated record refers to an abbreviated genus and full species picked up by taxize. For example, *Apis mellifera* would be abbreviated in the form *A. mellifera*. We also encountered problems regarding accepted and synonymous species names. In attempting to mitigate this issue, we substituted any record picked up by the COL as a synonym with its corresponding accepted name. For any further analysis, we then worked only from accepted names.

For all studies related to pollination that mention an animal species, we initially calculated the frequency of taxonomic mentions at the level of genera (Supplementary material Appendix 1 Fig. A2). We opted to investigate the frequency of mentions at the level of genera given the hypothesized dominance of studies on *Apis* and *Bombus* species. We included only those genus names associated with a Latin binomial, given the increased ambiguity with just genus names. For example, *Prunella* is both a genus of plants and a genus of birds. We presented taxonomic mentions as a raw count rather than a proportion since mentions of different genera are not necessarily independent of one another: one study abstract may mention multiple genera.

We also investigated change over time in mentions of the animal genera *Apis*, *Bombus*, and all other pollination-related animal genera (henceforth 'other genera'). We opted to cluster all other genera to test the hypothesis that publications concerning *Apis* and *Bombus* species are largely responsible for the rapid increase in pollination-related papers.

Geoparsing

Following the verification of animal species, we then anchored each abstract mentioning an animal species to a geographic location, using an approach called geoparsing. Geoparsing refers to the resolution of ambiguous free-text place name descriptions as specific geographic coordinates. Not all abstracts will mention a location, but we assume that those that did were representative of the geographic distribution of the animal pollination literature as a whole.

We chose to use the open-source geoparser CLIFF-CLAVIN, due to the high accuracy of its focus implementation, relative to commercial geoparsers such as Yahoo Placespotter and OpenCalais. The main focal country for a given text will henceforth be referred to as 'major' mentions, and any specific locations found in an abstract as 'minor' mentions. 'Minor' mentions can therefore be of any geographical scale, from the continental to the local level. We used vagrant – a software tool for leveraging virtual environments – and the GitHub repository CLIFF-up to host CLIFF-CLAVIN (<https://github.com/ahalterman/CLIFF-up>).

After geoparsing the pollination-related abstracts, we carried out a series of verification steps to improve the quality of the data. First, we plotted the data on a global map to check for any unusual-looking patterns, which revealed that continental 'minor' mentions were distorting the apparent geographic distribution of studies. For example, the continental 'minor' mention 'Europe' appeared in a number of abstracts, which CLIFF-CLAVIN had assigned to a single coordinate in central Europe. We were also not interested in oceanic 'minor' mentions, since these would not relate to the study of terrestrial animal pollinators. Before proceeding with any further analysis, we therefore removed any continental or oceanic 'minor' mentions. Second, in initial runs of CLIFF-CLAVIN we also noticed that the geoparser was picking up geographic information associated with copyright details, typically included at the end of the Scopus abstract following a copyright symbol. We therefore removed any characters following the copyright, before rerunning the geoparser.

Third, after removing any low-resolution geographic locations, we then visually inspected the whole raw dataset, searching for any place names that were either questionable or clearly wrong. For example, place names that had been incorrectly disambiguated by CLIFF-CLAVIN, such as 'Ivory' and 'Hay Meadows', as well as locations that seem overly specific or strange, such as 'Blue Ridge Parkway Milepost 234 ca 886 m'. Fourth, given CLIFF-CLAVIN was originally trained on news articles, we were aware that performance might be reduced when applied to academic texts, particularly those mentioning Latin binomial species names. For example, the 'Linnaeus Terrace', a rock terrace in the Antarctic, was incorrectly identified from Linnaeus, the species authority, while taxonomic names such as 'Peia' and 'Pavonia' were also incorrectly identified as place names. We therefore also manually inspected any locations that could have been referred to in text as a species or genera, and removed any that we identified as mistakes. After the removal of text following the copyright symbol, CLIFF-CLAVIN identified geographic locations in 2087/3974 (53%) of the pollination-related abstracts containing an animal species (Supplementary material Appendix 1 Fig. A1). After further verification of the geoparsed data, geographic locations were identified in 2072 abstracts (Supplementary material Appendix 1 Fig. A1), meaning 2072/2087 (~99%) of those abstracts with a location contained a usable sub-continental geographical location.

We calculated a study count through counting the number of 'major' mentions coordinates within each set of country polygons, and a study density by dividing this value by the area of those polygons (Supplementary material Appendix 1 Fig. A3). Beforehand we removed duplicated study-country combinations for both 'minor' and 'major' mentions, accounting for abstracts mentioning a given location more than once. Given focus describes the algorithms attempt to identify the main geographic focus of the text, we reasoned that 'major' mentions would provide an indication of primary study location. Due to the highly right-skewed distribution of the country study counts, we $\log_{10}$ transformed the values. 'Minor' mentions were also plotted onto this map, with the size of each point representing the number of unique study-location combinations.

For abstracts mentioning an animal species and a geographic location, we examined the way in which taxonomy and geographic location interact. We assumed that all geographic locations mentioned within a given abstract related directly to all animal species mentioned within that same abstract. Consistent with our investigation of overall taxonomic distribution, we examined taxonomy-geography interaction for Latin binomial species at the level of genera.

Temporal distribution

Over time, the number of studies on pollination has increased substantially, with a particularly rapid increase beginning in the mid-1990s and 2000s (Fig. 1), occurring in conjunction with widespread incidences of colony collapse disorder (CCD) in the early 2000s (van Engelsdorp et al. 2008, Genersch et al. 2010). Much of this increase can be attributed to studies of *Apis* and *Bombus* species (Fig. 2, Supplementary material Appendix 1 Table A1). From 1980–2017, the number of studies mentioning *Apis* and *Bombus* species increased non-linearly. The rapid increase for *Apis* coincided with the introduction of the parasite *Varroa destructor* to the United States in the 1980s (Oldroyd 1999), and for *Bombus* with the first commercialisation of *Bombus* pollination in the late 1980s (Velthuis and van Doorn 2006). Over this period, the general trend in publication number for other pollination-related genera increased marginally, with a slight non-linear increase from the year 2000. Given that *Apis* and *Bombus* are often referred to by their common name, it is likely that we underestimate the disparity in publication rate between *Apis*, *Bombus*, and other genera ('Limitations' and Supplementary material Appendix 1 Fig. A4).

Taxonomic distribution

Of the abstracts related to pollination mentioning an animal species (3974), the Hymenoptera were overwhelmingly the most frequently mentioned of all taxonomic orders, in approximately 65% of all abstracts (Fig. 3). Of the 13 most frequently mentioned genera, 11 were hymenopteran genera, all but one (*Ceratosolen*) of which were bees. This is to be expected, since bees are regarded as the most important pollinating group (Potts et al. 2010b, Ollerton 2017). Unsurprisingly, *Apis* and *Bombus* figure highly. Approximately 1/4 of abstracts mention a species of honey bee (*Apis*) and 1/5 a species of bumble bee (*Bombus*) (Fig. 3), which is likely an underestimate given we do not consider common names ('Limitations' and Supplementary material Appendix 1 Fig. A4). The disparity for *Apis* and *Bombus* is consistent with more anecdotal descriptions in the literature, describing honey bees and bumble bees as the main study groups (Ghazoul 2015). This taxonomic pattern probably to a large extent reflects commercial value: both *Apis* and *Bombus* are economically important commercial pollinators (Goulson 2003, Klein et al. 2007), with bumble bees in particular providing a unique contribution in the form of buzz-pollination (Goulson 2003). Moreover, 4 out of 13 other top-mentioned genera (*Osmia*, *Megachile*, *Melipona*, *Trigona*) are also managed commercially to some extent, either for pollination services or honey production (Vit et al. 2004, Velthuis and van Doorn 2006, National Research Council 2007).

More generally, the insects are dominant in the pollination literature. Five of the top eight orders are insects (in decreasing order: Hymenoptera, Lepidoptera, Diptera, Coleoptera and Hemiptera), the first four of which are well-known pollinating groups. The greater number of studies mentioning lepidopteran than dipteran species probably reflects a bias in study effort. Although flies are likely the second most important pollinators today – behind only the Hymenoptera (Ssymank et al. 2008) – and evolved as one of the first angiosperm pollinators (Endress 2001), lepidopteran flower-visitors are often deemed more conspicuous and attractive (New 2004), making them likely study

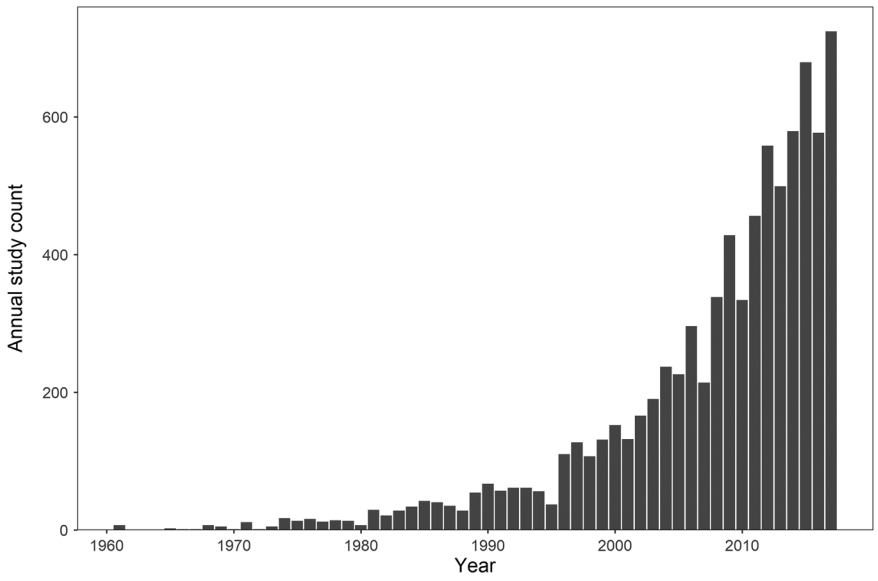

Figure 1. Annual study count for pollination-related studies mentioning an animal species, years 1961–2017.

candidates. The smaller number of studies mentioning beetles is more likely a true reflection of pollination importance. Although beetles are important ecosystem service providers on the whole (Noriega et al. 2018) and evolved as some of the earliest gymnosperm pollinators (Labandeira et al. 2007, Ollerton 2017), modern beetles are widely recognised as less important pollinators relative to the Diptera, Hymenoptera and Lepidoptera (Ollerton 2017). A surprising result was the appearance of the moth *Manduca* in the top 13 genera, given the reputation of the moths as understudied relative

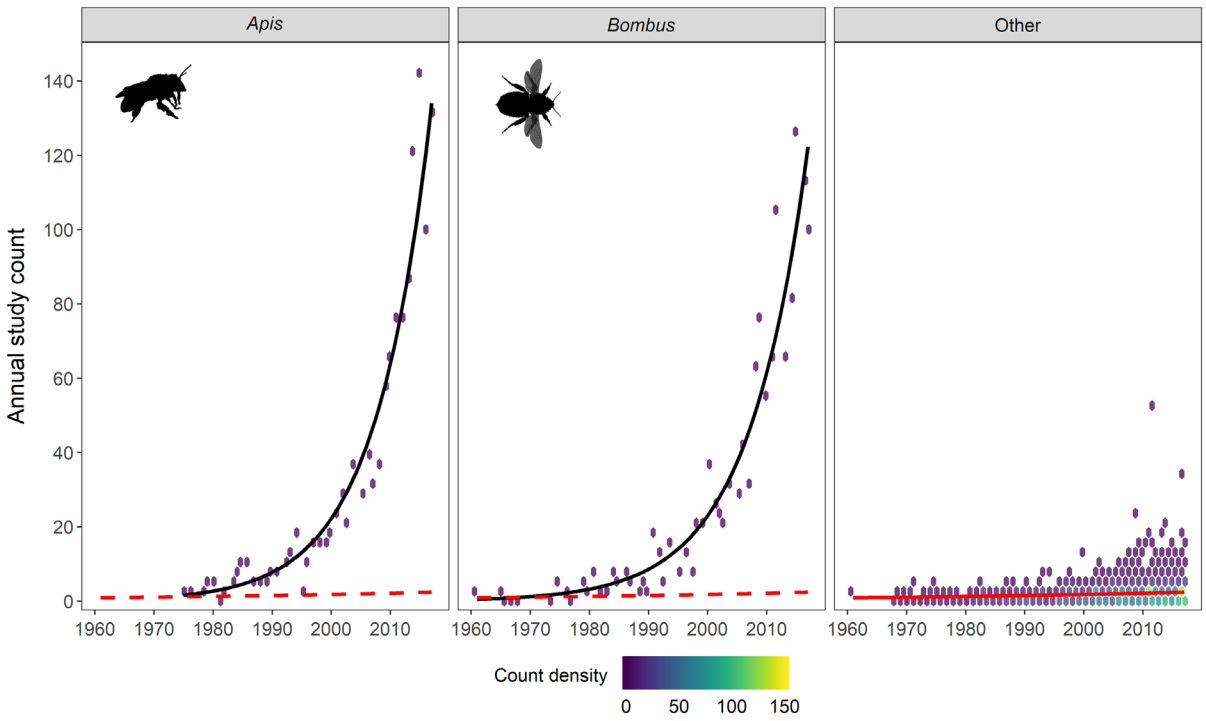

Figure 2. Annual study count for *Apis*, *Bombus*, and all other genera 1961–2017. Black lines represent the fit of a generalized linear model with Poisson errors for *Apis* and *Bombus*, relating study count to year. Red lines represent the output of the same model for all other genera, presented in dotted form in the *Apis* and *Bombus* facets. Counts have been binned as a density to represent multiple counts at the same study-year combination, from dark blue to yellow (150 studies). The model for both *Apis* and *Bombus* deviates from all other pollination-related animal genera in the 1980s, with a rapid and non-linear increase.

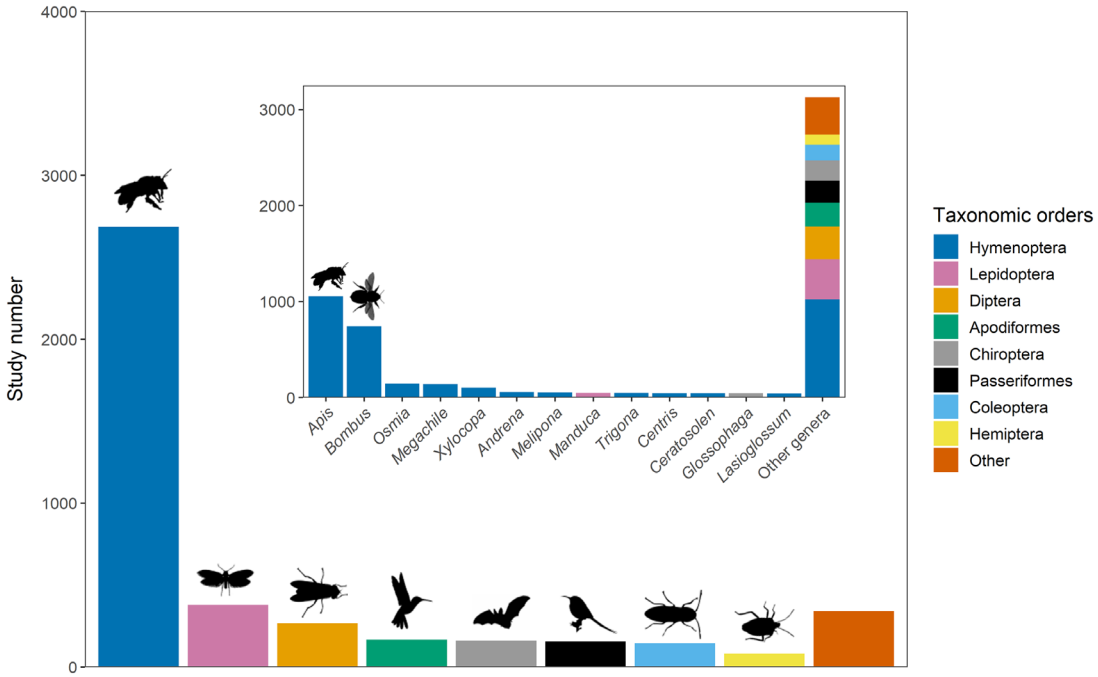

Figure 3. Order-level distribution of animal species in the pollination literature across 3974 studies. Given a study may mention multiple genera or orders, each bar is not independent, meaning study count values will not sum to the total of 3974. 'Other' orders are represented by 55 orders, with 29/55 of these appearing in 2 or fewer abstracts. Inset shows genus-level distribution of animal species in the pollination literature. 'Other genera' are represented by 1000 genera across the Hymenoptera, Lepidoptera, Diptera, Apodiformes, Chiroptera, Passeriformes, Coleoptera, Hemiptera and 'other' orders. Colours are the same in the main panel and the inset.

to other lepidopterans (Hahn and Brühl 2016). Most likely this is an artefact of model taxa rather than pollinator importance. Although *Manduca* species can be agricultural pests and pollinators, with the larval stage feeding on a variety of plant species in the Solanacae family (Kessler and Baldwin 2002), and the adult stage a generalist nectar feeder (Raguso and Willis 2002), mentions are driven by *Manduca sexta*, an important model species for molecular and genetic studies (Riddiford et al. 2003). The hemipterans are represented primarily by aphid genera: ~1/4 of the hemipteran genera in the pollination literature are aphids. Despite being flower visitors, hemipterans more often feed on plant stem sap, making them incidental pollinators (Wardhaugh 2015). Broadly, the extent to which absolute distribution of mentions might predict pollinator importance for the five insect orders is an interesting point. Although the Hymenoptera are likely the most important pollinating order, it seems unlikely that this would be by a factor of ~7 globally. Probably there will be a signal of importance, but confounded by geography and study biases.

Vertebrates are also mentioned relatively frequently in pollination studies. Three vertebrate orders fall in the top eight: two avian, Apodiformes and Passeriformes; and one mammalian, Chiroptera. The Apodiformes are entirely represented by the Trochilidae (hummingbirds), a well-known nectar-feeding (and thus pollinating) family. The Passeriformes are represented more diversely, with approximately 75% of identified species coming from six nectar-feeding families: the sun birds (Nectarinidae), honey-eaters (Meliphagidae), Icteridae, honey-creepers (Thraupidae), white-eyes (Zosteropidae) and sugar birds (Promeropidae) (Proctor et al. 1996). Many species of bats are known to feed on fruit or nectar (Fleming et al. 2009). The bat genus *Glossophaga*, the only vertebrate genus falling in the top 13, is a common lowland nectar-feeding group distributed in central and South America (Fleming et al. 2009).

Interestingly, some groups are associated with pollination through their nature as pollinator parasites and predators. For example, the Mesostigmata (an order containing the *Varroa* mites, 41 abstracts) and Araneae (spiders, 27 abstracts) are primarily parasites and predators respectively. *Varroa* mites parasitize honey bees, and are implicated in colony collapse disorder (van Engelsdorp et al. 2009). Araneae, such as crab spiders, prey on pollinators through hiding on the flower and ambushing at visitation (Dukas and Morse 2003).

Geographical distribution

We also investigated the geographical distribution of the animal pollinator literature, inferred by extracting place names with the geoparser CLIFF-CLAVIN. The top five countries for animal pollination studies are the United States, Brazil, Australia, Canada and China (Fig. 4), together representing ~50% of all studies. Previous systematic reviews of the distribution of pollinator data identified Australia, Brazil, the United States, Germany and Spain, as the top five

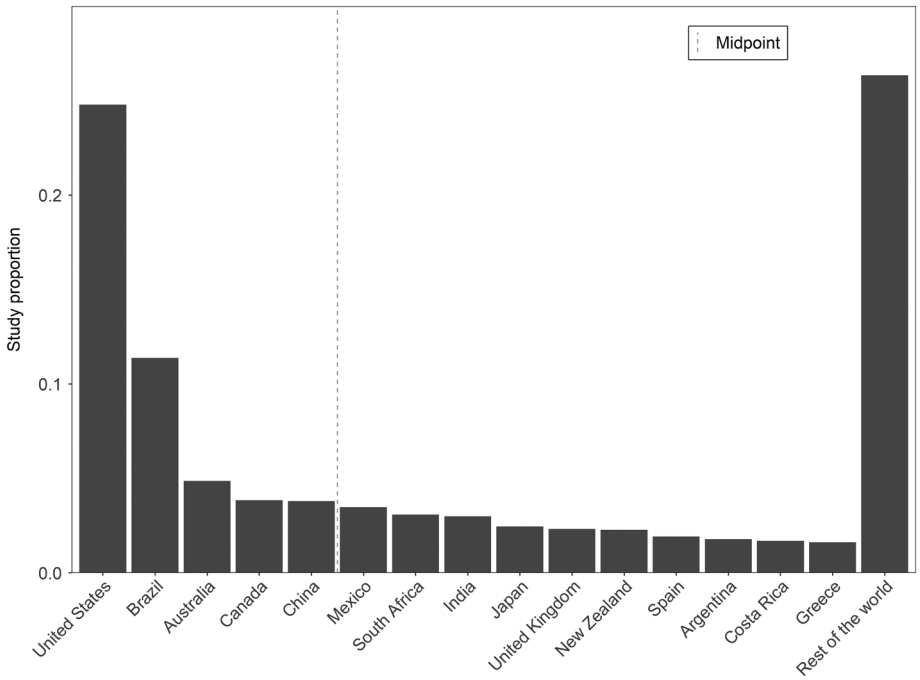

Figure 4. Distribution of the animal pollinator literature among countries. The 'major' focus of each abstract, as resolved by CLIFF-CLAVIN, was used as an indicator of the geographical location likely representing the main study area. The red dotted line represents the study proportion midpoint. Half of all animal pollinator related studies fall in only five countries. 'Rest of the world' is represented by 238 countries.

contributors (Archer et al. 2014). Germany is notable by its absence in our analysis. However, exact character string matches of the term 'Germany' with each of the abstracts indicate that Germany is indeed less strongly associated with the animal pollination literature (Supplementary material Appendix 1 Fig. A8). Potentially lower representation of Germany is explained by the confounding effect of study subject. Archer et al. (2014) found that German studies were frequently represented among studies of pollinator perturbation, but relatively infrequently among general pollination studies. Habitat perturbation studies represented a relatively small percentage of the pollination-related papers analysed here (Supplementary material Appendix 1 Fig. A6), while general pollination studies accounted for a much higher proportion. In general, our results suggest that overall, pollinator information is less restricted to western Europe and North America than was previously thought (Mayer et al. 2011), although we recognise that our analysis likely underestimates geographic disparities (see 'Limitations'). Indeed, only three European countries appear in the top 15 (United Kingdom, Spain and Greece). However, although study count in European countries is relatively low, density is higher since European countries tend to have small areas (Supplementary material Appendix 1 Fig. A3).

The global distribution of study counts reveals geographic disparities in animal pollination literature (Fig. 5). Study counts are particularly low across large regions of Africa, with the exception of Kenya, South Africa and Madagascar. Central Asia is also underrepresented, with no studies returned for any of Afghanistan, Turkmenistan, Uzbekistan, Kazakhstan and Tajikstan. It is probable that some central Asian pollination studies were published in Russian, meaning they were missed in the initial download. Interestingly, and as you might expect, the geographical distribution of animal pollinator-related studies to some extent reflects the global crop production, as shown by the high research effort in eastern Brazil, India, Europe and North America (Potts et al. 2016). Indeed, the largely unproductive region of north Africa has low study density with the exception of the Nile Delta, a fertile region of the Sahara Desert (Elbasiouny et al. 2014).

Interactions between taxonomy and geography

We also investigated the geographical distribution of pollination studies across taxa (Fig. 6). We assumed that a taxonomic record in an abstract was related to any geographical location in the same abstract, and then plotted all unique abstract-species-location combinations (Fig. 6) for the top five taxonomic orders. Our analysis shows clear spatial patterns in pollination studies for different taxa, which probably reflects some combination of the actual distribution of pollinator species, study biases, and methodology-induced biases (see 'Limitations' for more details). Given the inaccuracies inherent in CLIFF-CLAVIN, and the problem in assuming all taxonomic names are associated with all locations in an abstract, our results should be interpreted with caution.

Genus-level study distributions for the Hymenoptera are associated with North America, South America and Europe,

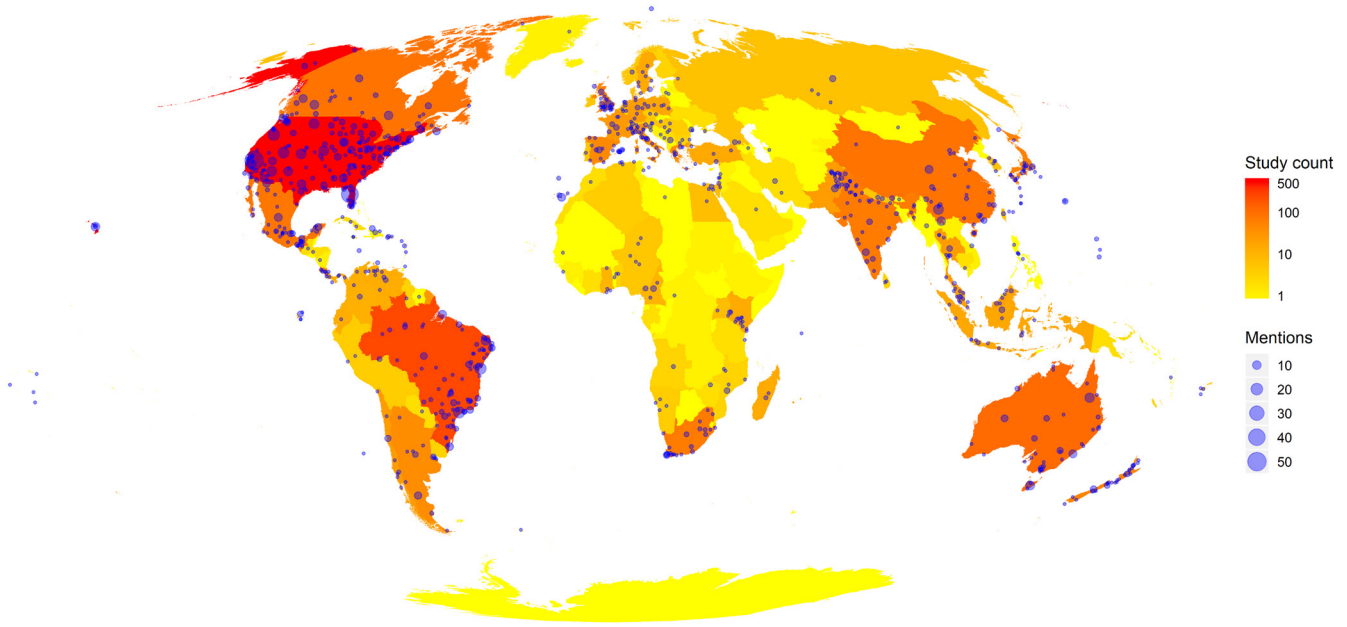

Figure 5. Global study counts for animal pollinator related studies, aggregated at country level. Study counts were derived from the number of abstracts with their 'major' focus in each country. All oceanic and otherwise obviously incorrect mentions, as well as mentions that could only be resolved to a unit larger than a country, were removed. Study counts were $\log_{10}$-transformed. Partially transparent blue points ('minor' mentions) represent the number of unique abstracts in which CLIFF-CLAVIN resolved that location. 'minor' mentions include all specific geographical locations geoparsed by CLIFF-CLAVIN, with the exception of continents, oceans, and incorrectly geoparsed locations.

reflecting some signal of actual distribution, although in our analysis Africa is conspicuous by its absence. Our analysis indicates *Apis* has the largest study distribution, associated with North America, Europe, south and south-east Asia, Australia and eastern South America, with some 'minor' mentions in Kenya, South Africa and Ghana. These results are consistent with the almost-global distribution of *Apis* (Han et al. 2012), an important pollinating genus non-native to large portions of its current range (Whitfield et al. 2006). *Bombus* also appears to be associated with a global study distribution, albeit reduced in Africa, central and south-east Asia and Australia. This is to some extent concordant with the actual distribution of *Bombus* as a genus of the temperate regions, with anthropogenic introductions to New Zealand and Tasmania in the late 1800s and early 1990s respectively (Semmens et al. 1993, Velthuis and van Doorn 2006). However, its association with studies in Africa and Australia is surprising, given it has no known distribution in either region. We found that all three *Bombus* abstracts associated with Africa, and all three with mainland Australia were false positives. Although each mentions a species of *Bombus*, the locations were inaccurate (either being mistakenly georeferenced taxonomic names or incorrect identification of a location). In all other top hymenopteran genera, our results show some signal of actual distribution: *Centris* and *Melipona* are found naturally in the Neotropic and Neartic realms, *Trigona* in the Neotropic and Indo-Australian, *Andrena* and *Osmia* in the Holarctic and North America, and *Megachile* the Western Hemisphere and Palearctic (Michener 2007).

Order-level trends likely indicate spatial patterns of pollinator importance, in part confounded by geographical study biases. For example, study records for dipterans are absent from Brazil – in a region highly populated with hymenopteran studies – but concentrated in Europe and North America. This would suggest some signal of lower ecological importance for dipterans in Brazil relative to hymenopterans. Indeed, previous studies have suggested an opposing latitudinal relationship for dipteran and hymenopteran pollination importance, with fly visitation decreasing at low latitude and hymenopteran increasing at low latitude (Ssymank et al. 2008). Similarly, chiropteran studies are concentrated in Central America, and the Apodiformes in South America, both regions within part of their respective native distribution (Fleming and Muchhala 2008). Some localities however are again conspicuous by their absence. Although hummingbirds went extinct in Africa in the Miocene, nectar-feeding fruit bats do occur in Africa and much of the tropics (Fleming and Muchhala 2008), which appears not to be represented in our analysis. Potentially also our outputs are influenced by the taxonomic spread of vertebrate pollinators in the Old and New World. The New World vertebrate pollinators are more diverse, but this diversity is concentrated in the hummingbirds and leaf-nosed bats, whereas in the Old World diversity is represented across multiple avian orders (Fleming and Muchhala 2008). For example, Old World pollinators include nectar-feeding Psittaciformes and Passeriformes, both of which are not considered major pollinators in the New World (Fleming and Muchhala 2008).

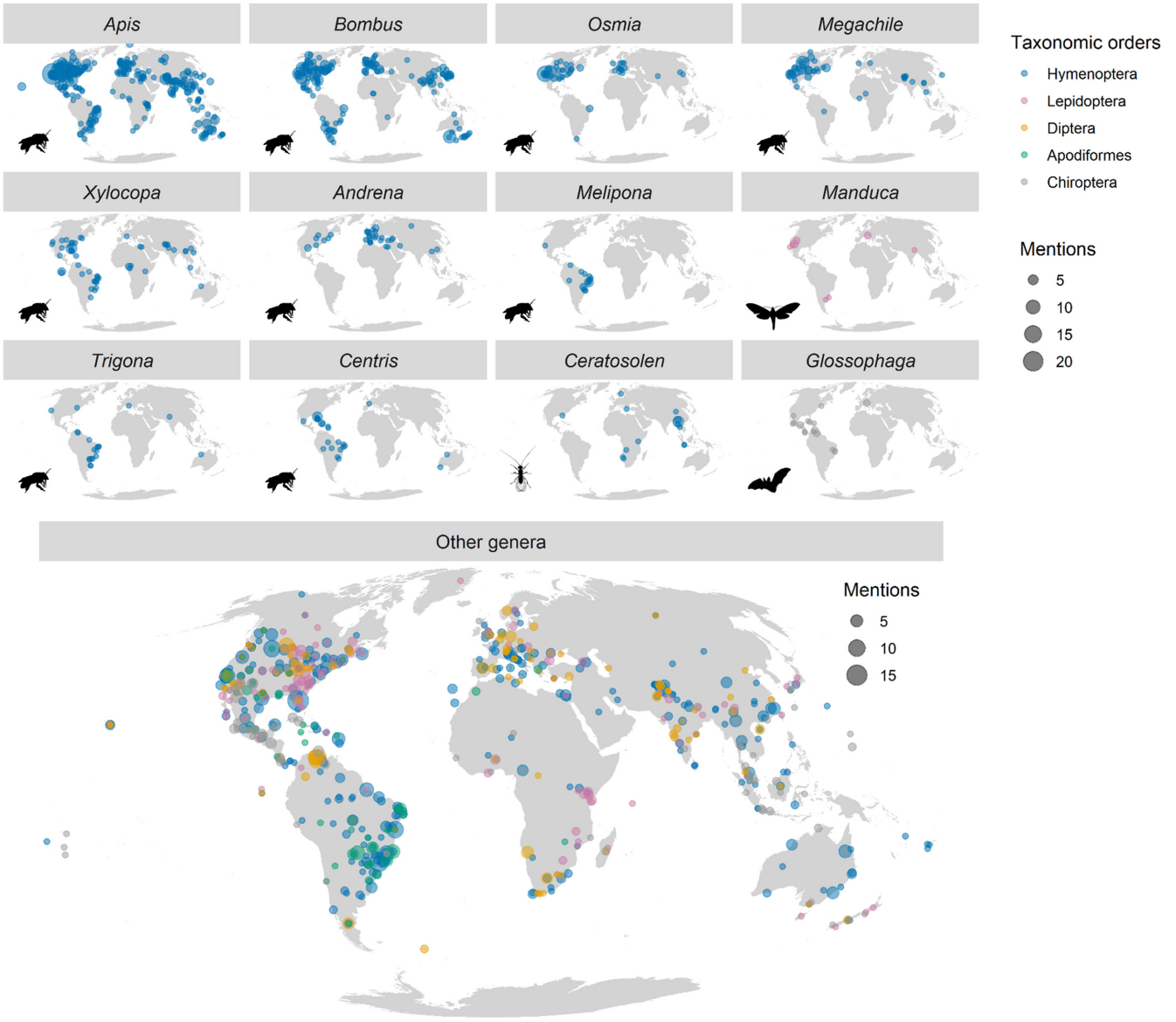

Figure 6. Distribution of the animal pollinator literature broken down into taxonomic groups, for the top 12 genera and top 5 orders. Taxonomic orders are indicated here by fill colour, consistent across top and bottom panels: Hymenoptera (blue), Lepidoptera (pink), Diptera (orange), Apodiformes (green) and Chiroptera (grey). Point size represents the frequency of unique abstract-genera-location combinations. 'Other genera' here consists of 1001 genera across the Hymentopera, Lepidoptera, Diptera, Apodiformes and Chiroptera.

Limitations

Although taxonomic and geographical entity extraction are exciting developments, and show promise in both synthesising systematic review findings and prioritising the search path, these technologies are not without their limitations.

We acknowledge that applying a single academic indexing tool (Scopus) will mean our initial search likely overlooked relevant literature. Notably, relative to indexing tools such as Web of Science (WoS) or Scopus, Google Scholar has been shown to return a greater proportion of the relevant literature for a given search term (Beckmann and von Wehrden 2012). Both Scopus and Web of Science have also been shown to exhibit geographical and language biases (Mongeon and Paul-Hus 2016), which could potentially influence our outputs. However, we reasoned that although an academic indexer might favour some geographical regions, taxonomic groups, or years over another, and Google Scholar would give greater coverage, an academic indexer would return a text corpus we could be confident had undergone peer-review (Mongeon and Paul-Hus 2016). Given the underlying motive – to identify studies more efficiently for systematic review and meta-analysis – we decided that as a proof of concept, there was greater value in excluding the grey literature. In minimising the potential for biases, we opted for Scopus over WoS given its greater coverage (Mongeon and Paul-Hus 2016), and a

single indexer to minimise the potential for duplication, a non-trivial risk in systematic reviews (Rathbone et al. 2015).

Potentially, our restriction to abstracts from English-language articles biased the outputs of our results. Articles of any language can be indexed in Scopus, with the caveat that an English version of the abstract must be included (Scopus 2018). Choice of language should not significantly change our outputs, for two key reasons: firstly, English is the dominant language of the scientific literature (Tardy 2004, Hamel 2008), independent of the nationality of researcher, meaning articles published in English are representative of a geographical distribution greater than just native English speaking countries; and secondly, options for other languages within Scopus are minimal – meaning their inclusion would likely not significantly influence the distribution of our results – with the greatest contributors after English (for the term pollinat*, returned on 16/08/18) being Portuguese (~1.7%), Chinese (~1.6%), Spanish (~0.9%), German (~0.5%) and Russian (~0.4%). However, we cannot exclude the possibility that the exclusion of some key languages may have biased the outputs, particularly given that 35.6% of the biodiversity conservation literature is not written in English (Amano et al. 2016).

Biases may also have been introduced by the taxonomic entity extraction algorithms, Neti Neti and Taxonfinder. Because the scrape for taxonomic information only picks up Latin binomial names, any species more often referred to by its common name, or any species more often referred to in the abstract through a higher taxonomic level, will likely be underrepresented. This may be the case for the western honey bee, *Apis mellifera*, which is often referred to by its common name, and possibly also for bumble bees. We briefly explored this limitation through investigating the frequency of common names for each of the top 4 genera (*Apis*, *Bombus*, *Osmia* and *Megachile*), finding that the taxonomic disparity between honey bees, bumble bees and other taxa is likely even greater than suggested by our results, potentially by a factor of ~2 (Supplementary material Appendix 1 Fig. A4). Given the strong association of *Apis* and *Bombus* studies with North America, Europe and east Asia, we also expect that our analysis underestimates geographic disparity. Groups such as hummingbirds, which are more often mentioned in the abstract without an accompanying Latin binomial species name, may also be underrepresented. We explored this limitation through investigating the frequency of family names for 5 families with well-known common names (fig wasps, hawk-moths, hoverflies, hummingbirds and leaf-nosed bats), selected from each of the top 5 orders (Hymenoptera, Lepidoptera, Diptera, Apodiformes and Chiroptera). We found that three of these families (hummingbirds, fig wasps and hoverflies) were likely under-represented by considering only Latin binomials (Supplementary material Appendix 1 Fig. A5).

Our findings were also influenced by the approach used to verify animal species records. In particular, in counting only Latin binomials we will have missed records. However, we reasoned that the unambiguity of the Latin binomial would help to reduce noise. Moreover, we assumed that, with the possible exception of taxa referred to by widely accepted common names, the frequency of mentions for the full species record would likely correlate with higher taxonomic levels (Correia et al. 2017).

Spelling mistakes and failure to resolve as an accepted name are two different although closely-related limitations. Failure to resolve as an accepted name could potentially be caused by a spelling mistake in a synonym or accepted name, or by that record being absent from the COL as either an accepted name or synonym. Although we investigated fuzzy-matching for non-matched records, we opted not to include this implementation here because taxonomic name resolution became more ambiguous as a result. Fuzzy-matching returned multiple potential matches for a given record, requiring significant input to exclude false positives. Moreover, spelling mistakes would only be problematic for our conclusions if unevenly distributed among taxonomic groups, which is unlikely to be the case.

CLIFF-CLAVIN may also have introduced geographical biases in the distribution of our outputs, through the way in which it is trained and its probabilistic nature. Given that CLIFF-CLAVIN is trained on news articles, its effectiveness on academic texts is unclear. During our analysis, we noticed that CLIFF-CLAVIN would occasionally mistake Latin taxonomic entities for geographic locations. For example, the genus *Pavonia* was mistaken for a geographical location. There may be instances in which the algorithm's training interacts with taxonomy to bias our outputs. Relatedly, since CLIFF-CLAVIN is trained on news outlets based primarily in the US, US-based studies may be overestimated (Imani et al. 2017). However, our results for the representation of US pollination studies are consistent with Archer et al. (2014), which applied a different methodology. The probabilistic nature of CLIFF-CLAVIN may also have influenced our results to a small degree (indeed, running the algorithm a second time led to a reduced number of 'minor' mentions in Brazil).

Future directions

Here we have used two text-analysis tools to quantify the geographical and taxonomic distribution of the animal pollinator literature. We showed that the literature is heavily concentrated in the honey bees and bumble bees of North America, albeit less biased than some authors have implied (Mayer et al. 2011).

The skewed taxonomic and geographical distribution of pollinator literature is a problem for the robustness of animal pollinator biodiversity models. Unfortunately, solving this problem is hard. Well-designed, long-term and resource-intensive studies on little known taxonomic and geographical regions are needed. However, such studies are logistically difficult, expensive, and may not be achievable in time to

inform important decisions. Another option – although not mutually exclusive – could be mitigating the problem through fully engaging with the available literature. Here we explore how this could be achieved by using the same text-analysis tools to yield a more representative and comprehensive set of studies for systematic reviews and meta-analyses. We briefly describe the conventional literature search path, as used for example in systematic reviews, before introducing a new search process. Finally, we conclude by highlighting two other ways in which text-analysis could contribute towards key research in the field of pollination ecology, and how these relate to decline models.

Closing the synthesis gap for pollinator biodiversity modelling

The conventional literature search process for a systematic review can be conceptualised as three key phases, with the first two concerning literature retrieval (Fig. 7): the search-term phase, in which key words in an online database are optimised to return literature deemed representative of the given research question; and the manual filtering phase, in which each article is assessed according to a series of specific criteria, and then excluded if it is deemed irrelevant. This manual filtering phase can be long and labour-intensive (Haddaway and Westgate 2018); some authors will assess > 10 000 papers (Lavoie et al. 2014, O'Mara-Eves et al. 2015), with one review as high as 800 000 papers (Shemilt et al. 2014). The manual filtering phase is followed by a third phase: the appraisal of each selected article, in which data are extracted to quantify the main findings of the study (Pullin and Stewart 2005). Although article appraisal can in part be addressed through text analysis (Lajeunesse 2016), here we focus on optimising the manual filtering and data-extraction phases.

Text-analysis has been introduced as a useful tool in optimising the literature filtering phase, but uptake in ecology is still low (Westgate et al. 2015). This is in part a symptom of unintuitive text-analysis tools, and insufficient technical skills required to use them (Westgate 2018). However, arguably the bigger barrier is the lack of understanding as to how text-analysis tools relate practically to the literature search process. Although text-analysis approaches in ecology have advanced (Nunez-Mir et al. 2016, Roll et al. 2017, Westgate et al. 2018), as far as we know there are no clear recommendations as to how ecological researchers should implement these approaches in the literature search. Here, taking inspiration from the 'revtools' package (Westgate 2018), we propose a text-analysis search path in the context of the systematic review (Fig. 7). The technology is available for this path, but not yet the specific intuitive tools or validation in the context of pollination ecology. Our proposed synthesis path can be conceptualised as five key phases: search term, topic similarity, taxonomic and geographic identification, manual filtering and appraisal. Below we briefly describe each of the first three modified phases.

1) Initial search terms should be used to return a broad body of literature for a given field. Fewer and less specific search terms should be used across multiple databases, aiming to return all of the relevant literature irrespective of a potentially high false positive rate. This less restrictive initial search will require less researcher input, thus reducing the time required. Such an approach will also reduce the likelihood of overlooking relevant literature through overly specific search terms.

2) The key filtering step should be shifted downstream to a text-analysis filter. Topic-clustering algorithms should be used to exclude irrelevant articles. For example, in the context of the potential pollination literature returned by Scopus, topic-clustering can be used to exclude papers on the flower-pollination algorithm, an area of computer science unrelated to pollination ecology. Topic clustering is more reproducible than database search terms, less subject to researcher methodology, more representative of overall content, and not subject to differences across database.

3) Taxonomic and geographical entity extraction algorithms should be used to indicate the geography and taxonomy of each study. Taxonomic group and geography can be used to prioritise for underrepresented taxa or regions, as well as to identify likely literature for regional or taxonomic systematic reviews.

Future text-analysis applications in the field of pollination ecology

Text-analysis tools could also be applied to answer other research questions in the field of animal pollination, such as the identification of likely plant–pollinator interactions and the estimation of the number of pollinating species. Here we briefly explore how these research questions might be addressed.

Likely plant–pollinator interactions could be identified through investigating the strength of animal–plant associations across the pollination literature. For example, a plant and animal species frequently occurring in the same abstract would imply a closely-related pair, while an animal–plant combination occurring infrequently would imply a weak or non-existent interaction. These animal–plant networks could then be validated with observational plant–pollinator interaction data, using an approach similar to Tamaddoni-Nezhad et al. (2013), who showed that text-derived food web networks approximated empirical data. Such a database of likely interacting plant–pollinators could be invaluable for pollinator biodiversity modelling. For example, we might be able to predict better the likelihood of co-extinctions, as well as improve estimations of global and regional pollinator importance.

We also envisage that an accurate lower bound for the number of pollinating species could be estimated through text-analysis. This would build on the work of Ollerton (2017) and Wardhaugh (2015), who estimated there to be ~350 000 described species of animal pollinator, based

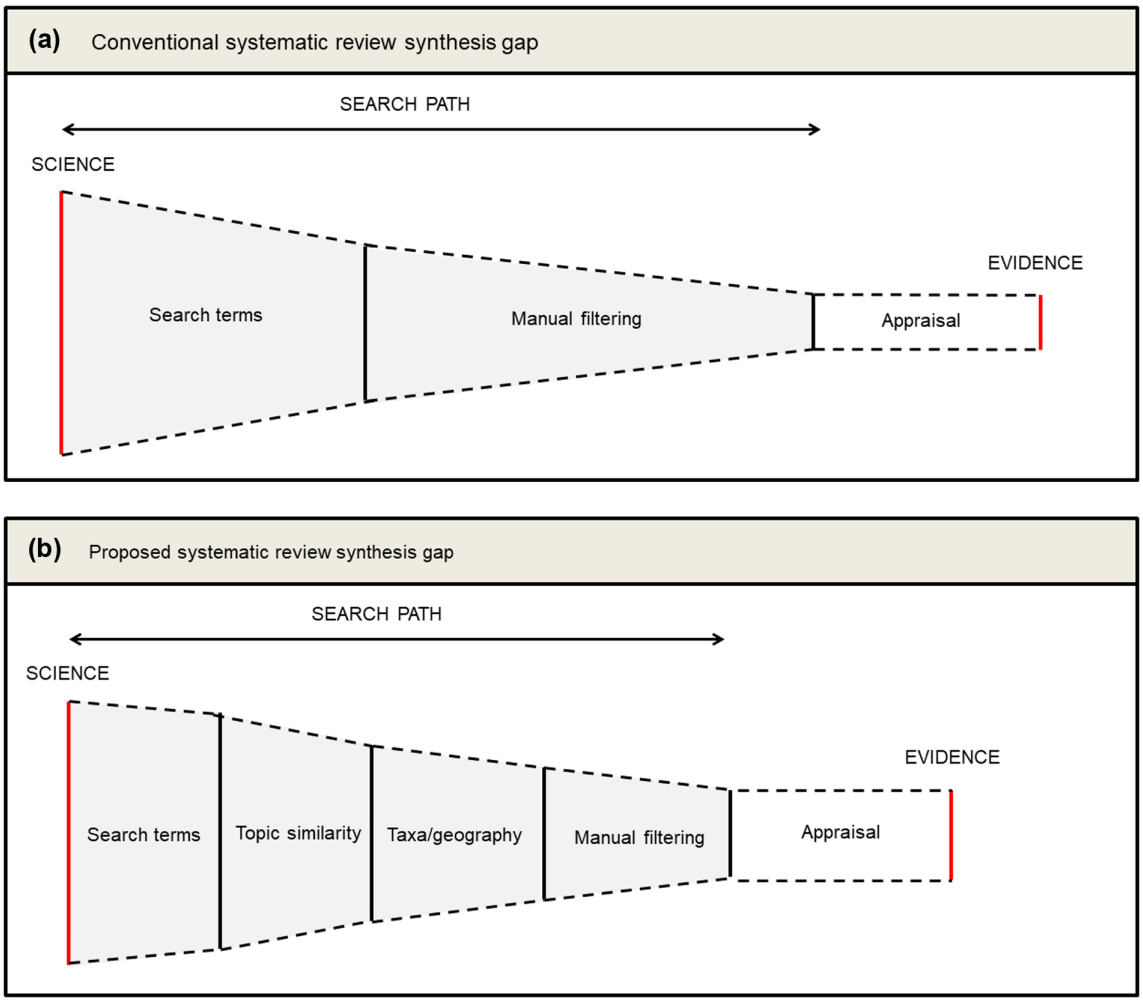

Figure 7. Reducing the length of the 'synthesis gap' for ecological systematic reviews. The conventional synthesis gap (a) proceeds through three primary steps: the search term phase, manual filtering and appraisal. Our proposed, more efficient synthesis gap (b) proceeds through five steps: a less-exclusionary search term phase; text-analysis prioritisation (topic similarity and taxonomic/geographic filtering or prioritization); manual filtering and appraisal, with both the search term and manual filtering stages shortened. Solid lines represent the research volume between each stage, which in our proposed process is greater, owing to the increased taxonomic and geographic representativeness. Dotted lines represent the change in research volume for each selection step. Red lines represent the beginning and ends of the synthesis gap, from the practice of science to the synthesis of evidence.

on the number of species in key pollinating groups. We propose that within taxonomic groups, the number of pollinating species could be estimated through quantifying the rate at which unique animal species accumulate as pollination text volume increases. This would be analogous to the species-effort rarefaction curve; a mathematical relationship between pollinator text volume (effort) and species number, revealing how the number of unique pollinators increases as a function of research volume (effort). Such work would likely require scraping of full articles, rather than just abstracts as presented here. Full-text scraping presents additional problems regarding rapid article download and permissions, file format, text volume quantity, but the benefits relative to abstract scraping are known (Westergaard et al. 2018). Similar methodologies have been applied to quantify the number of species on Earth, both for the flowering plants (Joppa et al. 2011) and for all Eurkaryotes (Mora et al. 2011). In the context of pollinator biodiversity models, good estimates of the diversity and number of pollinating species are fundamental for understanding how ecosystems will respond to future environmental change (Ollerton 2017).

Summary

Here we have shown – using a novel combination of informatics tools – how text-analysis can be used to quantify the taxonomic and geographical distribution of the animal pollinator literature. In doing so we have confirmed that the literature is heavily focused on the honey bees and bumble bees of North America and Europe, although many studies also exist for other taxa and regions. This skewed taxonomic

and geographical distribution likely has a large impact on the robustness of systematic reviews and meta-analyses of animal pollinator decline. We have shown how text-analysis might to some degree mitigate these data biases. Text-analysis could be used to make the literature search process more efficient, as well as increase the taxonomic and geographic representativeness of the studies fed into systematic reviews and meta-analyses. To this end, we briefly outlined a new literature search process, using an ecological systematic review as an example for how text-analysis might contribute. We have also explored some potential broader applications of text-analysis in pollination ecology, such as the identification of likely plant–pollinator interactions and the estimation of the number of pollinating species, both of which could feed into more robust systematic reviews and meta-analyses in the future. Text-analysis undoubtedly shows promise in increasing our understanding of the rapidly growing pollination ecology literature, and in turn the robustness of studies estimating pollinator decline.

Data deposition

Key data and code are currently hosted on Github (<https://github.com/Joemillard/pollinator_taxonomic_geographic_dist_text-analysis>) and Figshare (<https://doi.org/10.6084/m9.figshare.8326973.v1>).

Acknowledgements – Thanks to Richard D. Gregory and Adrienne Etard for comments on an earlier draft, and the RSPB for financial support. We also thank three anonymous reviewers for their constructive criticism and comments.
Funding – JM was funded by the London NERC DTP, award number NE/R012148/1, and the RSPB on a CASE studentship. TN was supported by a Royal Society University Research Fellowship and a grant from the UK Natural Environment Research Council (grant number: NE/R010811/1).
Author contributions – JM, RF and TN conceived and designed the review; JM carried out all analysis, produced the figures, and wrote the initial draft of the manuscript; TN reviewed and edited the manuscript; JM, TN and RF read and approved the final version of the manuscript.
Conflicts of interest – The authors declare no conflicts of interest.

References

Aizen, M. A. and Harder, L. D. 2009. The global stock of domesticated honey bees is growing slower than agricultural demand for pollination. – Curr. Biol. 19: 915–918.
Akella, L. et al. 2012. NetiNeti: discovery of scientific names from text using machine learning methods. – BMC Bioinform. 13: 211.
Amano, T. et al. 2016. Languages are still a major barrier to global science. – PLoS Biol. 14: e2000933.
Archer, C. R. et al. 2014. Economic and ecological implications of geographic bias in pollinator ecology in the light of pollinator declines. – Oikos 123: 401–407.
Bartomeus, L. R. et al. 2018. Historical collections as a tool for assessing the global pollination crisis. – Phil. Trans. R. Soc. B 374: 1763.
Beaman, R. S. and Conn, B. J. 2003. Automated geoparsing and georeferencing of Malesian collection locality data. – Telopea 10: 43–52.
Beckmann, M. and von Wehrden, H. 2012. Where you search is what you get: literature mining – Google Scholar versus Web of Science using a data set from a literature search in vegetation science. – J. Veg. Sci. 23: 1197–1199.
Biesmeijer, J. et al. 2006. Parallel declines in pollinators and insect-pollinated plants in Britain and the Netherlands. – Science 313: 351–353.
Butchart, S. H. M. et al. 2010. Global biodiversity: indicators of recent declines. – Science 328: 1164–1168.
Chiang, Y.-Y. 2017. Unlocking textual content from historical maps – potentials and applications, trends, and outlooks. – In: Santosh, K. C. et al. (eds), International conference on recent trends in image processing and pattern recognition. Springer, pp. 111–124.
Cohen, K. B. and Hunter, L. 2008. Getting started in text mining. – PLoS Comput. Biol. 4: e20.
Correia, R. A. et al. 2017. Internet scientific name frequency as an indicator of cultural salience of biodiversity. – Ecol. Indic. 78: 549–555.
Correia, R. A. et al. 2018. Nomenclature instability in species culturomic assessments: why synonyms matter. – Ecol. Indic. 90: 74–78.
CRAN 2018. taxize: taxonomic information from around the Web. – R package taxize ver. 0.9.4.
De Palma, A. et al. 2016. Predicting bee community responses to land-use changes: effects of geographic and taxonomic biases. – Sci. Rep. 6: 31153.
D'Ignazio, C. et al. 2014. CLIFF-CLAVIN: determining geographic focus for news articles. – NewsKDD: Data Science for News Publishing.
Dukas, R. and Morse, D. H. 2003. Crab spiders affect flower visitation by bees. – Oikos 101: 157–163.
Elbasiouny, H. et al. 2014. Spatial variation of soil carbon and nitrogen pools by using ordinary Kriging method in an area of north Nile Delta, Egypt. – CATENA 113: 70–78.
Endress, P. K. 2001. The flowers in extant basal angiosperms and inferences on ancestral flowers. – Int. J. Plant Sci. 162: 1111–1140.
Ferreira, C. et al. 2015. The evolution of peer review as a basis for scientific publication: directional selection towards a robust discipline? – Biol. Rev. 91: 3.
Fleming, T. H. and Muchhala, N. 2008. Nectar-feeding bird and bat niches in two worlds: pantropical comparisons of vertebrate pollination systems. – J. Biogeogr. 35: 764–780.
Fleming, T. H. et al. 2009. The evolution of bat pollination: a phylogenetic perspective. – Ann. Bot. 104: 1017–1043.
Genersch, E. et al. 2010. The German bee monitoring project: a long term study to understand periodically high winter losses of honey bee colonies. – Apidologie 41: 332–352.
Ghazoul, J. 2005. Buzziness as usual? Questioning the global pollination crisis. – Trends Ecol. Evol. 20: 367–373.
Ghazoul, J. 2015. Qualifying pollinator decline evidence. – Science 348: 981–982.
Goulson, D. 2003. Conserving wild bees for crop pollination. – J. Food Agric. Environ. 1: 142–144.
Goulson, D. et al. 2008. Decline and conservation of bumble bees. – Annu. Rev. Entomol. 53: 191–208.
Goulson, D. et al. 2015. Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. – Science 347: 6229.

Griffiths, T. L. et al. 2004. Finding scientific topics. – Proc. Natl Acad. Sci. USA 101: 5228–5235.
Grimmer, J. and Stewart, B. M. 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. – Polit. Anal. 21: 267–297.
Gritta, M. et al. 2018. What's missing in geographical parsing? – Lang. Resour. Eval. 52: 603–623.
Guralnick, R. and Hill, A. 2009. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. – Bioinformatics 25: 421–428.
Gurevitch, J. et al. 2018. Meta-analysis and the science of research synthesis. – Nature 555: 175–182.
Haddaway, N. R. and Westgate, M. J. 2018. Predicting the time needed for environmental systematic reviews and systematic maps. – Conserv. Biol. 33: 434–443.
Hahn, M. and Brühl, C. A. 2016. The secret pollinators: an overview of moth pollination with a focus on Europe and North America. – Arthropod-Plant Interact. 10: 21–28.
Hamel, R. E. 2008. The dominance of English in the international scientific periodical literature and the future of language use in science. – AILA Rev. 20: 53–71.
Han, F. et al. 2012. From where did the western honeybee (*Apis mellifera*) originate? – Ecol. Evol. 2: 1949–1957.
Imani, M. B. et al. 2017. Focus location extraction from political news reports with bias correction. – In: 2017 IEEE International Conference on Big Data (BIGDATA), pp. 1956–1964.
IPBES 2017. The assessment report on pollinators, pollination and food production. – Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, Bonn, Germany.
Joppa, L. N. et al. 2011. How many species of flowering plants are there? – Proc. R. Soc. B 278: 554–559.
Kessler, A. and Baldwin, I. T. 2002. *Manduca quinquemaculata*'s optimization of intra-plant oviposition to predation, food quality, and thermal constraints. – Ecology 83: 2346–2354.
Klein, A.-M. et al. 2007. Importance of pollinators in changing landscapes for world crops. – Proc. R. Soc. B 274: 303–313.
Kosior, A. et al. 2007. The decline of the bumble bees and cuckoo bees (Hymenoptera: Apidae: Bombini) of western and central Europe. – Oryx 41: 79.
Labandeira, C. C. et al. 2007. Pollination drops, pollen, and insect pollination of Mesozoic gymnosperms. – Int. Assoc. Plant Taxon. 56: 663–695.
Lajeunesse, M. J. 2016. Facilitating systematic reviews, data extraction and meta-analysis with the METAGEAR package for R. – Methods Ecol. Evol. 7: 323–330.
Lautenbach, S. et al. 2012. Spatial and temporal trends of global pollination benefit. – PLoS One 7: e35954.
Lavoie, M.-C. et al. 2014. Devices for preventing percutaneous exposure injuries caused by needles in healthcare personnel. – Cochrane Database Syst. Rev. 3: 1–76.
Leary, P. R. et al. 2007. uBioRSS: tracking taxonomic literature using RSS. – Bioinformatics 23: 1434–1436.
Leidner, J. L. and Lieberman, M. D. 2011. Detecting geographical references in the form of place names and associated spatial natural language. – SIGSPATIAL Special 3: 5–11.
Loh, J. et al. 2005. The Living Planet Index: using species population time series to track trends in biodiversity. – Proc. R. Soc. B 360: 289–295.
Lortie, C. J. 2014. Formalized synthesis opportunities for ecology: systematic reviews and meta-analyses. – Oikos 123: 897–902.
Mayer, C. et al. 2011. Pollination ecology in the 21st century: key questions for future research. – J. Pollinat. Ecol. 3: 8–23.
Michener, C. D. 2007. The bees of the world. – Johns Hopkins Univ. Press.
Mongeon, P. and Paul-Hus, A. 2016. The journal coverage of Web of Science and Scopus: a comparative analysis. – Scientometrics 106: 213–228.
Mora, C. et al. 2011. How many species are there on Earth and in the Ocean? – PLoS Biol. 9: e1001127.
National Research Council 2007. Status of pollinators in North America. – National Academies Press.
New, T. R. 2004. Moths (Insecta: Lepidoptera) and conservation: background and perspective. – J. Insect Conserv. 8: 79–94.
Newbold, T. et al. 2015. Global effects of land use on local terrestrial biodiversity. – Nature 520: 45–50.
Noriega, J. A. et al. 2018. Research trends in ecosystem services provided by insects. – Basic Appl. Ecol. 26: 8–23.
Nunez-Mir, G. C. et al. 2016. Automated content analysis: addressing the big literature challenge in ecology and evolution – Methods Ecol. Evol. 7: 1262–1272.
Oldroyd, B. P. 1999. Coevolution while you wait: *Varroa jacobsoni*, a new parasite of western honeybees. – Trends Ecol. Evol. 14: 312–315.
Ollerton, J. 2017. Pollinator diversity: distribution, ecological function, and conservation. – Annu. Rev. Ecol. Evol. Syst. 48: 353–376.
Ollerton, J. et al. 2011. How many flowering plants are pollinated by animals? – Oikos 120: 321–326.
O'Mara-Eves, A. et al. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. – Syst. Rev. 4: 5.
Parr, C. S. et al. 2012. Evolutionary informatics: unifying knowledge about the diversity of life. – Trends Ecol. Evol. 27: 94–103.
Pereira, H. M. et al. 2010. Scenarios for global biodiversity in the 21st century. – Science 330: 1496–1501.
Potts, S. G. et al. 2010a. Declines of managed honey bees and beekeepers in Europe. – J. Apic. Res. 49: 15–22.
Potts, S. G. et al. 2010b. Global pollinator declines: trends, impacts and drivers. – Trends Ecol. Evol. 25: 345–353.
Potts, S. G. et al. 2016. Safeguarding pollinators and their values to human well-being. – Nature 540: 220–229.
Proctor, M. et al. 1996. The natural history of pollination. – Harper Collins.
Pullin, A. S. and Stewart, G. B. 2005. Guidelines for systematic review in conservation and environmental management. – Conserv. Biol. 20: 1647–1656.
Raguso, R. A. and Willis, M. A. 2002. Synergy between visual and olfactory cues in nectar feeding by naïve hawkmoths, *Manduca sexta*. – Anim. Behav. 64: 685–695.
Rathbone, J. et al. 2015. Better duplicate detection for systematic reviewers: evaluation of systematic review assistant-deduplication module. – Syst. Rev. 4: 6.
Rech, A. R et al. 2016. The macroecology of animal versus wind pollination: ecological factors are more important than historical climate stability. – Plant Ecol. Distrib. 9: 253–262.
Regan, E. C et al. 2015. Global trends in the status of bird and mammal pollinators. – Conserv. Lett. 8: 397–403.
Riddiford, L. M. et al. 2003. Insights into the molecular basis of the hormonal control of molting and metamorphosis from *Manduca sexta* and *Drosophila melanogaster*. – Insect Biochem. Mol. Biol. 33: 1327–1338.
Roll, U. et al. 2017. Using machine learning to disentangle homonyms in large text corpora. – Conserv. Biol. 32: 716–724.

Roskov Y. et al. 2017. Species 2000 and ITIS Catalogue of Life, 2017 annual checklist. – Naturalis, Leiden.
Sarkar, I. N. 2007. Biodiversity informatics: organizing and linking information across the spectrum of life. – Brief. Bioinform. 8: 347–357.
Scopus 2018. Scopus content policy and selection. – <www.elsevier.com/solutions/scopus/how-scopus-works/content/content-policy-and-selection>.
Semmens, T. D. et al. 1993. *Bombus terrestris* (L.) (Hymenoptera: Apidae) now established in Tasmania. – Aust. J. Entomol. 32: 346–346.
Shemilt, I. et al. 2014. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. – Res. Synth. Methods 5: 31–49.
Ssymank, A. et al. 2008. Pollinating flies (Diptera): a major contribution to plant diversity and agricultural production. – Biodiversity 9: 86–89.
Steffan-Dewenter, I. and Westphal, C. 2007. The interplay of pollinator diversity, pollination services and landscape change. – J. Appl. Ecol. 45: 737–741.
Stokstad, E. 2007. The case of the empty hives. – Science 316: 970–972.
Tamaddoni-Nezhad, A. et al. 2013. Construction and validation of food webs using logic-based machine learning and text mining. – Adv. Ecol. Res. 49: 225–289.
Tardy, C. 2004. The role of English in scientific communication: lingua franca or *Tyrannosaurus rex*? – J. English Acad. Purposes 3: 247–269.
Thessen, A. E. et al. 2012. Applications of natural language processing in biodiversity science. – Adv. Bioinform. 2012: 1–17.
Tittensor, D. P. et al. 2014. A mid-term analysis of progress toward international biodiversity targets. – Science 346: 241–244.

Supplementary material (Appendix ECOG-04532 at <www.ecography.org/appendix/ecog-04532>). Appendix 1.

van Engelsdorp, D. et al. 2008. A survey of honey bee colony losses in the U.S., fall 2007 to spring 2008. – PLoS One 3: e4071.
van Engelsdorp, D. et al. 2009. Colony collapse disorder: a descriptive study. – PLoS One 4: e6481.
Velthuis, H. H. W. and van Doorn, A. 2006. A century of advances in bumblebee domestication and the economic and environmental aspects of its commercialization for pollination. – Apidologie 37: 421–451.
Vit, P. et al. 2004. Quality standards for medicinal uses of Meliponinae honey in Guatemala, Mexico and Venezuela. – Bee World 85: 2–5.
Wardhaugh, C. W. 2015. How many species of arthropods visit flowers? – Arthropod-Plant Interact. 9: 547–565.
Westergaard, D. et al. 2018. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. – PLoS Comp. Biol. 14: 2.
Westgate, M. J. 2018. revtools: bibliographic data visualization for evidence synthesis in R. – bioRxiv: 262881, <https://doi.org/10.1101/262881>.
Westgate, M. J. and Lindenmayer, D. B. 2016. The difficulties of systematic reviews. – Conserv. Biol. 31: 1002–1007.
Westgate, M. J. et al. 2015. Text analysis tools for identification of emerging topics and research gaps in conservation science. – Conserv. Biol. 29: 1606–1614.
Westgate, M. J. et al. 2018. Software support for environmental evidence synthesis. – Nat. Ecol. Evol. 2: 588–590.
Whitfield, C. W. et al. 2006. Thrice out of Africa: ancient and recent expansions of the honey bee, *Apis mellifera*. – Science 314: 642–645.
Winfree, R. et al. 2011. Native pollinators in Anthropogenic habitats. – Annu. Rev. Ecol. Evol. Syst. 42: 1–22.
Woodcock, B. A. et al. 2016. Impacts of neonicotinoid use on long-term population changes in wild bees in England. – Nat. Commun. 7: 12459.