Ecography

**ECOG-04532**
Millard, J. W., Freeman, F. and Newbold, T. 2019. Text-analysis reveals taxonomic and geographic disparities in animal pollination literature. – Ecography doi: 10.1111/ecog.04532

## Supplementary material

# Appendix 1

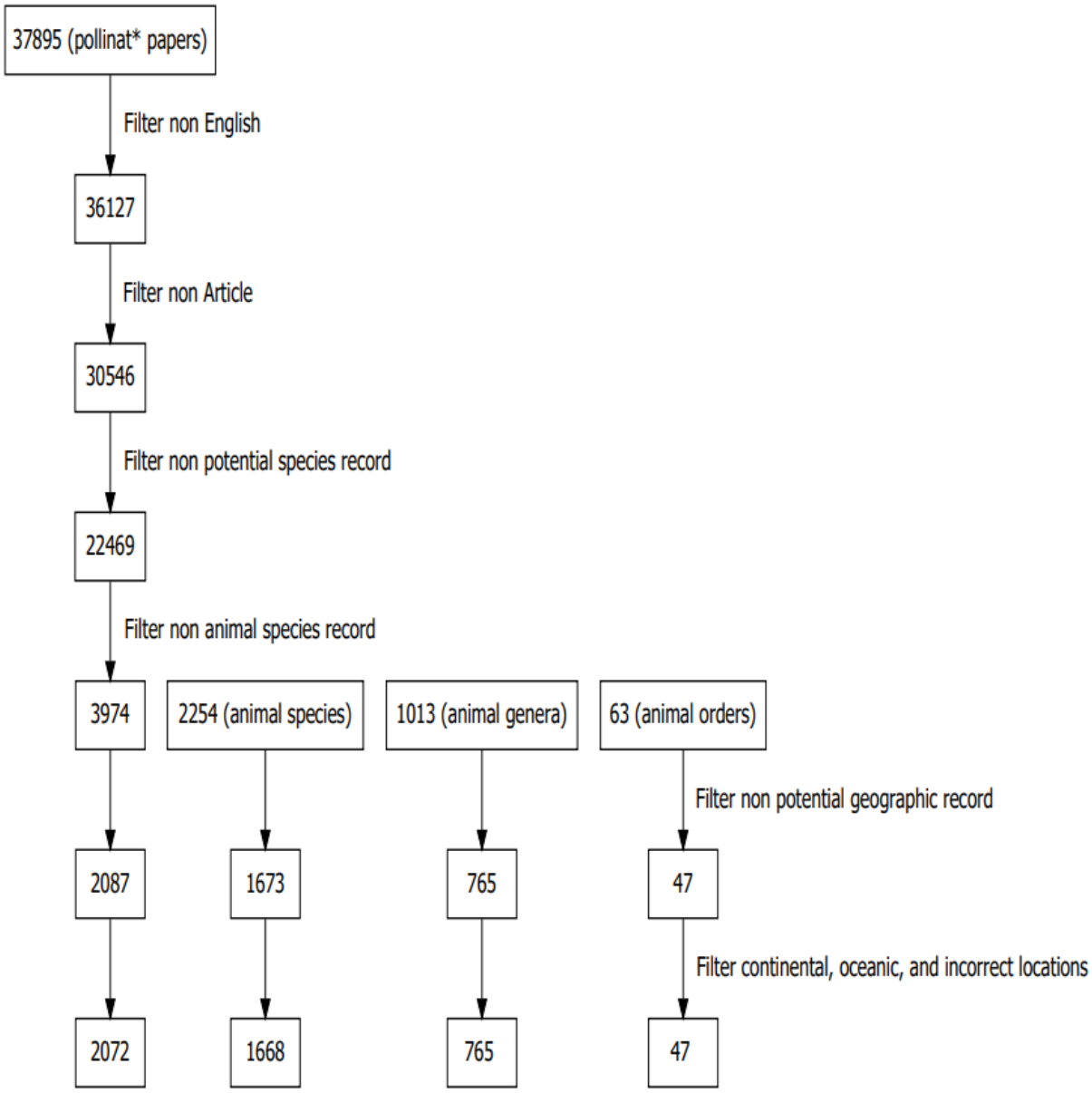


**Figure A1. PRISMA diagram for pollination paper selection. 37895 abstracts were returned through entering the search term pollinat* in Scopus. These abstracts were filtered for English language ("Filter non English"), primary research articles ("Filter non Article"), any potential species records ("Filter non potential species record"), confirmed animal species ("Filter non animal species record"), geographic locations ("Filter non potential geographic record"), and those that do not contain only a continental, oceanic, and incorrect locations.**

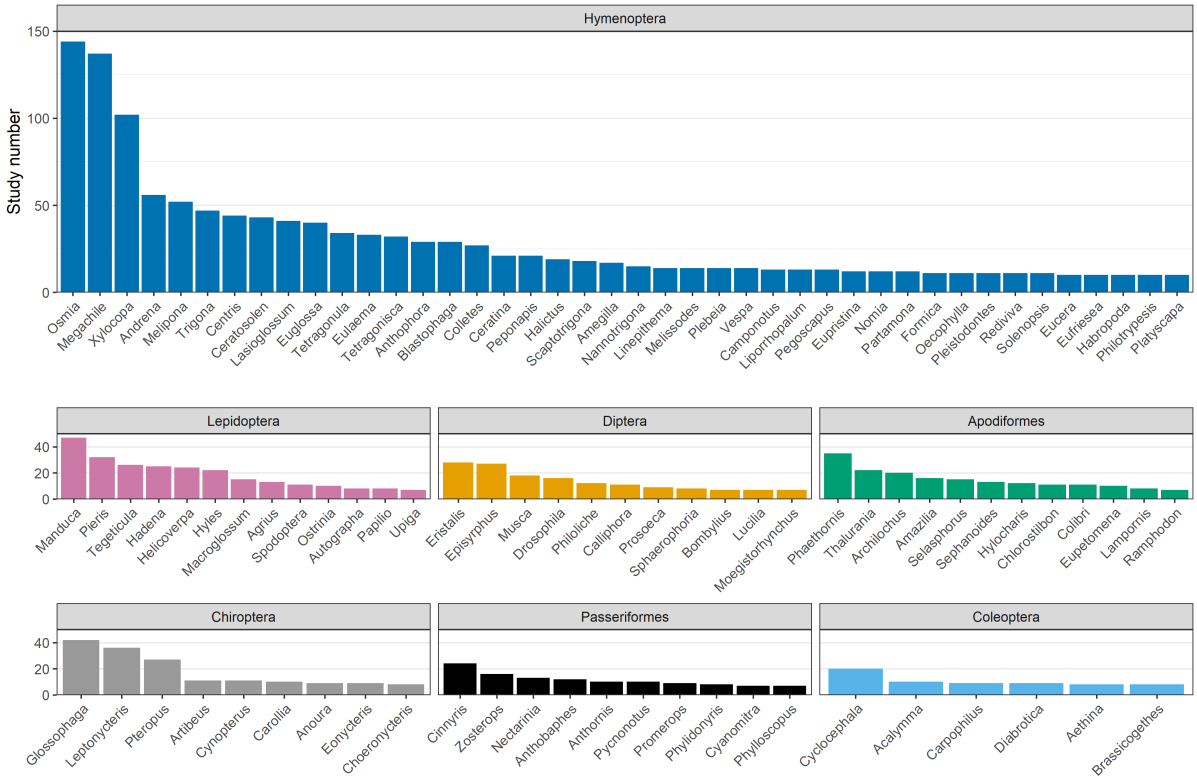


**Figure A2. Distribution of all animal genera, with the exception of *Apis* and *Bombus*, occurring in 10 or more studies related to pollination. *Apis* and *Bombus* have been excluded here to better represent distribution for less well studied genera (see Figure 3 for comparable values for *Apis* and *Bombus*).**

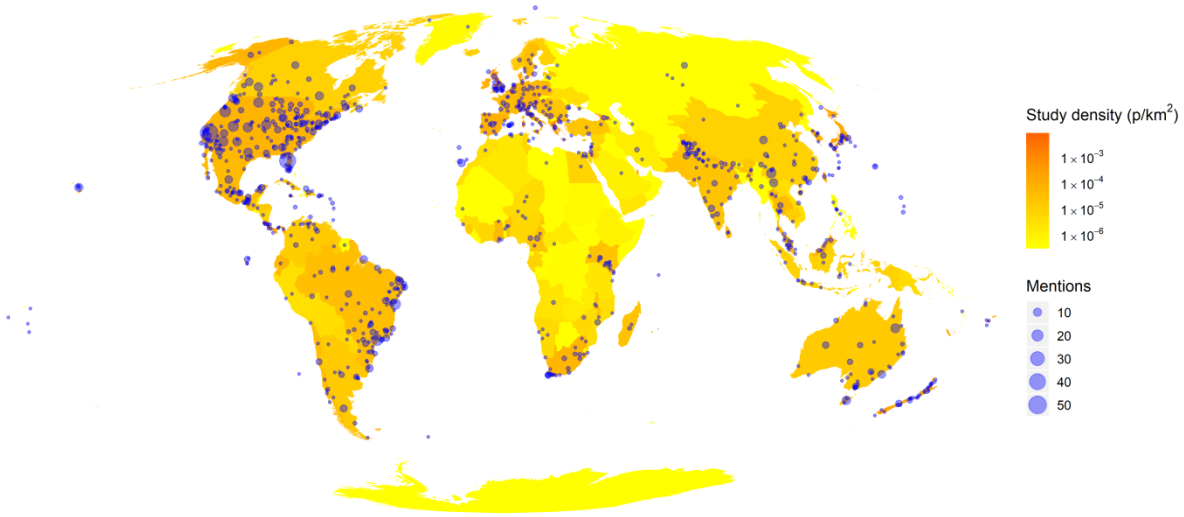


**Figure A3. Global study density of animal pollinator related studies, aggregated at country level and adjusted for country area. Study densities were calculated by counting the number of abstracts with their "major" focus in each country, and then dividing this value by country area. All oceanic and otherwise obviously incorrect "minor" mentions, as well as "minor" mentions that could only be resolved to a unit larger than a country, were removed. Densities were $\log_{10}$-transformed. Partially transparent blue points ("minor" mentions) represent the number of abstracts in which CLIFF-CLAVIN resolved that location.**

**Table A1. Summary output for a poisson generalised linear model, predicting annual study count against year (1961-2017) and taxonomic genera (Apis, Bombus, Other).**

| | **Estimate** | **Std. Error** | **P value** |
|---|---|---|---|
| Intercept | -1.069e+02 | 2.561e+00 | <2e-16 |
| Year | 5.357e-02 | 1.274e-03 | <2e-16 |
| Apis | 3.178e+00 | 3.166e-02 | <2e-16 |
| Bombus | 3.130e+00 | 3.224e-02 | <2e-16 |

**Validation**

We carried out a series of checks of the validity of the outputs of our text analysis methods. We validated our outputs at three levels: first, the accuracy of the extraction of animal species names; second, at the level of abstract subject area, to determine whether we had selected abstracts that are typically related to animal pollination ecology; and third, the accuracy and potential bias of the geographic locations as determined by CLIFF-CLAVIN.

We estimated the completeness and accuracy of the animal species extraction by sampling and manually searching approximately 1% of the original full set of abstracts (300 in total). For any random samples, we used the R function set.seed() to seed the random number generator, and then sampled abstracts at random. Given that the taxonomic extraction algorithm attempts to resolve each animal as its accepted name, in order to fairly judge its effectiveness, any manual searching would have to attempt to resolve accepted names in a similar way. For each potential animal species record identified manually, we searched for the species in the COL hoping to confirm as an accepted name. If the species was not accepted but recognised by the COL as a synonym, we changed the species record for that abstract to the accepted name. If the COL did not recognise the potential animal species as either an accepted name or a synonym, we then searched the website Discover Life (http://www.discoverlife.org/) for the accepted name and changed the record if appropriate. We removed any potential species records that we could not confirm as either an accepted name or a synonym verified by either the COL or Discover Life.

After manually confirming accepted animal species, we then compared these outputs to the performance of the algorithm. 79.5% of the animal species records manually extracted were found by the automated algorithm (i.e. a 20.5% omission error). Precision on the other hand was high at 100%, meaning that the algorithm found no animals which were not in that given abstract (i.e. a 0% commission error).

We also conducted a validation to investigate whether considering only Latin binomial names influenced the taxonomic bias in the pollination literature (Fig. A4). This analysis shows that the abstract count for honey bees and bumble bees is underestimated by searching for Latin binomials, potentially by a factor of ~2. This is because honey bees and bumblebees are referred to by their common name more often than other species. We counted the number of "pollinat*" abstracts mentioning two typical common name spellings for Apis and Bombus, and then to control for string number, analogous strings for both *Osmia* and *Megachile*: *Apis* - "honey bee" and "honeybee"; *Bombus* - "bumble bee" and "bumblebee"; *Osmia* - "mason bee" and "mason-bee"; *Megachile* - "leafcutter bee" and "leaf-cutter bee". We included *Osmia* and *Megachile* as a control, to investigate common name frequency for less well-known species. For *Apis* and *Bombus*, including abstracts mentioning a common name doubles their respective abstract count. For Osmia and Megachile, the inclusion of common names increases abstract count by 4.6% and 7.2% abstracts respectively. These results would indicate that whilst *Apis* and *Bombus* study count is underestimated, for other less well-known taxa the Latin binomial will be an effective indicator of study effort. We also reasoned that, whilst our analysis might underestimate for *Apis* and *Bombus,* including only the Latin binomial would help to reduce false positive rate.

We conducted an additional validation to investigate whether excluding taxonomic names above the level of species influenced the taxonomic bias in the pollination literature. This analysis indicates that although representation for some families (hummingbirds, fig wasps, and hoverflies) may be underestimated, the overall trend is likely similar (Fig. A5). We counted the number of "pollinat*" abstracts mentioning family names for each of 4 well-known pollination-related families (hummingbirds, fig wasps, hoverflies, hawk-moths, and leaf-nosed bats), selected from each of the top 5 orders (Hymenoptera, Lepidoptera, Diptera, Apodiformes, and Chiroptera). For each family, we searched for four common, Latin, and pluralised family names: hummingbirds ("humming-bird", "hummingbird", "Hummingbird", "Trochilidae"); fig wasps ("fig wasp", "Fig wasp", "fig-wasp", "Agaonidae"); hoverflies

("Hoverfly", "hoverflies", "hoverfly", "Syrphidae"); hawk-moths ("Hawk-moth", "hawk moth", "hawk-moth", "Sphingidae") and leaf-nosed bats ("Phyllostomidae", "leaf-nosed bat", "leaf nosed bat", "Leaf-nosed bat"). The number of abstracts for hummingbirds, hoverflies, and fig wasps all increased by more than a factor of ~2 with the inclusion of family names, with the leaf-nosed bats making only a marginal increase in total abstract number (Fig. A5).

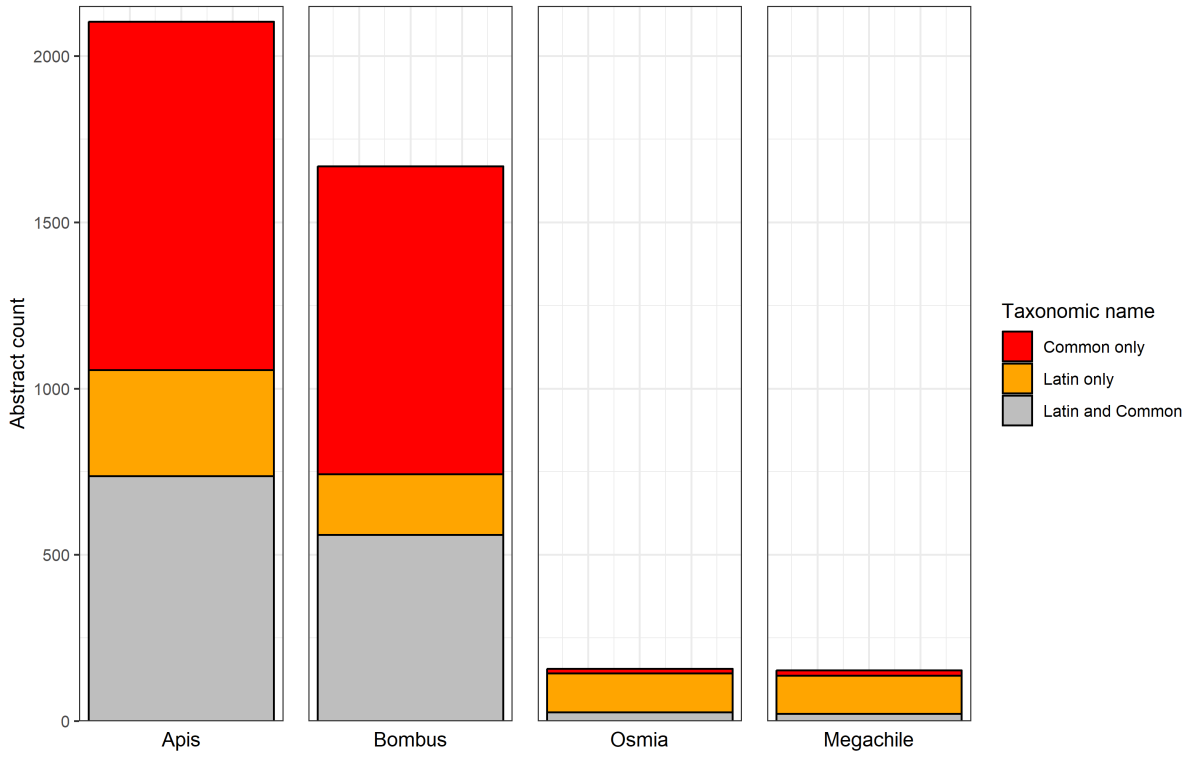


**Figure A4. Frequency of "pollinat" Scopus abstracts containing a common name and Latin binomial for each of *Apis* (honey bee), *Bombus* (bumblebee), *Osmia* (mason bee), and *Megachile* (leafcutter bee). Grey bars represent abstracts containing both a Latin binomial and common name for that genus. Orange bars represent abstracts containing only a Latin binomial for a species in that genus. Red bars represent abstracts containing only a common name for that genus.**

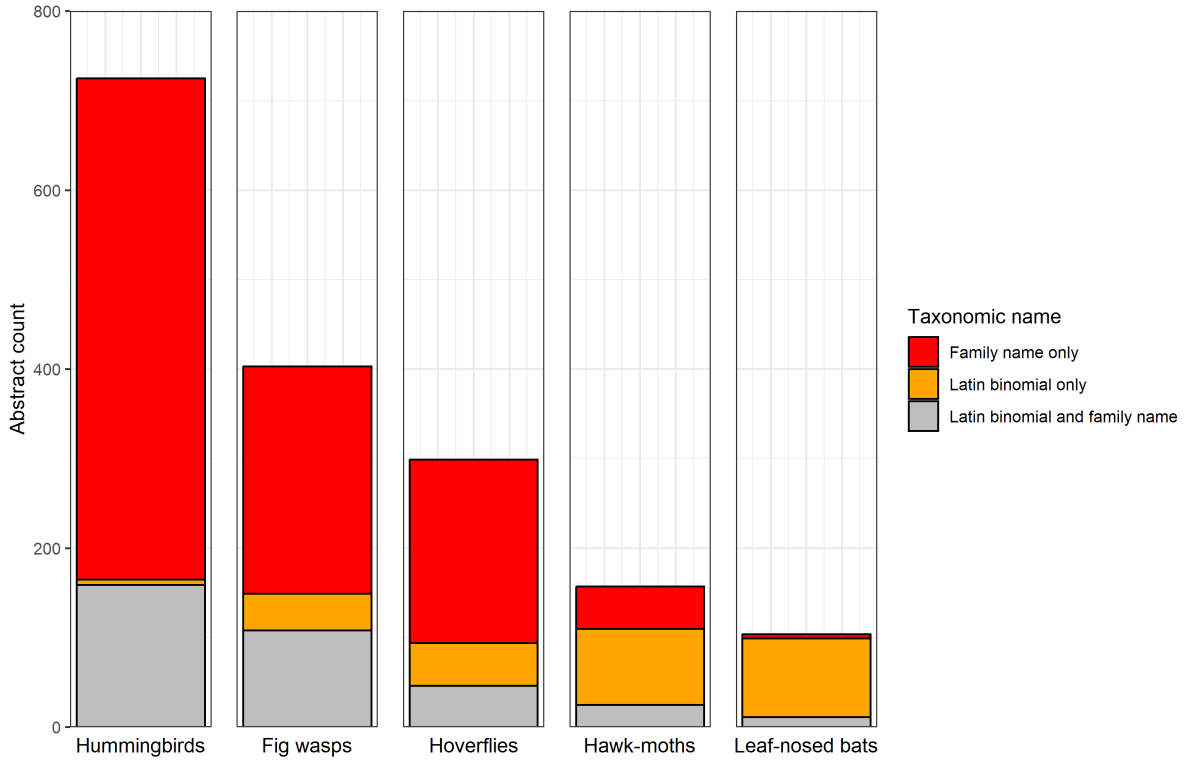


**Figure A5. Frequency of "pollinat" Scopus abstracts containing a family name (either common or Latin) and Latin binomial for each of the hummingbirds (Trochilidae), fig wasps (Agaonidae), hoverflies (Syrphidae), hawk-moths (Sphingidae), and leaf-nosed bats (Phyllostomidae). Grey bars represent abstracts containing both a Latin binomial and the family name for that family. Orange bars represent abstracts containing only a Latin binomial for a species in that family. Red bars represent abstracts containing only the family name for that family.**

To validate the subject areas of the identified abstracts, we randomly sampled 100 abstracts from the subset of original abstracts that also mentioned an animal species (approximately 2.5% of the total). We then read each abstract and title, assigning the subject area as any of three categories: general pollination ecology, pollinator status or habitat disturbance, and other pollinator related literature (Fig. A6). No abstracts were totally unrelated to pollination. Abstracts on general pollination ecology included any studies on visitation, efficiency, pollinator movement ecology, pollinator foraging behaviour, pollination syndromes, plant-pollinator networks, and pollination dependent crop yields. Abstracts on pollinator status included any studies on pollinator population trends, diversity, abundance, ecological impacts, and habitat disturbance. All “other” abstracts concerned analyses of population genetics, pest/disease management, pollinator predation, invasive species management, animal floral mimicry, pollinator mating behaviour, pollinator awareness, pollinator learning behaviour, and pollinator nesting behaviour.

**Figure A8. Geographic distribution of the pollination literature, resolved through exact character string matches and CLIFF-CLAVIN. We deemed exact character string matches a coarse and imperfect check on CLIFF-CLAVIN. The red dotted line represents the study proportion midpoint. Consistent with CLIFF-CLAVIN, exact character string matches also return Germany outside of the top 15 countries, at 22$^{nd}$ for CLIFF-CLAVIN and 17$^{th}$ for exact character string matches.**