Title
Leveraging Large Biological Interaction Data to Quantify Plant Specialization by Bees

Authors
Bachelder, Nicholas R.
Chen, Angel
Zoe, Fang
et al.

Publication Date
2021-07-01

Leveraging Large Biological Interaction Data to Quantify Plant Specialization by Bees

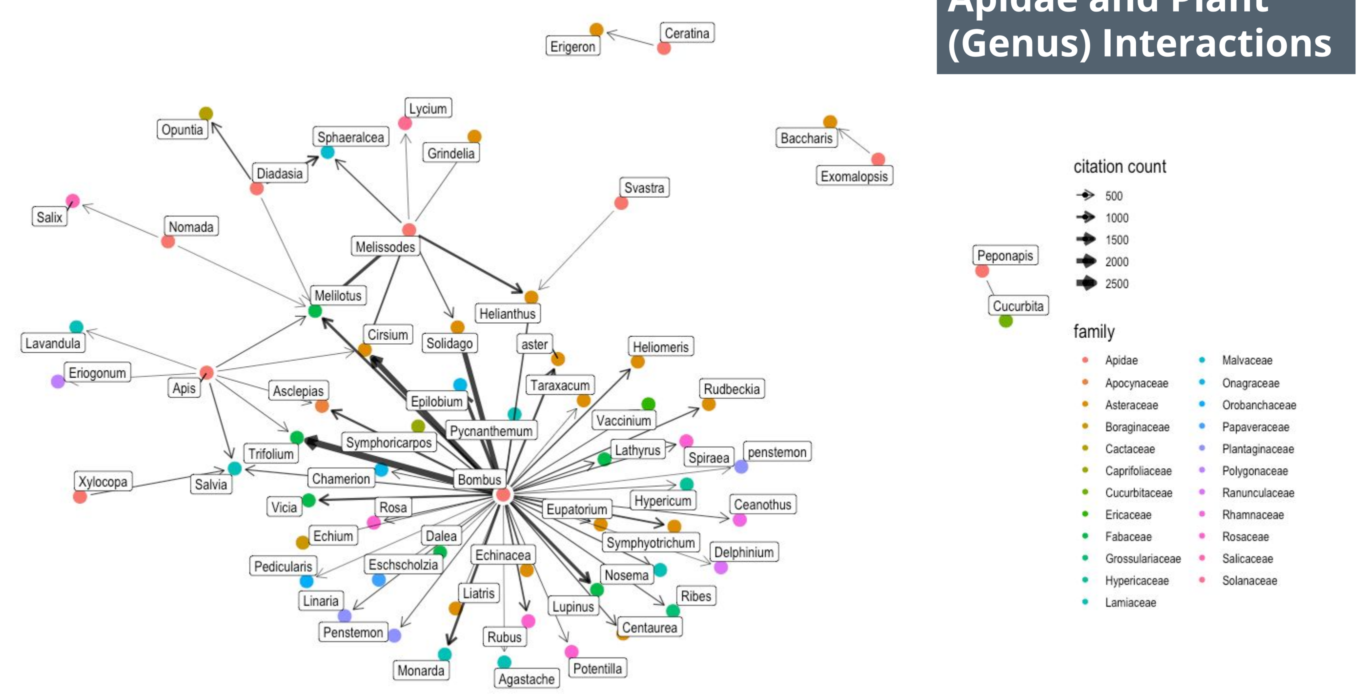
Nick Bachelder[1], Angel Chen[1], Zoe Fang[2], Mitchell Rapaport[1], Samantha Solomon[1], Joshua Bang[1], Michelle Lee[3] & Katja C. Seltmann[4]

[1]Statistics and Applied Probability, University of California, Santa Barbara,[2]Department of Linguistics, University of California, Santa Barbara, [3]Ecology Evolution and Marine Biology, University of California, Santa Barbara, [4]Cheadle Center for Biodiversity & Ecological Restoration, University of California, Santa Barbara

INTRODUCTION

Large, open-access biological datasets, like those hosted by Global Biotic Interactions (GloBI) at https://www.globalbioticinteractions.org/data, have become increasingly accessible due to greater data collection, compilation, and improved storage. These data providers serve to better inform our understanding of species occurrences, interactions, and ecosystem structure. **We sought to develop a better understanding of bee specialization of floral resources, an evolutionary trait in bees that underscores the stability and structure of pollinator interaction networks** (Fig. 1). GloBI and expert-compiled data were compared to better understand patterns in resource specialization. Within our analysis, it became clear that there was a significant correlation between the number of recorded observations and number of unique interactions for any specific bee family, genus, or species.

Fig. 1: Apidae and Plant (Genus) Interactions



How can we leverage big data, and in particular biological interaction data, to better understand plant-pollinator specialization?
How well defined are the terms we use to describe plant-pollinator specializations?

METHODOLOGY

- Downloaded GloBI bee data (n = 294,914 bee interactions).
- Determined what types of bees were 'specialists' or 'generalists' using networks and heatmaps.
- Degree of specialization (degree value) is defined as the number of unique plant family interactions a bee species had within the GloBI data.
- Set different cut points (no cut, 5 interaction cut) for how many interactions a bee species had to have with a plant family in order to contribute to the degree count.
- Identified an expertly defined list of specialist bees in America authored by Jarrod Fowler (Fowler, 2020).
- Transformed degree value to separate the 'specialist' and 'generalist' groups for better classification accuracy. Fig. 2 is the formula describing how the transformed degree was generated.
- Trained supervised machine learning models (Decision Trees, Support Vector Machine, Logistic Regression) using the number of recorded observations, degree values, and the expert defined list to create a quantitative classification system to define specialist and non-specialist bees.
- In order to obtain a consistent conclusion, differences in results between various taxonomic levels of bees were analysed.

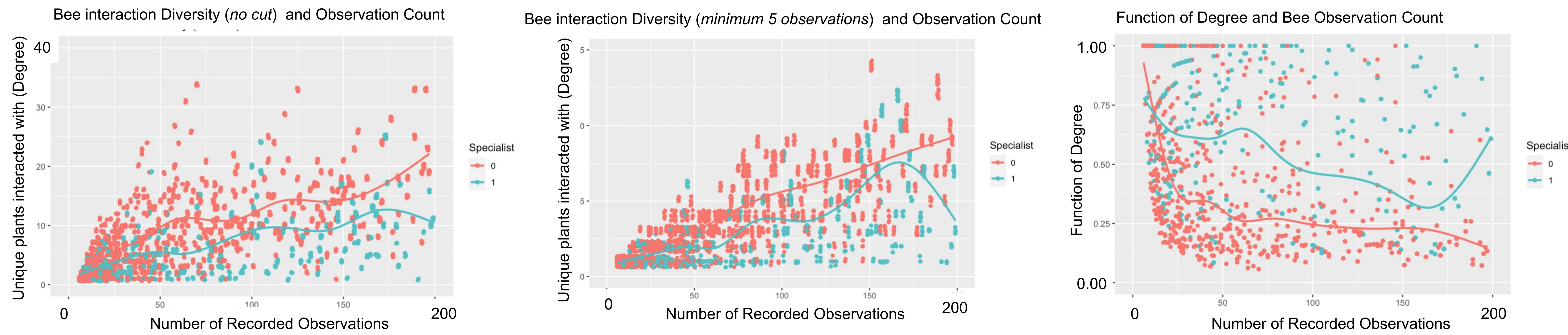
N = *Number of Bee Species Citations*, I_i = *Number of Interaction Citations with Plant i*

$$\text{Transformation Function of Degree for Bee Species} = F(N, I_i) = \sum_{i=1}^{I} (I_i / N)^2$$

Fig. 2: Transforming Degree Value

RESULTS

Fig. 3: Modeling & Validation Accuracy



Degree w/ No Cut	Degree w/ Citation Cut	Transformed Degree
Random Forest: 0.625	Random Forest: 0.630	Random Forest: 0.762
Decision Tree: 0.658	Decision Tree: 0.739	Decision Tree: 0.783
Logistic: 0.716	Logistic: 0.748	Logistic: 0.784

Figs. 3 and 4 demonstrate our results including:

- Strong positive correlation between number of recorded observations and the number of plant families visited
- Progressive steps of feature engineering for degree of specialization made modeling better
- After transforming our degree, modeling achieves below 80% validation accuracy
- Binary classification of bee pollinator specialization not consistent

(red box) = Area where bee citation sample size is too limited for a significant prediction

(dashed lines / arrow) = Decision Boundary with some margin of uncertainty

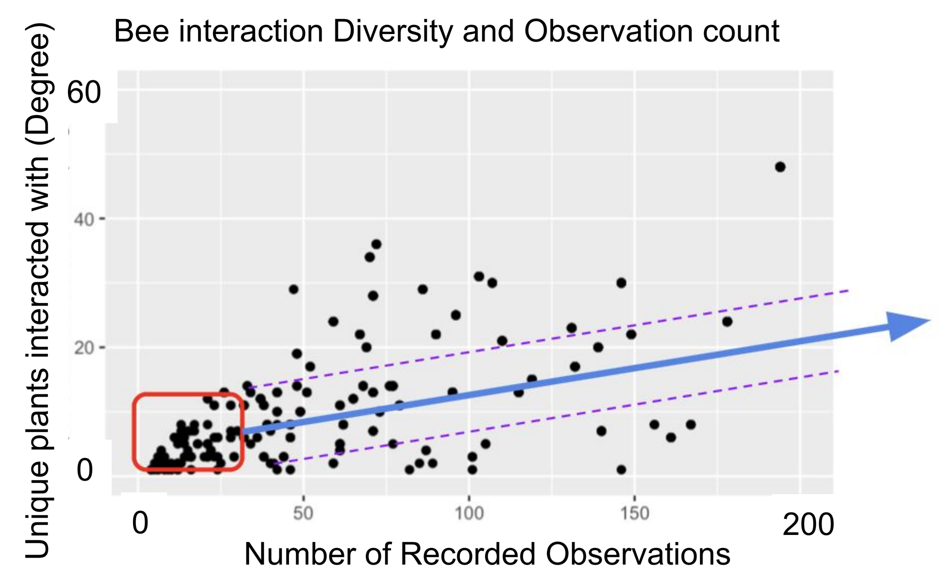


Fig. 4: Modeling & Validation Accuracy

Fig 5 are some summary statistics from the bee-plant interactions found in the GloBI dataset including the top 10 bee species and the top 10 plant families recorded. Recorded observations or records in GloBI come from many sources including published literature, observations (including community science) and natural history collection records.

	beeFam	beeSpecies	numObservations
1	Apidae	Apis mellifera	11692
2	Apidae	Bombus impatiens	6159
3	Andrenidae	Andrena wilkella	5574
4	Andrenidae	Andrena crataegi	4678
5	Halictidae	Halictus ligatus	4286
6	Apidae	Bombus bifarius	3242
7	Halictidae	Halictus confusus	2964
8	Apidae	Bombus griseocollis	2820
9	Andrenidae	Andrena miserabilis	2668
10	Andrenidae	Andrena nasonii	2473

	plantFam	numObservations
1	Asteraceae	76668
2	Fabaceae	42710
3	Rosaceae	20867
4	Brassicaceae	13936
5	Lamiaceae	13903
6	Salicaceae	11518
7	Apiaceae	7627
8	Boraginaceae	6218
9	Ericaceae	5974
10	Malvaceae	5822

Total Number of Bee Species: 2504
Total Number of Plant Families: 249

Fig. 5: Top 10 Most 'Common' Bee Species & Plant Families Within GloBI

DISCUSSION

We found that while bees expertly classified as bee specialists constantly visited fewer plant families than other bees in the GloBI dataset, there are clusters of species that diverge from the expected trend. These findings indicate that observer bias, on a global scale, can skew our definition of resource specialization or generalization. Moreover, large, open-access datasets like GloBI can change our previous understanding of biological interactions and systems by accessing novel data sources and aggregation.

Although our models weren't precisely accurate in predicting a bee species as a specialist or generalist, they pave a possible future path of predicting a rank (spectrum) of specialization by using number of observations and number of different plant taxa visited. This degree of classification allows for some biases unlike the binary classifier method we attempted.

References:
Jorrit H. Poelen, James D. Simons and Chris J. Mungall. (2014). Global Biotic Interactions: An open infrastructure to share and analyze species-interaction datasets. Ecological Informatics. https://doi.org/10.1016/j.ecoinf.2014.08.005

Fowler, J. (2020). Pollen Specialist Bees of the Western, Central and Eastern United States. https://jarrodfowler.com/pollen_specialist.html

CONCLUSION

We found several sources of bias such as how rare bees have fewer recorded observations, making them harder to classify.
We found a strong positive correlation between the number of recorded observations and the number of plant families visited by that same bee species.
Defining specialization of bee pollinators as binary (as generalists or specialists) is not an effective method for describing bee - plant interactions.
In observing the supposed bee specialists from the Fowler dataset, the GloBI dataset showed us that these bees interacted with more than just their presumed specialized plant family.
With all of these findings, it is evident that big data can help challenge our assumptions, but also should be carefully utilized to avoid leaving unsound assumptions unacknowledged.

This project was supported by the National Science Foundation projects: Central Coast Data Science Partnership: Training a New Generation of Data Scientists (DSC:1924205) and Digitizing collections to trace parasite-host associations and predict the spread of vector-borne disease (DBI:1901926)