Leveraging Biological Interaction Datasets to Quantify Plant Specialization of Bees

Nick Bachelder, Angel Chen, Zoe Fang, Mitchell Rapaport, Samantha Solomon

Sponsors: Katja Seltmann and Michelle Lee
from Cheadle Center for Biodiversity & Ecological Restoration (CCBER)
Mentor: Joshua Bang

1. Introduction & Abstract

1.1 Abstract

Large, open-access biological datasets, like those hosted by Global Biotic Interactions (GloBI), have become increasingly accessible due to greater data collection, compilation, and storage. These databases serve to better inform our understanding of species occurrences, interactions, and ecosystem structure, broadly. In this work, we leverage GloBI data to better understand patterns of pollination, a biologically and economically essential biotic interaction between plants and pollinators. Specifically, we sought to develop a better understanding of bee specialization of pollen, an evolutionary trait in bees that underscores the stability and structure of pollinator interaction networks. We compared GloBI and expert-compiled data to better understand patterns in resource specialization. We then trained various machine learning models (Decision Trees, Support Vector Machine, Logistic Regression) to create a defining line between specialist and non-specialist bees. In addition, data transformation was also useful for more clearly differentiating our specialization groups.

Through our exploration of GloBI, we found several sources of bias, including the limitations of community data collection and scarcity of rare bees. We found a strong positive correlation between the number of sources (i.e. literature, natural history collection) citing the interactions of a bee species and the number of plant families visited by that same bee species. We also found that while expert classification of bee specialists visit fewer plant families than other bees in the GloBI dataset, there are clusters of species that diverge from the expected trend. Our trained models suggested that binary classification was not completely effective in determining the label of “specialist” or “non-specialist” for all bee species. 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.

1.2 Introduction

The organization that is supporting our project is UCSB's CCBER (Cheadle Center for Biodiversity and Ecological Restoration). CCBER has various areas of research related to ecological restoration, species endangerment, and biodiversity. For our Data Science Capstone Project, we are focusing on the research area of biodiversity informatics. This entails the usage of taxonomic databases and geographic information systems. We study digital interaction records to assess the state of our environment and its change over time. We can extract significant insight through the analysis of this open source data.

The primary goal of our project is to take advantage of big data (biological interaction data) to better understand plant-pollinator specialization, specifically in bee species. Additionally, we use exploratory data analysis to interpret our raw data and develop initial assumptions. We eventually reevaluate our initial assumptions through visualization, transformations, and modeling.

Our work is relevant to the field of big data analytics and the growing importance of biodiversity. The vast library of biological interaction records provides a universal outlook on biodiversity at the global scale due to the high volume of observations. At the same time, however, there are limitations to this big data, such as human error and observation bias. As a result, we become more aware of the assumptions we need to make in these cases, serving to guide the work of future researchers. Our findings will affect downstream research related to the stability and structure of plant-pollinator relationships, hopefully helping to advance the field of biological data science.

We seek to question the utility of classifying bees strictly into specialization groups, and potentially suggest more useful alternatives. We hypothesize whether or not there is significant benefit in classifying certain bee species as specialists or non-specialists, or if we can recognize that there may be more effective methods.

The domain knowledge required to tackle this project is related to the understanding of the taxonomic hierarchy (Kingdom, Phylum, Class, Order, Family, Genus, Species) and the vital role bees play in global biotic interactions. It was also necessary to know the collection methods and sources of our data, and how these were transferred into digital formats. Through our sponsors' knowledge and experience, we were able to get context on how this data has been used historically and ways to measure the quality and usefulness of it.

We can define the terms that we use in this report. First of all, we frequently refer to the specialist versus the generalist. A specialist is an organism that may have a more specific diet or ecological constraints while a generalist can thrive in more diverse conditions. Next, the type of data that we work with is interaction data. So, for each observation, there is a source, a target, and an interaction type. The data is nodal and directional, meaning that the nodes are the sources and targets, while the links are the interactions. We can weigh these interactions by citation count, or the amount of interactions a specific bee has with another plant. We can also think of citation count as the frequency. Further, the degree of specialization is the number of unique plant families a bee interacts with.

Considering all of these concepts and definitions, we can look more into the nature of plant-pollinator networks. For the health and well-being of the ecological community, it is vital that there is a proper balance between specialization and generalization in plant-pollinator relationships.

Furthermore, in the field of biodiversity informatics, various machine learning methods have been developed to promote research and to advance findings. In our case, we will be using machine learning models to more robustly define what 'specialization' is.

The two datasets we have used in our analysis are our GloBI data and our Fowler data. GloBI refers to Global Biotic Interactions. This data comes from literature reports, human observations, and natural history specimens. This is aggregated data using open source software. In general, it provides a wide supply of data to have a more comprehensive understanding of plant-pollinator networks. Our second dataset(s) come from Jarrod Fowler. This serves as somewhat of an "expert" specialist dataset. There are three tables for each of three regions in the continental United States: West, Central, East. We have used this as a standard way to quantify specialization.

Two of our main questions of interest over our project have been: How do we quantify specialization? Is there a binary result, or is there a spectrum? In addition, we would like to look at which forms of bias are significant in our analysis. What assumptions do we need to make to account for these biases, and how much should we trust these assumptions?

2. Data

2.1 Data Overview

There are two datasets we use during our research. The first dataset is a 15G csv file downloaded directly from the GloBI website. There are around 8 million rows and 84 columns, where each row is a recorded biotic interaction, and each column contains the participant information of the interaction. Most of the columns contain taxonomic information, namely the IDs and names of their kingdom, family, genus, etc. There are two columns that describe the type of interaction. The rest of columns contain information about the locations, time and reference of the interactions. The reference of the data comes from three main sources: scientific journals, museum records, and human observations. The second dataset is three lists of specialist bees in eastern, central, and western America provided by a bee expert Jarrod Fowler. Each list contains the taxonomic information of the specialist bee.

2.2 Data Processing

Since we only want to focus on interactions between bees and plants, we create a new data frame from the GloBI data. We filtered the data based on the taxonomic information. We want to have either a bee as a source that interacts with a plant target, or a plant as a source that interacts with a bee target. For example, Apis Mellifera, more commonly known as honeybees, eat Sixalix Atropurpurea, which is a type of sweet flower. We filter the bees by controlling the taxon family name to be one of the seven families of bees (Andrenidae, Apidae, Colletidae, Halictidae, Megachilidae, Melittidae, Stenotritidae) and filter the plants by controlling taxon kingdom name to be Plantae. There are duplicates in the data frame, which we later found out are caused by merging errors on GitHub, so we ended up dropping the duplicates. After cleaning the data, we have around 300 thousand rows, which is around 3.8% of the original data.

2.3 Data Quality

During our process of exploring and cleaning the GloBI data, we identified several imperfections. As mentioned before, there are some duplicate entries, which we ended up dropping. And there are some missing values in the genus and species columns, in which case the bees or plants are less specifically defined. Since our objective is to identify generalist and specialist bees, these missing values would pose a problem when we want to make our predictions on more specific levels. The geospatial and temporal information is also sparse, so it is hard to determine the relationship between these factors and bee behavior. Another major issue with the data is that some entries have flipped source and target. For instance, there are several interactions cited as a plant eats a bee. After looking into these instances, we found out the source and target information were mistakenly flipped and decided to reverse their roles in the analysis.

It is our assumption that the Fowler expert list is true – yet not exhaustive – and that GloBI data roughly represents global bees' interactions with plants. One of the potential biases we discovered in the data is underrepresentation of uncommon bees. For example, Bombus is the largest genus in the Apidae family, but there is one species in Bombus in Northern America, and there are plenty of recorded interactions of that species, whereas other spices in Bombus are far less common and therefore have very few recorded interactions. In addition, bee interactions that happen in less observable places of time may be cited less despite equal occurrence. Another interesting bias we learned from our sponsors is that there might even be a negative bias against common bees in human observations, as some of the bee researchers choose to neglect the occurrence of a common bee just because how common that bee is. All of these biases would play an important role in our process of deciding the pollination preference of the bees, since we heavily rely on the number of interactions the bees of interest have with different plants.

3. Methods

In order to get a better understanding of the data we were dealing with, we calculated basic statistics of the GloBI dataset. By late April 2021, there were 295,938 rows and 89 columns in the dataset after filtering the observations to just bees and plants. The top three bee families with the most species are Apidae (823), Megachilidae (536), and Halictidae (461). Meanwhile, the top three bee families with the most citations were Apidae (86,487), Andrenidae (54,446), and Halictidae (42,270). This intuitively makes sense because the bee families with more citations should be more diverse in species. We also calculated the bee species with the most citations, and found that the Apis mellifera bee from the Apidae family had the most citations by far, with 11,692 occurrences. Additionally, the plant family that had the most citations by far is Asteraceae, with 76,668 citations. Interestingly, the bee-plant interactions in the GloBI dataset seem to be dominated by the Apis mellifera bee and the Asteraceae plants.

Continuing on with our exploratory analysis, we also created heatmaps and network graphs with the GloBI data to help us figure out what types of bees are specialists or generalists.

These visualizations will also allow us to see if there are any patterns in bee-plant interactions. Under our capstone sponsors' guidance, we investigated these relationships at the bee genera to plant genera level. On the next page is an example of a network graph that we created.

Figure 1 shows genera from the Apidae bee family interacting with plant genera, color-coded at the family level. The cut-off point that we chose was interactions that have occurred more than 200 times because otherwise, the network graph would be too complex and hard to understand. We can tell from Figure 1 that the Bombus bee interacts with many different plant genera and families, making the Bombus bee appear to be a generalist genus. On the other hand, Peponapis, Exomalopsis, Ceratina, Svastra, and Xylocopa bees all interact with one specific plant genus each, which may indicate that they are specialist bees. Since this network graph only shows the interactions for one bee family, we have included visualizations for other bee families in the appendix.

One intriguing observation that we noticed from our visualizations is that classifying bees as generalists or specialists is a complex problem that cannot be solved by categorizing these bees into two simple categories. For instance, referring back to Figure 1, Melissodes is a bee genus that has over 200 interactions each with the Sphaeralcea, Lycium, Grindelia, Melilotus, Helianthus, Solidago, and Cirsium plant genera. At first glance, one might think that Melissodes bees are generalists since they interact with several different plant genera. However, upon closer inspection, out of the seven plant genera they interact with, four of them belong to the same plant family, Asteraceae. Melissodes bees may interact with lots of different plant genera but it seems that they mainly interact with one plant family. We learned that we must take into consideration the taxonomic levels when trying to classify bees and that perhaps specialization and generalization lie in a spectrum rather than binary categories.

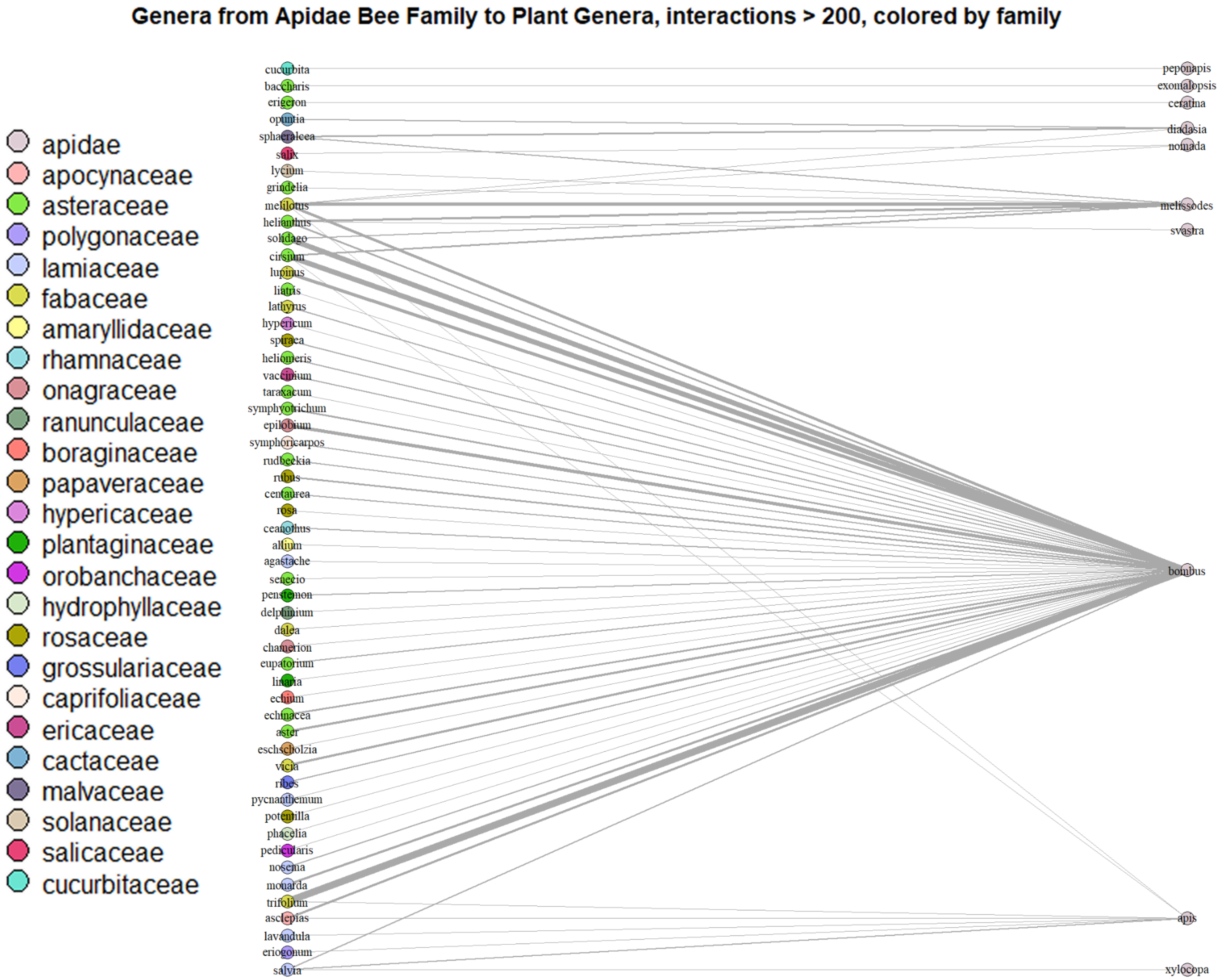


Figure 1: The interactions between the genera from the Apidae bee family and all plant genera, presented as a network graph. The nodes are color-coded by family. The left nodes represent plants while the right nodes represent the genera from the Apidae bee family.

By doing some exploratory analysis and making these visualizations, we also realized that there are multiple sources of biases in the GloBI and Fowler datasets. As mentioned before, some bees may be over or underrepresented in the GloBI dataset due to imperfect detection. For the Fowler datasets, one important source of bias is that the bees listed in there are from the United States only, so conflicts may come up as we compare the Fowler datasets with the global data in GloBI. Fowler's data entry is not consistent, as well. In the column for plant information, not every plant has its family, tribe, or genera all listed, making it difficult to extract plant information from Fowler. All of these biases play an important role in our process of deciding the pollination preference of the bees so we need to take these into account for our next steps.

We could not determine for certain which types of bees are specialists or generalists from visualizations alone, so we needed statistical methods and models to support our hypotheses. We wanted to confirm our suspicion that the number of times a bee appears in the dataset (frequency) was correlated with the number of different plant families it interacts with (degree). To examine this relationship, we attempted to calculate the Spearman's rank correlation coefficient between these two variables, which turned out to be pretty strongly correlated at 0.782. In order to further our investigation, we created a new dataset that merged information from both GloBI and Fowler. Since the bees in the Fowler datasets were considered specialists by an expert, we were able to create labels for the bees, categorizing them as either specialists or generalists. Then we used machine learning techniques to create models such as logistic regression, random forest, and decision tree to try to classify bees as specialists or generalists/non-specialists based on degree and frequency.

4. Analysis / Results / Interpretation

Since a large majority of our project was spent developing the questions we wanted to answer and organizing methods to effectively answer them, the interpretation of our results was developed progressively through our exploratory analysis. Specifically, to recap, we feel that our results lead to conclusions related to two large motivating questions. (1) *How well defined are the terms we use to describe plant-pollinator specialization, and specifically how useful are the terms 'specialist' and 'generalist' in accurately communicating the pollination tendencies of bees?* (2) *How can we leverage big data, and in particular biological interaction data, to better understand plant-pollinator specialization?*

The results of our machine learning (ML) methods, as explained previously, led to insights on both these questions. Moreover, while the final results of our machine learning accuracy led to our conclusion that the binary classification of bees into generalist and specialist groups is not specific enough, the ML obstacles we encountered before coming to this conclusion are equally useful.

Initially, when training our model to classify each bee as a specialist or a generalist using citation count and degree, we had defined degree simply as the raw number of interactions a bee species had with unique plant families in the GloBI set. When we did this, we found that many bees, including many that were identified as specialists by Fowler, had very high degree counts. This was odd, since in theory specialists would interact with very few unique plant families. We realized that we had neglected to account for the possibility that in a large dataset like GloBI there would be some misidentified interactions. In other words, for many of the bee species there were interactions with unique plant families with very few citations. These unique interactions were majorly inflating the degree counts for all bees, so we decided to redefine degree to reduce this error.

To do this, we found that creating a minimum citation cut point for each unique plant-pollinator interaction to contribute to the degree helped to account for misclassification. After testing different cut points, we found that a cut point of five citations per unique interaction was the best at reducing problems caused by misclassification and still accounting for all legitimate unique interactions. When retraining models on this degree, our specialist and generalist groups became further separated.

Still, though, there remained many seemingly misclassified specialists and generalists. After discovering this, we began trying to find ways to separate the data to minimize these inconsistencies. One such way was to create a function of the degree specialization to transform the data in such a way that would further separate specialist and generalist groups. By summing the square of the percentage of citations an interaction accounts for, we were able to better separate the clusters. This data transformation is noted below, and was used as a feature to replace degree in our modeling, and even further separated the groups which helped with model accuracy.

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

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

Figure 2: Above is the formula for the data (degree) feature transformation we used to best separate specialist and generalist groups in preparation for modeling/classification.

Below is the model training dataset for modeling with the three methods from above.

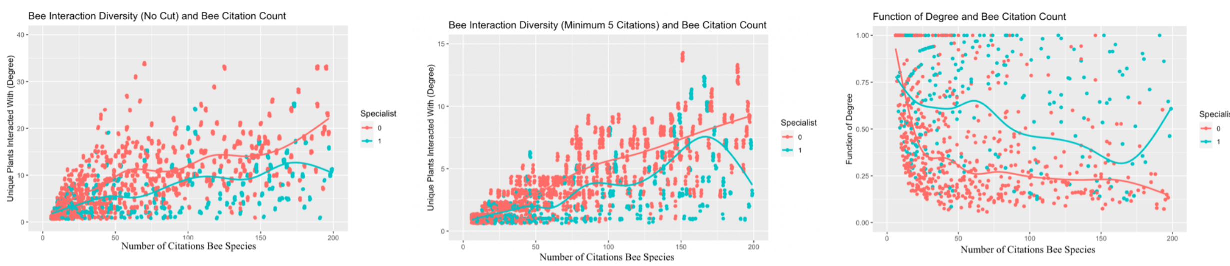


Figure 3: Above are the training datasets of specialists and generalists. In red are Fowler defined generalists and in blue are the specialists. (Left) Training data with degree defined with no citation cut. (Middle) Training data with degree defined with a citation cut of five. (Right) Degree transformed data.

Notice that with each progressive step, we were further able to separate the two groups. Below are the best cross validation modeling scores for each of the three methods for logistic, random forests, and decision tree modeling.

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

Figure 4: Results of machine learning techniques (Random Forest, Decision Tree, Logistic Regression) on training sets listed above.

While it's clear that each progressive step of our feature engineering made our modeling better, our biggest observation from the above accuracies are (1) even after transforming our degree we are still achieving below 80% validation accuracy and (2) the increases in accuracy from our degree with citation cut to our transformed degree are trivial for our purposes.

While our data transformation allowed us to further separate the clusters and achieve a few percent increase in validation accuracy, the feature is far less practical to understanding how the degree of specialization affects specialist labels. Since the main objective of modeling is gaining an intuitive understanding of how specialization is defined, this trade off is not worth it.

Moreover, these somewhat low validation accuracies led us to potentially our biggest conclusion of the project: *the classification of plant-pollinators into binary groups of either specialists or generalists does not accurately describe the variety of specialization that characterize various species.* Since, even through modeling, we couldn't consistently match Fowler's classification of bees to some definition of number of unique plant interactions within real GloBI biological data, we feel that this conclusion is sound.

Regarding our original motivation, we found that the process of training models to classify bee specialization helped us to identify many ways we can leverage big data to better understand plant-pollinator specialization. Using data from GloBI we were able to explore how real-life citations compared to scientific literature (ie. Fowler) and subsequently identify inconsistencies/inaccuracies within such literature. For example, our machine learning models, while not completely successful in determining a binary cutpoint of specialization, may be quite useful for finding new bees that should be reclassified as more specialized than they are in academic literature like Fowler's.

Finally, to complete our analysis we have to mention the various assumptions we made about GloBI and Fowler data in our conclusions. When reviewing our results two noteworthy assumptions were identified. Recall that Fowler is a list of identified bee specialists. In other words, while surely the data set is as complete as possible, it is almost impossible that it is an exhaustive list of all specialists. One clear assumption we made that is likely a limitation is that all 'non-specialists', as shown in red on the plots above, are generalists. In other words, since we assumed all bees that were excluded from Fowler's list to be generalists, we may be misclassifying some bees as generalists if they are lesser known specialists that were omitted from Fowler's database. Accounting for this assumption, we defined our 'weird bees' as specialists with high degree counts and worried less about 'generalists' with low degree counts.

The second assumption that affects our analysis regards geography. Since the Fowler dataset was originally split into three regions of the United States, we initially had very little issue finding which bees were specialists in different regions and making a comprehensive list of specialists. In the process, though, we found some bees that were identified as specialists exclusively in some regions and not others. To make a comprehensive list, we had to assume that if a bee was a specialist in one region, it was also a specialist in another. While we certainly have this in mind while making conclusions, and consider it a potential possibility for the number of 'weird bees' we identified through modeling, we believe this assumption is fairly sound. Our sponsor leaders, who have backgrounds in biology and bee studies, believed this assumption likely held true.

5. Conclusion and Future Work

Through our exploration of GloBI, we found several sources of bias, including the limitations of community data collection and scarcity of rare bees. We found a strong positive correlation between the number of sources (i.e. literature, natural history collection) citing the interactions of a bee species and the number of plant families visited by that same bee species. We also found that while expert classification of bee specialists visit 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.

Our main finding in our analysis of the specialization of bee pollinators is that binary classification of bee pollinators as generalists or specialists is not a completely effective method. Our model could not consistently and accurately predict a bee species as a generalist or specialist. This is due to the loose definition of what a specialist is in the Fowler dataset. 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. In simple terms, specialization of bee plant interactions is a spectrum. A bee species is more of a specialist if it goes to a specific plant family more than others, but we didn't find any bee species that are absolute specialists.

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 of specialization by using number of observations and number of plant families visited. This degree of classification allows for some biases unlike the binary classifier method we attempted. We started looking for ways to start this classification method but didn't put much time into it. Such a model could be applied to more than just bee-pollinator networks but rather any large biological dataset. This type of model would help biologists in our overarching theme of preserving biodiversity by providing information on what species would be in need of human intervention based on their specialization of resources.

Bibliography

GloBI (2021). *Data*. https://www.globalbioticinteractions.org/data.

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

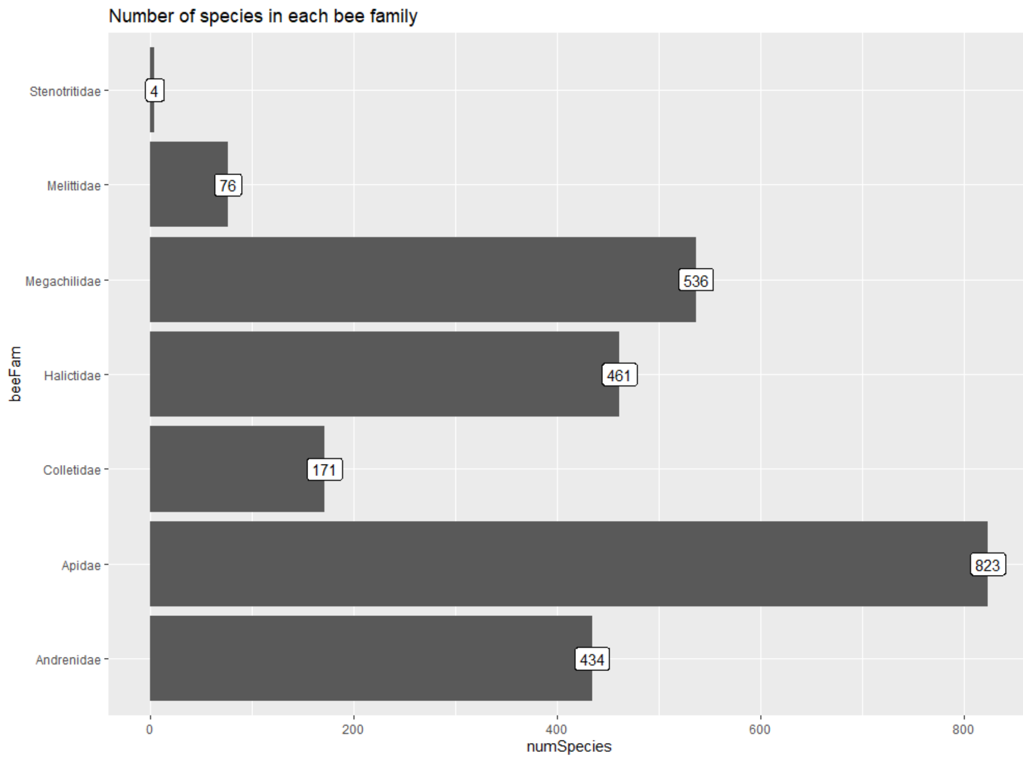
Fowler, J. (2020). *Pollen Specialist Bees of the Central United States*. https://jarrodfowler.com/bees_pollen.html

Fowler, J. & Droege, S. (2020). *Pollen Specialist Bees of the Eastern United States.* http://jarrodfowler.com/specialist_bees.html

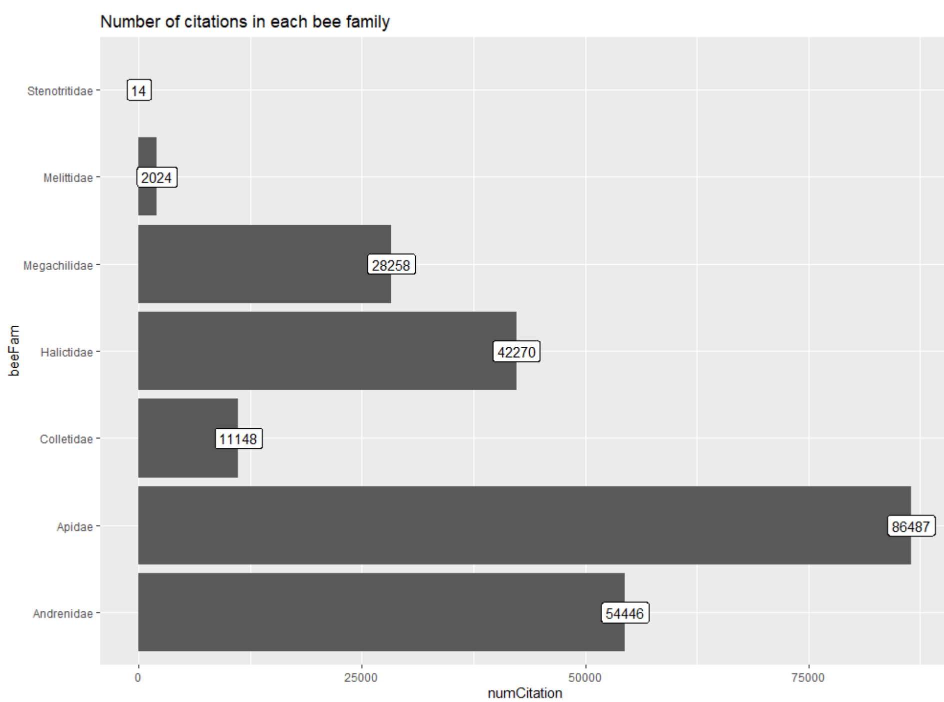
Brosi, B.J. (2016). *Pollinator specialization: from the individual to the community.* New Phytol, 210: 1190-1194. https://doi.org/10.1111/nph.13951

Cheadle Center for Biodiversity and Ecological Restoration. (2020). *Open Geospatial Data*. https://www.ccber.ucsb.edu/research-areas

Appendix



Appendix Figure 1: Here is a barplot showing the number of species in each bee family.



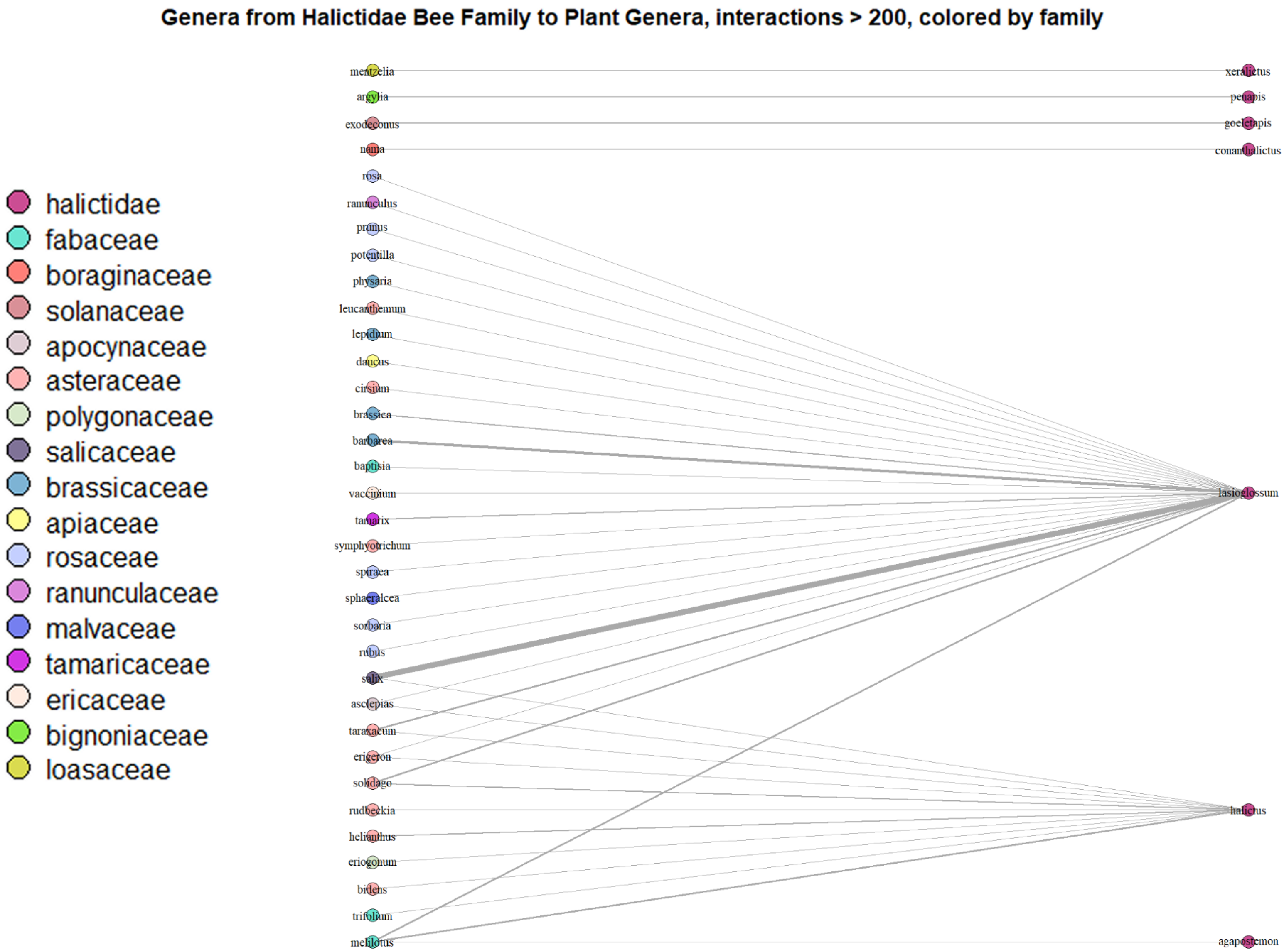
Appendix Figure 2: This is another barplot showing the number of citations/interactions for each bee family.

Top 10 bee species and plant families with most citations

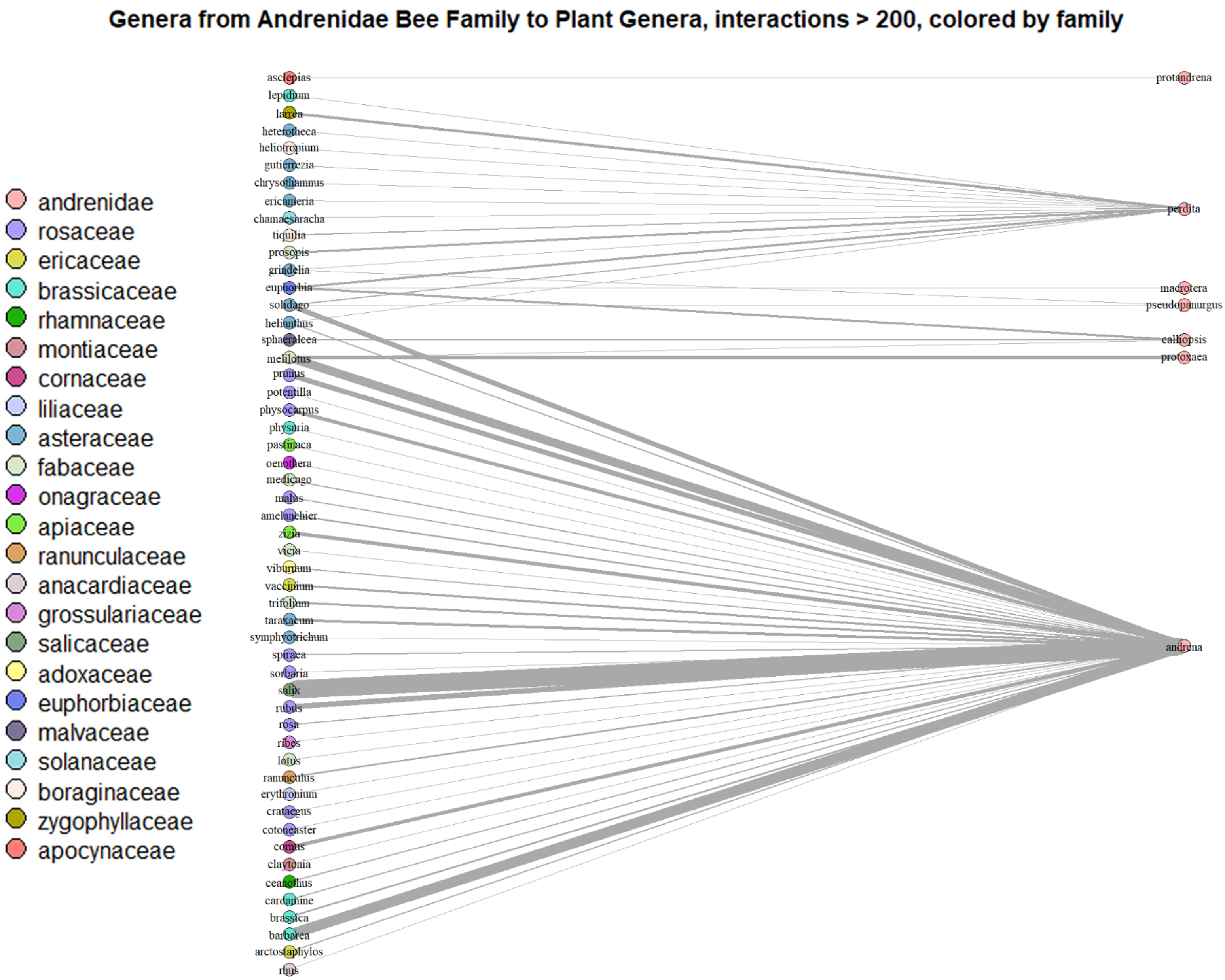
	beeFam	beeSpecies	numCitation
1	Apidae	Apis mellifera	11692
2	Apidae	Bombus impatiens	6159
3	Andrenidae	Andrena wilkella	5574
4	Andrenidae	Andrena crataegi	4678
5	Halictidae	Halictus ligatus	4286
6	Apidae	Bombus bifarius	3242
7	Halictidae	Halictus confusus	2964
8	Apidae	Bombus griseocollis	2820
9	Andrenidae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

	plantFam	numCitation
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

Appendix Figure 3: The top 10 bee species with the most citations (left) and the top 10 plant families with the most citations (right).



Appendix Figure 4: A network graph showing the interactions between the genera from the Halictidae bee family and all plant genera, presented as a network graph. The nodes are color-coded by family. The left nodes represent plants while the right nodes represent the genera from the Halictidae bee family. The cut-off point is at 200 citations/interactions.



Appendix Figure 5: Another network graph showing the interactions between the genera from the Andrenidae bee family and all plant genera, presented as a network graph. The nodes are color-coded by family. The left nodes represent plants while the right nodes represent the genera from the Andrenidae bee family. The cut-off point is at 200 citations/interactions.

Regional Specialist Frequency by Month

All Fowler Freq. Table by Month

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
0.20%	0.70%	6.60%	12.30%	14.60%	16.00%	15.80%	15.70%	11.80%	4.80%	1.50%	0.20%

Central Freq. Table by Month

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
0.10%	0.40%	6.00%	11.20%	13.80%	15.60%	16.50%	16.90%	12.60%	5.10%	1.50%	0.20%

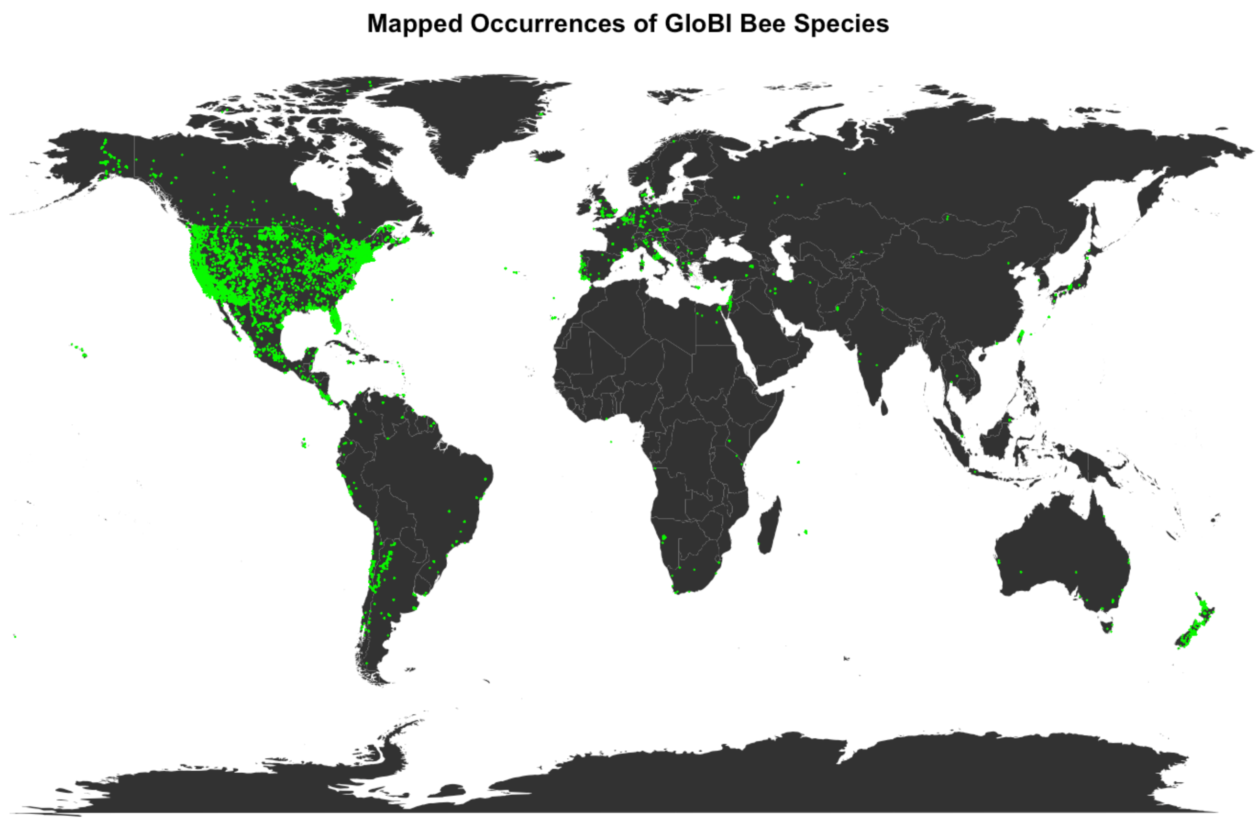
Eastern Freq. Table by Month

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
0.20%	0.20%	4.60%	10.90%	14.30%	15.50%	15.70%	14.90%	13.60%	7.30%	2.50%	0.20%

Western Freq. Table by Month

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
0.20%	1.00%	7.80%	13.60%	15.40%	16.50%	15.30%	15.10%	10.40%	3.50%	1.10%	0.20%

Appendix Figure 6: These tables show the spread by month of specialist observations from the Fowler datasets. It is apparent that most of the bee specialist species appear during the spring and summer months. This provides greater temporal understanding of the data.



Appendix Figure 7: The map shows the locations of the GloBI bee to plant citations in the world. This is based on the provided latitude and longitude data in GloBI. We can see that most of the observations come from North America.