Globi Database:

**Running title (< 45 characters):** XXX

Graziella V. DiRenzo1\*, Michelle Lee2, XXX….., & Katja Seltmann

1 U.S. Geological Survey, Massachusetts Cooperative Fish and Wildlife Research Unit, University of Massachusetts Amherst, USA

2 ….

\* Corresponding author; email gdirenzo@umass.edu

**Target journal:** Methods in Ecology & Evolution?

Type of article: Standard article

Number of words (6000 - 7000 words):

Number of words abstract (max = 350 words):

Number of figures & tables: X figs. & Y table

**Competing interest statement**

We have no competing interests.

**Abstract**

1. Species

**Keywords:** imperfect

**Introduction**

* Occupancy models are powerful tools that take into account imperfect detection
  + Require standardized data collection and meeting specific assumptions
  + If violated- grave consequences on parameter estimation and ecological inference
  + However, researchers have come up with ways of using occupancy models for non-standardized methods typically used for opportunistic collections
  + For example, to use historical data, authors directly re-survey sites that were surveyed in historical periods or constrained analyses to locations where 2 or more sampling events occurred within a calendar year
  + Replication ensures that occupancy and detection can be estimated
* Opportunistic community science sampling, atlas data, historical museum records contain presence-only data, species non-detections must be inferred
  + One way to do this infer a non-detection for a particular species if a different species was observed at that same site on the same date
  + Other problems:
    - Sampling effort differs by time interval
      * Solution: add covariates
      * Solution: filter the data by removing vistis with only a single observation
      * Solution: contrain analyses to species that meet some minimum number of observations
  + Other problems with historical dataset:
    - How to define a site?
      * Spatial resolution?
    - A museum record is only a subset of a field collection
      * When and where collected?
      * Sometimes common species are not curated
  + Scientists have been trying to use these data sets to learn about species distributions across space and time.
    - This requires a lot of sampling
    - Here, we are interested in exploring species interactions (bee-plant) and what kind of bias there is in documenting these relationships.
      * What are the next steps for these aggregated databases?
      * What holes are there left in our biological understanding?
      * Which bees do we not have enough data for
      * Only have data for large and common bees?
      * Taxonomic biases
      * Looking at the number of collectors in the data- looking to see what they predominately collected
* Other problems of opportunistic datasets
  + 1. Sampling (taxonomic) bias
    - Particular species may be sampled more frequently than others because more is known about them or inference is desired on that species
    - Apidae is over-represented with many more records for this family than others
  + 2. Detection bias
    - Species detectability changes over time and space as a result of observers or number of surveys
    - Number of observers, quality of observers, length of survey, survey conditions
  + 3. Spatial bias
    - Particular locations are more easily visited
    - Some areas are never visited
  + As a result, patterns may be masked (or there are false patterns) because of observation effort
  + So, we need to account for species non-detections in the database across bee species
* Objectives
  + We are interested in determining if social bees are more generalist (and visit more flowers) than solitary bees.
  + We will be using a multi-species occupancy model for the analysis
    - We will estimate:
      * psi = The probability a bee species interacts with a plant species
      * p = The probability that a study documented the bee-plant interaction
  + We will be exploring hypotheses related to bee and flower characteristics that relate to the ecological process (i.e., what characteristics relate to a species degree of generalism?) and the detection process (i.e., what characteristics relate to species detectability?)
    - Bee
      * Size
      * Sociality
      * rare/common, pilosity, seasonality, coloration
      * Phenology
      * Physiology (the change in thermal position, thermal range)
      * Short vs long tongued
    - Flower
      * Shape
      * Size
      * Density
      * Color
* Citation characteristics
  + Number of unique collectors per citation
  + List of collectors and determinors
    - Katja can identify who is a good taxonomist – like Robin and others- some confidence scale in the determination of the bees
  + Figure out how long they were in the island
  + Event date- when the observation occurred
    - if everyone is collecting in the spring, there is a chance they would miss winter and summer bees
* We can also ask the same question but from a plants perspective: which plants support the bee diversity?
  + What are the keystone plants?
* Forbidden links- Bartomeaous model
  + Some of those interactions that would never occur no matter what
  + Do some bees and flowers never interact because of their phenology?
  + If we see they are phenologically separated, we would have a huntch that would be the case

“they inferred non-detections for time intervals where there is no evidence that the site was visited”

* Guzman et al. 2021

“Bumble bees are a charismatic 99 insect species that have been collected by researchers and independent naturalists for hundreds of years, but, as with any taxon, it is possible that agriculturally important species (e.g. *B. terrestris* and *B. impatiens*), common species, and larger or more recognizable species have greater likelihoods of detection.”

* Soroye et al. 2020

Simulation study scenarios:

Simulate the true ecological process and then generate the Globi dataset

* Then analyze the data with a variety of models
  + Accounting for different levels of detection, observer, etc.
* How does inferring non-detection for time intervals where there is no evidence that the site was visited

**Methods**

*Data:*

* To do this, we are using the Globi interactions database in combination with site checklists.
  + Observations are obtained from museum collections and research studies worldwide.
  + Studies vary in terms of objectives, study design, etc.
  + The Globi database consists of presence-only data.
* We have compiled bee & flower checklists for X number of sites, and subsetted the Globi database to those species.
  + Note that the Globi data does not need to geographically match with the bee & flower checklists of our sites.
  + Our questions relate to determining possible bee-flower interactions (but not specific to a site).
  + We use the checklists to retroactively assign species non-detections (similar to the Kery et al. 2009 paper where they ad hoc assigned bird non-detections using site checklists).
  + We are using different collections/source citations as the 'repeated survey' for each interaction.
  + In globi this will be, ecological studies (published papers), meta-analyses, different museums, and singleton observations from websites like iNaturalist.
  + We collapsed all opportunistic records by source citation- such that if a bee-plant interaction was ever documented, then it received a 1 (present) in the database, and all others were marked 0.
  + We are assuming that an observer COULD have encountered the bee-plant interaction.
  + We subset the data to a rectangular region around Santa Cruz Island, CA.

*Assumptions:*

* We do not explicitly consider space or time in the model
  + Data are collected worldwide across time, and these are records of any possible plant-bee interaction anywhere
  + Implicit assumption that bee-plant interactions are static and do not vary across space or time (i.e., no extinction or colonization events- a bee species always interacts with a plant species regardless of location and time)
* We assume that all bee and plant species occur across our entire geographic window, and that not all bee and plant species interact with one another
  + We do not explicitly consider phenology
  + We assume that the range of each bee and plant species is within the defined area
* We assume that each source citation had the opportunity to document all bee-plant interactions.
  + we inferred non-detections for bee-plant interactions (even when there is no evidence that each plant species was visited)
* What is the closed period of time over which surveys are occurring in this study? There is no space or time in the model
  + All bee-plant interactions remain consistent across all citations
  + i.e., a generalist bee remains generalist

*Model:*

*Covariates*:

Detection part of the model

* Number of unique collectors per citation
* Robin Thorpe as a covariate 1/0
* List of collectors and determinors
  + Katja can identify who is a good taxonomist – like Robin and others- some confidence scale in the determination of the bees
* How much time did these people spend in the field?
  + Katja can get it from the RAMS records
  + Figure out how long they were in the island
  + Katja will reach out to the director- J Ready- to see if they have the information
  + Not sure how long they have been keeping track
* When the observation occurred
  + Event date- when the observation occurred
  + Verbatium event date- for literature records might be filled out more- usually summer, not exact dates in their papers
  + Most will have this
  + The ones we will lose
  + Katja- if everyone is collecting in the spring, there is a chance they would miss winter and summer bees
* We won’t have any weather data for the historic data- like cloudy, or temperature, etc.
  + There is a functional weather station
  + So we could get temperature, precipitation that year, and other environmental factors
* We likely don’t have enough data- but an interesting question- are bees using different floral resources in dry years than in wet years?
  + In dry years, going to second or third favorite things
  + In wet years, might be able to go to favorite things
  + Specialization increases in dy years- related to nutrients rather than precipitation
  + Likely very dynamic – how strong the competition becomes
* Weather idea
  + Smooshing space and time together
  + All interactions are in 1 space bin
  + Wondering how we would fit the yearly, temperature precipitation
* Flower size, shape, & color
  + Bartomeous paper- flower type- tube, inflorecent- related to flower shape (clusters vs small)

Ecological part of the model

* Forbidden links- Bartomeaous model
  + Some of those interactions that would never occur no matter what
  + They try to take that into account by looking at compatiability of bee and flower biology
    - Type of bee and flower
    - Short vs long tongued
    - Do some bees and flowers never interact because of their phenology?
      * Katja says yes- there are bees that emerge really early in the year than miss flowers later in the year
      * No data for that
      * There isn’t a list of things that would never interact
      * We could persue understanding that
      * If we see a group of bees never on a plant and a group of bees on the plant- use a different dataset to look at the phenology of bees and plants- go to GBIF or something like that
      * If we see they are phenologically separated, we would have a huntch that would be the case
      * There might be a step where we ask that question
    - Looking at the majority of ultimate questions- not sure how much the forbidden links will be important
    - In terms of our discussion of data quality, and advancing biological thinking- good place
    - Might not be as important for our bigger questions
* Flower covariates
  + Predominate flower color
  + Flower size
    - Could be size or density
  + Plant trait database with a lot of that information
  + There are about 600 plants or so after they removed grasses
  + Big bees go to white flowers
  + Bartomeous paper- flower type- tube, inflorecent- related to flower shape (clusters vs small)
* Flower size, shape, & color
  + Flower size- bees don’t visit it as much, people don’t pay as much attention
  + Asters is associated with many interactions- importance of taking into account

*Simulation study*:

* We examine how our assumptions impact our inference.

**Results**

…..

**Discussion**

…..

**Acknowledgments**-Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.We thank the USGS Advanced Research Computing, USGS Yeti Supercomputer: U.S. Geological survey (<https://doi.org/10.5066/F7D798MJ>).

**Author contributions**

GVD contributed to project development, wrote the model, analyzed data, and wrote the first draft of the paper.

XXX contributed to project and model development.

XXX contributed to project and model development.

All co-authors edited the manuscript.

**Data Accessibility**

All data and code for analyses can be reproduced and accessed at either the github repository:

[https://github.com/Grace89/](https://github.com/Grace89/Bd-IL) or a USGS server following publication.

**References**

…..

**Figures & Tables**