**Project Name:** BioIndicator\_bees\_checklists

**Project Description:** Predict bee richness using bird species abundance, across eastern half of USA

**Analysis completed by:** Josée Rousseau, with guidance from Ali Johnston and Amanda Rodewald

**Project leader:** Amanda Rodewald

**Start and end date:** 2021 to 2024

**Program(s) used:** R, qGIS

**File location:** ../BioD\_Pollination/Analysis/

**ANALYSIS FLOW CHART**

|  |  |  |
| --- | --- | --- |
| **Data file** | **Program / Tool used** | **Comments (details, file locations, etc)** |
| **Extract bee data and summarize richness / abundance per grid cell** | | |
| SCAN data :  list.files(path = here("Data", "BeeData\_Conservation\_paper"), pattern = "occurrences.csv", recursive = TRUE) | ./Programs/BioIndicator/Extract\_BeeBirdLand\_data/ 1\_FilterExplore\_SCAN\_Bee\_datasets\_20231004.Rmd | This step involves filtering the SCAN data:   * Selected records from 2021   Export data to:   * here("Data", "BeeData\_Conservation\_paper", "SCAN\_Bees\_2021\_USA\_20231004.csv") |
| GBIF data :  occ\_download\_get(key = "0431196-210914110416597", overwrite = TRUE) %>%  occ\_download\_import(here("Data", "BeeData\_Conservation\_paper", "GBIF\_Bees6Families\_USA\_1992to2021\_20220824"))  SCAN data:  here("Data", "BeeData\_Conservation\_paper", "SCAN\_Bees\_2021\_USA\_20231004.csv")  Chesshire et al. (2023) data:  read.csv(here("Data/BeeData\_Conservation\_paper/Chesshire\_2023/contiguousRecords\_high\_Only.csv") | ./Programs/BioIndicator/Extract\_BeeBirdLand\_data/2\_CombineBeeDatasets\_ChesshireSCANGBIF\_20231004.rmd | Formatted all three sources:   * Removed duplicates across datasets using unique combinations of institutionCode and catalogNumber * Selected records in USA * Made sure points on offshore islands and close to shore were snapped to the USA * Added meter uncertainty to records with low lat / long decimal degree precision * Formatted trap types based on info in each record. * Selected records with a species name other than Apis mellifera * Selected records with meter uncertainty lacking or less than 15 km * Includes a little bee species editing based on edits recommendations published by Chesshire et al (2023) and Bryan Danforth. * Did some formatting of the columns to append all data together. Kept as many as possible.   Output:   * here("Data", "BioIndicator", "bees\_Chesshire2021\_USA\_20231011.csv") |
| Bee data:  here("Data", "BioIndicator ", " bees\_Chesshire2021\_USA\_20231011.csv")  Raster template: st\_read("/Users/jsr293/Documents/GIS\_Data/eBird\_StatusTrend\_Grids/statusGrid\_USA\_4326.shp")  This was created from: srd\_mask\_land\_hr.tif | ./Programs/BioIndicator/Extract\_BeeBirdLand\_data/3\_BeeSummaries\_perStatusCell\_20230718.r  Updated October 17, 2023 | Associate a grid cell ID to each bee record. The grid is the same as the 3 x 3 km status grid.  Output is:  - here("Data", "BioIndicator", "allBees\_statusGrid\_EastHalfUSA\_2007to2021\_20231017.rds")  Selected records from 2007 to 2021.  Selected records with location accuracy of at least 3 km.  Remove records without day, month, or year information  Create summaries of number of bees and number of species, per grid cell:   * Calculate the standardized number of species, by first calculating the number of species per survey, then averaging number of species/survey across each grid cell   Outputs:   * here("Data", "BioIndicator", "bees\_standardized\_meanPerSurvey\_no1beeSurv\_2007to2021\_20231017.rds") |
| **Extract bird checklists data and associated prevalence** | | |
| Ebird observations and checklists:   * "/Volumes/BirdsBees/Data/eBird/World/ebd\_relApr-2022/ebd\_relApr-2022.txt" * "/Volumes/BirdsBees/Data/eBird/World/ebd\_sampling\_relApr-2022/ebd\_sampling\_relApr-2022.txt"   Grid cells with bee data:   * here("Data", "BioIndicator", " bees\_standardized\_meanPerSurvey\_no1beeSurv\_2007to2021\_20231017.rds") | ./Programs/BioIndicator/Extract\_BeeBird\_data/ 4\_Extract\_eBirdObsChecklists\_InGrid3km\_20230822.Rmd  Updated October 17, 2023 | 1. Extracted the bird checklists for states and years of interest. 2. Selected checklists covering the breeding season. 3. Extracted checklists that were located in grid cells with bee data and added grid cell ID to checklist info. Only grid cells with at least 30 bee records OR 2 surveys were selected (i.e., grids with 1 survey and less than 30 bees were removed). 4. Extracted bird observations for states and years of interest. 5. Minimized the ebd obs size by selecting only a few columns. 6. Use the “chunked” r package to extract (one chunk at the time) the obs that were associated with the checklists above. 7. Identify the prevalence level of each species within the grid cells with bee data for 2007 to 2021. 8. Zero-filled the checklists to have info about places where a species was not detected. 9. Keep only species with at least 20% prevalence within field data grids and species for which their distribution covered at least 40% of our study area.   Several mid-steps outputs were created and are not listed here. Main outputs are:   * here("Data", "BioIndicator", "eBirdObsSampl\_clean\_atGridWithBees\_EastUSA\_2007to2021\_20231017.csv")   For each bird species with minimum prevalence, exported a zero-filled file:   * paste0 /Volumes/BirdsBees/Data/BioIndicator/eBirdObsSampl\_2007to2021\_min30beesCells/eBirdObsSampl\_10prev\_zf\_", .y$common\_name, \_atGridWith30Bees2Surv\_EastUSA\_2007to2021\_20231017.csv") |
| Bird data:   * here/Volumes/BirdsBees/Data/BioIndicator/eBirdObsSampl\_2007to2021\_min30beesCells/eBirdObsSampl\_10prev\_zf\_", species, \_atGridWith30Bees2Surv\_EastUSA\_2007to2021\_20231017.csv") | ./Programs/BioIndicator/ 5\_Residuals\_TravStaCombined\_CountPerMinutesPerSpp\_20230720.r  Updated October 17, 2023 | This is a two-step program.    First assess at what survey duration a species is associated with maximum abundance (abundance ~ survey duration)  Second, use only surveys with max 300 minutes or max time from above and assess if each species, in each survey, has higher or lower abundance compared to expected (gam curve), based on survey length.  Output one file per species:   * paste0("/Volumes/BirdsBees/Results/BioIndicator/Standardize/CountResidual\_EastUSA\_perSpecies", species, "\_gamModel\_TravStaCombined\_20231017.rds") |
| list.files(path = "/Volumes/BirdsBees/Results/BioIndicator/Standardize\_2007to2021/", pattern = "\_gamModel\_TravStaCombined\_20231017.rds", recursive = TRUE) | ./Programs/BioIndicator /Standardize/ 6\_StandardizeBirdObs\_EastHalfUSA\_20230720.r  Updated October 17, 2023 | Calculate the mean residual per species and grid cell (mean of all bird surveys).  Output:   * here("Results", "BioIndicator", "birdSpp\_20prev\_meanResidPerCell\_2007to2021\_min30beesCells\_20231017.csv") |
| **Extract land cover percentage per cell** | | |
|  | ./ Programs/BioIndicator\_USGSabund\_KohAbund\_Status/Extract\_BeeBirdLand\_data/ 7\_Create\_YearlyCropScape Layers\_20230120.r | The programs downloads the year and state cropscape data from the web.  Output folder:   * "/Volumes/BirdsBees/Data/CropScape/2021   The steps to create one layer representing one year of cropscape for eastern USA are listed in the r program, but completed using the terminal and gdal (this is much much faster than through R).  Yearly layer output:  Data/CropScape/ 1\_Yearly\_USA\_CropScapeLayers\_CropInfo/CropScape\_USA\_2021\_20230120.tif |
| CropScape\_USA\_2021\_20230120.tif  Grid:  /Users/jsr293/Documents/GIS\_Data/eBird\_StatusTrend\_Grids/statusGrid\_USA.shp | ./Programs/BioIndicator/Extract\_BeeBirdLand\_data/ 8\_Calculate\_pctLandCovers\_perGridCell\_20230120.r | Use yearly cropscape .tif layer and a grid. For each cell of the grid, summarizes the percentage cover of the different land covers for each grid cell. It uses the 45 land cover categories published by Koh et al. (2016).  Output:   * here("Data", "BioIndicator\_USGS\_SCAN\_BirdStatus", "pctLandCovers\_CropScape2021\_EastHalfUSA\_20230120.rds") * "pctLandCovers\_CropScape2013\_EastHalfUSA\_20230130.rds"   Calculate the percentage of grid cells that have the presence of each land cover, i.e., calculate land cover prevalence in Eastern half of USA.  Output:   * here("Data", "BioIndicator\_USGS\_SCAN\_BirdStatus", "prevalence\_CropScape2021\_landCovers\_EastHalfUSA\_20230214.csv") |
| **Build model to extract bird and land cover indicators** | | |
| Bee data:  here("Data", "BioIndicator", " bees\_standardized\_meanPerSurvey\_no1beeSurv\_2007to2021\_20231017.rds")  Bird data:  "Results", "BioIndicator", " birdSpp\_20prev\_meanResidPerCell\_2007to2021\_min30beesCells\_20231017.csv")  Land cover data:  here("Data", "BioIndicator", "pctLandCovers\_CropScape2021\_EastHalfUSA\_20230120.rds") | ./Programs/BioIndicator /Extract\_BeeBirdLan\_data/ 9\_CompileDatasetsForModels\_20230720.r  Updated October 17, 2023 | Combine the bee response with the bird and land cover variables in one data frame, ready to model.  Output:   * here("Data", "BioIndicator", " BeesMin30\_Birds20pct\_LandCovers20pct\_2007to2021\_no1beeSurv\_Resid30BeesCells\_20231017.csv") |
| Bees, Birds, Lands dataset:  here("Data", "BioIndicator", " BeesMin30\_Birds20pct\_LandCovers20pct\_2007to2021\_no1beeSurv\_Resid30BeesCells\_20231017.csv") | ./Programs/BioIndicator/Modelling/ 10\_LeapsModel\_BirdsAndLands\_20231107.rmd  Updated Nov. 30, 2023 | First step is to reduce the number of predictors from 100 to a smaller number of the most relevant predictors. This is accomplished by subsetting the data in ‘n’ subset of predictors and extracting the best 3 predictors per subset. Then shuffling the dataset predictor list and subsetting again, 1000 times. This step was done only for the birds & land covers, and birds only analysis (not for the lands only because it had only 21 covariates).  Second step is to use the subset of the best predictors and create 100 models of bee richness. Each of the 100 models are slightly different from each other, partly because of the randomness of the Bayesian process.  We are predicting bee richness at each grid cell using each of the 100 models, then averaging bee richness from the 100 maps to obtain our final bee richness map.  Uncertainty at each location is calculated using a parametric bootstrap process which accounts both for predictor selection uncertainty and predictor coefficient uncertainty.  We are also calculating R-squared (model fit) for our models and five-fold validation using locations where we have observed (field) bee data.  Outputs for birds & land covers:   * results.30, here("Results", "BioIndicator", paste0("leapModels\_ BeesMin30\_Birds20pct\_LandCovers20pct\_2007to2021\_no1beeSurv\_Resid30BeesCells \_20231017.rds") * mapsMean, here("Results", "BioIndicator", "Maps", " meanBeePreds\_20231120.tif") * CI.bees, here("Results", "BioIndicator", " beeRich\_uncertainty\_BirdsLCovers\_atLocs\_20231130.csv") |
| Bees, Birds, Lands dataset:  here("Data", "BioIndicator", " BeesMin30\_Birds20pct\_LandCovers20pct\_2007to2021\_no1beeSurv\_Resid30BeesCells\_20231017.csv") | ./Programs/BioIndicator/Modelling/  11\_ LeapsModel\_BirdsOnly\_20230905.rmd | Same tasks as with the modeling using birds and land covers only, but here we used birds only (no land cover predictors)  Outputs for birds only:   * results.30, here("Results", "BioIndicator", paste0("leapModels\_birdsOnly\_ BeesMin30\_Birds20pct\_LandCovers20pct\_2007to2021\_no1beeSurv\_Resid30BeesCells \_20231018.rds") * mapsMean, here("Results", "BioIndicator", "Maps", " mean100Birds\_Oct19.tif") * CI.bees, here("Results", "BioIndicator", "beeRich\_uncertainty\_atLocs\_birdsOnly\_20231019.csv") |
| Bees, Birds, Lands dataset:  here("Data", "BioIndicator", " BeesMin30\_Birds20pct\_LandCovers20pct\_2007to2021\_no1beeSurv\_Resid30BeesCells\_20231017.csv") | ./Programs/BioIndicator/Modelling/  12\_ LeapsModel\_LandsOnly\_20230829.rmd | Outputs for lands only:   * results.30, here("Results", "BioIndicator", paste0("leapModels\_landsOnly\_ BeesMin30\_Birds20pct\_LandCovers20pct\_2007to2021\_no1beeSurv\_Resid30BeesCells \_20231019.rds") * mapsMean, here("Results", "BioIndicator", "Maps", " mean100Lands\_Oct19.tif") * CI.bees, here("Results", "BioIndicator", "beeRich\_uncertainty\_atLocs\_landsOnly\_20231020.csv") |
|  |  |  |