Chapter 1 Introduction and Setup

1.1 Introduction

Citizen science data are increasingly making important contributions to ecological research and conservation. One of the most common forms of citizen science data is derived from members of the public recording species observations. eBird (Sullivan et al. 2014) is the largest of these biological citizen science programs. As of January 2019, the database contained nearly 600 million bird observations from every country in the world, with observations of nearly every bird species on Earth. The eBird database is valuable to researchers across the globe, due to its year-round, broad spatial coverage, high volumes of open access data, and applications to many ecological questions. These data have been widely used in scientific research to study phenology, species distributions, population trends, evolution, behavior, global change, and conservation. However, robust inference with eBird data requires careful processing of the data to address the challenges associated with citizen science datasets. This book, and the associated paper, outlines a set of best practices for addressing these challenges and making reliable estimates of species distributions from eBird data.

There are two key characteristics that distinguish eBird from many other citizen science projects and facilitate robust ecological analyses: the checklist structure enables non-detection to be inferred and the effort information associated with a checklist facilitates robust analyses by accounting for variation in the observation process (La Sorte et al. 2018; Kelling et al. 2018). When a participant submits data to eBird, sightings of multiple species from the same observation period are grouped together into a single **checklist**. **Complete checklists** are those for which the participant reported all birds that they were able to detect and identify. Critically, this enables scientists to infer counts of zero individuals for the species that were not reported. If checklists are not complete, it's not possible to ascertain whether

the absence of a species on a list was a non-detection or the result of a participant not recording the species. In addition, citizen science projects occur on a spectrum from those with predefined sampling structures that resemble more traditional survey designs, to those that are unstructured and collect observations opportunistically. eBird is a **semi-structured** project, having flexible, easy to follow protocols that attract many participants, but also collecting data on the observation process (e.g. amount of time spent birding, number of observers, etc.), which can be used in subsequent analyses (Kelling et al. 2018).

Despite the strengths of eBird data, species observations collected through citizen science projects present a number of challenges that are not found in conventional scientific data. The following are some of the primary challenges associated these data; challenges that will be addressed throughout this book:

- **Taxonomic bias:** participants often have preferences for certain species, which may lead to preferential recording of some species over others (Greenwood 2007; Tulloch and Szabo 2012). Restricting analyses to complete checklists largely mitigates this issue.
- **Spatial bias:** most participants in citizen science surveys sample near their homes (Luck et al. 2004), in easily accessible areas such as roadsides (Kadmon, Farber, and Danin 2004), or in areas and habitats of known high biodiversity (Prendergast et al. 1993). A simple method to reduce the spatial bias that we describe is to create an equal area grid over the region of interest, and sample a given number of checklists from within each grid cell.
- **Temporal bias:** participants preferentially sample when they are available, such as weekends (Courter et al. 2013), and at times of year when they expect to observe more birds, notably during spring migration (Sullivan et al. 2014). To address the weekend bias, we recommend using a temporal scale of a week or multiple weeks for most analyses.
- **Spatial precision:** the spatial location of an eBird checklist is given as a single latitude-longitude point; however, this may not be precise for two main reasons. First, for traveling checklists, this location represents just one point on the journey. Second, eBird checklists are often assigned to a **hotspot** (a common location for all birders visiting a popular birding site) rather than their true location. For these reasons, it's not appropriate to align the eBird locations with very precise habitat covariates, and we recommend summarizing covariates within a neighborhood around the checklist location.
- **Class imbalance:** bird species that are rare or hard to detect may have data with high

class imbalance, with many more checklists with non-detections than detections. For these species, a distribution model predicting that the species is absent everywhere will have high accuracy, but no ecological value. We'll follow the methods for addressing class imbalance proposed by Robinson et al. (2018).

- **Variation in detectability:** detectability describes the probability of a species that is present in an area being detected and identified. Detectability varies by season, habitat, and species (Johnston et al. 2014, 2018). Furthermore, eBird data are collected with high variation in effort, time of day, number of observers, and external conditions such as weather, all of which can affect the detectability of species (Ellis and Taylor 2018; Oliveira et al. 2018). Therefore, detectability is particularly important to consider when comparing between seasons, habitats or species. Since eBird uses a semi-structured protocol, that collects variables associated with variation in detectability, we'll be able to account for a larger proportion of this variation in our analyses.

The remainder of this book will demonstrate how to address these challenges using real data from eBird to produce reliable estimates of species distributions. In general, we'll take a two-pronged approach to dealing with unstructured data and maximizing the value of citizen science data: imposing more structure onto the data via data filtering and including covariates in models to account for the remaining variation.

The next two chapters show how to access and prepare eBird data and land cover covariates, respectively. The remaining three chapters provide examples of different species distribution models that can be fit using these data: encounter rate models, occupancy models, and abundance models. Although these examples focus on the use of eBird data, in many cases they also apply to similar citizen science datasets.

1.2 Prerequisites

To understand the code examples used throughout this book, some knowledge of the programming language R is required. If you don't meet this requirement, or begin to feel lost trying to understand the code used in this book, we suggest consulting one of the excellent free resources available online for learning R. For those with little or no prior programming

experience, Hands-On Programming with R is an excellent introduction. For those with some familiarity with the basics of R that want to take their skills to the next level, we suggest R for Data Science as the best resource for learning how to work with data within R.

1.3 Setup

1.3.1 Data package

The first two chapters of this book focus on obtaining and preparing eBird and land cover data for the modeling that will occur in the remaining chapters. These steps can be time consuming and laborious. If you'd like to skip straight to the analysis, **download this package of prepared data**. Unzip this file so that the contents are in the `data/` subdirectory of your RStudio project folder. This will allow you to jump right in to the modeling and ensure that you're using exactly the same data as was used when creating this book. This is a good option if you don't have enough hard drive space to store the full eBird data set, which is more than 200GB, or don't have a fast enough internet connection to download it.

1.3.2 Software

The examples throughout this website use the programming language **R** (R Core Team 2018) to work with eBird data. If you don't have R installed, download it now, if you already have R, chances are you're using an outdated version, so update it to the latest version now. R is updated regularly, and **it is important that you have the most recent version of R** to avoid headaches when installing packages. We suggest checking every couple months to see if a new version has been released.

We strongly encourage R users to use **RStudio**. RStudio is not required to follow along with this book; however, it will make your R experience significantly better. If you don't have RStudio, download it now, if you already have it, update it because new versions with useful additional features are regularly released. Pro tip: immediately go into RStudio preferences

(Tools > Global Options) and on the General pane uncheck “Restore .RData into workspace at startup” and set “Save workspace to .RData on exit” to “Never”. This will avoid cluttering your R session with old data and save you headaches down the road.

Due to the massive size of the eBird dataset, working with it requires the Unix command-line utility AWK. You won’t need to use AWK directly, since the R package `auk` does this hard work for you, but you do need AWK to be installed on your computer. Linux and Mac users should already have AWK installed on their machines; however, Windows user will need to install Cygwin to gain access to AWK. Cygwin is free software that allows Windows users to use Unix tools. Cygwin should be installed in the default location (`C:/cygwin/bin/gawk.exe` or `C:/cygwin64/bin/gawk.exe`) in order for everything to work correctly. Note: there’s no need to do anything at the “Select Utilities” screen, AWK will be installed by default.

1.3.3 R packages

The examples in this book use a variety of R packages for accessing eBird data, working with spatial data, data processing and manipulation, and model fitting. To install all the packages necessary to work through this book, run the following code:

```
install.packages("remotes")
remotes::install_github("mstrimas/ebppackages")
```

Note that several of the spatial packages require dependencies. If installing these packages fails, consult the instructions for installing dependencies on the `sf` package website. Finally, **ensure all R packages are updated** to their most recent version by clicking on the Update button on the Packages tab in RStudio.

1.3.4 Tidyverse

Throughout this book, we use packages from the Tidyverse, an opinionated collection of R packages designed for data science. Packages such as `ggplot2` , for data visualization, and `dplyr` , for data manipulation, are two of the most well known Tidyverse packages; however, there are many more. In the following chapters, we often use Tidyverse functions without

explanation. If you encounter a function you're unfamiliar with, consult the documentation for help (e.g. `?mutate` to see help for the `dplyr` function `mutate()`). More generally, the free online book R for Data Science by Hadley Wickham is the best introduction to working with data in R using the Tidyverse.

The one piece of the Tidyverse that we will cover here, because it is ubiquitous throughout this book and unfamiliar to many, is the pipe operator `%>%`. The pipe operator takes the expression to the left of it and "pipes" it into the first argument of the expression on the right, i.e. one can replace `f(x)` with `x %>% f()`. The pipe makes code significantly more readable by avoiding nested function calls, reducing the need for intermediate variables, and making sequential operations read left-to-right. For example, to add a new variable to a data frame, then summarize using a grouping variable, the following are equivalent:

```
library(dplyr)

# pipes
mtcars %>%
  mutate(wt_kg = 454 * wt) %>%
  group_by(cyl) %>%
  summarize(wt_kg = mean(wt_kg))
#> # A tibble: 3 x 2
#>     cyl wt_kg
#>   <dbl> <dbl>
#> 1     4 1038.
#> 2     6 1415.
#> 3     8 1816.

# intermediate variables
mtcars_kg <- mutate(mtcars, wt_kg = 454 * wt)
mtcars_grouped <- group_by(mtcars_kg, cyl)
summarize(mtcars_grouped, wt_kg = mean(wt_kg))
#> # A tibble: 3 x 2
#>     cyl wt_kg
#>   <dbl> <dbl>
```

```
#> 1     4 1038.
#> 2     6 1415.
#> 3     8 1816.

# nested function calls
summarize(
  group_by(
    mutate(mtcars, wt_kg = 454 * wt),
    cyl
  ),
  wt_kg = mean(wt_kg)
)
#> # A tibble: 3 x 2
#>     cyl wt_kg
#>   <dbl> <dbl>
#> 1     4 1038.
#> 2     6 1415.
#> 3     8 1816.
```

Once you become familiar with the pipe operator, we believe you'll find the the above example using the pipe the easiest of the three to read and interpret.

1.3.5 Getting eBird data access

The complete eBird database is provided via the eBird Basic Dataset (EBD), a large text file. To access the EBD, begin by creating an eBird account and signing in. Then visit the eBird Data Access page and fill out the data access request form. eBird data access is free; however, you will need to request access in order to download the EBD. Filling out the access request form allows eBird to keep track of the number of people using the data and obtain information on the applications for which the data are used.

Once you have access to the data, proceed to the download page. Download both the World EBD (~ 42 GB compressed, ~ 210 GB uncompressed) and corresponding Sampling Event Data (~ 3.5 GB compressed, ~ 11 GB uncompressed). The former provides observation-level data, while the latter provides checklist-level data; both files are required for species distribution modeling. If limited hard drive space or a slow internet connection make dealing with these large files challenging, consult Section 2.6.1 for details on a method for downloading a subset of EBD.

The downloaded data files will be in `.tar` format, and should be unarchived. The resulting directories will contain files with extension `.txt.gz`, these files should be uncompressed (on Windows use 7-Zip, on Mac use the default system uncompression utility) to produce two text files (e.g., `ebd_relAug-2019.txt` and `ebd_sampling_relAug-2019.txt`). Move these two large, uncompressed .txt files to a sensible, central location on your computer. In general, we suggest creating an `ebird/` folder nested in a `data/` folder within your home directory (i.e. `~/data/ebird/`) to store these files, and throughout the remainder of this chapter we'll assume you've placed the data there. If you choose to store the EBD elsewhere, you will need to update references to this folder in the code. If the files are too large to fit on your computer's hard drive, they can be stored on an external hard drive.

Each time you want to access eBird data in an R project, you'll need to reference the full path to these text files, for example `~/data/ebird/ebd_relAug-2019.txt`. In general, it's best to avoid using absolute paths in R scripts because it makes them less portable–if you're sharing the files with someone else, they'll need to change the file paths to point to the location at which they've stored the eBird data. The R package `auk` provides a workaround for this, by allowing users to set an environment variable (`EBD_PATH`) that points to the directory where you've stored the eBird data. To set this variable, use the function `auk_set_ebd_path()`. For example, if the EBD and Sampling Event Data files are in `~/data/ebird/`, use:

```
# set ebd path
auk::auk_set_ebd_path("~/data/ebird/")
```

After **restarting your R session**, you should be able to refer directly to the EBD or Sampling Event Data files within `auk` functions (e.g., `auk_ebd("ebd_relAug-2019.txt")`). Provided your collaborators have also set `EDB_PATH`, your scripts should now be portable.

You now have access to the full eBird dataset! Note, however, that **the EBD is updated monthly**. If you want the most recent eBird records, be sure to **regularly download an updated version**. Finally, **whenever you update the EBD, always update the `auk` package as well**, this will ensure that `auk` will be able to handle any changes to the EBD that may have occurred.

1.3.6 GIS data

Throughout this book, we'll be producing maps of species distributions. To provide context for these distributions, we'll need GIS data for political boundaries. Natural Earth is the best source for a range of tightly integrated vector and raster GIS data for producing professional cartographic maps. The R package, `rnaturalearth` provides a convenient method for accessing these data from within R. We'll also need Bird Conservation Region (BCR) boundaries, which are available through Bird Studies Canada.

These GIS data layers are most easily accessed by **downloading the data package** for this book. These data were generated using the following code. If you intend to run this code yourself, first create an RStudio project, so the files will be stored within the `data/` subdirectory of the project. This will allow us to load these data in later chapters as they're needed. Note that calls to `ne_download()` often produce warnings suggesting that you've used the incorrect "category"; these can safely be ignored.

```
library(sf)
library(rnaturalearth)
library(dplyr)

# file to save spatial data
gpkg_dir <- "data"
if (!dir.exists(gpkg_dir)) {
  dir.create(gpkg_dir)
}
f_ne <- file.path(gpkg_dir, "gis-data.gpkg")
```

```r
# download bcrs
tmp_dir <- normalizePath(tempdir())
tmp_bcr <- file.path(tmp_dir, "bcr.zip")
paste0("https://www.birdscanada.org/research/gislab/download/",
       "bcr_terrestrial_shape.zip") %>%
  download.file(destfile = tmp_bcr)
unzip(tmp_bcr, exdir = tmp_dir)
bcr <- file.path(tmp_dir, "BCR_Terrestrial_master_International.shp") %>%
  read_sf() %>%
  select(bcr_code = BCR, bcr_name = LABEL) %>%
  filter(bcr_code == 27)
# clean up
list.files(tmp_dir, "bcr", ignore.case = TRUE, full.names = TRUE) %>%
  unlink()

# political boundaries
# land border with lakes removed
ne_land <- ne_download(scale = 50, category = "cultural",
                       type = "admin_0_countries_lakes",
                       returnclass = "sf") %>%
  filter(CONTINENT == "North America") %>%
  st_set_precision(1e6) %>%
  st_union()
# country lines
# downloaded globally then filtered to north america with st_intersect()
ne_country_lines <- ne_download(scale = 50, category = "cultural",
                                type = "admin_0_boundary_lines_land",
                                returnclass = "sf") %>%
  st_geometry()
ne_country_lines <- st_intersects(ne_country_lines, ne_land, sparse = FALSE) %
  as.logical() %>%
  {ne_country_lines[.]}
# states, north america
```

```r
ne_state_lines <- ne_download(scale = 50, category = "cultural",
                              type = "admin_1_states_provinces_lines",
                              returnclass = "sf") %>%
  filter(adm0_a3 %in% c("USA", "CAN")) %>%
  mutate(iso_a2 = recode(adm0_a3, USA = "US", CAN = "CAN")) %>%
  select(country = adm0_name, country_code = iso_a2)

# output
unlink(f_ne)
write_sf(ne_land, f_ne, "ne_land")
write_sf(ne_country_lines, f_ne, "ne_country_lines")
write_sf(ne_state_lines, f_ne, "ne_state_lines")
write_sf(bcr, f_ne, "bcr")
```

References

Courter, Jason R., Ron J. Johnson, Claire M. Stuyck, Brian A. Lang, and Evan W. Kaiser. 2013. “Weekend Bias in Citizen Science Data Reporting: Implications for Phenology Studies.” *International Journal of Biometeorology* 57 (5): 715–20. https://doi.org/10.1007/s00484-012-0598-7.

Ellis, Murray V., and Jennifer E. Taylor. 2018. “Effects of Weather, Time of Day, and Survey Effort on Estimates of Species Richness in Temperate Woodlands.” *Emu-Austral Ornithology* 118 (2): 183–92.

Greenwood, Jeremy J. D. 2007. “Citizens, Science and Bird Conservation.” *Journal of Ornithology* 148 (1): 77–124. https://doi.org/10.1007/s10336-007-0239-9.

Johnston, Alison, Daniel Fink, Wesley M. Hochachka, and Steve Kelling. 2018. “Estimates of Observer Expertise Improve Species Distributions from Citizen Science Data.” *Methods in Ecology and Evolution* 9 (1): 88–97.

Johnston, Alison, Stuart E. Newson, Kate Risely, Andy J. Musgrove, Dario Massimino, Stephen R. Baillie, and James W. Pearce-Higgins. 2014. “Species Traits Explain Variation in Detectability of UK Birds.” *Bird Study* 61 (3): 340–50.

Kadmon, Ronen, Oren Farber, and Avinoam Danin. 2004. “Effect of Roadside Bias on the Accuracy of Predictive Maps Produced by Bioclimatic Models.” *Ecological Applications* 14 (2): 401–13.

Kelling, Steve, Alison Johnston, Daniel Fink, Viviana Ruiz-Gutierrez, Rick Bonney, Aletta Bonn, Miguel Fernandez, et al. 2018. “Finding the Signal in the Noise of Citizen Science Observations.” *bioRxiv*, May, 326314. https://doi.org/10.1101/326314.

La Sorte, Frank A., Christopher A. Lepczyk, Jessica L. Burnett, Allen H. Hurlbert, Morgan W. Tingley, and Benjamin Zuckerberg. 2018. “Opportunities and Challenges for Big Data Ornithology.” *The Condor* 120 (2): 414–26.

Luck, Gary W., Taylor H. Ricketts, Gretchen C. Daily, and Marc Imhoff. 2004. “Alleviating Spatial Conflict Between People and Biodiversity.” *Proceedings of the National Academy of Sciences* 101 (1): 182–86. https://doi.org/10.1073/pnas.2237148100.

Oliveira, Camilo Viana, Fabio Olmos, Manoel dos Santos-Filho, and Christine Steiner São Bernardo. 2018. “Observation of Diurnal Soaring Raptors in Northeastern Brazil Depends on Weather Conditions and Time of Day.” *Journal of Raptor Research* 52 (1): 56–65.

Prendergast, J. R., S. N. Wood, J. H. Lawton, and B. C. Eversham. 1993. “Correcting for Variation in Recording Effort in Analyses of Diversity Hotspots.” *Biodiversity Letters*, 39–53.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, Orin J., Viviana Ruiz-Gutierrez, Daniel Fink, Robert J. Meese, Marcel Holyoak, and Evan G. Cooch. 2018. “Using Citizen Science Data in Integrated Population Models to Inform Conservation Decision-Making.” *bioRxiv*, 293464.

Sullivan, Brian L., Jocelyn L. Aycrigg, Jessie H. Barry, Rick E. Bonney, Nicholas Bruns, Caren B. Cooper, Theo Damoulas, et al. 2014. “The eBird Enterprise: An Integrated Approach to Development and Application of Citizen Science.” *Biological Conservation* 169 (January): 31–40. https://doi.org/10.1016/j.biocon.2013.11.003.

Tulloch, Ayesha IT, and Judit K. Szabo. 2012. “A Behavioural Ecology Approach to Understand Volunteer Surveying for Citizen Science Datasets.” *Emu-Austral Ornithology* 112 (4): 313–25.