The eBird Reference Dataset, Version 5.0

M. Arthur Munson†, Kevin Webb‡, Daniel Sheldon†,
Daniel Fink‡, Wesley M. Hochachka‡, Marshall Iliff‡,
Mirek Riedewald*, Daria Sorokina**, Brian Sullivan‡,
Christopher Wood‡, Steve Kelling‡

† Cornell University Computer Science Department, Ithaca, NY 14853
‡ Cornell Lab of Ornithology, Ithaca, NY 14850
* Northeastern University, Boston, MA 02115
** Carnegie Mellon University, Pittsburgh, PA 15213

September 30, 2013

Abstract

This document describes the eBird reference data set and the processing steps taken during creation. We hope this data will be a useful resource for studying avian dynamics and for developing new ecological modeling techniques.

1 Usage and Copyright

The eBird reference data is freely available for all usages. eBird[1] is run by the National Audubon Society[2] and the Cornell Lab of Ornithology[3], and the data is copyrighted by both organizations. A primary goal in publishing this data is to provide a common data resource for studying and comparing ecological models; as such, derivative versions of the eBird reference data set must not be distributed without explicit permission from the copyright holders.

The data set is a snapshot of submitted observations for years prior to 2013 that were submitted to eBird and reviewed by January 1, 2013. Observations with an observation date later than December 31, 2012 were not included.

Published results using this data should cite this document as follows:

> M. Arthur Munson, Kevin Webb, Daniel Sheldon, Daniel Fink, Wesley M. Hochachka, Marshall Iliff, Mirek Riedewald, Daria Sorokina, Brian Sullivan, Christopher Wood, and Steve Kelling. *The eBird Reference Dataset, Version 5.0.* Cornell Lab of Ornithology and National Audubon Society, Ithaca, NY, January 2013.

Please direct any questions to:

M. Arthur Munson at `mmunson@cs.cornell.edu` and
Steve Kelling at `stk2@cornell.edu`

[1] `http://ebird.org`
[2] `http://www.audubon.org`
[3] `http://www.birds.cornell.edu`

2 Introduction

This dataset contains count data for bird species observed by novice and experienced bird observers (a.k.a. *birders*). The data was submitted by volunteers to the eBird Citizen Science Project, run by the Cornell Lab of Ornithology and the National Audubon Society. A record in this dataset corresponds to a *checklist* that a birder uses to mark the number of birds of each species detected; one checklist is submitted per sampling event (i.e. birding session). Each checklist submitted from locations in the Western Hemisphere is additionally annotated with predictor variables (called covariates below) that are derived from the location of the sampling event.

One pragmatic note: Excel is unable to handle the larger data files in this dataset. The data from year 2008 contains more records than Excel supports (rows were truncated around 175,000); previous years contain fewer records. In our experiments with Excel, columns were truncated from the US48 checklist files and the extended covariates files.

Finally, the *stratified_random_design* (SRD) dataset is a companion dataset to the eBird reference dataset with a similar format. The SRD data contains the predictor values for random locations chosen using a stratified random design. There are SRD data sets of differing spatial extent and resolution; one of the point sets, roughly based on a 30km grid, contains around 130,000 random locations covering the lower 48 U.S. States, a second data set containing approximately 933,700 random locations is roughly based on a 3km grid of the lower 48 states. The third data set covering the entire Western Hemisphere contains 4,686,274 data points and was created using an approximate 1.5km grid. The SRD data can be combined with an ecological model to generate maps of predicted surfaces. In brief, use the model to predict at each location in the SRD data, and then plot the predicted values overlaid with geographical boundaries. Helpful tools for plotting maps include Matlab and R.

3 Dataset Organization

The eBird reference data set actually consists of two separate datasets:

US48 — all checklists from the 48 states in the contiguous US, with detailed location-based covariate information.

WESTERN HEMISPHERE — all checklists from the Western Hemisphere, between -30 degrees longtiude to -180 degrees longtiude, with detailed location-based covariate information.

The two datasets are organized in a similar fashion, and all checklists included in the US48 dataset are also included in the WESTERN HEMISPHERE dataset. The WESTERN HEMISPHERE dataset replaces the AMERICAS dataset included in previous revisions.

3.1 Column Summaries

The data is organized by checklist, with one row per checklist. The columns are split into three groups:

- Checklist description. 17 columns describing the sampling event (date, time, lat, long, duration, distance traveled, protocol, observer id, etc.), plus variables for all the species in the data set. There is one column per species, with each column listing the number of birds observed.
- Core covariates (US48 only). Approximately 46 columns that we believe a priori are most important for most of the species. These are suggested as a starting point for analyses. Details are in section 4.4.
- Extended covariates (US48). These include extensive statistics about habitat configuration, fine-grained climate measurements, and observer expertise. See sections 4.5 and 4.6, and tables 3, 4, and 5.
- Extended covariates (WESTERN HEMISPHERE). These include statistics about habitat configuration, and elevation at the location. See section 4.6, and tables 4 and 5.

In the both datasets, all of the files are row aligned. That is, the covariates for the sampling event in line 2 of `checklists.csv` are in line 2 of `extended_covariates.csv` and line 2 for US48 `core_covariates.csv`.

3.2 File Organization and Directory Structure

The checklists are ordered chronologically by year; within a year they are ordered by the observation location. Files are split into years for manageability. Data for a given year are organized in a single directory containing three row-aligned files, each containing one of the column groups described above. For example, the directory structure for 2 years from the US48 dataset looks like this:

```
us48/2011:

  checklists.csv
  core_covariates.csv
  extended_covariates.csv

us48/2012:

  checklists.csv
  core_covariates.csv
  extended_covariates.csv
```

The structure of the WESTERN HEMISPHERE dataset is similar except it does not include core-covariates.csv.

```
western_hemisphere/2011:

  checklists.csv
  extended_covariates.csv

western_hemisphere/2012:

  checklists.csv
  extended_covariates.csv
```

In addition, there is also a `docs` folder containing this document and text files describing the names and types of all data columns:

```
docs:

  checklists.names
  core_covariates.names      [US48 only]
  extended_covariates.names  [US48 only]
  taxonomy.csv
  speciesfreq.csv
```

See section 4.1 for the names files' format.

The data set uses scientific names (i.e. latin names) to name species. The file `taxonomy.csv` lists all of the scientific names and the corresponding common names in english, as well as taxonomic codes for the species. The first line of the file contains the names of the columns.

Finally, how often each species is observed is given in `speciesfreq.csv`. The numbers are the percentage of checklists containing the species for all years up through December 31, 2012. These frequencies determine the order of the species' columns in `checklists.csv`.

4 Data Set Description

The data is written in CSV format with one line (data record) per checklist. All data files include the column SAMPLING_EVENT_ID, a unique identifier for each checklist. This can be used to verify that all column groups are correctly aligned. The first line of every file contains the names of the columns.

Two special values are used in the dataset. When the value for a variable is missing / unknown, it is represented as `?`. Second, when a covariate measurement is not applicable to the context of an observation, it is represented as `NA`. How to deal with these special cases depends on how the data is used and the analysis tools, and is best decided by the data analyst.

The rest of this section describes the names file format and the different variable groups.

4.1 Names File Format

Each names file lists the names and types of all variables in the group. One variable is listed per line, and the order of variables matches the order of data columns in the corresponding CSV file. Variables can be `string`, `continuous`, or `nominal` valued. Example variable descriptions are

```
SAMPLING_EVENT_ID: string.
ELEVATION: continuous.
BCR: 1,2,3,...,35.
```

The variable name precedes the colon, and type information comes after. Continuous variables are real- or integer-valued measurements. Nominal variables are categorical, and can take one of a small set of values; the allowed values are listed after the variable name. String variables are high-arity nominal-valued variables where a legal value is a sequence of letters or numbers (no whitespace). Generally string variables should not be used for predictive modeling, but they are useful for fitting random effects and for data provenence.

4.2 Species Counts and Checklist Types

There is one column per species containing the count of how many were observed. If the species was unobserved or unreported the count value is 0, or if the species was reported as present without a count the count is replaced with an `X`. Birders often use present-without-count if they do not have a good idea how many they detected, perhaps because there were many individuals of that species, or if they did not want to bother with counting the species if the species is not particularly interesting to the birder. Present-without-count records contain useful information about the presence of a species, but should probably be discarded when modeling species abundance.

In most cases, a model should be fit using a single species variable as the response variable. Unless the research question involves relationships between bird species, the remaining species variables should be ignored during the modeling.

Researchers wondering which species to choose for modeling may wish to refer to the species matrix table included with the documentation `docs/MatrixV1.xls`. This table lists a few dozen species and categorizes them along multiple axes (e.g. migration and population trends) based on domain expert opinion.

4.2.1 Checklist Types

Casual count checklists, and random count checklists are included in this release. *Casual counts* are observations made while birding was not the submitter's primary activity. *Random counts* are observations made at a randomly selected location over a period of at least five minutes. The presence of *complete*, *casual counts*, and *random counts* provides a varied set of observation protocols for research.

4.2.2 Group Checklists

Included in the datasets are group checklists representing observation events where multiple participants submit identical checklists in eBird to record their sightings. Shared group checklists are identified with a common GROUP_ID value. For research that requires unique checklists without repetition, the PRIMARY_CHECKLIST_FLAG (Boolean) can be used to select single representative checklists from group submissions. Singleton non-group checklists are by default identified with PRIMARY_CHECKLIST_FLAG = 1.

4.3 Sampling Event Covariates

Each data record contains information describing when, where, and how the observations were made, as well as a unique identifier for the checklist. Table 1 lists the covariates tied to the sampling event.

Table 1: Summary of sampling event covariates.

Variable Name	Comments
SAMPLING_EVENT_ID	Unique identifier for each data sample / checklist.
LATITUDE	Decimal latitude. Location is tied to starting position of traveling counts. Datum = WGS 84.
LONGITUDE	Decimal longitude. Datum = WGS 84.
COUNT_TYPE	What kind of observation the sample is: stationary (P21), traveling (P22, P34), area (P23, P35), casual (P20), or random (P48). Protocol P34 is a small amount of data contributed from the Rocky Mountain Bird Observatory that we believe is high quality. Protocol P35 data are back-yard area counts made on consecutive days (see `http://www.birds.cornell.edu/MyYardCounts`).
COUNTRY	Full name of country / political unit where observation took place. Useful for extracting sub-portions of the data.
STATE_PROVINCE	Name of the state, province, or region where observation was made. Useful for extracting sub-portions of the data.
YEAR	
MONTH	Month of the year, ranging from 01 through 12. Useful for extracting sub-portions of the data.
DAY	Day of the year, ranging from 1 through 366.
TIME	Time when observation started, ranging over $[0, 24)$. Fractional hours represent minutes (e.g. $13.5 = 1{:}30$PM). Times are local times (including daylight savings when/where appropriate).
EFFORT_HRS	Duration of observation for the checklist, in hours.
EFFORT_DISTANCE_KM	Distance traveled during observation period, in kilometers. Equals 0 for non-traveling counts.
EFFORT_AREA_HA	Size of survey area for area counts, in hectares. Equals 0 for non-area counts.
OBSERVER_ID	Identifier for the person who submitted the data.
NUMBER_OBSERVERS	Number of observers in the birding party.
GROUP_ID	ID string indicating a group checklist. Group/shared checklists contain summations of the highest counts for each species seen by members of the group.
PRIMARY_CHECKLIST_FLAG	Boolean value to indicate singleton checklists and the primary checklist for a shared set of group checklists. This flag should be used to select unique checklists.

Location is highly correlated with many covariates describing the environment. Further, the pair LATITUDE-LONGITUDE is highly correlated with the sampling event. For these reasons, the decision to include latitude and longitude as predictor variables should be made carefully.

4.4 Core Covariates

Table 2 summarizes the core covariates that we feel are generally useful for most species. Two elevation variables are included because a) the different resolutions serve complementary purposes, and b) sometimes one will have a measurement while the other's value is missing.

Table 2: Summary of core covariates.

Variable Name	Comments
SAMPLING_EVENT_ID	Unique identifier for each data sample / checklist.
BCR	Bird conservation region (numeric identifier).
BAILEY_ECOREGION	Ecoregions defined by common climatic and vegetation characteristics. Details at `http://www.nationalatlas.gov/mld/ecoregp.html`
OMERNIK_L3_ECOREGION	Ecoregions defined by vegetation, animal life, geology, soils, water quality, climate, and human land use. Details at `http://www.nationalatlas.gov/mld/ecoomrp.html`
CAUS_PREC†	Mean total precipitation for month in which observation made. For value mapping see Table 8.
CAUS_SNOW†	Mean snow depth for month in which observation made. Always missing (coded as '?') for observations in May through Sept. since no data available on snow depth from the climate atlas for these months. For value mapping see Table 10.
CAUS_TEMP_AVG†	Mean daily average temperature for month in which observation made. For value mapping see Table 11.
CAUS_TEMP_MIN†	Mean daily minimum temperature for month in which observation made. For value mapping see Table 13.
CAUS_TEMP_MAX†	Mean daily maximum temperature for month in which observation made. For value mapping see Table 15.
ELEV_GT	Elevation in meters. Horizontal resolution is roughly 1km by 1km. Source: GTOPO30 elevation dataset, acquired from USGS in 2004. Details at `http://eros.usgs.gov/products/elevation/gtopo30.php`
ELEV_NED	Elevation in meters. Horizontal resolution is roughly 30m by 30m. Source: National Elevation Dataset, acquired from USGS. Described at `http://seamless.usgs.gov/website/seamless/products/1arc.asp` and `http://www.usgsquads.com/elevationdata.htm#NED_Info`
HOUSING_DENSITY‡	Number of housing units per square mile (2000 census) for the census blockgroup containing the location.
HOUSING_PERCENT_VACANT‡	Percentage of housing units in census blockgroup vacant in 2000.
NLCDYYYY_FS_CTT_7500_PLAND	Percent of surrounding landscape that is habitat class TT. Data values extracted from source landcover rasters for years YYYY = 2001 and YYYY = 2006 are included. See section 4.5 and Table 5 for more details.
POP00_SQMI‡	Population per square mile (2000 census) for the census blockgroup containing the location.

† Source: Climate Atlas of the US, v2 (1961–1990), from NOAA-NCDC. Grid cell resolution is 4km by 4km. Note that these are *climate* variables averaged over 30 years, not weather variables for the year the observation was made. Described at `http://www.ncdc.noaa.gov/oa/about/cdrom/climatls2/info/atlasad.html`

‡ Source: US 2000 census; acquired from ESRI summer 2004. `http://www.census.gov/geo/www/tiger/glossary.html`

All of these covariates are static. The climate variables use the month when the observation is made to select the appropriate climate value from those listed in Table 3.

4.5 Static Environment Covariates

Based on an observation's location, covariate information about climate and habitat is extracted from GIS databases and joined to the checklist records. These measurements are static, in that they are derived from environmental snapshots tied to a time frame independent of when observations are made.

Table 3: Summary of static environment covariates.

Variable Name	**Comments**
SAMPLING_EVENT_ID	Unique identifier for each data sample / checklist.
CAUS_PRECMM†	Mean total precipitation for month MM. Value map in Table 8.
CAUS_SNOWMM†	Mean snow depth for month MM. Value map in Table 10.
CAUS_TEMP_AVGMM†	Mean daily average temperature for month MM. Value map in Table 11.
CAUS_TEMP_MINMM†	Mean daily minimum temperature for month MM. Value map in Table 13.
CAUS_TEMP_MAXMM†	Mean daily maximum temperature for month MM. Value map in Table 15.
CAUS_LAST_SPRING_32F_MEAN†	Last 32 F day in spring (mean). Value map in Table 17.
CAUS_LAST_SPRING_32F_MEDIAN†	Last 32 F day in spring (median). Value map in Table 18.
CAUS_LAST_SPRING_32F_EXTREME†	Last 32 F day in spring (extreme). Value map in Table 19.
CAUS_FIRST_AUTUMN_32F_MEAN†	First 32 F day in autumn (mean). Value map in Table 20.
CAUS_FIRST_AUTUMN_32F_MEDIAN†	First 32 F day in autumn (median). Value map in Table 21.
CAUS_FIRST_AUTUMN_32F_EXTREME†	First 32 F day in autumn (extreme). Value map in Table 22.
DIST_FROM_FLOWING_FRESH‡	Distance from flowing fresh water. Value map in Table 6.
DIST_IN_FLOWING_FRESH‡	Distance inside flowing fresh water. Value map in Table 6.
DIST_FROM_STANDING_FRESH‡	Distance from standing fresh water. Value map in Table 6.
DIST_IN_STANDING_FRESH‡	Distance inside standing fresh water. Value map in Table 6.
DIST_FROM_WET_VEG_FRESH‡	Distance from wet vegetation, fresh water. Value map in Table 6.
DIST_IN_WET_VEG_FRESH‡	Distance inside wet vegetation, fresh water. Value map in Table 6.
DIST_FROM_FLOWING_BRACKISH‡	Distance from flowing brackish water. Value map in Table 6.
DIST_IN_FLOWING_BRACKISH‡	Distance inside flowing brackish water. Value map in Table 6.
DIST_FROM_STANDING_BRACKISH‡	Distance from standing brackish water. Value map in Table 6.
DIST_IN_STANDING_BRACKISH‡	Distance inside standing brackish water. Value map in Table 6.
DIST_FROM_WET_VEG_BRACKISH‡	Distance from wet vegetation, brackish water. Value map in Table 6.
DIST_IN_WET_VEG_BRACKISH‡	Distance inside wet vegetation, brackish water. Value map in Table 6.
NCLDYYYY_FS_*	Landscape and landcover statistics describing the habitat in a square neighborhood around the location. Computed from the National Land Cover Data from MRLC (`www.mrlc.gov/nlcd.php`) source data years YYYY = 2001 and YYYY = 2006 using the FRAGSTATS program. See text and Table 5 for more information.
SUBNATIONAL2_CODE	String encoding the state and county of the location. Useful for extracting sub-portions of the data.

† Source: Climate Atlas of the US, v2 (1961–1990), from NOAA-NCDC. Grid cell resolution is 4km by 4km. Month code 13 denotes the annual aggregate statistic. Described at `http://www.ncdc.noaa.gov/oa/about/cdrom/climatls2/info/atlasad.html`

‡ Source: Shapefiles provided by Idaho National Gap Analysis Program `http://www.gap.uidaho.edu/bulletins/10/idaho_gap.htm`. Refer to Appendix Table 6 for the description of distance values.

4.6 MODIS/ASTER Covariates

Variables based on MODIS landcover imagery (MCD12Q1) and ASTER elevation imagery (ASTGTM) supplement the US48 covariate dataset and are the basis of the covariate data included in the WESTERN HEMISPHERE dataset. In addition to providing discrete location based data, the MODIS UMD landcover classification is also used to generate additional Fragstat covariates.

Table 4: Summary of MODIS/ASTER environment covariates.

Variable Name	Comments
ASTER2011_DEM	30 meter resolution digital elevation developed jointly by the U.S. National Aeronautics and Space Administration (NASA) and Japans Ministry of Economy, Trade, and Industry (METI). `https://lpdaac.usgs.gov/products/aster_products_table/astgtm`
UMD2011_LANDCOVER	500 meter resolution landcover classification using the UMD classification scheme `https://lpdaac.usgs.gov/products/modis_products_table/mcd12q1`

4.7 Fragstat Covariates

Ecologists generally agree that both the type of habitat and its configuration are important factors in ecological processes.[4] Consequently, we processed the raw MODIS landcover data, and the 2001 and 2006 NLCD habitat data using the FRAGSTATS program [MCNE] to generate covariates describing landscape configurations. The configuration settings for FRAGSTATS can be found in Appendix A.1. We included the subset of FRAGSTATS statistics we felt were most likely to be informative for a variety of species across the continent.

More specifically, we extracted the landcover information from the MODIS and NLCD imagery for a grid centered on each checklist location, creating a landcover matrix. Each landcover matrix was given as input to FRAGSTATS, which returned an array of landscape statistics summarizing the habitat neighborhood for the corresponding checklist.

Since the ideal neighborhood size depends on the species under consideration as well as the spatial resolution of the source data, we repeated the FRAGSTAT generation process for a number of different spatial extents. MODIS based FRAGSTATS included with the WESTERN HEMISPHERE and US48 datasets were generated using a bounding box of approximately 900 hectares. NLCD based FRAGSTATS included with the US48 dataset use bounding boxes of 2.25 hectares, 225 hectares, and 22,500 hectares. These extents were selected to cover local ecological processes at small, medium, and large ranges and to accomodate the native resolution of the source data. We decided not to include spatial extents large enough to cover entire migration areas because a) the computation costs would be considerable, and b) habitat configurations becomes less distinct as they are averaged over larger and larger areas. The scale of each covariate is indicated by the "radius" of the neighborhood in meters, and appears in the name of each covariate. *Radius* is actually a misnomer. The neighborhoods are square regions centered on the location. The length of the neighborhood square side is twice the "radius". In other words, the radius number is the radius for a circle inscribed within the neighborhood square.

We post-processed the FRAGSTATS covariates to recode most not-applicable (NA) values as numeric values, in almost all cases a recoding to numeric zero. This was done because the NA values actually do have a biological meaning when recoded to numeric values. For example, in a landscape that is entirely composed of grassland, FRAGSTATS would return NA values for metrics describing landcover types such as forest that were not present. However, absence of forest really does mean that, for example, there is no forest edge; zero-values are justified. Brief descriptions of when NA values were recoded are given in Table 5.

[4] A list of relevant literature can be found in the FRAGSTATS online documentation (background section): `http://www.umass.edu/landeco/research/fragstats/documents/ConceptualBackground/LiteratureCited/LiteratureCited.htm`

Table 5: Summary of habitat statistics.

Variable Name	Comments
	Class Level Statistics
NLCDYYYY_FS_CTT_RR_ED	Edge density for patches of habitat type TT. Ratio of total edge length to landscape area (meters per hectare). NA occurs if habitat TT not in landscape. Recoded as 0 since no edges implies 0 density.
NLCDYYYY_FS_CTT_RR_LPI	Largest patch index. Percentage of landscape area comprised by the largest patch of habitat type TT. NA occurs if habitat TT not in landscape. Recoded as 0.
NLCDYYYY_FS_CTT_RR_PD	Patch density. Number of patches of habitat class TT per 100 hectares in surrounding landscape. NA values occurred if no TT patches in landscape. Recoded as 0.
NLCDYYYY_FS_CTT_RR_PLAND	Percent of surrounding landscape that is habitat class TT. NA values occurred if no TT patches in landscape. Recoded as 0.
UMD2011_FS_CTT_RR_ED	Edge density for patches of habitat type TT. Ratio of total edge length to landscape area (meters per hectare). NA occurs if habitat TT not in landscape. Recoded as 0 since no edges implies 0 density.
UMD2011_FS_CTT_RR_LPI	Largest patch index. Percentage of landscape area comprised by the largest patch of habitat type TT. NA occurs if habitat TT not in landscape. Recoded as 0.
UMD2011_FS_CTT_RR_PD	Patch density. Number of patches of habitat class TT per 100 hectares in surrounding landscape. NA values occurred if no TT patches in landscape. Recoded as 0.
UMD2011_FS_CTT_RR_PLAND	Percent of surrounding landscape that is habitat class TT. NA values occurred if no TT patches in landscape. Recoded as 0.
	Landscape Level Statistics
NLCDYYYY_FS_L_RR_ED	Edge density for the landscape. Ratio of sum of all edges between patches over total landscape area (meters per hectare).
NLCDYYYY_FS_L_RR_LPI	Percentage of landscape area occupied by the largest patch (any habitat type).
NLCDYYYY_FS_L_RR_PD	Patch density (number of patches per 100 hectares).
UMD2011_FS_L_RR_ED	Edge density for the landscape. Ratio of sum of all edges between patches over total landscape area (meters per hectare).
UMD2011_FS_L_RR_LPI	Percentage of landscape area occupied by the largest patch (any habitat type).
UMD2011_FS_L_RR_PD	Patch density (number of patches per 100 hectares).

* All habitat statistics summarize square neighborhood around location with "radius" RR.
** YYYY = 2001 and 2006

A second class of corner case was caused by the fully automated application of FRAGSTATS. A checklist location near the edge of the NLCD map has unknown values in its landscape matrix, corresponding to the grid cells that extend past the NLCD map edge. We handled this by setting these cells to -9999, and configuring FRAGSTATS to treat this value as background that is omitted from computations. Essentially, this truncates the extent of the landscape to the parts of the matrix with known values.

Acknowledgements

Development of eBird was supported in part by the National Science Foundation under Grant ESI-0087760. Additional support for the analysis of eBird and Avian Knowledge Network data came from the Leon Levy Foundation, Wolf Creek Foundation, and the National Science Foundation (Grants ITR-0427914, DBI-0542868, DUE-0734857, IIS-0612031, IIS-0748626, and IIS-0832782). Any opinions, findings, and conclusions

or recommendations expressed in this manuscript are those of the authors and do not necessarily reflect the views of the National Science Foundation, the Leon Levy Foundation, or the Wolf Creek Foundation.

The authors warmly thank Tim Levatich and Jeff Gerbracht (Cornell Lab of Ornithology) for patiently answering questions about eBird data and how it is warehoused; Ken Rosenberg (Cornell Lab of Ornithology) for assistance in compiling the species matrix table accompanying this dataset; and Ben Zuckerberg (Cornell Lab of Ornithology) for assistance in configuring FRAGSTATS. Thank you also to Thomas Finley (Microsoft) for early work done to clean data. Finally, thank you to Giles Hooker (Cornell University), Rebecca Hutchison (Oregon State University), and Thomas Dietterich (Oregon State University) for useful feedback on this dataset.

References

[MCNE] K. McGarigal, S. A. Cushman, M. C. Neel, and E. Ene. *FRAGSTATS: Spatial Pattern Analysis Program for Categorical Maps.* University of Massachusetts, Amherst. Version 3. Available from: `http://www.umass.edu/landeco/research/fragstats/fragstats.html`.

A Covariate Processing Details

Covariate processing depends on whether the raw, unprocessed data are stored as shape files or raster data. Shape files are lists of polygons and their locations, and a covariate's value is constant within a polygon's region. Raster data represent the covariate data as a regular grid superimposed on a spatial area; each grid cell contains the covariate value for the corresponding region.

The census data (covariates HOUSING_DENSITY, HOUSING_PERCENT_VACANT, and POP00_SQMI) and climate data (all covariates named CAUS_) start as shape files that are subsequently processed using the Geospatial Data Abstraction Library (GDAL)[5]. A C++ program calls the library to convert the decimal latitude and decimal longitude for each sampling event into the native coordinate system for a particular shape file and then assigns each event the appropriate covariate value based on the polygon containing each location. Locations that fall outside of all polygons are assigned NODATA values that are later converted into missing values (represented as `?` in the dataset). Locations can fall outside all polygons because the location is outside the boundaries of the covariate data source, or because there are quirks with the data source (e.g., the climate data polygons vary in their boundaries from month to month).

The covariates BCR, COUNTRY, STATE_PROVINCE, and SUBNATIONAL2_CODE also started as shape files from ESRI that were then loaded into a spatial database (Oracle Spatial) that directly supports location-based queries.

The raw elevation, canopy, and impervious surface data are stored as raster images. A second C++ program calls the GDAL library to associate these data with sampling event locations. Decimal latitude and longitude are converted from the WGS 84 coordinate system to the coordinate system of the raster data. Next, the location is re-projected to a flat surface and converted to pixel coordinates to facilitate a fast data lookup. For the canopy and impervious surface covariates, a GDAL-based C++ program computes the mean data value for a square neighborhood around the location.

The landcover-based (habitat) covariates are also derived from raster data; these covariate values are statistics computed by the FRAGSTATS program [MCNE] using MODIS (MCD12Q1), NLCD-2001, and NLCD-2006 raster data as input. Version 1.0 of the ERD and SRD datasets contain FRAGSTATS covariate statistics generated from v3.3 of the FRAGSTATS program. Version 2.1 onward of the ERD and SRD datasets contain FRAGSTATS covariate values generated from v2.0 of the FRAGSTATS program. Source code for version 2.0 of the Fragstats program is publicly available and provides for much higher processing throughput with the drawback of generating a less rich suite of statistics compared to those generated by the 3.3 release. Processing locations using the v2.0 release of the Fragstats code makes use of the GDAL library to extract a square neighborhood mini-raster around the location being processed, and then passing this mini-raster buffer to the Fragstats processing code built into the program. The processing pipeline for the v3.3 Fragstats is described as follows. A python wrapper script takes the decimal latitude and longitude

[5] `http://www.gdal.org`

coordinates for a sampling event and prepares a mini-raster data file describing the landscape neighborhood around the location. The script then passes the mini-raster to the FRAGSTATS executable. To create the mini-raster, the script reprojects the coordinates using an Albers Equal Area projection (using the GDAL utility program `cs2cs`) and calls the ArcGIS scripting library to extract the neighborhood around the location into the mini-raster. The mini-raster includes a landscape border around the neighborhood data (1 extra row / column on all sides) as recommended in the FRAGSTATS user manual.[6] The resolution of the mini-raster matches the original landcover data's resolution (30m x 30m).

A.1 FRAGSTATS Configuration

FRAGSTATS is run with the following configuration settings:

- Patch neighbors use the 8 cell rule (i.e., cells with touching corners are adjacent).
- Boundary does not count as edge (since there is a landscape border outside the landscape extent).
- Cells with unknown (missing) values are coded as background (value 9999).
- A similarity index file is used to customize the similarity of the different habitat classes defined in the National Land Cover Database. By default, each habitat class has 0 similarity with the other classes and perfect similarity with itself (1). We use the following custom similarities:
  - deciduous forest (class 41) and evergreen forest (class 42): 0.2
  - deciduous forest and mixed forest (class 43): 0.7
  - evergreen forest and mixed forest: 0.2
  - pasture, hay (class 81) and cultivated crops (class 82): 0.7
  - grassland, herbaceous (class 71) and pasture, hay: 0.3
  - grassland, herbaceous and cultivated crops: 0.3

  We believe these settings are adequate for most bird species.
- Similarly, we use an edge contrast file to soften the edges (dissimilarity) between the above habitat class pairs. Specifically, the edge contrast for the above pairs is set to 1 minus the similarity setting.

[6] See the section *Backgrounds, Borders, and Boundaries* in the FRAGSTATS user guidelines.

A.2 Hydrography Covariate Distance Values

Each hydrography feature type listed in Static Covariate Table 3 has a distance-in and distance-from value. If a checklist lat/lon location is inside a particular water feature, the distance-in value is described by the range bins listed in the Hydrography Distance Values Table 6; the matched distance-from value for this location will be represented as a missing value. A lat/lon location can only be inside a single hydrography feature type, so other distance-in variables for this location will be represented as missing values with the remaining distance-from values described by the range bins listed in the Hydrography Distance Values Table 6 as appropriate.

Similarly, if a checklist lat/lon location is outside all water features, the distance-from values are described by the range bins listed in the Hydrography Distance Values Table 6, and all distance-in values for this location will be represented as a missing value.

Table 6: Hydrography Distance Values

Value	**Distance Range (meters)**
1	0 - 30
2	30 - 60
3	60 - 120
4	120 - 250
5	250 - 500
6	500 - 1,000
7	1,000 - 2,000
8	2,000 - 4,000
9	>4,000

A.3 Protocol Listing

Table 7: Protocol Codes and Names

Protocol Code	**Protocol Name**
P20	eBird - Casual Observation
P21	eBird - Stationary Count
P22	eBird - Traveling Count
P23	eBird - Exhaustive Area Count
P34	RMBO Early Winter Waterbird Count
P35	eBird My Yard Count
P39	eBird Vermont - LoonWatch
P40	My Yard eBird - Standardized Yard Count
P41	eBird–Rusty Blackbird Blitz
P44	eBird California - Yellow Billed Magpie General
P45	eBird California - Yellow Billed Magpie Traveling
P46	eBird Caribbean - CWC Stationary Count
P47	eBird Caribbean - CWC Area Search
P48	eBird Random Location Count
P49	eBird Peru–Coastal Shorebird Survey
P50	Caribbean Martin Survey
P51	Audubon NWR Protocol
P52	eBird - Oiled Birds
P55	eBird–Heron Stationary Count
P56	eBird–Heron Area Count

A.4 Climate Covariate Value Mappings

The following set of tables describe the value mapping scheme used for the respective covariates.

Table 8: CAUS_PREC and CAUS_PRECMM (MM = 2 digit month of year)

Value	**Range in INCHES**
1	< 0.51
2	0.51 - 1.00
3	1.01 - 1.50
4	1.51 - 2.00
5	2.01 - 3.00
6	3.01 - 5.00
7	5.01 - 10.00
8	10.01 - 20.00
9	> 20.00

Table 9: CAUS_PREC*13* (*13* = ANNUAL)

Value	**Range in INCHES**
1	< 5.01
2	5.01 - 12.00
3	12.01 - 20.00
4	20.01 - 30.00
5	30.01 - 40.00
6	40.01 - 50.00
7	50.01 - 70.00
8	70.01 - 100.00
9	> 100.00

Table 10: CAUS_SNOW and CAUS_SNOW*MM* (*MM* = 2 digit month of year)

Value	**Range in INCHES**
1	< 0.5
2	0.5 - 2.4
3	2.5 - 5.4
4	5.5 - 10.4
5	10.5 - 15.4
6	15.5 - 20.4
7	20.5 - 30.4
8	30.5 - 60.4
9	> 60.4

Table 11: CAUS_TEMP_AVG and CAUS_TEMP_AVG*MM* (*MM* = 2 digit month of year)

Value	**Temperature Range in DEGREES FAHRENHEIT**
1	< 20.0
2	20.0 - 32.0
3	32.1 - 40.0
4	40.1 - 50.0
5	50.1 - 60.0
6	60.1 - 70.0
7	70.1 - 80.0
8	80.1 - 90.0
9	> 90.0

Table 12: CAUS_TEMP_AVG*13* (*13* = ANNUAL)

Value	**Temperature Range in DEGREES FAHRENHEIT**
1	< 32.0
2	32.0 - 40.0
3	40.1 - 45.0
4	45.1 - 50.0
5	50.1 - 55.0
6	55.1 - 60.0
7	60.1 - 65.0
8	65.1 - 70.0
9	> 70.0

Table 13: CAUS_TEMP_MIN and CAUS_TEMP_MIN*MM* (*MM* = 2 digit month of year)

Value	**Temperature Range in DEGREES FAHRENHEIT**
1	< 0.1
2	0.1 - 15.0
3	15.1 - 25.0
4	25.1 - 32.0
5	32.1 - 40.0
6	40.1 - 50.0
7	50.1 - 60.0
8	60.1 - 70.0
9	> 70.0

Table 14: CAUS_TEMP_MIN*13* (*13* = ANNUAL)

Value	**Temperature Range in DEGREES FAHRENHEIT**
1	< 20.1
2	20.1 - 32.0
3	32.1 - 40.0
4	40.1 - 45.0
5	45.1 - 50.0
6	50.1 - 55.0
7	55.1 - 60.0
8	60.1 - 70.0
9	> 70.0

Table 15: CAUS_TEMP_MAX and CAUS_TEMP_MAX*MM* (*MM* = 2 digit month of year)

Value	**Temperature in DEGREES FAHRENHEIT**
1	< 32.1
2	32.1 - 40.0
3	40.1 - 50.0
4	50.1 - 60.0
5	60.1 - 70.0
6	70.1 - 80.0
7	80.1 - 90.0
8	90.1 - 100.0
9	> 100.0

Table 16: CAUS_TEMP_MAX*13* (*13* = ANNUAL)

Value	**Temperature Range in DEGREES FAHRENHEIT**
1	< 40.1
2	40.1 - 50.0
3	50.1 - 60.0
4	60.1 - 65.0
5	65.1 - 70.0
6	70.1 - 75.0
7	75.1 - 80.0
8	80.1 - 85.0
9	> 85.0

Table 17: CAUS_LAST_SPRING_32F_MEAN

Value	**Date Range**
-1	RARE OR NO FREEZE
1	RARE OR NO FREEZE
2	JAN 1 - FEB 28
3	MAR 1 - MAR 31
4	APR 1 - APR 15
5	APR 16 - APR 30
6	MAY 1 - MAY 15
7	MAY 16 - MAY 31
8	JUN 1 - JUN 30
9	JUL 1 - JUL 31

Table 18: CAUS_LAST_SPRING_32F_MEDIAN

Value	Date Range
-1	RARE OR NO FREEZE
1	RARE OR NO FREEZE
2	JAN 1 - FEB 28
3	MAR 1 - MAR 31
4	APR 1 - APR 15
5	APR 16 - APR 30
6	MAY 1 - MAY 15
7	MAY 16 - MAY 31
8	JUN 1 - JUN 30
9	JUL 1 - JUL 31

Table 19: CAUS_LAST_SPRING_32F_EXTREME

Value	Date Range
1	RARE OR NO FREEZE
2	JAN 1 - FEB 28
3	MAR 1 - MAR 31
4	APR 1 - APR 15
5	APR 16 - APR 30
6	MAY 1 - MAY 15
7	MAY 16 - MAY 31
8	JUN 1 - JUN 30
9	JUL 1 - JUL 31

Table 20: CAUS_FIRST_AUTUMN_32F_MEAN

Value	Date Range
1	RARE FREEZE
2	NO FREEZE
3	AUG 1 - AUG 31
4	SEP 1 - SEP 30
5	OCT 1 - OCT 15
6	OCT 16 - OCT 31
7	NOV 1 - NOV 15
8	NOV 16 - NOV 30
9	DEC 1 - DEC 31

Table 21: CAUS_FIRST_AUTUMN_32F_MEDIAN

Value	**Date Range**
1	RARE FREEZE
2	NO FREEZE
3	AUG 1 - AUG 31
4	SEP 1 - SEP 30
5	OCT 1 - OCT 15
6	OCT 16 - OCT 31
7	NOV 1 - NOV 15
8	NOV 16 - NOV 30
9	DEC 1 - DEC 31

Table 22: CAUS_FIRST_AUTUMN_32F_EXTREME

Value	**Date Range**
1	RARE FREEZE
2	NO FREEZE
3	AUG 1 - AUG 31
4	SEP 1 - SEP 15
5	SEP 16 - SEP 30
6	OCT 1 - OCT 15
7	OCT 16 - OCT 31
8	NOV 1 - NOV 30
9	DEC 1 - DEC 31

B Change History

2013/9/30 Dataset release and document update (version 5.0) Release of Western Hemisphere data set, addition of MODIS/ASTER covariates, documentation of climate variable mapping - K. Webb

2012/10/1 Dataset release and document update (version 4.0) - K. Webb

2012/6/5 Dataset release and document update (version 3.1) Includes taxonomy updates and corrected core climate variables referenced in section 4.4 - K. Webb

2011/4/19 Dataset release and document update (version 3.0) - K. Webb

2010/12/21 Dataset release and document update (version 2.1) - K. Webb

2010/6/28 Dataset release and document update (version 2.0) - K. Webb

2010/2/4 Updated documentation to describe covariate processing and include SRD dataset.

2009/8/5 Release of covariates for 130,000 random locations in contiguous USA (SRD dataset).

2009/6/8 Initial dataset release (version 1.0).