

The eBird Reference Dataset

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

June 5, 2009

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. The observational data included in the data set have data access level 5 in the Avian Knowledge Network (AKN) data warehouse and are published here in compliance with the AKN data sharing policy.¹ eBird² is run by the National Audobon Society³ and the Cornell Lab of Ornithology⁴, 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 2009 that were submitted to eBird and reviewed by January 21, 2009. 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*. Cornell Lab of Ornithology and National Audobon Society, Ithaca, NY, June 2009.

Please direct any questions to:

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

© 2009 by Cornell Lab of Ornithology and the National Audobon Society.

¹See <http://www.avianknowledge.net/content/about/akn-data-sharing-policy> for details.

²<http://ebird.org>

³<http://www.audobon.org>

⁴<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 the 48 states in the contiguous United States is additionally annotated with hundreds of 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.

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.

AMERICAS — all checklists from North, Central and South America, including the Caribbean islands. No location-based covariates are provided.

The primary difference is availability of covariates. Many researchers will want to start with the US48 dataset. This includes roughly 85% of all eBird observations and comes with detailed location-based covariates describing habitat, climate and human population. However, it is limited to the contiguous US due to availability of the datasets (e.g., US census) that provide the covariates.

Some researchers may feel restricted by political borders and want access to all of the eBird reports. The AMERICAS dataset includes all eBird reports from the Americas. Researchers may reference external datasets for location-based covariates that extend outside of the United States.

The two datasets are organized in a similar fashion, and observations are *repeated* between them. That is, all checklists included in the US48 dataset are also included in the AMERICAS dataset.

3.1 Column Summaries

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

- Checklist description. About 15 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. In the US48 data set there is one column per species, with each column listing the number of birds observed. In the AMERICAS data set the observation count data uses a sparse representation, and only the observed species are listed (all within a single column called SPECIESLIST). See sections 4.2 and 4.3 for more details.
- Core covariates (US48 only). Approximately 25 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 only). These include extensive statistics about habitat configuration, fine-grained climate measurements, and observer expertise. See section 4.5 and Tables 3 and 4.

In the US48 dataset, all of the files are row aligned. That is, the covariates for the sampling event in line 2 of `checklists.csv` are in lines 2 of `core_covariates.csv` and `extended_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 of the US48 dataset looks like this:

us48/2007:

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

us48/2008:

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

The structure of the AMERICAS dataset is similar except it excludes the covariates.

americas/2007:

```
checklists.csv
```

americas/2008:

```
checklists.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 2008. In US48, 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 provenance.

4.2 Species Counts

All species reported in complete checklists to eBird are included in the reference dataset. Species not reported within the contiguous US are not included in the US48 dataset. When a birder submits a *complete checklist*, they are reporting on all birds that they were able to identify. By only including complete checklists in the dataset, one can assume that if a species is not reported it was not present.⁵ In the US48 dataset, there is one column per species, containing the count of how many were observed or a 0 if the species was unobserved (the zero-filling relies on the above assumption).

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 (i.e. if the species is not particularly interesting to the birder). These 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 available at <http://www.avianknowledge.net/content/files/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. A copy of the file is also included in the `docs/` folder.

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.

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

⁵Casual count checklists are also excluded, as we believe these checklists are much less thorough. *Casual counts* are observations made while birding was not the submitter's primary activity.

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), or area (P23, P35). 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 = 1:30PM). 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.

one will have a measurement while the other's value is missing. See section 4.5 for details about the NLCD2001 covariates.

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).
CAUS_PREC†	Mean total precipitation for month in which observation made.
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.
CAUS_TEMP_AVG†	Mean daily average temperature for month in which observation made.
CAUS_TEMP_MIN†	Mean daily minimum temperature for month in which observation made.
CAUS_TEMP_MAX†	Mean daily maximum temperature for month in which observation made.
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.
NLCD2001_FS_CYY_7500_PLAND	Percent of surrounding landscape that is habitat class YY. See text and Table 4 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/climat1s2/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, and are tied to the location of the observation. 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 summarizes the static environment covariates. Information about landcover / habitat comes from raster data that can be downloaded from <http://www.mrlc.gov/nlcd.php>.

Ecologists generally agree that both the type of habitat and its configuration are important factors in

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.
CAUS_SNOWMM†	Mean snow depth for month MM.
CAUS_TEMP_AVGMM†	Mean daily average temperature for month MM.
CAUS_TEMP_MINMM†	Mean daily minimum temperature for month MM.
CAUS_TEMP_MAXMM†	Mean daily maximum temperature for month MM.
CAUS_LAST_SPRING_32F_MEAN†	Last 32 F day in spring (mean). Numbers indicate date ranges.
CAUS_LAST_SPRING_32F_MEDIAN†	Last 32 F day in spring (median).
CAUS_LAST_SPRING_32F_EXTREME†	Last 32 F day in spring (extreme dates).
CAUS_FIRST_AUTUMN_32F_MEAN†	First 32 F day in autumn (mean). Numbers indicate date range.
CAUS_FIRST_AUTUMN_32F_MEDIAN†	First 32 F day in autumn (median).
CAUS_FIRST_AUTUMN_32F_EXTREME†	First 32 F day in autumn, (extreme dates).
NCLD2001_FS_*	Landscape and landcover statistics describing the habitat in a square neighborhood around the location. Computed from the 2001 National Land Cover Data from MRLC (www.mrlc.gov/nlcd.php) using the FRAGSTATS program. See text and Table 4 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>

ecological processes.⁶ Consequently, we processed the raw 2001 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. 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 NLCD database 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, we repeated this process for three different spatial extents: 2.25 hectares, 225 hectares, and 22,500 hectares. These extents were selected to cover local ecological processes at small, medium, and large ranges. 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 4.

⁶A 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 4: Summary of habitat statistics.

Variable Name	Comments
<i>Class Level Statistics</i>	
NLCD2001_FS_CYY_RR_AI	Aggregation index. Measures the clumpiness of patches of habitat type YY. Maximized at 100 when class is clumped into a single, compact patch; 0 when none of the YY raster pixels are adjacent. NA values occur if no pixel adjacencies are possible; this happens if the landscape contains exactly 1 or 0 raster pixels of type YY. The former case is conceptually maximally aggregated and is recoded as 100. The latter case is recoded as 0, since one cannot aggregate nothing.
NLCD2001_FS_CYY_RR_AREA_AM	Area-weighted mean patch area of habitat type YY. NA values occur if no YY patches in landscape. Recoded as 0 (b/c mean patch size is 0).
NLCD2001_FS_CYY_RR_AREA_CV	Coefficient of variation in patch area for habitat patches of type YY. NA occurs if no patches of type YY. Recoded as 0.
NLCD2001_FS_CYY_RR_AREA_SD	Standard deviation in patch area for habitat patches of type YY. NA occurs if no patches of type YY. Recoded as 0.
NLCD2001_FS_CYY_RR_ECON_AM	Area-weighted mean edge contrast. Average of edge contrast index scores for habitat YY patches, with each patch's contribution weighted by its area. Edge contrast is how different the habitat is from adjacent patches of different habitat type. NA occurs if landscape does not contain YY patches. Recoded as 0.
NLCD2001_FS_CYY_RR_ED	Edge density for patches of habitat type YY. Ratio of total edge length to landscape area (meters per hectare). NA occurs if habitat YY not in landscape. Recoded as 0 since no edges implies 0 density.
NLCD2001_FS_CYY_RR_ENN_AM	Average nearest neighbor distance between patches of habitat type YY; each patch's contribution is weighted by its area. NA occurs if landscape contains 0 or 1 patches of type YY. <i>There is not an obviously logical numeric recoding, so this is left as "NA" in the data.</i>
NLCD2001_FS_CYY_RR_ENN_CV	Coefficient of variation of nearest neighbor distance between patches of type YY. NA occurs if landscape contains 0 or 1 patches of type YY. <i>NA is not recoded.</i>
NLCD2001_FS_CYY_RR_ENN_SD	Standard deviation of nearest neighbor distance between patches of type YY. NA occurs if landscape contains 0 or 1 patches of type YY. <i>NA is not recoded.</i>
NLCD2001_FS_CYY_RR_FRAC_AM	Average fractal dimension of habitat YY patches, weighted by their areas. Ranges from 1 (simple shapes) to 2 (complex patch shapes). NA occurs if habitat YY not in landscape. Recoded as 0, based on the reasoning that complete absence is a <i>very</i> simple shape.
NLCD2001_FS_CYY_RR_LPI	Largest patch index. Percentage of landscape area comprised by the largest patch of habitat type YY. NA occurs if habitat YY not in landscape. Recoded as 0.
NLCD2001_FS_CYY_RR_PD	Patch density. Number of patches of habitat class YY per 100 hectares in surrounding landscape. NA values occurred if no YY patches in landscape. Recoded as 0.
NLCD2001_FS_CYY_RR_PLAND	Percent of surrounding landscape that is habitat class YY. NA values occurred if no YY patches in landscape. Recoded as 0.

* All habitat statistics summarize square neighborhood around location with "radius" RR.

continued on next page

Table 4: Summary of habitat statistics — *continued*

Variable Name	Comments
NLCD2001_FS_CYY_RR_PROX_AM	Area-weighted mean proximity distance between patches of habitat YY. Zero if only one YY patch. NA occurs if habitat YY not in landscape. Recoded as 0 since a non-existent patch is not proximal to anything.
NLCD2001_FS_CYY_RR_SIML_AM	Area-weighted mean similarity between patches of habitat YY and <i>all</i> other patches (incl. other habitat types). Zero if all patches close to YY patches have 0 similarity coefficients. NA occurs if habitat YY not in landscape. Recoded as 0 since non-existent patch is not proximal to anything.
<i>Landscape Level Statistics</i>	
NLCD2001_FS_L_RR_AI	Area-weighted mean class aggregation index. Average of class-specific aggregation index scores, weighted by their total area in landscape.
NLCD2001_FS_L_RR_AREA_AM	Area-weighted mean patch area, for all habitat types present in landscape.
NLCD2001_FS_L_RR_AREA_CV	Coefficient of variation of patch area, over all habitat types.
NLCD2001_FS_L_RR_AREA_SD	Standard deviation of patch area, over all habitat types.
NLCD2001_FS_L_RR_ECON_AM	Area-weighted mean edge contrast for all habitat patch types.
NLCD2001_FS_L_RR_ED	Edge density for the landscape. Ratio of sum of all edges between patches over total landscape area (meters per hectare).
NLCD2001_FS_L_RR_ENN_AM	Area-weighted mean nearest neighbor distance between patches in landscape.
NLCD2001_FS_L_RR_ENN_CV	Coefficient of variation of distances between patches.
NLCD2001_FS_L_RR_ENN_SD	Standard deviation of distances between patches.
NLCD2001_FS_L_RR_FRAC_AM	Average fractal dimension of patches in landscape, weighted by the area of each patch. Ranges from 1 (simple patch shapes) to 2 (complex patch shapes).
NLCD2001_FS_L_RR_LPI	Percentage of landscape area occupied by the largest patch (any habitat type).
NLCD2001_FS_L_RR_PD	Patch density (number of patches per 100 hectares).
NLCD2001_FS_L_RR_PROX_AM	Area-weighted mean proximity between all patches in landscape.
NLCD2001_FS_L_RR_SIML_AM	Area-weighted mean similarity between all patches in landscape.

* All habitat statistics summarize square neighborhood around location with “radius” RR.

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

The authors warmly thank Tim Levatch 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 FRAGSTATS Configuration

We are still documenting the methodology used for producing the covariates. This section will be completed shortly.