Simulating avian body mass measurements using `birdsize`

Introduction

Different currencies of measurement - e.g. total number of individuals, total biomass, or total metabolic flux or energy use - provide linked, but qualitatively very different, perspectives on the structure and function of ecological systems (White et al. (2007)). The study of the interrelated dynamics of size structure, species composition, individual abundance, and biomass and energy use is well-established for systems for which data on both individuals' body sizes and individual organismal abundance are widely available, including aquatic systems, terrestrial forest systems, and, to a lesser extent, small mammal systems (Kerr and Dickie (2001), White et al. (2007)). Work in these systems has yielded important insight into - for example - how ecological degradation can manifest in the relationship between total abundance and total biomass (Warwick and Clarke (1994)), or how shifts in community-wide mean body size can buffer total energy use against apparent changes in total individual abundance (White et al. (2004)). Efforts to generalize these efforts to terrestrial vertebrate systems have been constrained due to the lack of body size measurements for these communities (White et al. (2007), Thibault et al. (2011)). Sampling methodologies for avian communities often rely on visual or auditory point-counts, which provide information about species abundance and diversity but do not directly capture information about body size or energy use.

The `birdsize` R package offers a way around this limitation by estimating individual-level (and, from there, population or community-wide) body size measurements for birds given either species identity or a species' mean and/or standard deviation of body size. Birds exhibit determinate growth, and `birdsize` assumes that intraspecific body size distributions for birds are, to a first approximation, well-described by normal distributions parameterized with a species-specific mean and standard deviation (see also Thibault et al. (2011)). Moreover, there is a strong scaling relationship between a species' mean body size and its standard deviation of body size, meaning that, for species for which the standard deviation is not known, the standard deviation can be estimated from the mean (see also Thibault et al. (2011)). Estimates obtained in this way are, of course, considerably less precise than those that could be obtained through exhaustive field sampling, and may not be appropriate for all use cases. However, given the logistical constraints on field operations of this scale (and the even harsher constraint of time, which prevents us from retroactively taking these measurements for ecological timeseries), `birdsize` makes it possible to conduct macroecological-scale analyses of avian communities that would not otherwise be possible. This approach was first used at scale by Thibault et al. (2011) and subsequently by Diaz and Ernest (2022) (in review). `birdsize` formalizes this method and makes it accessible via a straightforward user interface, in order to facilitate use by other research groups with diverse use cases.

The estimation procedure in `birdsize`

The core functionality of `birdsize` is to generate estimates of individual body size for populations of birds by drawing from a normal distribution parameterized with a species-level mean and standard deviation of body size. It includes built-in values for these parameters for 443 species found in the North American Breeding Bird Survey (Pardieck et al. (2019)), and can accept user-supplied parameter values for additional species.

For the 443 species included with `birdsize`, mean and standard deviation values were manually obtained from the CRC Handbook of Avian Body Masses (Dunning (2008)). These species are listed in the data frame `birdsize::known_species`. Many species in Dunning (2008) have multiple records from different time periods, locations, and subspecies. In these instances, parameter values are averaged across records to obtain a single species-wide value. For records in Dunning (2008) with mean, but no standard deviation, reported, the standard deviation is estimated via a scaling relationship between the mean and standard deviation of body mass (see also Thibault et al. (2011)). Specifically, a linear model of the form `log(variance(body_size)) ~ log(mean(body_size))` has a model R^2 of 0.89, and produces the scaling relationship of `variance(body_size) = 0.0047(body_size) ^ 2.01`. This scaling relationship is used to generate estimated standard deviations for records without standard deviation recorded, affecting 353 of 928 raw records.

A user may also manually supply parameter values, in order to generate estimates for species not included in `birdsize::knownspecies`, or to use different parameter values than those included with `birdsize`. This may be of particular interest for users wishing to explore questions related to (for example) intraspecific variation in body size across different populations of the same species, or extending to species not common to North America. In this case, if both mean and standard deviation are supplied, they will be used, and if only the mean is provided, the standard deviation is estimated via the scaling relationship explained above.

Population and community-wide summaries

While `birdsize` generates estimated body size measurements at the level of individual birds, in many instances the quantity of interest is actually the population or community-wide total biomass or metabolic rate. Indeed, given the several layers of estimation involved in obtaining measurements via `birdsize`, it is likely to generally be more appropriate to focus on these aggregate properties than on estimates for "individuals". Accordingly, `birdsize` includes functions to compute these summaries, grouping by species, year, or other variables supplied by the user. These are demonstrated in the package vignettes and use cases, below.

Integration with the Breeding Bird Survey

The methodology in `birdsize` was first developed and applied to the North American Breeding Bird Survey, and `birdsize` is built to naturally accommodate Breeding Bird Survey data obtained from ScienceBase (Pardieck et al. (2019)) or tools such as the Data Retriever (Senyondo et al. (2017)). There is no actual data from the Breeding Bird Survey included in the `birdsize` package, and users are encouraged to access the most up-to-date data from the creators directly. To facilitate this, the `bbs-data` and demonstration vignettes illustrate how to access these data and use them with `birdsize`, and the example data tables in `birdsize` (i.e. `demo_route_raw` and `demo_route_clean`) contain synthetic data matching the format of the Breeding Bird Survey.

However, `birdsize` is not constrained to work *only* with Breeding Bird Survey data. It accepts any dataset, real or synthetic, that includes population sizes and species identity and/or body size parameters (see above); see Use case #3, below.

Use case 1: Simulation over the Breeding Bird Survey timeseries

A common anticipated use case for `birdsize` is to generate estimates of species- and community- level biomass and metabolic rate for a Breeding Bird Survey route over time. Here, we generate these estimates using the `demo_route_raw` dataset, which has the same shape and structure as data from the Breeding Bird Survey, but contains simulated values for the actual data.

First, it is recommended to clean the raw data to remove species poorly sampled via Breeding Bird Survey methods and remove records not identified to species. This is accomplished using the `filter_bbs_survey` function:

```
clean_data = filter_bbs_survey(demo_route_raw)
head(clean_data)
```

```
##   record_id   routedataid countrynum statenum route rpid year  aou count10
## 1    900000 9009911011994        900       99     1  101 1994 4730       8
## 2    900001 9009911011995        900       99     1  101 1995 4730      13
## 3    900002 9009911011996        900       99     1  101 1996 4730       8
## 4    900003 9009911011997        900       99     1  101 1997 4730       9
## 5    900004 9009911011998        900       99     1  101 1998 4730      10
## 6    900005 9009911011999        900       99     1  101 1999 4730      12
##   count20 count30 count40 count50 stoptotal speciestotal
## 1      12      15      12      15         5           62
## 2       9      11      10      10         5           53
## 3      11       9      13      15         5           56
## 4      13      16       9      12         5           59
## 5       6      12       8       7         5           43
## 6      13       5       9       5         5           44
```

For the purposes of simulating body size and metabolic rate, the relevant columns in these data are `year`, `aou`, and `speciestotal`, which refer to the year of the survey, the species identity, and the total number of individuals of that species recorded on that route in that year, respectively.

Given a dataframe like this, `birdsize::community_generate` iterates over rows and draw `speciestotal` individuals of the appropriate species (identified by the `aou`, or species code). The resulting data frame has one row per simulated individual. It retains all columns from the original data frame, and adds columns for `sim_species_id`, `genus`, `species`, `individual_mass`, `individual_bmr`, `mean_size`, `sd_size`, `abundance`, and `sd_method`. Most of these are bookkeeping columns explained in the package documentation (see `?birdsize::community_generate`). Of particular relevance are the `individual_mass` and `individual_bmr` columns, which include the estimated body mass (in grams) and estimated basal metabolic rate for each simulated "individual". The `sd_method` column notes which method (see above) was used to obtain parameters for the species' mean and standard deviation body size. In this instance, it is `AOU lookup`, meaning parameters were obtained based on the `aou` column.

```
simulated_community <- community_generate(clean_data)

head(simulated_community)
```

```
##   record_id   routedataid countrynum statenum route rpid year count10 count20
## 1    900000 9009911011994        900       99     1  101 1994       8      12
## 2    900000 9009911011994        900       99     1  101 1994       8      12
## 3    900000 9009911011994        900       99     1  101 1994       8      12
## 4    900000 9009911011994        900       99     1  101 1994       8      12
## 5    900000 9009911011994        900       99     1  101 1994       8      12
## 6    900000 9009911011994        900       99     1  101 1994       8      12
##   count30 count40 count50 stoptotal speciestotal  aou sim_species_id  genus
## 1      15      12      15         5           62 4730           4730 Alauda
## 2      15      12      15         5           62 4730           4730 Alauda
## 3      15      12      15         5           62 4730           4730 Alauda
## 4      15      12      15         5           62 4730           4730 Alauda
## 5      15      12      15         5           62 4730           4730 Alauda
## 6      15      12      15         5           62 4730           4730 Alauda
##    species individual_mass individual_bmr mean_size  sd_size abundance
## 1 arvensis        38.47392       141.7141    37.475 3.300613        62
```

```
## 2 arvensis        39.43680       144.2338    37.475 3.300613        62
## 3 arvensis        34.41491       130.8849    37.475 3.300613        62
## 4 arvensis        36.32271       136.0180    37.475 3.300613        62
## 5 arvensis        40.87909       147.9755    37.475 3.300613        62
## 6 arvensis        37.79325       139.9219    37.475 3.300613        62
##    sd_method
## 1 AOU lookup
## 2 AOU lookup
## 3 AOU lookup
## 4 AOU lookup
## 5 AOU lookup
## 6 AOU lookup
```

These individual-level estimates can be condensed into year and species totals using `birdsize::community_summarize`. Summarizing by `"species_and_year"` will produce species-level totals for each year surveyed:

```
annual_species_summaries <- community_summarize(simulated_community, level = "species_and_year")

head(annual_species_summaries)
```

```
## # A tibble: 6 x 21
##   routed~1 count~2 state~3 route  rpid  year   aou sim_s~4 genus species mean_~5
##   <chr>      <dbl>   <dbl> <dbl> <dbl> <int> <int>   <int> <chr> <chr>     <dbl>
## 1 9009911~     900      99     1   101  1994  3000    3000 Bona~ umbell~    532
## 2 9009911~     900      99     1   101  1994  3151    3151 Stre~ chinen~    159
## 3 9009911~     900      99     1   101  1994  3152    3152 Stre~ roseog~    155
## 4 9009911~     900      99     1   101  1994  3280    3280 Elan~ leucur~    346
## 5 9009911~     900      99     1   101  1994  3460    3460 Buteo plagia~    528.
## 6 9009911~     900      99     1   101  1994  3550    3550 Falco mexica~    734
## # ... with 10 more variables: sd_size <dbl>, species_designator <chr>,
## #   total_abundance <int>, total_biomass <dbl>, total_metabolic_rate <dbl>,
## #   total_richness <int>, mean_individual_mass <dbl>, sd_individual_mass <dbl>,
## #   mean_metabolic_rate <dbl>, sd_metabolic_rate <dbl>, and abbreviated
## #   variable names 1: routedataid, 2: countrynum, 3: statenum,
## #   4: sim_species_id, 5: mean_size
```

Summarizing by only `"year"` will produce community-wide totals (over all species) for each year::

```
annual_summaries <- community_summarize(simulated_community, level = "year")

head(annual_summaries)
```

```
## # A tibble: 6 x 15
##   routedataid  count~1 state~2 route  rpid  year speci~3 total~4 total~5 total~6
##   <chr>          <dbl>   <dbl> <dbl> <dbl> <int> <chr>     <int>   <dbl>   <dbl>
## 1 90099110119~     900      99     1   101  1994 aou        1361 157199. 352926.
## 2 90099110119~     900      99     1   101  1995 aou        1443 162763. 366541.
## 3 90099110119~     900      99     1   101  1996 aou        1413 168319. 372057.
## 4 90099110119~     900      99     1   101  1997 aou        1381 159684. 358338.
## 5 90099110119~     900      99     1   101  1998 aou        1415 156643. 356783.
## 6 90099110119~     900      99     1   101  1999 aou        1412 164306. 368244.
## # ... with 5 more variables: total_richness <int>, mean_individual_mass <dbl>,
```

```
## #   sd_individual_mass <dbl>, mean_metabolic_rate <dbl>,
## #   sd_metabolic_rate <dbl>, and abbreviated variable names 1: countrynum,
## #   2: statenum, 3: species_designator, 4: total_abundance, 5: total_biomass,
## #   6: total_metabolic_rate
```

Similarly, summarizing by only `"species"` will produce species-level totals over all years:

```
species_summaries <- community_summarize(simulated_community, level = "species")

head(species_summaries)
```

```
## # A tibble: 6 x 19
##   countrynum statenum route  rpid   aou sim_spec~1 genus species mean_~2 sd_size
##        <dbl>    <dbl> <dbl> <dbl> <int>      <int> <chr> <chr>     <dbl>   <dbl>
## 1        900       99     1   101  3000       3000 Bona~ umbell~    532     38.7
## 2        900       99     1   101  3151       3151 Stre~ chinen~    159     11
## 3        900       99     1   101  3152       3152 Stre~ roseog~    155     11.0
## 4        900       99     1   101  3280       3280 Elan~ leucur~    346     23.3
## 5        900       99     1   101  3460       3460 Buteo plagia~    528.    37.8
## 6        900       99     1   101  3550       3550 Falco mexica~    734     51.0
## # ... with 9 more variables: species_designator <chr>, total_abundance <int>,
## #   total_biomass <dbl>, total_metabolic_rate <dbl>, total_richness <int>,
## #   mean_individual_mass <dbl>, sd_individual_mass <dbl>,
## #   mean_metabolic_rate <dbl>, sd_metabolic_rate <dbl>, and abbreviated
## #   variable names 1: sim_species_id, 2: mean_size
```

Finally, `community_summarize` can group by other variables as specified by setting `level = "custom"` and supplying column names via the `id_vars` argument. Here, we group by `genus` and `year`:

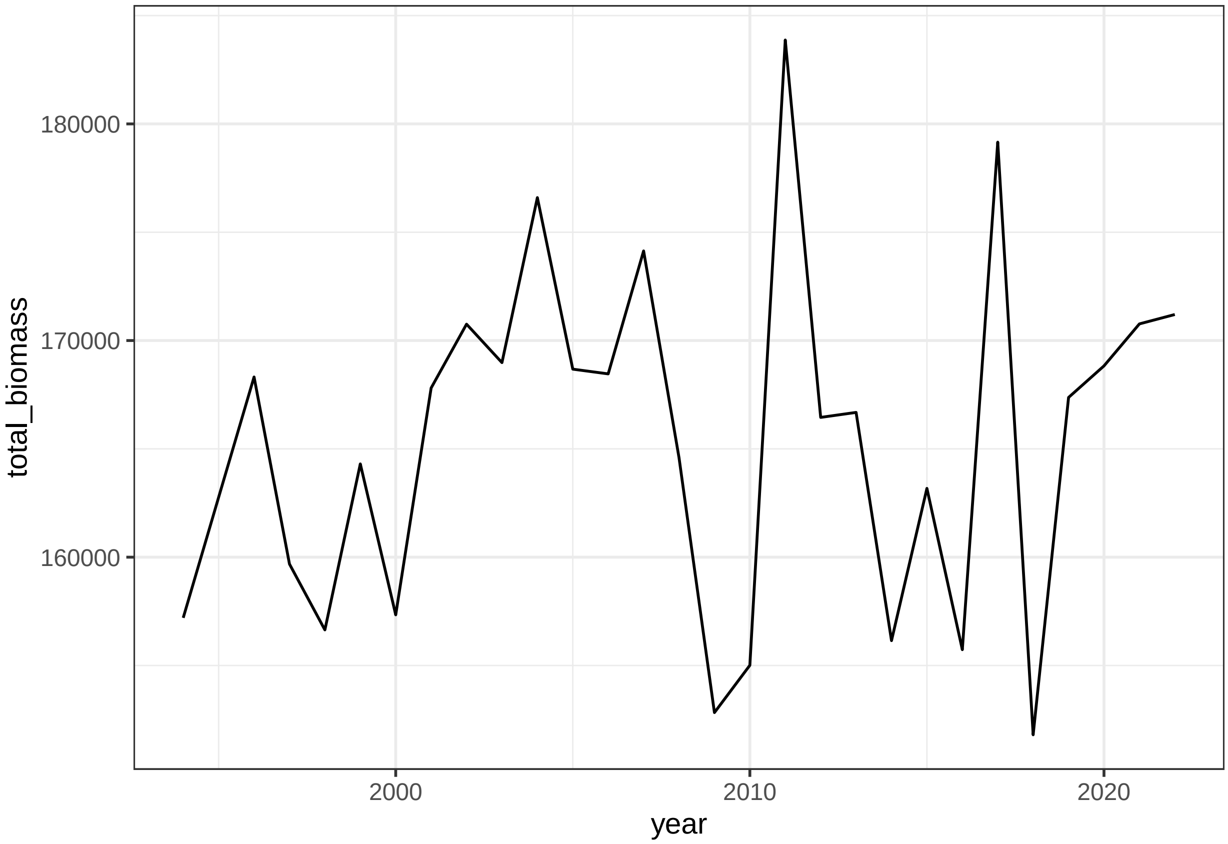
```
genus_year_summaries <- community_summarize(simulated_community, level = "custom", id_vars = c("year",

head(genus_year_summaries)
```

```
## # A tibble: 6 x 11
##    year genus    speci~1 total~2 total~3 total~4 total~5 mean_~6 sd_in~7 mean_~8
##   <int> <chr>    <chr>     <int>   <dbl>   <dbl>   <int>   <dbl>   <dbl>   <dbl>
## 1  1994 Acridot~ aou          13   1516.   4058.       1   117.   10.4     312.
## 2  1994 Alauda   aou          62   2342.   8665.       1    37.8   3.73    140.
## 3  1994 Amphisp~ aou          52    941.   4302.       1    18.1   0.202    82.7
## 4  1994 Bonasa   aou          45  23849.  41358.       1   530.   38.8     919.
## 5  1994 Buteo    aou          69  36993.  63943.       1   536.   37.6     927.
## 6  1994 Carduel~ aou          50    805.   3805.       1    16.1   1.03     76.1
## # ... with 1 more variable: sd_metabolic_rate <dbl>, and abbreviated variable
## #   names 1: species_designator, 2: total_abundance, 3: total_biomass,
## #   4: total_metabolic_rate, 5: total_richness, 6: mean_individual_mass,
## #   7: sd_individual_mass, 8: mean_metabolic_rate
```

These functions can be used to generate plots of species or community level biomass over time. For example, here we plot community-wide biomass in each year surveyed:

```
ggplot(annual_summaries, aes(year, total_biomass)) +
  geom_line()
```



Use case 2: Simulating temporal shifts in body size based on the Breeding Bird Survey

Use case 3: Simulating imaginary birds

References

Diaz, R. M., and S. K. M. Ernest. 2022, November. Temporal changes in the individual size distribution decouple long-term trends in abundance, biomass, and energy use of North American breeding bird communities. bioRxiv.

Dunning, J. B. 2008. CRC handbook of avian body masses. CRC handbook of avian body masses. 2nd ed. CRC Press, Boca Raton.

Kerr, S. R., and L. M. Dickie. 2001. The Biomass Spectrum: A Predator-Prey Theory of Aquatic Production. Pages 352 Pages. Columbia University Press.

Pardieck, K. L., D. J. Ziolkowski, M. Lutmerding, V. Aponte, and M.-A. Hudson. 2019. North American Breeding Bird Survey Dataset 1966 - 2018, version 2018.0. U.S. Geological Survey.

Senyondo, H., B. D. Morris, A. Goel, A. Zhang, A. Narasimha, S. Negi, D. J. Harris, D. G. Digges, K. Kumar, A. Jain, K. Pal, K. Amipara, and E. P. White. 2017. Retriever: Data Retrieval Tool. Journal of Open Source Software 2:451.

Thibault, K. M., E. P. White, A. H. Hurlbert, and S. K. M. Ernest. 2011. Multimodality in the individual size distributions of bird communities. Global Ecology and Biogeography 20:145–153.

Warwick, R. M., and K. R. Clarke. 1994. Relearning the ABC: Taxonomic changes and abundance/biomass relationships in disturbed benthic communities. Marine Biology 118:739–744.

White, E. P., S. K. M. Ernest, A. J. Kerkhoff, and B. J. Enquist. 2007. Relationships between body size and abundance in ecology. Trends in Ecology & Evolution 22:323–330.

White, E. P., S. K. M. Ernest, and K. M. Thibault. 2004. Trade-offs in Community Properties through Time in a Desert Rodent Community. The American Naturalist 164:670–676.