**Data analysis protocol**

Author: Jon F. Went, [jwent@ethz.ch](mailto:jwent@ethz.ch)

Date: 27.06.2023

Data sets used:

* North American Breeding Bird Survey
* National Land Cover Database
* Climatic data
* Elevation data
* Human footprint data
* Life history data (Bird *et al.* 2020)
* Ecoregion data
* Functional diversity data?
* Taxonomic diversity data?

# Methods

All statistical analyses were conducted using the programming language R version (4.2.3), within the application RStudio version (2023.06.0+421).

Data wrangling was done using the packages: dplyr, tidyverse, magrittr, lubridate, reshape2, vegan, tidyr, pvclust, …

Data visualizations were done using the packages: ggplot2, hrbrthemes, viridis, ggridges, mapview, …

# Data sources and preprocessing

## Biodiversity data

We used the North American Breeding Bird Survey (BBS) as the source of avian diversity data, given its extensive temporal and spatial coverage, as well as its well-established record in research. The North American Breeding Bird Survey is an ongoing long-term survey that has been conducted since 1966, covering Canada, the United States, and Mexico. It comprises records of bird species abundances. Our study focuses on the routes within the geographical boundaries of the contiguous United States, as the available land use data at high resolution and spanning multiple years encompasses this region.

The routes of the survey are situated along secondary roads and are approximately 40 km (24.5 miles) long, with 50 point counts conducted at 0.8 km (0.5 mile) intervals. At each stop, all birds seen or heard in a 400 m (0.25 mile) radius are counted during a 3-min period. The surveys take place within the 3-hours following sunrise from late April to July, coinciding with the peak of the breeding period of the region where the route is located. The routes are surveyed once per year by volunteers skilled in avian identification. The volunteers record the abundance data on provided data sheets. Additionally, supplementary information, such as weather variables and disturbance indicators such as numbers of passing cars, is collected. The same stop locations are surveyed in subsequent years to ensure consistency, although relocations can occur duet to access or safety reasons. By 2021, the database includes X routes, of which X were surveyed.

Prior to analysis, we filtered the data to include only routes with complete survey length (RouteTypeDetailID value greater than 2) and favorable weather conditions (Route Run Protocol ID equal to 0). Each route was then divided into five segments, consisting of 10 point count stops each, and only the first third and fifth segment were retained for analysis. This approach aimed to reduce spatial autocorrelation, avoid pseudoreplication, and therefore improve the association between land use data and bird diversity data. Further, we narrowed the bird species subset to only include those observed on a segment in 2000, 2001 and 2002. The abundance per bird species and segment was averaged for analysis. This selection was was motivated by the need to reduce stochastic effect in species abundance, address detectability issues, and focus on species with stable populations. The choice of these three years was influenced by the availability of land use data, with 2001 being the first year and the two adjacent years providing additional context.

Caveats:

* Observer effects: observer ID as random effects in model
* Spatial complexity due to unknown exact location of the stops along the routes 🡪 following Haddou et al: splitting the route into 5 segments and using only first, third and fifth segment could reduce the pseudoreplication and spatial autocorrelation, plus only routes used that follow protocol accurately (RPID 101)
* Temporal complexity as first stops are always earliest 🡪 time of day include as covariate in model
* Spatial autocorrelation: Unbalanced density of routes, biased towards the eastern USA,
* Detectability: very simplified way we handled it
* Population trends and stability: very simplified way of handling it by pooling data together

## Taxonomic, functional and life history data

To explore the effects of life history traits on the legacy effects resulting from land use change, we used the generation lengths reported by (Bird *et al.* 2020) to categorize bird species into groups characterized by fast versus slow life histories.

* Functional data?
* Taxonomic data? Jetz, which paper to cite?

## Landcover and environmental data

### Land use

We used the products of the National Land Cover Database (NLCD) developed by the U.S. Geological Survey (USGS). The NLCD provides high resolution (30 m pixels) classified land use data for the United States. At the time of the study, the latest product available covered the years 2001, 2004, 2006, 2008, 2011, 2013, 2016 and 2019. The NLCD uses 16 land cover classes, which we aggregated into eight: “urban” composed of the classes 21 to 24 (developed-open space, developed-low intensity, developed-medium intensity and developed-high intensity), “forest” composed of the classes 41 to 43 (deciduous forest, evergreen forest, mixed forest), “barren” composed of class 31 (barren land), “grass” composed of class 71 (grassland-herbaceous), “pasture” composed of class 81 (pasture-hay), “crop” composed of class 82 (cultivated crops), “wet” composed of class ?, and “other” for pixels within our study area unclassified by the NLCD classes mentioned above.

To be able to associate the avian diversity data to the land use data, we followed the methods outlined in (Haddou *et al.* 2022). Specifically, we built 500 m buffers around each BBS segment line, which was validated by the authors and is also close to the 400 m radius used by the BBS surveyors at a point count stop. In these buffer zones we extracted the land use classes from the pixels. Our eight land cover classes were calculated as percentage of the buffer zone. Subsequently, we calculated the land cover change as the difference in percentage within a segment buffer between available years.

Historically, land cover data is not directly comparable between years due to changes in interpretation methods. However, (Haddou *et al.* 2022) validate the use of the land cover data by showing the correlation between the change in land cover change calculated with the above mentioned method and the Land Cover Change Index product provided by the USGS.

To group the segments with similar habitats together, we used the U.S. level III ecoregion data provided by the U.S. Environmental Protection Agency. This was motivated by the potential difference of a habitat depending on its location in the U.S. For example, a forest in the Southern Florida Coastal Plain might have markedly different habitats compared to a forest in the Pudget Lowlands in Washington state. Consequently, we expect also markedly different bird communities and varying impacts on the community if the habitat is changed (SOURCES HERE?).

To find out which habitats are prominent in an ecoregion, we clustered the segments with similar land use together. We used the R package ‘pvclust’ to cluster the land use data of segments in ecoregions. Agglomerative hierarchical clustering is useful to cluster small data sets with continuous variables. In short, Ward’s method of minimal variance is used to merge clusters iteratively together while minimizing the within-cluster variance, in our case calculated by the Euclidean distance between centroids. Subsequently, bootstrapping is used to investigate cluster stability at each node and AU (Approximately Unbiased) p-values for all nodes calculated. We used a conservative cut-off AU p-value at 0.80 to pick the clusters.

Include: rarefaction of clusters, longitude and latitude, landscape heterogeneity calculations

Caveats:

* Historical differences in land cover classes, comparability between years
* validation of the 500m buffer
* land scape heterogeneity could be important
* validation of the scale of avian diversity metric 🡪 of segments within a cluster
* longitude and latitude?

### Climatic data

We downloaded climatic data from CHELSA version 2. The selected monthly variables were: precipitation, climate moisture index, maximum temperature and minimum temperature. Additionally, the annual soil water balance was included. For each variable, we aggregated the data across the years 1996-2001 and 2014-2019 to better capture climatic trends.

We calculated the mean annual precipitation sum for both time periods as a climatic variable. Additionally, we calculated the summer-winter difference, where the precipitation sum from October through March was substracted from the precipitation sum from April through September.

We calculated the mean annual climate moisture index for both time periods. Additionally,

# Model development