Does protected area connectivity moderate the efficacy of protection on tropical biodiversity? Evidence from a replication of Brodie et al. 2023

Peter Kedron, Lei Song, Wenxin Yang, Amy Frazier

2024-08-23

1 Introduction

This study is a *computational reproduction* of:

> Brodie, J.F., Mohd-Azlan, J., Chen, C. et al. Landscape-scale benefits of protected areas for tropical biodiversity. Nature 620, 807–812 (2023). https://doi.org/10.1038/s41586-023-06410-z

Using a causal framework that controls for forest structure, site accessibility, and geographic location through matching, Brodie et al. (2023) find evidence that protected areas (PA) preserve vertebrate biodiversity within their boundaries and in the adjacent unprotected landscape.

In this reproduction, we attempt to identically reproduce the primary results of the original study. The results of interest include 1) six measures of the effect of PA status on bird and mammal biodiversity, 2) six measures of the effect of adjacency to a large PA on bird and mammal biodiversity at unprotected sites, and 3) six measures of the effect of distance from a PA on bird and mammal biodiversity at unprotected sites. We attempt the reproduction twice. First, we use the data and code published with author correction in April 2024. Second, for completeness and because we had already begun our reproduction attempt before the correction was published, we attempt the reproduction using the code and data shared with the original publication in August 2023. A successful reproduction should recreate the numerical results published in Brodie et al. (2023), or the those published with the correction in April 2024.

All materials and procedures used in this reproduction are publicly available at GitHub **(LINK)** with the identifier **PROJECT DOI**. We implemented the reproduction in MacBook Pros using R programming language (Version 4.2.2). Structural causal modeling was performed using the DAGITTY package (Version #.#.#). We used the MatchIt package (Version #.#.#) to perform propensity score matching and the NLME package (Version #.#.#) to fit the linear mixed effects regression models.

2 Study design

We attempt to reproduce the statistical results of the original authors by implementing their workflow as identically as possible. This effort allows us to assess study design and conclusion validity of the original study. The spatial extent of this reproduction attempt is Southeast Asia, matching the area studied by the original authors. The spatial scale of our statistical analysis is the observation site level with national scale adjustments for human development levels. Our primary data source is the data files publicly shared by Brodie et al. (2023). All data acquisition from original sources followed the procedures presented by the authors. Both the original study and our reproduction attempt were conducted in R.

2.1 Original Study Design

The original study uses a quasi-experimental design with the objective of identifying the causal effect of PAs on tropical biodiversity in Southeast Asia while deconfounding for the influence of site accessibility and habitat quality. The authors use three measures of bird and mammal biodiveristy as their response variables - Species richness (SR), Functional richness (FR), and Phylogenetic diversity (PD). Bird observations were gathered from 1,079 sampling locations in the eBird database spanning the period between January 2015 and August 2021. Mammal observations were assembled by the authors from camera traps 1,365 camera stations deployed across the region.

The primary predictors of interest were 1) a binary measure indicating whether a observation site was located inside or outside a PA, 2) a binary indicator that identified of the PA closest to an unprotected site as larger than 500 sq km, and 3) a binary indicator that identified if the PA closest to an unprotected site was within 2km. Primary predictors measuring PA status were derived from the World Database of Protected Areas. Additional predictors used to deconfound for site accessibility, forest structure, understory density, and human development pressure were derived from the NASA Global Ecosystem Dynamics Investigation (GEDI) mission, and the UN Development program Human Development Index. Site accessibility was measured using circuit theory-based metrics of proximity to human development[1]. Forest structure and understory density were measured use three-dimensional metrics derived from GEDI mission. Code to produce these predictors was not provided by the authors.

Using propensity score matching to control for the confounds of location, site accessibility, and forest structure, Brodie et al. fit a linear mixed-effects models to estimate three PA related effects. The authors completed two sets of statistical analyses. First, the authors estimated the the effect of a site being located inside or outside a PA on bird and mammal biodiversity. In total, the authors fit six separate regression models - two response variable taxons (birds, mammals) for three biodiversity measures (SR, FR, PD) - testing the null hypothesis that:

> OR-Ho-1: The protected area status of a site has no effect on the level of mammalian of avian biodiversity observed at that site when adjusting for the confounds of site accessibility, habitat condition, and the socioeconomic development.

Brodie et al. find evidence that the legal designation of PAs provides substantial and significant benefits to Southeast Asian bird biodiversity. The authors did not find the same effect for mammals. None of the three measures of mammalian biodiversity was significantly different inside v. outside or PAs.

Second, Brodie et al. tested whether the biodiversity preserving effects of PA status have positive (spillover) or negative (leakage) effects on the biodiversity of unprotected areas surrounding PAs. Within the subset of observation sites outside PA, the authors tested for these effects using the same propensity score matching procedure and statistical framework presented above, but replaced the binary PA status predictor with the size of the nearest PA or the distance to the nearest PA in separate models. In total, the authors fit 12 separate regression models - two response variable taxons (birds, mammals), by three biodiversity measures (SR, FR, PD), by two PA measures (area, distance). The null hypothesis tested were:

> OR-Ho-2a: Being located within 2km of a protected area of at least 500ˆkm in size has no effect on the level of mammalian of avian biodiversity observed at an unprotected site when adjusting for the confounds of site accessibility, habitat condition, the socioeconomic development, and distance to that protected site.

> OR-Ho-2b: The distance to the protected area located closest to an unprotected observation site has no effect on the level of mammalian of avian biodiversity observed at that site when adjusting for the confounds of site accessibility, habitat condition, and the socioeconomic development.

[1] "circuit theoretical models parameterized with human travel speeds across different terrains and the locations of populations centers and transportation networks" (Brodie et al. 2023)

Brodie et al find evidence that large PAs are associated with higher biodiversities for mammals and birds in surrounding unprotected areas, but that the effects for birds are smaller than those for mammals. The authors found that distance to the nearest PA was significantly associated with only mammal functional richness.

2.2 Study-level metadata

- `Key words`: Biodiversity, Conservation, Protected Areas, Connectivity, 30x30
- `Subject`: Ecology and Evolutionary Biology, Natural Resources and Conservation,
- `Date created`: August 23, 2023
- `Date modified`: August 23, 2023
- `Spatial Coverage`: Southeast Asia
- `Spatial Resolution`: Species observations - GPS located point data, GEDI - derived forest structural covariates - 1 km raster, HDI - country-level, Protected Areas - PA Polygons
- `Spatial Reference System`: WGS84, UTM
- `Temporal Coverage`: 01-2015 to 08-2021
- `Temporal Resolution`: Varies with data set

2.3 Data-level Metadata

We use two datasets provided by the Brodie et al. to conduct our reproduction. Our primary analysis uses the author corrected dataset available at **INSERT DATASET LINK**. We also attempted to reproduce study results using the dataset originally provided by the authors, which is available at https://doi.org/10.6084/m9.figshare.22527298.v1. The authors' originally published dataset is missing the country-level measure of the Human Development Index used in biodiversity modeling. We gathered this data from the United Nations Development program and merge it with the authors' original dataset. For completeness, we gathered and include here metadata on the datasets used by the original authors to construct their shared analytical datasets.

2.3.1 eBird

Brodie et al. use the eBird database to construct the biodiversity measures for birds that are then used as response variables in statistical modeling. The authors did not provide scripts on how they derived the biodiversity metrics.

- `Title`: eBird.
- `Abstract`: A community science platform for reporting bird sightings.
- `Spatial Coverage`: Tropical region (overlapping countries of Brunei, Cambodie, China, Indonesia, Laos, Malaysia, Singapore, Thailand, and Vietnam).
- `Spatial Resolution`: Vector data model with point observations of species occurrence
- `Spatial Reference System`: Not specified.
- `Temporal Coverage`: 2015/01 - 2021/08.
- `Temporal Resolution`: Not applicable.
- `Lineage`: Brodie et al. (2023) queried and subset data directly from eBird website or its R package or API.
- `Distribution`: eBird webpage and other download methods.
- `Constraints`: Non-commercial use.
- `Data Quality`: Although a direct data quality layer is not associated, Brodie et al. (2023) stated that they followed recommendations from existing studies to filter out data points.

2.3.2 The World Database on Protected Areas

The original study used protected area boundaries to derive the three treatment variables (Table 3). Brodie et al. (2023) did not specify a data processing procedure or provide code from the preparation of protected area boundaries.

- `Title`: The World Database on Protected Areas (WDPA).
- `Abstract`: A global database on protected areas (PAs) and other effective conservation measures (OECM).
- `Spatial Coverage`: Tropical region (overlapping countries of Brunei, Cambodia, China, Indonesia, Laos, Malaysia, Singapore, Thailand, and Vietnam).
- `Spatial Resolution`: Vector.
- `Spatial Reference System`: WGS 84.
- `Temporal Coverage`: Accessed sometime in 2023
- `Temporal Resolution`: Updated monthly.
- `Lineage`: Brodie et al. (2023) subset from the dataset, but procedures are unknown.
- `Distribution`: WDPA webpage.
- `Constraints`: Non-commercial use.
- `Data Quality`: Unknown.

2.3.3 GEDI L2 metrics

The Global Ecosystem Dynamics Investigation (GEDI) is a spaceborne light detection and ranging (LiDAR) mission monitoring forest structure on earth. The original study derived both ground elevation and forest structure metrics from the Level 2 dataset of GEDI. Level 2 GEDI data are at footprint level, so Brodie et al. (2023) used kriging interpolation to create wall-to-wall layers at 1-km resolution.

Level 2 GEDI data includes elevation data. Brodie et al. computed slope and topographic position index (TPI) to represent topographic traits at each site. The authors originally gathered five L2B metrics, canopy height (relative height at 95%), plant area volume density (PAVD) between 0 and 5 m (represents understory density), cumulative plant area index from ground to canopy top, foliage height diversity of plant area index, and proportional canopy cover. The authors found the five forest structure metrics to be highly correlated and retained only canopy height and understory density in statistical models.

Raster files at 1-km resolution for GEDI derived metrics and circuit-based accessibility were shared through a weblink. The authors did not provide code for the calculation of GEDI metrics.

- `Title`: The Global Ecosystem Dynamics Investigation Level 2 Elevation and Height Metrics.
- `Abstract`: Global footprint level observations from GEDI on ground elevation and forest structure.
- `Spatial Coverage`: Tropical region (overlapping countries of Brunei, Cambodie, China, Indonesia, Laos, Malaysia, Singapore, Thailand, and Vietnam).
- `Spatial Resolution`: Footprints are of 25-m resolution and extrapolated into 1-km resolution.
- `Spatial Reference System`: WGS 84.
- `Temporal Coverage`: 2019/04/17 to 2022/04/12
- `Temporal Resolution`: Not applicable.
- `Lineage`: ???
- `Distribution`: Original GEDI L2 metrics can be derived from NASA website and Brodie et al. (2023) shared krigged results on a webpage.
- `Constraints`: Non-commercial use.
- `Data Quality`: Original GEDI L2 metrics have quality and degrade flags and Brodie et al. (2023) kept only data points of satisfying quality.

2.3.4 Human Development Index (HDI)

Data on this measure are missing from the analysis file originally shared by Brodie et al.. We gathered HDI values for each country from the Human Development Report 2020 following the citation provided by the authors. This data omission was corrected in the authors update. Our HDI addition to the original file matched the authors updated inclusion, other than cases where the authors noted that they hand corrected certain measures. Clear reasoning for those adjustments was not provided in the note on author corrections.

- `Title`: Human Development Index
- `Abstract`: An index on the level of human development by country.
- `Spatial Coverage`: Tropical region (overlapping countries of Brunei, Cambodie, China, Indonesia, Laos, Malaysia, Singapore, Thailand, and Vietnam).
- `Spatial Resolution`: Not applicable.
- `Spatial Reference System`: Not applicable.
- `Temporal Coverage`: 2020.
- `Temporal Resolution`: Not applicable.
- `Lineage`: Direct query though the official website.
- `Distribution`: Acquired directly through Human Development Report 2020.
- `Constraints`: Non-commercial use.
- `Data Quality`: Unknown.

Table 1 Response, treatment, and predictor variables generated by Brodie et al (2023).

Label	Alias	Definition	Type	Accuracy	Domain	Missing Data Value(s)	Missing Data Frequency
eBird							
SR.mean	Species richness	Number of species	Float	Unknown	Equal or greater than 0	Not applicable	Unknown
maxFRic	Functional richness	Diversity of species functional traits	Float	Unknown	Equal or greater than 0	Not applicable	Unknown
asymptPD	Phylogenetic diversity	Cumulative evolutionary time of the species assemblage	Float	Unknown	Equal or greater than 0	Not applicable	Unknown
WDPA							
PA	Within or outside PAs	Whether the point is inside a PA or not	Binary	Not applicable	1 for inside and 0 for outside	Not applicable	Unknown
PA_size_km2	Functional richness	Diversity of species functional traits	Float	Unknown	Equal or greater than 0	Not applicable	Unknown
dist_to_PA	Phylogenetic diversity	Cumulative evolutionary time of the species assemblage	Float	Unknown	Equal or greater than 0	Not applicable	Unknown

Label	Alias	Definition	Type	Accuracy	Domain	Missing Data Value(s)	Missing Data Frequency
GEDI L2 Metrics							
elev	Elevation	Ground elevation at the site (krigged)	Integer	Unknown	Equal to or greater than 0 (terrestrial observations)	Unknown	Unknown
slope	Slope	Slope of topography	Float	Unknown	0 to 90	Unknown	Unknown
TPI	Topographic Position Index	Difference between the elevation of a focal raster cell with those of its neighbors (not mentioned in paper)	Float	Unknwon	Not bounded	Unknown	Unknown
rh_95_a0.pred	Relative height at 95%	Roughly the top canopy height (krigged)	Float	Unknown	Equal to or greater than 0	Unknown	Unknown
pavd_0_5.pred	Plant area volume density from 0 to 5 m	A proxy of understory forest density (krigged)	Float	Unknown	Equal to or greater than 0	Unknown	Unknown
pai_a0.pred	Plant area index	Cumulative PAI from ground to canopy (krigged)	Float	Unknown	Equal to or greater than 0	Unknown	Unknown
fhd_pai_1m_a0.pred	Foliage height diversity	Shannon's diversity of PAI across heights (krigged)	Float	Unknown	Equal to or greater than 0	Unknown	Unknown
cover_a0.pred	Proportional coverage	Openness or closeness of canopy (krigged)	Float	Unknown	0 to 1	Unknown	Unknown
HDI							

Label	Alias	Definition	Type	Accuracy	Domain	Missing Data Value(s)	Missing Data Frequency
HDI	Human Development Index	Level of human development	Float	Unknown	0 to 1	Not applicable	Not applicable

Note: HDI was not included in the data file shared by the orignal study. We gathered HDI values for each country from the Human Development Report 2020.

2.4 Statistical Approach

Brodie et al. use propensity score matching to control for the potential confounding effects of site accessibility and habitat quality when estimating the efficacy of protected areas at improving bird and mammal biodiversity. In PA efficacy models, observations were matched based on their geographic locations (i.e., latitudes and longitudes), forest canopy height, accessibility, and HDI. In spillover models, observations were matched based on these same factors and either adjacent PA size or distance to the nearest PA. Weights produced by propensity score matching were then used in mixed-effects linear regression models that estimated the treatment effect - PA status, nearest PA size, nearest PA distance - while adjusting for forest canopy height, site accessibility, and HDI (Table 1).

Table 1. Variables used in statistical modeling

Name	Source	Usage
Biodiversity metrics - mammals	Authors	Outcome variable
Biodiversity metrics - birds	eBird	Outcome variable
Protected area boundaries	WDPA	Treatment variables (whether inside PAs)
Ground elevation	NASA GEDI L2B	Predictor - elevation and topography
Circuit-based site accessibility (log transformed)	Authors	Predictor - site accessibility
Human Development Index	Human Development Report 2020	Predictor - site accessibility
Forest structure metrics	NASA GEDI L2A	Predictor - forest structure

2.5 Observations Preceeding the Reproduction Attempt

Before beginning our reproduction attempt, we had observed the analysis file and code published by Brodie et al.. We noticed the following issues in the script and analysis file originally published by the authors.

1) The HDI measure was missing from the analysis file.
2) The procedure for computing biodiversity metrics, GEDI metrics, and the circuit-based accessibility metric was not provided in the scripts and not presented in detail in the methodological supplement.
3) The procedure for preparing PA boundaries for analysis was unclear.
4) The procedure for identifying and eliminating outliers was not included in the script. Only a hand coded list of outliers was provided and removed in the code.

The script and analysis file published by the authors with their correction in April 2024 addressed the first issue above. However, the other issues remain.

Other than adding the HDI variable to the original analytical file, we did not manipulate the data before beginning our reproduction attempt.

3 Reproduction Attempts

We completed two reproduction attempts - 1) using the analysis data and script originally published by the authors and 2) using the analysis data and script published with the author correction. Both analyses were conducted in R using the following packages

```
# library(groundhog)
pkgs <- c("tidyverse", "cowplot", "here", "dagitty", "ggdag", "Hmisc",
          "MatchIt", "modelsummary", "optmatch", "nlme")
# groundhog.library(pkgs, "2024-02-11")
lapply(pkgs, require, character.only=TRUE)
```

For completeness, we first reproduced the causal diagram underlying the original analysis. We made not substantive changes to the authors original code.

3.1 Data Preparation

Preparation of our analysis data followed the procedures outlined in Brodie et al. (2023). Complete details of the procedures are available in the called function headers and comments. All called function are available with the project repository. No additional preparation was needed for our reproduction attempt using the author corrected analysis file and script.

```
# Clean data and remove outliers for bird and mammal models
dat_clean_bird <- clean_data("bird",
                             conn_metrics,
                             src_dir = "data/derived/public",
                             conn_dir = "data/derived/public",
                             dst_dir = "data/derived/public")

dat_clean_mammal <- clean_data("mammal",
                             conn_metrics,
                             src_dir = "data/derived/public",
                             conn_dir = "data/derived/public",
                             dst_dir = "data/derived/public")

dat_analysis_bird <- identify_outliers(dat_clean_bird)

data_analysis_mammal <- identify_outliers(dat_clean_mammal)
```

3.2 Reproduction Using Author Corrected Data and Code

3.2.1 Analysis of PA Efficacy

To assess **OR-Ho-1** we reproduced the propensity score matching procedure and linear mixed-effects modeling conducted by the Brodie et al. for each the three biodiversity response variables. Following the authors, statistical significance of the treatment and predictor variables are indicated using p-value thresholds of 0.001 (***), 0.01 (**), and 0.05 (*).

« Insert Chunk calling model_pa_efficacy.R AND Table building script »

3.2.2 Analysis of PA Spillover Effects

« Insert Chunk calling model_pa_spillover.R AND Table building script »

3.2.3 Comparison of Statistical Estimates

Build comparison table in with simplified symbology in Rmarkdown

3.3 Reproduction Using Original Data and Code

To assess **OR-Ho-1** we reproduced the propensity score matching procedure and linear mixed-effects modeling conducted by the Brodie et al. for each the three biodiversity response variables. Following the authors, statistical significance of the treatment and predictor variables are indicated using p-value thresholds of 0.001 (***), 0.01 (**), and 0.05 (*).

« Insert Chunk calling model_pa_efficacy.R AND Table building script »

3.3.1 Analysis of PA Spillover Effects

« Insert Chunk calling model_pa_spillover.R AND Table building script »

3.3.2 Comparison of Statistical Estimates

Build comparison table in with simplified symbology in Rmarkdown

We identified significant PA effects on bird species diversity and non significant effects on mammal species diversity as reported in the paper. PA effect sizes for SR (OR:27.04, RPR:31.6), FR (OR:24.02, RPR:25.52), PD (OR:0.38, RPR:0.37) from the reproduction are similar to those reported in Brodie et al. (2023). Significance of other variables and their coefficient values were also similar to the original study.

```
#summary(mod_SR_efficacy_mam)
msummary(list('Bird SR'=mod_SR_efficacy_bird, 'Bird FR'=mod_FR_efficacy_bird, 'Bird PD'=mod_PD_efficacy
              ),stars = TRUE)

# Note: if rendering this part of the code returns errors for tex rendering
# Please consider going through steps on this debugging page:
# https://yihui.org/tinytex/r/#debugging

# Also: https://github.com/travis-ci/travis-ci/issues/10166
# sudo tlmgr install
# could be really helpful
```

4 Discussion & Conclusion

The goal of the report is to reproduce analysis and results from Brodie et al. (2023) on the effect of protected areas to preserve tropical bird and mammal biodiversity after removing confounding effects of site accessibility and forest structure. The original paper provided scripts and data which allowed for reproductions. The scripts were in general reproducible while missing data preprocessing steps such as cleaning protected area boundaries and computing secondary variables from raw data (e.g., GEDI metrics and circuit-based

accessibiltiy metrics). The data file contained most information but did not include raw observation data, a meta data file, or include HDI.

While there are ways to improve the reproducibility of the original work, our reproduction found results that were consistent with the main findings of Brodie et al. (2023). We found supporting evidence for OR-H1 (i.e., similar coefficient values and significance) as the original study but were unable to reproduce exact results as Brodie et al. (2023). Reasons for the differences could be HDI values, data version issues, or computation environment.

Through reconstructing the causal diagram and reproduction, we found that connectivity from sampling points to protected areas was not accounted for in the original study. We introduced hypothesis RPL-01 and tested it in our replication study.

4.1 Bias and threats to validity

Given the research design as described in the original paper and primary data shared, we find that potential spatial autocorrelation of sample points were not addressed. In addition, uncertainty issues were not discussed thoroughly in the paper, which includes 1) the representativeness of biodiversity measures, 2) the validity of forest structure measures created through krigging, and 3) accessibility represented by country-level Human Development Index.

// should we add a map/moran's I result here? // other concerns to add?

5 Acknowledgements

- `Funding Name`: name of funding for the project
- `Funding Title`: title of project grant
- `Award info URI`: web address for award information
- `Award number`: award number

This report is based upon the template for Reproducible and Replicable Research in Human-Environment and Geographical Sciences, DOI:%5B10.17605/OSF.IO/W29MQ](https://doi.org/10.17605/OSF.IO/W29MQ)

6 References

Brodie, J.F., Mohd-Azlan, J., Chen, C. et al. Landscape-scale benefits of protected areas for tropical biodiversity. Nature 620, 807–812 (2023). https://doi.org/10.1038/s41586-023-06410-z