Analyzing the Relationship Between Forest Cover Loss and Bird Species Richness in the Amazon Basin (2001 - 2024)

Nithyasri Paleti

December 2, 2025

Abstract

This project analyzes temporal and regional patterns in the richness of Amazon bird species using GBIF occurrence data (2000–2024) and forest loss from the Global Forest Watch Data API. After cleaning and filtering over 5.8 million GBIF records to an Amazon Basin bounding box, the following were computed: computed annual species richness, sampling effort, spatial sampling intensity, and regional trends (West, Central, East Amazon). A simple regression between raw richness and forest loss yielded a strong positive association ($R^2 \approx 0.48$), but effort-normalized richness showed a negative correlation with forest loss. Regional linear trends indicate consistently increasing observed richness ($\sim$ 22 species yr^{-1}) across all regions, driven primarily by shifting sampling effort. The results highlight how opportunistic data can mislead ecological inference if sampling bias is not explicitly addressed.

1 Introduction

The Amazon Basin is one of the most biodiverse regions on Earth but has undergone extensive deforestation over the last several decades. Open biodiversity databases such as the Global Biodiversity Information Facility (GBIF) now provide millions of species occurrence records that appear to enable large-scale analyses of biodiversity change. However, these data are strongly influenced by variation in observation effort, observer behavior, and data digitization, which can invert ecological signals.

The objectives of this project are to: (1) quantify temporal trends in bird species richness and observation effort, (2) map spatial and regional sampling intensity, (3) evaluate the effect of effort-normalization on relationships with forest loss, and (4) compare regional richness trends across West, Central, and East Amazon. All steps are implemented in a reproducible Python notebook.

2 Data and Methods

2.1 Biodiversity Data (GBIF)

A GBIF Simple Export of Amazonian bird records (tab-delimited CSV) with over 5.8 million rows and 9 columns was used, including species name, year, coordinates, country code, and basis of record. Only essential columns were read and then cleaned:

```
usecols = ["gbifID", "species", "scientificName", "year", "decimalLatitude", "
    decimalLongitude", "countryCode", "basisOfRecord", "issue"]

df = pd.read_csv("data/gbif_birds.csv", sep="\t",
                 usecols=lambda c: c in usecols,
                 low_memory=False)

# Drop missing fields and restrict years
df = df.dropna(subset=["species", "year",
                       "decimalLatitude", "decimalLongitude"])
df["year"] = df["year"].astype(int)
df = df[(df["year"] >= 2000) & (df["year"] <= 2024)]

# Drop duplicates and apply Amazon Basin bounding box
df = df.drop_duplicates(subset=["species", "year",
                                "decimalLatitude", "decimalLongitude"])

df = df[(df["decimalLatitude"].between(-20, 10)) &
    (df["decimalLongitude"].between(-85, -45))]
```

After cleaning and spatial filtering, the dataset contained 1,988,572 records.

2.2 Forest Loss Data (GFW API)

Annual forest loss (ha) for the Amazon Basin was obtained from the Global Forest Watch Data API (Hansen et al.). A GeoJSON polygon for the Amazon Basin was given as geometry and a SQL query requested the summed loss per year (2001–2024):

```
years = range(2001, 2025)
rows = []
for y in years:
    sql = f"""
    SELECT SUM(area__ha) AS loss_ha
    FROM results
    WHERE umd_tree_cover_loss__year = {y}
    """
    payload = {"sql": sql, "geometry": geometry}
    response = requests.post(url, headers=headers, json=payload).json()
    loss_val = response["data"][0]["loss_ha"]
    rows.append({"year": y, "loss_ha": loss_val})

forest = pd.DataFrame(rows).sort_values("year")
```

The final forest-loss table contained 24 annual records (2001–2024).

2.3 Annual Aggregation and Effort-Normalization

Annual bird species richness and observation effort were computed from the cleaned GBIF dataset:

```
richness = df.groupby("year")["species"].nunique().reset_index()
richness.columns = ["year", "species_richness"]

effort = df.groupby("year").size().reset_index(name="obs_count")

yearly = richness.merge(effort, on="year")
yearly.to_csv("data/bird_richness_by_year.csv", index=False)
```

To compare richness across years with very different numbers of records, an effort-normalized metric is calculated:

```
yearly["richness_per_10k_records"] = (yearly["species_richness"] /(yearly["
    obs_count"] / 10000.0))
```

This produced both raw and normalized richness time series for 2000–2024.

2.4 Regional Assignment and Trend Modeling

To explore spatial structure, each record was assigned to a broad Amazon region based on longitude:

```
def assign_region(lon):
    if lon <= -70:
        return "West Amazon"
    elif lon <= -60:
        return "Central Amazon"
    else:
        return "East Amazon"

df["species_clean"] = df["species"].astype(str).str.strip().str.lower()
df["region"] = df["decimalLongitude"].apply(assign_region)
```

Regional annual richness and effort were then computed:

```
regional_yearly = (df.groupby(["year", "region"])
.agg(species_richness=("species_clean", "nunique"),obs_count=("species_clean",
     "size"))
.reset_index().sort_values(["region", "year"]))
```

For each region, a linear model

$$\text{Richness}_y = \beta_0 + \beta_1 y$$

was fit using ordinary least squares:

```
from sklearn.linear_model import LinearRegression

region_trends = []
for region_name, sub in regional_yearly.groupby("region"):
    sub_sorted = sub.sort_values("year")
    X = sub_sorted["year"].values.reshape(-1, 1)
    y = sub_sorted["species_richness"].values

    model = LinearRegression()
    model.fit(X, y)

    region_trends.append({
        "region": region_name,
        "slope_richness_per_year": model.coef_[0],
        "intercept": model.intercept_,
        "r2": model.score(X, y),
    })

region_trends_df = pd.DataFrame(region_trends)
```

Region	Slope (species/yr)	Intercept	R^2
Central Amazon	23.228462	-45947.424615	0.857663
East Amazon	21.598462	-42928.344615	0.697014
West Amazon	22.656154	-44588.141538	0.820126

Table 1: Linear regression parameters for regional bird species richness trends (2000–2024).

2.5 Merging With Forest Loss and Correlation Analysis

Annual bird richness and effort were merged with forest-loss data:

```
birds_yearly = pd.read_csv("data/bird_richness_by_year.csv")
merged = birds_yearly.merge(forest, on="year", how="inner")
```

A simple linear regression of richness on forest loss was fit:

```
X = merged["loss_ha"].values.reshape(-1, 1)
y = merged["species_richness"].values

model = LinearRegression()
model.fit(X, y)

slope = model.coef_[0]
intercept = model.intercept_
r2 = model.score(X, y)
```

Slope: 0.00044741311247510846
Intercept: 811.2008520341236
R^2: 0.482438275213834

Pearson correlations for raw and effort-normalized richness were also computed:

```
merged_norm = yearly.merge(forest, on="year", how="inner")

corr_raw = merged_norm["species_richness"].corr(merged_norm["loss_ha"])
corr_norm = merged_norm["richness_per_10k_records"].corr(merged_norm["loss_ha
    "])
```

Correlation (raw richness vs forest loss): 0.6945777675781408
Correlation (normalized richness vs forest loss): -0.5442084733804119

3 Results

3.1 Dataset Characteristics and Sampling Effort

The cleaned GBIF dataset contains 1,988,572 bird records for 2000–2024 within the Amazon Basin bounding box. Records were highly uneven across space and time. Colombia and Ecuador contributed the largest numbers of records, followed by Brazil and Peru.

Annual observation counts increased dramatically from $\sim$ 8,500 records in 2001 to over 300,000 records by 2024, reflecting growth in citizen-science participation and digitization efforts rather than ecological change.

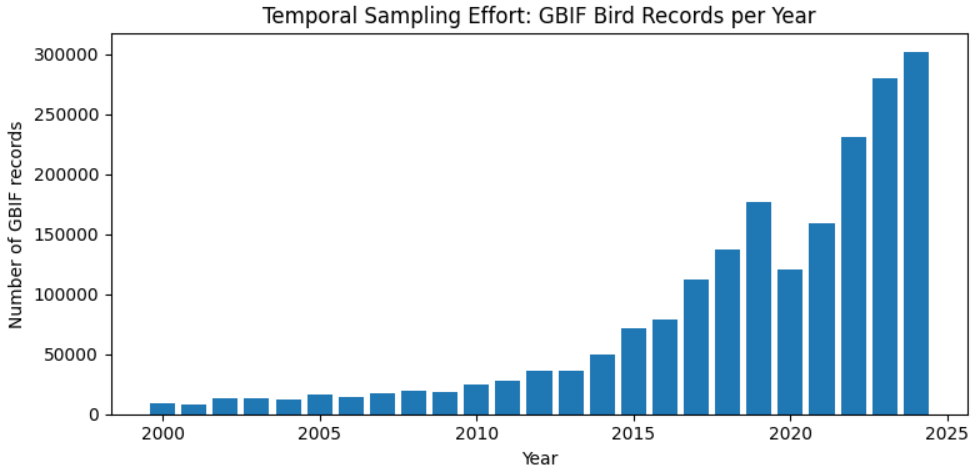

Figure 1: Temporal sampling effort: number of GBIF bird records per year in the Amazon Basin.

Spatially, sampling was clustered along major rivers, roads, and accessible regions. A hexbin map of record density shows strong hotspots and large unsampled areas.

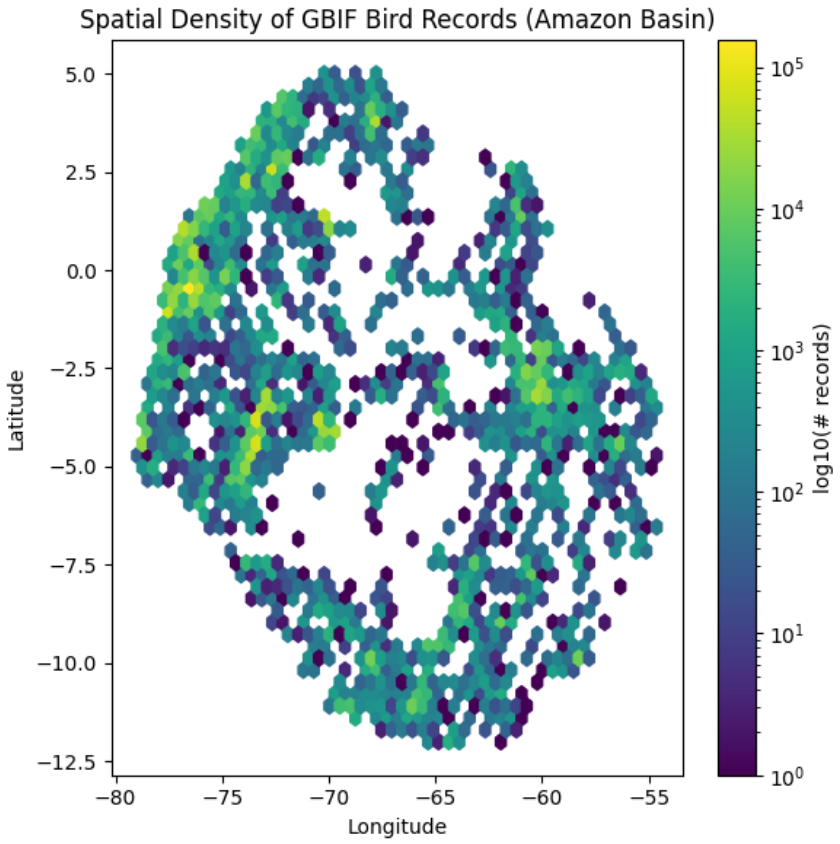

Figure 2: Spatial density of GBIF bird records in the Amazon Basin (hexbin of $\log_{10}$ record count).

3.2 Temporal Patterns in Richness and Effort

Observed annual bird richness (unique species) increased from roughly 770–950 species in the early 2000s to nearly 1,500 species by 2024. The shape of this curve closely mirrors the curve for observation effort.

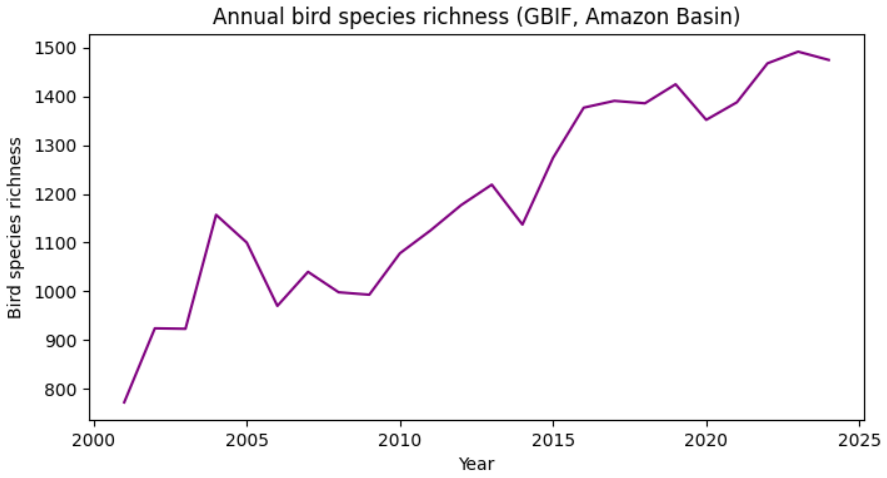

Figure 3: Annual bird species richness recorded in the Amazon Basin (GBIF, 2000–2024).

This parallel trend indicates that richer species lists in later years are largely explained by more intensive sampling, not necessarily by ecological change.

3.3 Effort-Normalization and Correlation With Forest Loss

The regression of raw richness on forest loss yielded:

$$\hat{\text{Richness}} = 811.2 + 4.47 \times 10^{-4} \times \text{Loss}_{\text{ha}},$$

with $R^2 \approx 0.48$. The Pearson correlation between raw richness and forest loss was:

$$\text{corr}(\text{richness}, \text{loss}) \approx 0.69.$$

After effort-normalization (species per 10,000 records), the correlation reversed:

$$\text{corr}(\text{richness per 10k records}, \text{loss}) \approx -0.54.$$

Thus, forest loss appears positively associated with observed richness when effort is ignored, but negatively associated once sampling intensity is accounted for. This reversal strongly suggests that the raw positive trend is driven by sampling effort.

3.4 Regional Patterns in Richness Trends

Record counts were heavily skewed among regions: West Amazon contained about 1.41 million records, Central Amazon $\sim$ 449,000, and East Amazon $\sim$ 125,000. Annual richness increased in all regions.

Fitting linear trends for species richness vs. year gave:

Region	Slope (species yr^{-1})	Intercept	R^2
Central Amazon	23.23	−45,947.42	0.86
East Amazon	21.60	−42,928.34	0.70
West Amazon	22.66	−44,588.14	0.82

Table 2: Linear trend parameters for regional bird species richness (2000–2024).

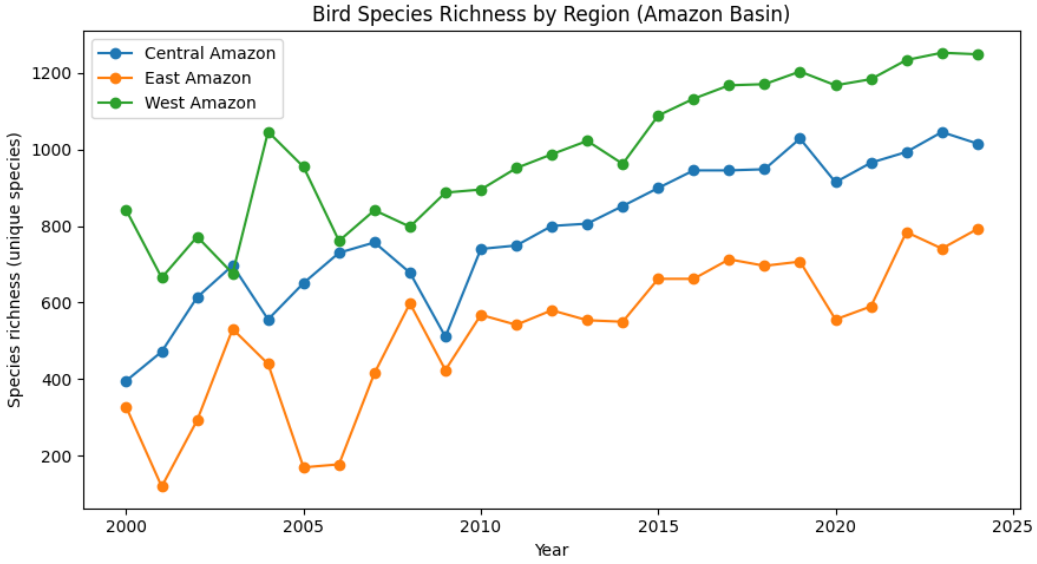

Figure 4: Regional bird species richness time series for West, Central, and East Amazon.

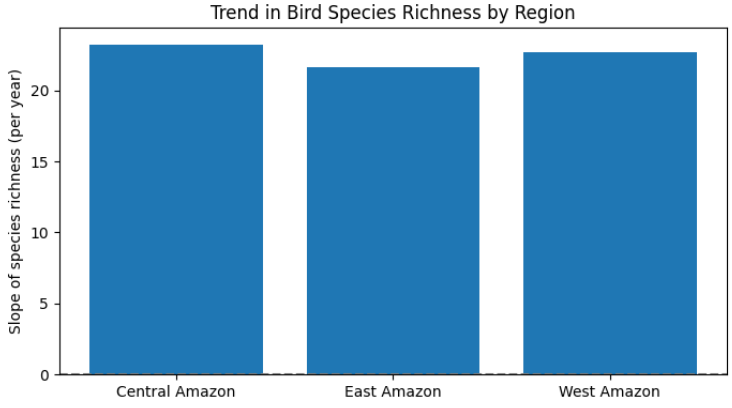

Figure 5: Slope of richness change per year for each Amazon region.

Similar slopes across regions suggest that increasing observation effort—rather than contrasting ecological trajectories—drives the observed patterns.

4 Discussion

The key findings from this analysis are:

- Raw bird species richness and forest loss are positively correlated ($r \approx 0.69$), but this relationship is reversed ($r \approx -0.54$) when richness is normalized by observation effort.
- Annual richness trends exhibit strong increases across all regions, closely tracking the growth in GBIF records over time.
- Spatial and regional patterns in the data primarily reflect where people go and where data are digitized, not necessarily where biodiversity is changing the most.

These results align with previous work showing that GBIF-based trends must be interpreted with caution when sampling effort is not explicitly modeled. Even simple effort-normalization can reveal that apparently strong and ecologically counterintuitive relationships are artifacts of the observation process.

5 Conclusions

This project demonstrates a reproducible, Python-based workflow for integrating GBIF bird occurrence data with satellite-derived forest loss and dissecting the role of sampling bias. Using simple linear models and effort-normalization, we show that:

1. Apparent increases in bird species richness in the Amazon Basin are largely driven by increased sampling effort.
2. The raw positive association between forest loss and richness reflects higher sampling in deforested and accessible areas.
3. Regional trends are similar in magnitude, consistent with shared underlying changes in observation effort.

More advanced models could further refine these conclusions, but overall, the findings highlight the importance of accounting for heterogeneous sampling effort when using opportunistic biodiversity datasets to infer ecological patterns.

6 Limitations

Several limitations inherent to GBIF data and the analytical design must be considered:

1. **Sampling bias.** GBIF data are spatially and temporally uneven, with higher sampling effort near roads, cities, tourist regions, and in countries with active birding communities (e.g., Colombia, Ecuador). Observation growth over time also inflates apparent richness trends.

2. **Trait approximation.** Endemic/generalist classifications based on geographic range are proxies and may not represent true ecological guilds or habitat specializations.

3. **Correlation vs. causation.** Forest loss and species richness correlations do not imply causality. Unmeasured drivers such as climate change, habitat fragmentation, and hunting may change or disguise relationships.

4. **Temporal resolution.** Annual aggregation hides seasonal or short-term responses to deforestation, which birds may detect at much finer temporal scales.

5. **Spatial mismatch.** GBIF point occurrences represent precise coordinates, whereas forest loss is measured in polygons. Birds are mobile and integrate landscape changes over broader areas than individual points.

6. **Detectability.** Species may be underdetected in degraded forest, producing false absences rather than reflecting true declines.

References

GBIF. (2024). *GBIF Occurrence Download*. Retrieved from: https://www.gbif.org/

Global Forest Watch. (2024). *GFW Data API Documentation*. World Resources Institute. https://data.globalforestwatch.org/

Hansen, M. et al (2013). High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160), 850–853. https://doi.org/10.1126/science.1244693

Pérez-Ruiz, A. et al (2020). Sampling bias and its influence on biodiversity research using GBIF data. *Ecology and Evolution*, 10(12), 12342–12356.