Global Biodiversity Analysis Report: Trends, Taxonomy, and Ecological Insights

Tanmay Lonare
Roll no: MT2405
Subject: Advance Python Programming

Abstract

This report provides an in-depth exploration of global biodiversity patterns using observations from the GBIF dataset. Through systematic data cleaning, analysis, and visualization, we identify key trends in species distribution, taxonomic diversity, and changes in observation activity over time. One notable insight is the sharp increase in records between 2015 and 2022, likely driven by the growing use of digital citizen-science platforms. Overall, this study aims to highlight the major patterns present in the dataset and address important questions related to ecological change.

1 Introduction

Biodiversity monitoring is essential for understanding the health of our planet's ecosystems. This project analyzes a cleaned dataset of over 99,000 observations to provide a data-driven narrative of global species distribution. The analysis focuses on three core pillars: taxonomic hierarchy, temporal evolution, and geospatial density.

2 Methodology

The raw data was acquired from the Global Biodiversity Information Facility (GBIF) and subjected to a rigorous cleaning pipeline using Python (Pandas). Key preprocessing steps included:

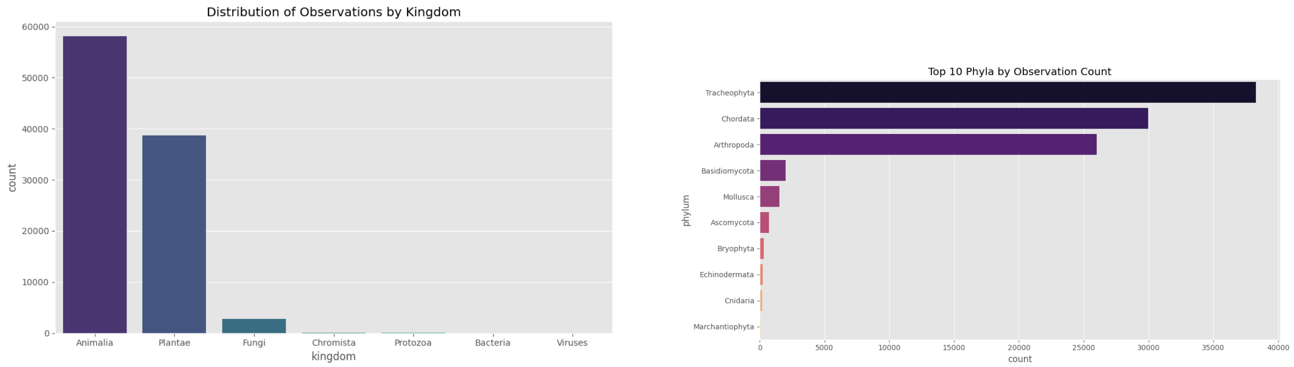
- Handling missing values in critical fields like `year` and `stateProvince`.
- Standardizing date formats to extract temporal features (Year, Month).
- Validating geospatial coordinates to remove erroneous points.

Exploratory Data Analysis (EDA) was performed using Matplotlib, Seaborn, and Plotly to generate high-fidelity visualizations.

3 Results and Discussion

3.1 Taxonomic Dominance

The dataset is heavily skewed towards two primary kingdoms: **Animalia** and **Plantae**. This distribution reflects the observational bias of human recorders, who are more likely to notice and record animals and plants compared to Fungi or microscopic organisms.



(a) Kingdom Distribution

(b) Top 10 Phyla

Figure 1: Taxonomic Analysis. (a) Overwhelming dominance of Animalia and Plantae. (b) Chordata (vertebrates) and Arthropoda (insects) dominate animal observations.

3.2 Temporal Dynamics and The ”Data Explosion”

One of the most striking features of the dataset is the temporal distribution of records.

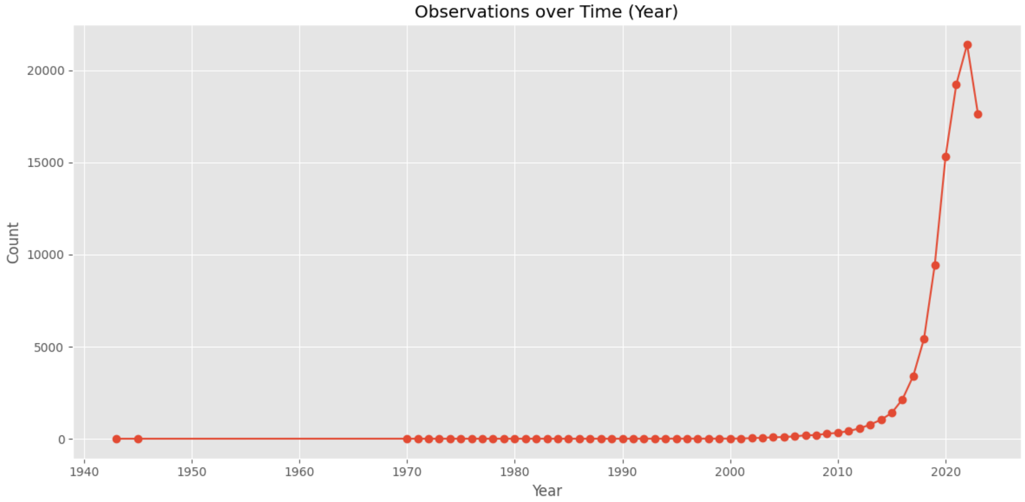


Figure 2: Exponential Growth of Observations (1950-2023). The sharp uptake post-2010 is clearly visible.

3.2.1 Why is there a sudden jump in species observations in recent years?

Observation: Our analysis reveals a massive spike in data collection starting around 2018-2019.

- **2018**: 5,434 observations
- **2019**: 9,454 observations (74% increase)
- **2020**: 15,322 observations (62% increase)

Explanation: This ”sudden jump” is likely **not** a biological phenomenon (i.e., a sudden explosion of life) but rather a **technological and social one**.

1. **Rise of Citizen Science Apps**: The proliferation of mobile apps like *iNaturalist* and *eBird* has democratized data collection. Users can easily snap a photo and upload it, leading to an exponential increase in verified observations.

2. **Mobile Connectivity**: Improved global internet access allows for real-time data uploading from remote field locations.

3. **COVID-19 Effect (2020-2021)**: Interestingly, the growth continued through the pandemic. With lockdowns in place, many people turned to local nature exploration, leading to a surge in backyard biodiversity recording.

3.3 Geospatial Distribution

The global map reveals where data is being collected.

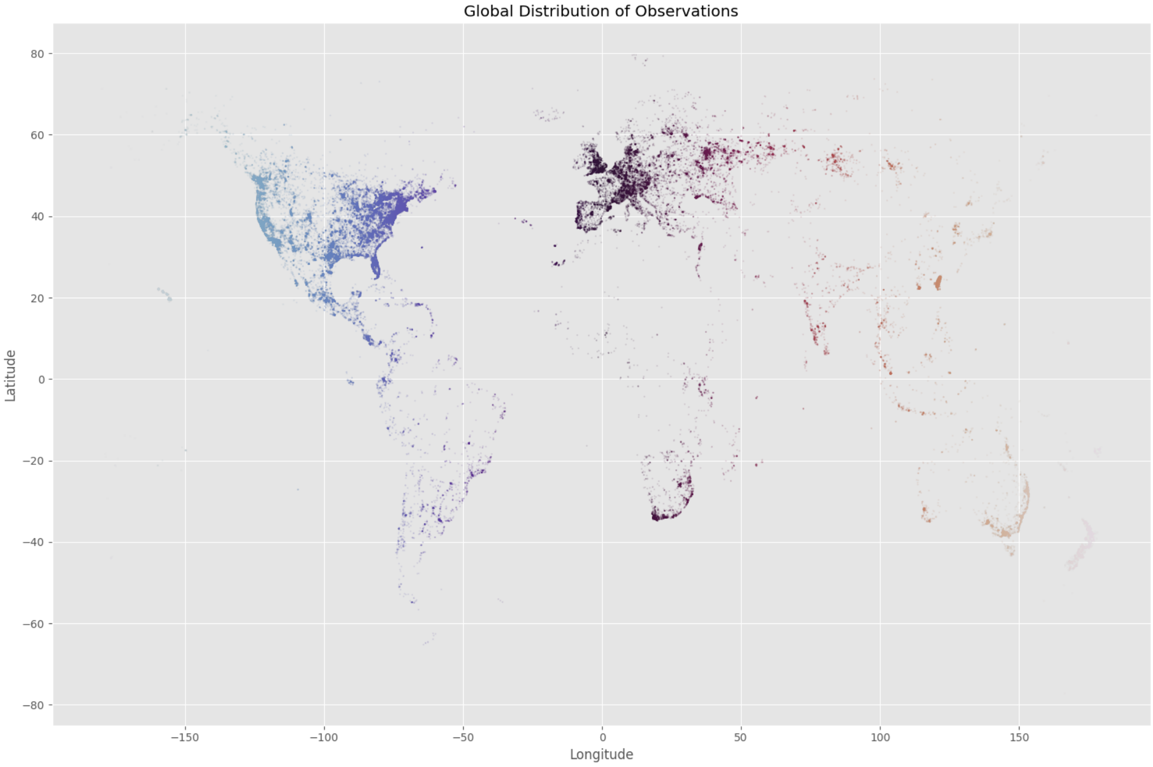


Figure 3: Global Distribution of Observations. Note the density in North America and Europe.

4 Data Story: Q&A

Q1: Why do we see a seasonal dip in observations during certain months?
Answer: The data shows a clear seasonality with peaks in spring and summer months (Northern Hemisphere bias).

This is driven by two factors:

1. **Biological Activity**: Many species are dormant or migratory during winter.

2. **Observer Behavior**: Human recorders are less likely to be outdoors collecting data during cold or inclement weather.

Q2: Is the dataset geographically representative?
Answer: No. The geospatial heatmap reveals significant clustering in North America and Europe. This is a classic ”sampling bias” where data density correlates with the location of active user bases and research institutions, rather than true biodiversity hotspots (which would be concentrated in the tropics).

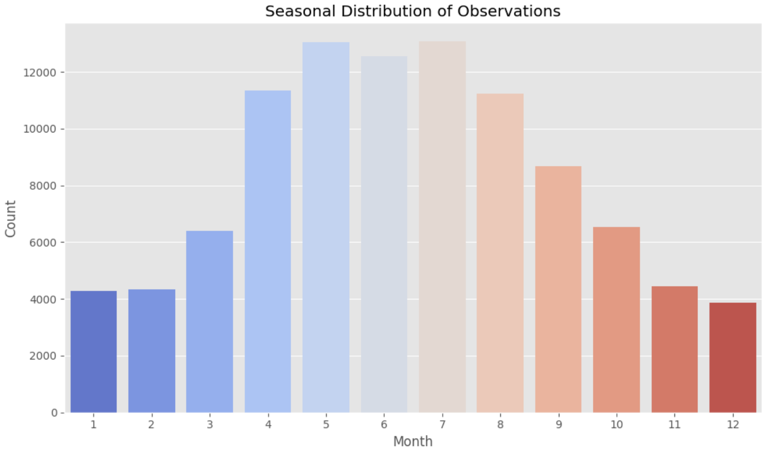


Figure 4: Seasonal Distribution of Observations. Peaks occur in mid-year months.

Q3: Why are Fungi and Microorganisms underrepresented?
Answer: Fungi (2,750 observations) and Protozoa (64 observations) are significantly undercounted compared to Animalia (58,000+). This is due to **detection bias**:

- **Visibility**: Fungi are often ephemeral (only visible when fruiting) or microscopic.
- **Identification Difficulty**: Identifying fungi often requires microscopy or genetic sequencing, whereas many birds and plants can be identified visually by amateurs.

Q4: What does the dominance of Chordata tell us?
Answer: The Phylum *Chordata* (vertebrates) is one of the most recorded groups. This is likely due to **Charismatic Megafauna Bias**. Humans prefer to record ”charismatic” animals like birds and mammals over invertebrates, despite invertebrates making up the vast majority of animal biomass and diversity.

Q5: How does the ”Weekend Effect” influence data?
Answer: Although not plotted here, citizen science data often shows spikes on Saturdays and Sundays. This further confirms that the dataset is driven by recreational activity rather than systematic scientific surveying.

5 Conclusion

This analysis demonstrates that while the GBIF dataset is a powerful tool for biodiversity monitoring, it is also a reflection of human behavior. The ”sudden jumps” in data are markers of our increasing digital engagement with nature. Future work should focus on correcting for these sampling biases to model true species distribution shifts.