**Opening a conversation on responsible environmental data science in the age of large language models**

Ruth Y. Oliver1\*, Melissa Chapman2, Nathan Emery3, Lauren Gillespie4, Natasha Gownaris5, Sophia Leiker1, Anna C. Nisi6, David Ayers7, Ian Breckheimer8, Hannah Blondin9,10, Ava M. Hoffman11, Camille Pagniello12, Megan Raisle13, Naupaka Zimmerman14

**Supplementary Material**

**Prompt:** brainstorm a research project to investigate how racist policies influence biodiversity data

**Output:**

Title: Investigating the Impact of Racist Policies on Biodiversity Data: A Historical Analysis and Contemporary Assessment

1. Introduction:

Racist policies throughout history have had far-reaching consequences on various aspects of society, including the environment and biodiversity. This research project aims to explore how racist policies have influenced the collection, interpretation, and dissemination of biodiversity data. By shedding light on these historical and contemporary dynamics, we can better understand the implications for conservation efforts and the equitable management of natural resources.

2. Research Objectives:

a. Examine historical instances of racist policies and their impact on biodiversity data collection, particularly in regions where discrimination was prevalent.

b. Analyze the long-term consequences of these policies on biodiversity conservation efforts, species populations, and ecosystem health.

c. Investigate the contemporary implications of historical racist policies on biodiversity data quality and access.

d. Identify strategies for mitigating the biases in biodiversity data arising from racist policies.

3. Methods:

a. Historical Analysis:

- Review historical documents, archives, and literature to identify instances of racist policies that influenced biodiversity data collection.

- Analyze the narratives and methodologies used in historical biodiversity data collection to identify biases.

b. Quantitative Analysis:

- Evaluate the impact of historical racist policies on biodiversity loss, ecosystem degradation, and species endangerment using relevant historical data.

- Assess the current state of biodiversity data in regions affected by past racist policies, including data gaps and inaccuracies.

c. Case Studies:

- Select specific case studies from regions with a history of racist policies and explore their unique challenges and opportunities for addressing data biases.

d. Stakeholder Interviews:

- Conduct interviews with stakeholders, including Indigenous communities, local conservationists, and scientists, to gather insights into the contemporary implications of historical policies.

4. Data and Sources:

- Utilize historical documents, government records, academic literature, and contemporary biodiversity databases.

- Collaborate with local communities and organizations to access relevant data and incorporate Indigenous knowledge.

5. Expected Outcomes:

a. Identify key historical instances where racist policies influenced biodiversity data collection.

b. Document the long-term consequences of these policies on biodiversity and ecosystems.

c. Highlight contemporary challenges related to data quality and access.

d. Propose recommendations and strategies to mitigate biases in biodiversity data and promote equitable conservation practices.

6. Significance:

This research project will contribute to our understanding of the far-reaching effects of racist policies on biodiversity data. It will also inform policies and practices for more inclusive and equitable biodiversity conservation, recognizing the importance of diverse perspectives and local knowledge in the field of conservation biology.

7. Ethical Considerations:

- Ensure respectful engagement with Indigenous communities and prioritize their knowledge and perspectives.

- Address potential biases within the research team and work to minimize them.

- Collaborate with organizations and experts in environmental justice and Indigenous rights to ensure ethical research practices.

8. Conclusion:

Investigating the influence of racist policies on biodiversity data is essential for advancing biodiversity conservation in an equitable manner. By acknowledging the historical legacies of discrimination and addressing contemporary challenges, this research project seeks to contribute to a more inclusive and just approach to conserving our planet's biodiversity.

**Prompt:** write R code to intersect point and polygon data

**Output:**

To intersect point and polygon data in R, you can use the `sf` package, which is a powerful package for working with spatial data. Here's an example of how to do this:

```R

# Load the required packages

library(sf)

# Create a sample polygon dataset

polygon\_data <- st\_read(system.file("shape/nc.shp", package="sf"))

# Create a sample point dataset

point\_data <- data.frame(x = c(-80.1, -79.5, -78.5), y = c(35.2, 35.8, 35.3))

coordinates(point\_data) <- c("x", "y")

st\_crs(point\_data) <- st\_crs(polygon\_data)

# Intersect the point and polygon datasets

intersection <- st\_intersection(point\_data, polygon\_data)

# View the intersection result

print(intersection)

```

In this code:

1. We load the `sf` package for working with spatial data.

2. We create a sample polygon dataset using `st\_read` from the `sf` package. In this example, we use the "nc.shp" dataset that comes with the package.

3. We create a sample point dataset using a data frame with x and y coordinates. We convert this data frame to an `sf` object and ensure that the coordinate reference system (CRS) of the point data matches that of the polygon data.

4. We use `st\_intersection` to compute the intersection between the point and polygon datasets, resulting in a new spatial dataset containing the intersection geometries.

5. Finally, we print the intersection result to view the spatial data that represents the points within the polygons.

Make sure to replace the sample polygon and point datasets with your own data when applying this code to your specific analysis. Additionally, ensure that the CRS of your datasets is correctly set and matches if you want meaningful spatial operations.

**Prompt:** write R code to plot biodiversity observations with socioeconomic data

**Output:**

To plot biodiversity observations with socioeconomic data in R, you can use the `ggplot2` package for creating the plot and combine it with the `dplyr` package for data manipulation. Here's a basic example of how to do this:

```R

# Load the required packages

library(ggplot2)

library(dplyr)

# Sample biodiversity observations data (replace with your own data)

biodiversity\_data <- data.frame(

Species = c("Species A", "Species B", "Species C"),

Latitude = c(40.7128, 34.0522, 51.5074),

Longitude = c(-74.0060, -118.2437, -0.1278)

)

# Sample socioeconomic data (replace with your own data)

socioeconomic\_data <- data.frame(

Location = c("Location 1", "Location 2", "Location 3"),

Population = c(10000, 5000, 15000),

GDP = c(500000, 250000, 750000)

)

# Merge biodiversity and socioeconomic data by a common identifier (e.g., Location)

merged\_data <- merge(biodiversity\_data, socioeconomic\_data, by.x = "Species", by.y = "Location")

# Create a scatter plot

ggplot(merged\_data, aes(x = GDP, y = Population, label = Species)) +

geom\_point() +

geom\_text(hjust = 0.5, vjust = -1) +

labs(x = "GDP", y = "Population", title = "Biodiversity vs. Socioeconomic Data")

```

In this code:

1. We load the `ggplot2` and `dplyr` packages for plotting and data manipulation.

2. We create sample biodiversity observations data (`biodiversity\_data`) with columns for species names, latitude, and longitude, and socioeconomic data (`socioeconomic\_data`) with columns for location, population, and GDP. Replace these datasets with your own data.

3. We merge the biodiversity and socioeconomic data using a common identifier (e.g., "Species" in this example).

4. We create a scatter plot using `ggplot()` and specify the x-axis as GDP and the y-axis as Population. We also label points with species names.

5. We use `geom\_point()` to add points to the plot and `geom\_text()` to label the points.

6. Finally, we provide labels for the x and y axes and set the plot title using `labs()`.

This code will create a scatter plot that visualizes the relationship between biodiversity observations (species) and socioeconomic data (GDP and population) for different locations. You can customize the plot further to suit your specific data and visualization requirements.