Data is the New Science - R Remix

## Learning Outcomes

Students completing this assigment will be able to:

* Access data from biodiversity digital data repositories
* Evaluate the research utility of occurrence data derived from different sources
* Create and interpret multiple graphs
* Calculate and compare basic summary statistics for a data set
* Use geo-spatial data to inform biological thinking

## License

This assignment is a remix of the **Data is the New Science** learning module[[1]](#footnote-23). It is licensed under CC Attribution-ShareAlike 4.0 International according to [these terms](https://creativecommons.org/licenses/by-sa/4.0/).

## Assignment Notes

Below you will find four Activities, each with multiple questions to answer. Answer the questions in this Rmd document, but italicize your answers. You can make any text italics by putting it in between two asterisk symbols, like *this*.

## Introduction

There is a changing landscape for those joining the 21st century workforce. Rapid advances in data research and technology are transforming how we conduct science. The volume and variety of data being generated, the increased accessibility of data for aggregation, the improved discoverability of data, and the increasingly collaborative and interdisciplinary nature of scientific research are driving the need for new skill sets.

The biodiversity sciences, for example, have experienced a rapid mobilization of data that has increased our capacity to investigate large-scale issues. **Biodiversity** is the variety of life on earth. Scientists study biodiversity at many different scales from the variation in the genes of a population to the diversity of species present in an ecosystem. Biodiversity can be described in many ways, including taxonomic, morphological, genetic, ecological, and functional. As we look around our world, all the variable forms of living organisms we see are the product of over 3.5 billion years of evolution in the form, function, and processes of life on earth. In order to address scientific issues of critical national and global importance, such as climate change, zoonotic disease, resource management, invasive species, and biodiversity loss, the 21st century biodiversity scientist must be fluent in integrative fields spanning evolutionary biology, systematics, ecology, geology, genetics, biochemistry, and environmental science and possess the quantitative, computational, and data skills to conduct research using large and complex data sets.

In this assignment, you will be introduced to some emerging biodiversity data resources. You will be asked to think critically about the strengths and utility of these data resources and then encouraged to think beyond the obvious to how these data could be used to answer big science questions. You will then use what you have learned about these data sources in an example case study, investigating the environmental conditions suited for a species of your choosing.

## Required R Packages

In this assignment you will need several new R packages. **Execute the install.packages code the first time you work through this module.** You do not need to re-install these packages after the first time you run the code.

### Install new packages

install.packages("maptools")  
install.packages("ridigbio")  
install.packages("rgbif")  
install.packages("raster")

### Load packages

Once the packages have been installed, and every time you want to use them, you need to load them into the R environment. Use the library function to do this.

library("maptools")

## Loading required package: sp

## Checking rgeos availability: TRUE

library("ridigbio")  
library("rgbif")

## Registered S3 methods overwritten by 'ggplot2':  
## method from   
## [.quosures rlang  
## c.quosures rlang  
## print.quosures rlang

library("raster")  
library("ggplot2")

## Activity 1: Examine Specimen-Based Occurrence Records

The data we collect about the where and when a species is found is called **occurrence data**. They are information on the presence of an individual from a defined species at a specific place and time. Information from occurrence records has proven to be highly valuable, allowing scientists to examine changes in distributions over time, perhaps in correlation with specific environmental factors, and to compare the distributions of different species. Occurrence data can be gathered by the collection of specimens, which are then preserved in a natural history collection, or may be based on records of observations.

**Specimen-based data** are based on archived biological specimens housed in a natural history collection. Scientists, naturalists, and explorers have been collecting and preserving specimens for hundreds of years. Today, we can interact directly with the specimens collected by Darwin, Lewis & Clark, and Teddy Roosevelt! Preserved with the specimens is a variety of information about the organism (e.g., collector, collection date, location, habitat, images, community assemblage, phenology). Specimens can be further examined to verify the data point or to yield additional information on the collected organism as scientists build on previous research or identify new questions to investigate. Preserved specimens from natural history collections are a treasure trove of data and information and have been used to study a variety of topics, such as evolutionary relationships, pesticide use, host-parasite evolution, and zoonotic disease transmission, etc. In 2010, museums and other collection holders initiated a massive data mobilization effort to provide digital databases of these archived specimens. This data is aggregated in searchable portals that provide broad access to information on individual specimens, images of the specimens, and associated data. These efforts have vastly increased the accessibility and utility of these data.

1. Go to the following website: <http://libguides.cmich.edu/lifesciencedata/organismal>. Here you will see several groupings of data sets. First, we will examine specimen-based data. We will be using the iDigBio portal for this activity. Click on the [iDigBio](https://www.idigbio.org/) link provided. Spend some time looking around the iDigBio site to find the answers to the questions below.
2. What is iDigBio short for?

* *Replace this text with your answer*

1. How many specimen records are currently available through this portal?

* *Replace this text with your answer*

1. Pick a species and search the iDigBio portal for occurrences records. Refer to the iDigBio User Guide as needed to help navigate the portal.
2. What species did you choose? **Note - you will need to search by the genus species name of your species of interest.**

* *Replace this text with your answer*

1. How many records of your species were found?

* *Replace this text with your answer*

1. What is the date that the oldest preserved specimen was collected?

* *Replace this text with your answer*

1. Click on the “must have media” button. Once the data set refreshes, “view” some of the specimen records. What types of media are available for researchers to use?

* *Replace this text with your answer*

1. Now, use the ridigbio package, as implemented in the code chunk below, to search for your species. **Replace the words “aralia elata” with your genus and species of interest.**

idigbio\_data <- idig\_search\_records(rq = list(scientificname = "aralia elata"))

1. Do the numbers of records found using the R function match what you found using the iDigBio portal?

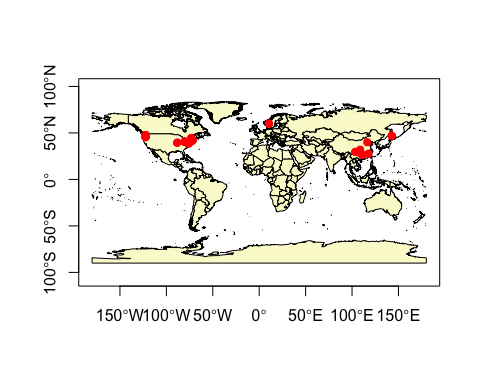
* *Replace this text with your answer*

1. Use the summary function to provide a summary of the idigbio\_data object.

# Put your code to report a summary of the data here

1. Use the code below to plot the geographic locations of your species on a map of the world.

# Call in the data for a simple map of the world  
data(wrld\_simpl)  
  
# Plot the world map with the points from the iDigBio search overlayed  
plot(wrld\_simpl, axes=TRUE, col="lightgoldenrodyellow")  
points(x = idigbio\_data$geopoint.lon, y = idigbio\_data$geopoint.lat,  
 col="red", pch = 19)



Describe the shape of the distribution of your species. The distribution is a description of where the species is geographically.

*Replace this text with your answer*

## Activity 2: Examine Observation-Base Occurrence Data

**Observation-based data** is information gathered by scientists, naturalists, and citizen scientists and represents an observation of an organism. Observation data is not linked directly to a physical specimen. While some data may be associated with just a location and date, other data types may be accompanied by a photo and detailed information (e.g., collection methods, environmental conditions, geographic location, associated species, weather, behavior, abundance, phenology). Similar to specimen-based data these data can provide a wealth of information on biodiversity and are now included in many online databases accessible to researchers, educators, and the public.

1. Return to this website: <http://libguides.cmich.edu/lifesciencedata/organismal>. The next set of questions relates to the GBIF data sets. These data sets include both specimen and observation-based occurrence records. We will be using GBIF for this activity. Click on the [GBIF link](https://www.gbif.org/dataset/search) provided. Spend some time looking around the GBIF site to find the answers to the questions below.
2. What does GBIF stand for?

* *Replace this text with your answer*

1. How many occurrence records are currently available through this portal?

* *Replace this text with your answer*

1. Search the GBIF portal for occurrences for the same species you used in the iDigBio exercise. Refer to the GBIF User Guide as needed to help navigate the portal.
2. How many records of your species were found?

* *Replace this text with your answer*

1. What other information and resources are provided with the results of your search?

* *Replace this text with your answer*

1. Now, use the rgbif package, as implemented in the code chunk below, to search for your species. **Replace the words “aralia elata” with your genus and species of interest.**

gbif\_data <- occ\_search(scientificName = "aralia elata", limit = 100000)

## Registered S3 method overwritten by 'crul':  
## method from  
## as.character.form\_file httr

gbif\_data <- gbif\_data$data

1. Do the numbers of records found using the R function match what you found using the GBIF data search web portal?

* *Replace this text with your answer*

1. Use the summary function to provide a summary of the gbif\_data object.

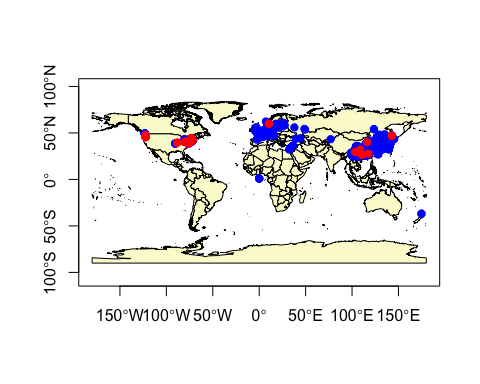
# Put your code to report a summary of the data here

1. Describe at least three differences you see between the data resulting from your iDigBio search and your GBIF search.

* *Replace this text with your answer*

1. Use the code below to add the GBIF geographic locations of your species on the map you created above. The blue points are the GBIF data and the red points are the iDigBio data. **Note - you may have to change the order in which you plot your data if one data set completely overlap ths other.**

# Plot the world map with the points from the iDigBio search overlayed  
plot(wrld\_simpl, axes=TRUE, col="lightgoldenrodyellow")  
points(x = gbif\_data$decimalLongitude, y = gbif\_data$decimalLatitude,  
 col = "blue", pch = 19)  
points(x = idigbio\_data$geopoint.lon, y = idigbio\_data$geopoint.lat,  
 col = "red", pch = 19)



1. How do the two distributions compare each other?

* *Replace this text with your answer*

1. What factors might contribute to any differences you found between the two distributions?

* *Replace this text with your answer*

## Activity 3: Examine Environmental Data Sets

**Environmental data** are available from multiple sources and include detailed information on a variety of abiotic variables (e.g., soil types, climate variables, hydrology, land use). Environmental data is available for different regions, with varying levels of resolution, and for different time frames. This data can be used to address long-term and short-term questions related to anthropogenic disturbance, climate variation, historical weather patterns, changing land use, environmental contaminants, etc.

Go to the following website: <http://libguides.cmich.edu/lifesciencedata/environmental>. Here you will see several environmental data sets. Pick a data set to investigate. This set of questions relates to the environmental data sets. Use that data link to address the questions below:

1. Which environmental data source did you review?

* *Replace this text with your answer*

1. What types of data are available at your selected data source?

* *Replace this text with your answer*

1. What types of questions could be addressed using this data? List three questions you think could be addressed with your environmental data set.
2. Question I:

* *Replace this text with your answer*

1. Question 2:

* *Replace this text with your answer*

1. Question 3:

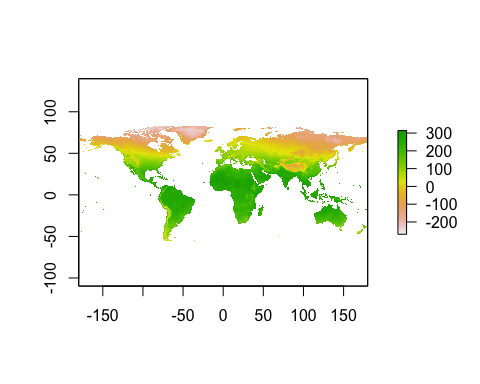
* *Replace this text with your answer*

1. Next we will download a set of global climate layers from the [WorldClim](https://www.worldclim.org) project.

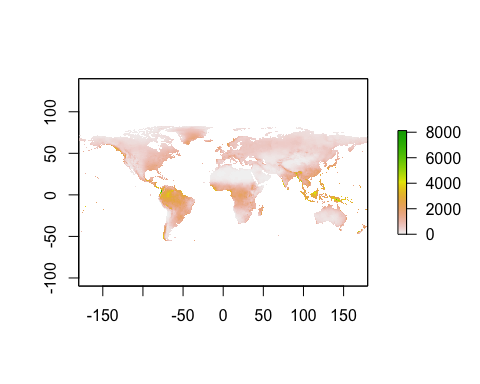
clim\_data <- getData('worldclim', var='bio', res=10)

These data represent a standard set of [bioclimatic variables](https://www.worldclim.org/bioclim). The code below plots **BIO 1: Annual Mean Temperature** and **BIO 12: Annual Precipitation**.

plot(clim\_data$bio1)



plot(clim\_data$bio12)



1. What do you think the values shown represent?

* *Replace this text with your answer*

1. Describe three aspects of the patterns you see in the figures above.

* *Replace this text with your answer*

## Activity 4: Connecting Species Occurrences and Envrionmental Data

Connecting species occurrences and the environmental data at those occurrences can help us understand what environmental conditions are conducive to a species survival. This information can be used to answer basic questions in organismal biology and ecology, and is also vital for conservation planning.

Using the code provided below, you will construct data sets with which you can use to make subsequent data visualizations to gain an understanding of the environmental requirements of your species of interest.

### Extracting Enviromental Data

1. The code in the chunk below extracts data from the climate data set we acquired above based on the locations we found using iDigBio.

idigbio\_env <- extract(x = clim\_data,   
 y = idigbio\_data[c("geopoint.lon", "geopoint.lat")])  
idigbio\_env <- as.data.frame(idigbio\_env)

1. Use the summary function to provide a summary of the idigbio\_env object.

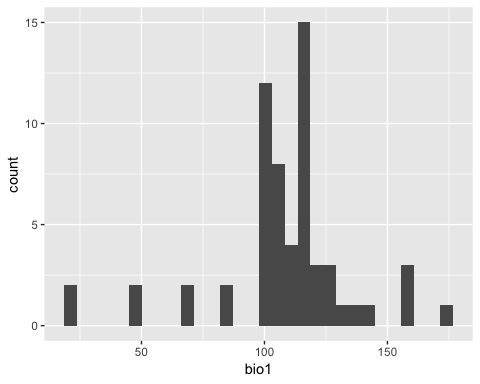
# Put your code to report a summary of the data here

1. Make two **histograms**, one showing the distribution of BIO 1 and the other the distribution of BIO 12, as found in the idigbio\_env object. **Note - below is the code to create a historgram for BIO 1. Use this as a template for your other histograms.**

ggplot(data = idigbio\_env, aes(x = bio1)) +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 91 rows containing non-finite values (stat\_bin).



# Put your code to make your second histogram here

1. Describe the shape of the two histograms.

* *Replace this text with your answer*

1. Calculate the mean, median, min, and max values for both BIO 1 and BIO 12, using R functions.

# Put your code to calculate mean, median, min, max values here.

1. Make a **scatter plot** showing the relationship between BIO 1 and BIO 12 at the occurrence locations for your species.

# Put your code to make your scatter plot here

1. Based on the scatter plot, describe the relationship between BIO 1 and BIO 12.

* *Replace this text with your answer*

1. The code in the chunk below extracts data from the climate data set we acquired above based on the locations we found using GBIF.

gbif\_env <- extract(x = clim\_data,   
 y = gbif\_data[c("decimalLongitude","decimalLatitude")])  
gbif\_env <- as.data.frame(gbif\_env)

1. Use the summary function to provide a summary of the gbif\_env object.

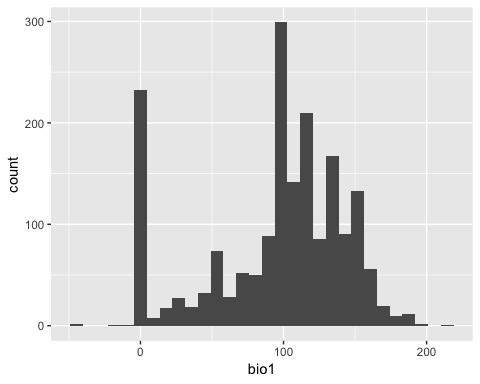
# Put your code to report a summary of the data here

1. Make two **histograms**, one showing the distribution of BIO 1 and the other the distribution of BIO 12, as found in the gbif\_env object. **Note - below is the code to create a historgram for BIO 1. Use this as a template for your other histograms.**

ggplot(data = gbif\_env, aes(x = bio1)) +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1572 rows containing non-finite values (stat\_bin).



# Put your code to make your second histogram here

1. Describe the shape of the two histograms.

* *Replace this text with your answer*

1. Calculate the mean, median, min, and max values for both BIO 1 and BIO 12, using R functions.

# Put your code to calculate mean, median, min, max values here.

1. Make a **scatter plot** showing the relationship between BIO 1 and BIO 12 at the occurrence locations for your species.

# Put your code to make your scatter plot here

1. Based on the scatter plot, describe the relationship between BIO 1 and BIO 12.

* *Replace this text with your answer*

1. Describe *at least* five differences you observed between the plots and summary statistics calculated using the idigbio\_env versus the gbif\_env data sets. Which do you think is better, in terms of gaining an accurate representation of the environmental conditions that are most appropriate for your species?

* *Replace this text with your answer*

1. Based on the figures and summary statistics observed, describe what you think are the ideal climatic conditions for your species.

* *Replace this text with your answer*

1. Monfils, A., Linton, D., Ellwood, L., Phillips, M. (2019). Data is the New Science. Biodiversity Literacy in Undergraduate Education, QUBES Educational Resources. <doi:10.25334/Q4RR0R> [↑](#footnote-ref-23)