Exploring the Global Mammal Parasite Database

Patrick Stephens, Richard Hall and Sonia Altizer

January 22, 2019

Goals of exercise

- To explore the data and structure of the publically available Global Mammal Parasite Database v. 2.0
- To get some experience working with a large dataset in R, including creating some simple plots

1. Getting started

For the most part we will be using the default R installation. You will also need to install the `worldmap` library and the following files:

GMPD_main.csv

Ecy1799-supp-0002-metadata.pdf

When you have the csv file in your workspace, load it to get started:

```
> gpdata <- read.csv("GMPD_main.csv")
```

2. Structure of the data

First, let's look at the structure of the dataset:

```
> dim(gpdata)
[1] 24323    28

> str(gpdata)
'data.frame':  24323 obs. of  28 variables:
 $ Group                : Factor w/ 3 levels "carnivores","primates",..: 1 1
1 1 1 1 1 1 1 1 ...
 $ HostReportedName     : Factor w/ 1081 levels "Acinonyx jubatus",..: 1 1 1
1 1 1 1 1 1 1 ...
 $ HostCorrectedName    : Factor w/ 462 levels "Acinonyx jubatus",..: 1 1 1 1
1 1 1 1 1 1 ...
 $ HostOrder            : Factor w/ 4 levels "Artiodactyla",..: 2 2 2 2 2 2 2
2 2 2 ...
```

.

.

The main file of the global mammal parasite database contains 24323 rows of data for 28 variables. The metadata file contains descriptions of all of these variables. The data are

from three different host groups. You can see the host groups by looking at the levels present in the `Group` variable.

```
> unique(gpdata$Group)
[1] carnivores ungulates  primates
Levels: carnivores primates ungulates
```

Next we want to figure out how many rows of data are available for each host group. There are several possible ways to do this (and to solve all of the exercise today). For example, there are at least three different ways to figure out how many rows of primate data are present:

```
> primates <- data[gpdata$Group == "primates",]
> dim(primates)
[1] 5070   28

> dim(data[gpdata$Group == "primates",])
[1] 5070   28

> length(gpdata$Group[gpdata$Group == "primates"])
[1] 5070
```

Exercise 1: Determine how many rows of carnivore and ungulate data are in the file.

Exercise 2: Determine which parasite groups are present in the data, and how many rows of data there are for each group.

H1: Parasite group variable is: gpdata$ParType

H2: table command can accomplish this more easily than any of the sample code listed above.

2. Visualizing spatial data from the GMPD

The `rworldmap` library makes it very easy to create some simple maps from coordinate data. For example, run the following lines of code:

```
> library(rworldmap)
Loading required package: sp
### Welcome to rworldmap ###
For a short introduction type :        vignette('rworldmap')

> newmap <- getMap(resolution = "low")

> plot(newmap)
```

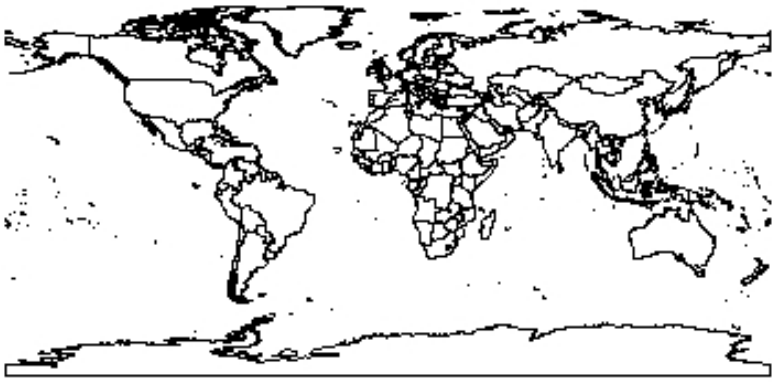

Fig. 1: a simple map of the world.

Now to add some coordinate data to the map. First create two vectors with latitude and longitude coordinates:

```
> #some lat long coordinates
> lats <- c(34.47, 41.30, 36.70, -14.27, 42.52, -08.83,
+           17.33, -36.50, 40.17, 12.53, -35.25, 48.20,
+           40.48)
> longs <- c(69.18, 19.82, 03.13, -170.72, 01.53, 13.25,
+           -61.80, -60.00, 44.52, -70.03, 149.13, 16.37,
+            49.93)
```

To plot these on the map, use the `points` command. Note that the order of variables is *longitude* then *latitude*.

```
> points(longs, lats, col = "red", cex = 1, pch = 19)
```

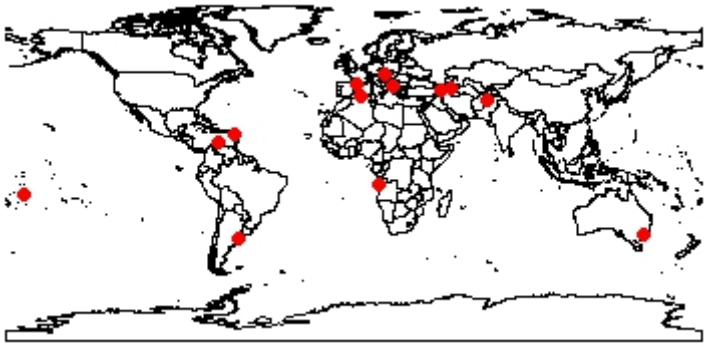

Figure 2: the same map with coordinates added.

It should be fairly easy to adapt this code to plot data from the GMPD.

Exercise 3: Create a world map showing the locality of all georeferenced data in the GMPD. *Hint:* note the `Latitude` and `Longitude` variables in the main data file.

Exercise 4: Create a map or series of maps that show localities for primate, carnivore and ungulate data.

3. On your own

The GMPD contains a mixture of discrete and continuous data. Note that missing data are designated by NA. The `hist` command can be used to create simple histograms in R.

Exercise 5: Use the `hist` command (or the histogram function of your choice) to create a histogram of any continuous variable from the GMPD. Most variables that you could choose will vary over several orders of magnitude. What data transformation do you need to use to create an informative summary of the data distribution? Does the graph look the same or different after using the `na.omit` command to get rid of missing data?

Exercise 6: Create a series of histograms showing how the variable differs between either different parasite groups or different host groups.