Mosquitoes and monkeys: Mapping cases to investigate yellow fever virus emergence

Dr. Sabrina Li

3/14/23

Introduction

Yellow fever virus is blah blah blah -

-Briefly discuss history

-Briefly discuss transmission cycle

-briefly discuss latest YFV outbreak

This tutorial will teach you how to make maps in RStudio to visualisase

Mosquitoes and non-human primates (NHP)

Let's load *Horto_MO_NHP.csv* into RStudio. This file contains data on confirmed cases of yellow fever virus extracted from mosquitoes and non-human primates sampled in Horto, a neighbourhood in Sao Paulo, and elsewhere. Samples were collected between 2017 and 2018.

```
horto<-read.csv("Horto_MO_NHP.csv")
```

Let's have a look at the data. We can run **View(horto)** to open the data frame. A preview of the data is shown below.

ID	source	Location	Accession_number	Country	State	Municipality	latitude	longitude	Host_species	date
New_mosquitoes[illegible] 01-05	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-23.46[illegible]	[illegible]	Haemagogus leucocelaenus	2018-01-05
New_mosquitoes[illegible] 01-05	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-23.46[illegible]	[illegible]	Haemagogus leucocelaenus	2018-01-05
New_mosquitoes[illegible] 01-11	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-23.46[illegible]	[illegible]	Haemagogus leucocelaenus	2018-01-11
New_mosquitoes[illegible] 01-11	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-23.46[illegible]	[illegible]	Haemagogus leucocelaenus	2018-01-11
New_mosquitoes[illegible] 01-11	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-23.46[illegible]	[illegible]	Haemagogus leucocelaenus	2018-01-11

What do you notice?

- From "ID", and "Location", data were collected in Horto (PEAL) but also in the north of Sao Paulo.
- In particular, we have a column called "Accession_number", which tells us whether the virus host is a mosquito or a non-human primate, and "Host_species", which refers to its species type.

How much data do we have on mosquitoes and on non-human primates? We can easily determine this by filtering the data to create a bar chart using **ggplot2**.

First, load **ggplot2** into our library in RStudio. We will also need to load **tidyverse** as it will help us filter our data for plotting. If you do not have these packages, you can install them by running *install.packages(" ")* in RStudio.

```
library(ggplot2)
library(tidyverse)
```

To create a bar chart, we need to determine the total number of observations per species and host type. We then plot this number by host type and species type. This takes a few steps, as shown in the code below. Each step ends with %, which is a pipe from the *tidyverse* package that tells us the sequence of steps we are taking.

Try the code below in RStudio to produce the bar chart.

```
horto %>%
  filter (Location == "SaoPaulo_PEAL") %>%
  group_by(Accession_Number,Host_species) %>%
  summarise(count=n()) %>%
  mutate(Accession_Number = reorder(Accession_Number,count, increasing = T)) %>%
  ggplot(aes(x = Accession_Number, count, fill = Host_species)) +
  geom_col(position = "dodge")
```

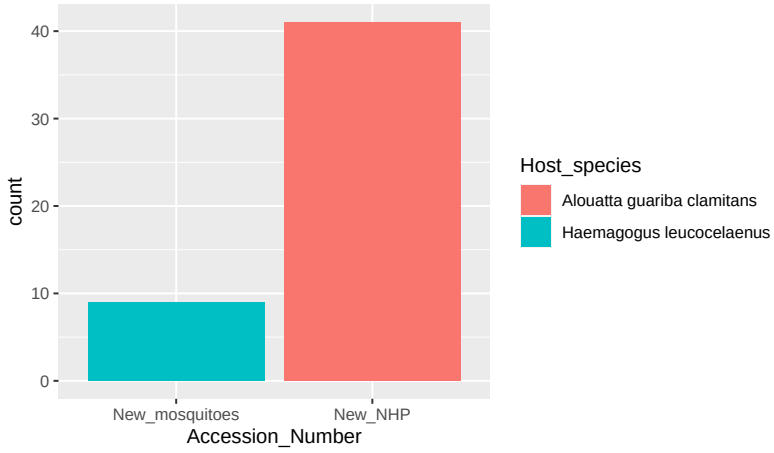


Figure 1: Bar chart showing distribution of species by host type.

We can see that we have one species of mosquitoes, *Haemagogus leucocelaenus*, and one species of NHP, *Alouatta*.

Let's now map these cases to understand its spatial distribution in Horto. Have a look again at the dataframe - what information here would be useful for mapping? ## Mapping basics

If we want to map a data set, we will need to look for location-related information. There are five columns in our data frame that tells us information about the location of the data set.

- Country
- State
- Municipality
- Location: place in Sao Paulo where sample was collected.
- Latitude: part of the coordinate system (Y-axis), location north or south of the equator.
- Longitude: part of the coordinate system (X-axis), location east or west of prime meridian at Greenwich.

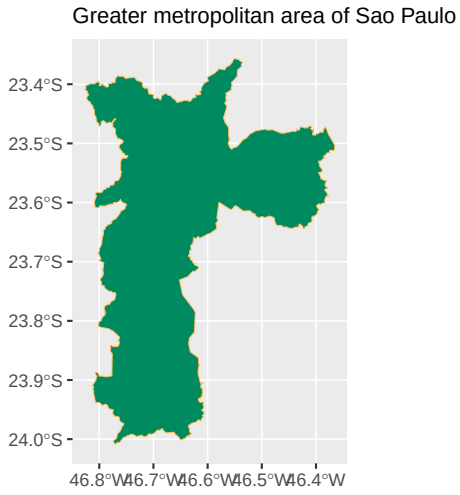
Let's now create a map showing Sao Paulo municipality, where Horto Florestal is located. The package *sf* and *ggspatial* offers spatial feature handling and mapping capabilities.

geobr offers maps of Brazil containing vector data on location, shape, and geographical attributes at various administrative levels. In this tutorial we will be extracting municipalities, conservation areas, and census tracts, all using *geobr*.

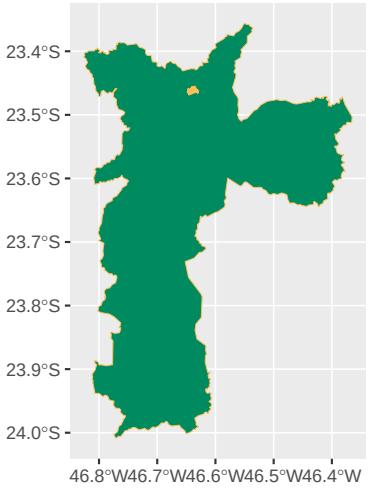
```
library(geobr)
library(sf)
library(ggspatial)

# Extract Sao Paulo municiaplity
sp_muni <- read_municipality(code_muni = "SP", year = 2020, showProgress = FALSE) %>%
  filter(code_muni == "3550308")
```

```
# head(sp_muni)
# Simple feature collection with 1 feature and 7 fields
# Geometry type: MULTIPOLYGON
# Dimension:     XY
# Bounding box:  xmin: -46.82619 ymin: -24.00826 xmax: -46.36531 ymax: -23.35629
# Geodetic CRS:  SIRGAS 2000
#   code_muni name_muni code_state abbrev_state name_state code_region name_region
# 1   3550308 São Paulo         35           SP  São Paulo           3     Sudeste
#                             geom
# 1 MULTIPOLYGON (((-46.54624 -...
```

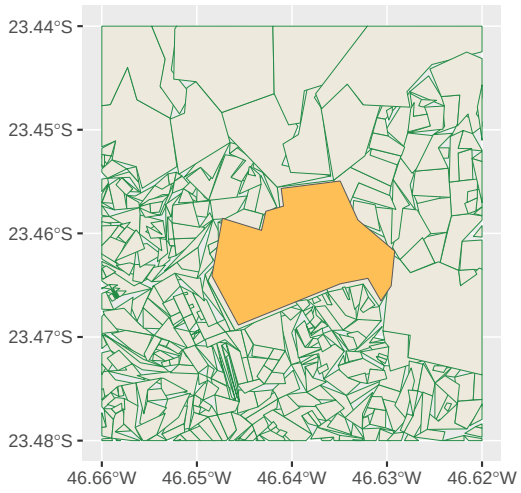
```
#create a map showing Sao Paulo municipality
ggplot() +
  geom_sf(data = sp_muni, fill = "#008A60", colour="#FEBF57", size=2,show.legend = FALSE) +
  labs(subtitle = "Greater metropolitan area of Sao Paulo", size=8)
```



Let's highlight the Horto Florestal area on the map. Because Horto Florestal is a conservation area, we can use **geobr** to extract the shapefile of the park from the function *read_conservation_units()*. The conservation area that Horto goes by is "PARQUE ESTADUAL ALBERTO LöFGREN".



Now let's create a zoomed - in version of our map showing Horto and its surrounding neighbourhoods. We will represent neighbourhoods at the census tracts administrative level.



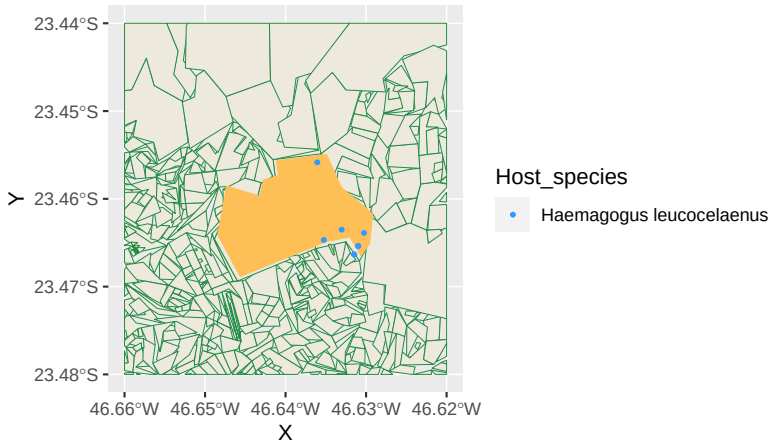
Mapping samples collected from Horto Florestal

Let's now create a map of Horto showing the locations of the *Haemagogus* samples from our *Horto* data set. First, we'll need to filter our data set for samples collected in Horto only. We will then convert the dataframe to a spatial feature for mapping.

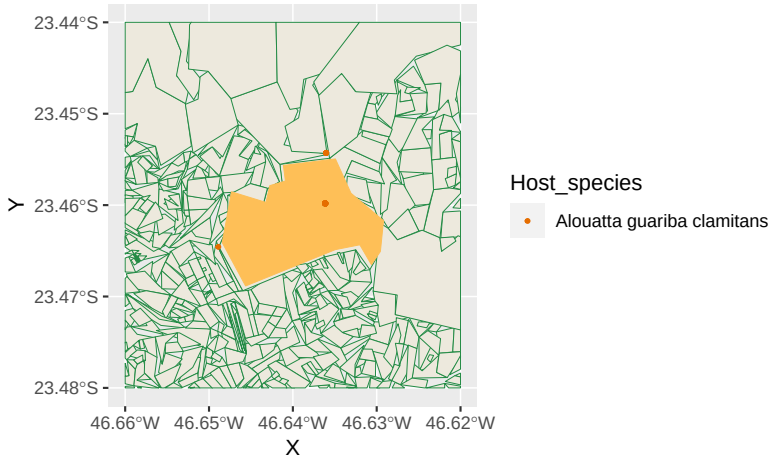
Only 9 samples were collected. What can you infer about sampling based on its spatial distribution?

```
#create a new data frame called horto_PEAL that contains data from Horto only

horto_PEAL <- horto %>%
  filter (Location == "SaoPaulo_PEAL")

# convert data frame to a shapefile and use a projected coordinate system
horto_data <- st_as_sf(horto_PEAL, coords = c("longitude","latitude"))

st_crs(horto_data) <- st_crs (sp_muni)

#create a points spatial feature with latitude and longitude coordinates attached
horto_lat_long <- cbind(horto_data, st_coordinates(horto_data))

#extract data on Haemagogus only
horto_hg <- subset(horto_lat_long, Host_species == "Haemagogus leucocelaenus")
```

```
#create a map showing distribution of Haemagogus
ggplot() +
 geom_sf(data = cts_crop, aes(geometry = geom), fill = "#ede9dd", colour = "#238B45", show.l
    geom_sf(data = horto_shp, fill = "#FEBF57", colour="#FEBF57", show.legend = FALSE) +
  geom_point (data = horto_hg, aes(x = X, y = Y, colour = Host_species), shape = 1, stroke =
  scale_color_manual(values=c("#3399FF"))
```



Can you revise the code and create one for *Alouatta*? Below are some hints...



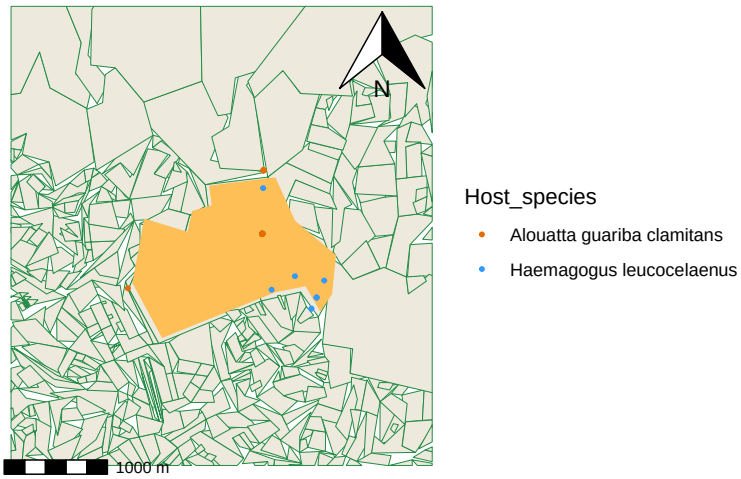
41 samples were collected for Alouatta. What can you infer about the spatial distribution of sampling? We can see that samples were taken within and at the borders of Horto Florestal.

Let's now create a map showing both host species.

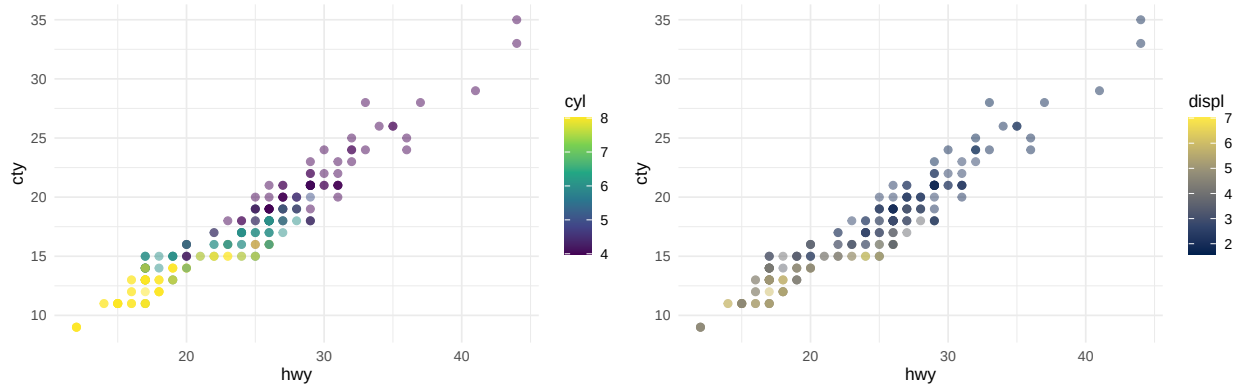
Let's now format our map a bit by adding a **north arrow**, **scale bar**, and remove the grid from our map. We will be using the *ggspatial* package to do this.

```
ggplot() +
 geom_sf(data = cts_crop, aes(geometry = geom), fill = "#ede9dd", colour = "#238B45", show.l
    geom_sf(data = horto_shp, fill = "#FEBF57", colour="#FEBF57", show.legend = FALSE) +
  geom_point (data = horto_lat_long, aes(x = X, y = Y, colour = Host_species), shape = 1, st
  scale_color_manual(values=c("#E56D00","#3399FF")) +
  annotation_scale(location = "bl") + # add scale
  annotation_north_arrow(location = "tr", which_north = "true",
                         pad_x = unit(0.5, "cm"),
                         pad_y = unit(0.5, "cm")) +
  theme(legend.position="bottom",
        panel.border = element_blank(),
        axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
```

```
  theme_void()
```



The plots in Figure 2 show the relationship between city and highway mileage for 38 popular models of cars. In Figure 2a the points are colored by the number of cylinders while in Figure 2b the points are colored by engine displacement.



(a) Color by number of cylinders

(b) Color by engine displacement, in liters

Figure 2: City and highway mileage for 38 popular models of cars.