Applied Ecological Statistics - ECO 636

Lab 2: Data Exploration

Introduction

I this lab we will reinforce some of the concepts presented Zuur *et al.* (2010) "A protocol for data exploration to avoid common statistical problems". Having already familiarized yourself with the protocol, here the focus is two-fold:

1. Dust off and expand your `R` skills by attempting to reproduce some of the figures presented in the paper
2. To give you some experience of the process of data exploration:
   - to figure out *how* to produce the visualizations
   - to reflect on *why* these visualizations are helpful

For each of the figures below, which are my attempts at reproducing the Zuur *et al* (2010) figures, your task is simply(?) to reproduce exactly the figures as they are (**Step 1**), and then discuss what characteristics of the data are being presented in the figure, and what issues it might help diagnose (**Step 2**). I'd like you to do this in small groups without referring to the paper. You can 'check your intuition' later.

All of the data you need to reproduce the figures are available on **Moodle**. If you care to do so, this would be a good opportunity to try out writing *dynamics documents* using `rmarkdown`, one of the nice features of **R Studio**. If you want to do that, ask Meg or Amanda and we can get you set up.

Wing Length (*Figure 2*)

To reproduce Figure 2, you will need the `Sparrows` data. You can download the data directly from **Moodle**. Note that the data are stored as a `.txt` file.

```r
Sparrows <- read.table(file = "Data/Sparrows.txt", header = TRUE)
par(mfrow= c (1,2), mar = c(5,4,2,1))
boxplot(Sparrows$wingcrd,  ylab = "Wing length (mm)")
dotchart(Sparrows$wingcrd, xlab = "Wing length (mm)", ylab = "Order of the data")
```

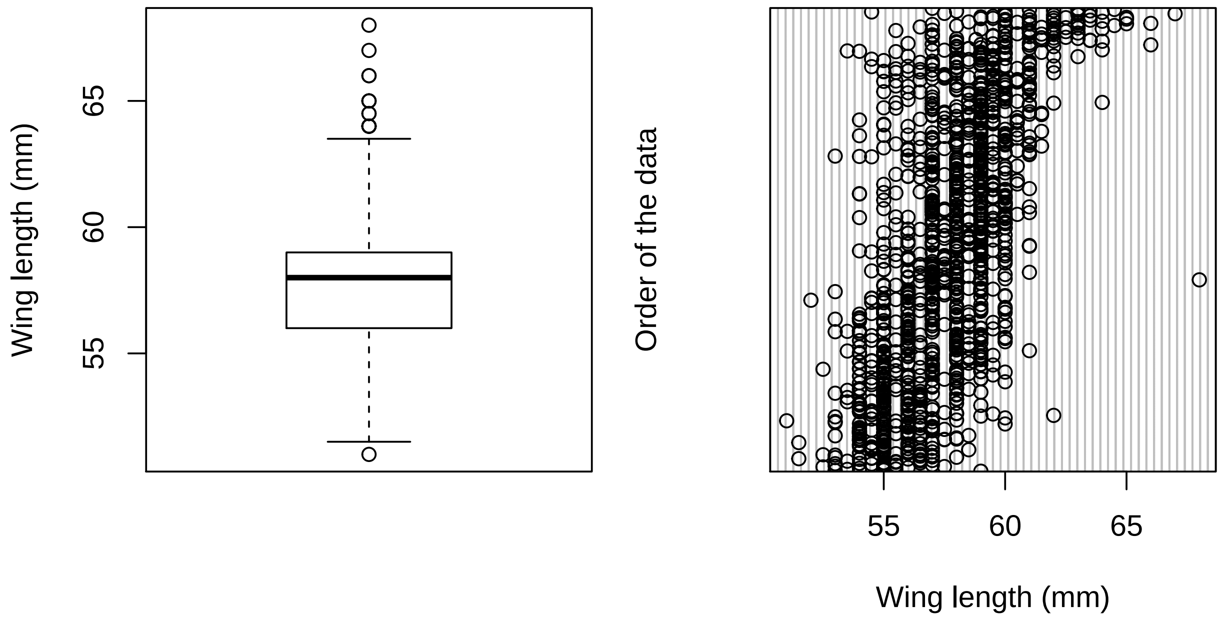

Fig. 2. (a) Boxplot of wing length for 1295 saltmarsh sparrows. The line in the middle of the box represents the median, and the lower and upperends of the box are the 25% and 75% quartiles respectively. The lines indicate 1.5 times the size of the hinge, which is the 75% minus 25% quartiles. (Note that the interval defined by these lines is not a confidence interval.) Points beyond these lines are (often wrongly) considered to be outliers. In some cases it may be helpful to rotate the boxplot 90 to match the Cleveland dotplot. (b) Cleveland dotplot of the same data. The horizontal axis represents the value of wing length, and the vertical axis corresponds to the order of the data, as imported from the data file (in this case sorted by the bird's weight).

1. Can you recreate the figure?
2. What issues are you trying to diagnose?
3. What can we learn from this visualization?

Godwit foraging data (*Figure 4*)

To reproduce Figure 4, you will need the `Godwits` data. You can download the data directly from **Moodle**. Note that the data are stored as a `.txt` file.

To get this looking exactly the same, you will need to do a few things:

- load the `lattice` library (*hint*: `library()`)
- convert `SEX` & `PERIOD` to factors with specific labels (*hint*: `?factor()`)
  - `SEX`: 1 is female, 2 is male
  - `PERIOD`: 1 is Summer, 2 is Pre-migration, 3 is Winter
- find out the `lattice` function for producing box and whisker plots

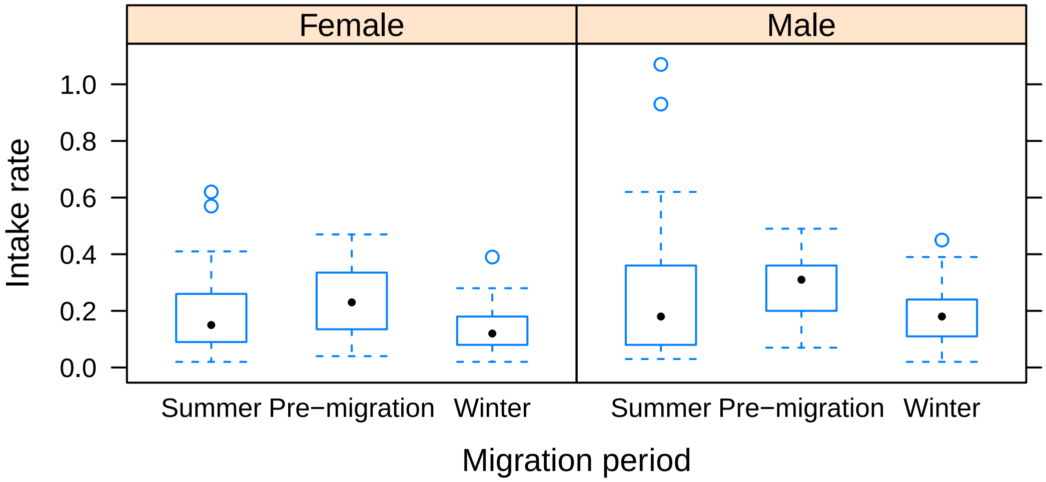

Fig. 4. Multi-panel conditional boxplots for the godwit foraging data. The three boxplots in each panel correspond to three time periods. We are interested in whether the mean values change between sexes and time periods, but need to assume that variation is similar in each group.

1. Can you recreate these figures?
2. What issues are you trying to diagnose?
3. What can we learn from this visualizations?

Summertime sparrow weight (*Figure 5*)

To reproduce Figure 5, you will need the `Sparrows` data again. You are on your own this time - no hints. . . !

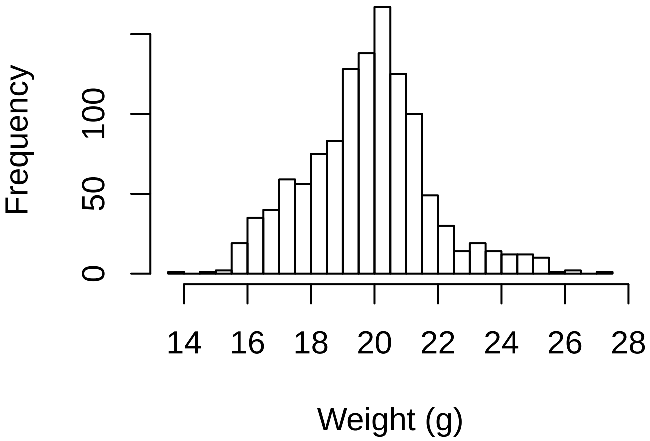

Fig 5a. Histogram of the weight of 1193 sparrows (only the June, July and August data were used). Note that the distribution is skewed.

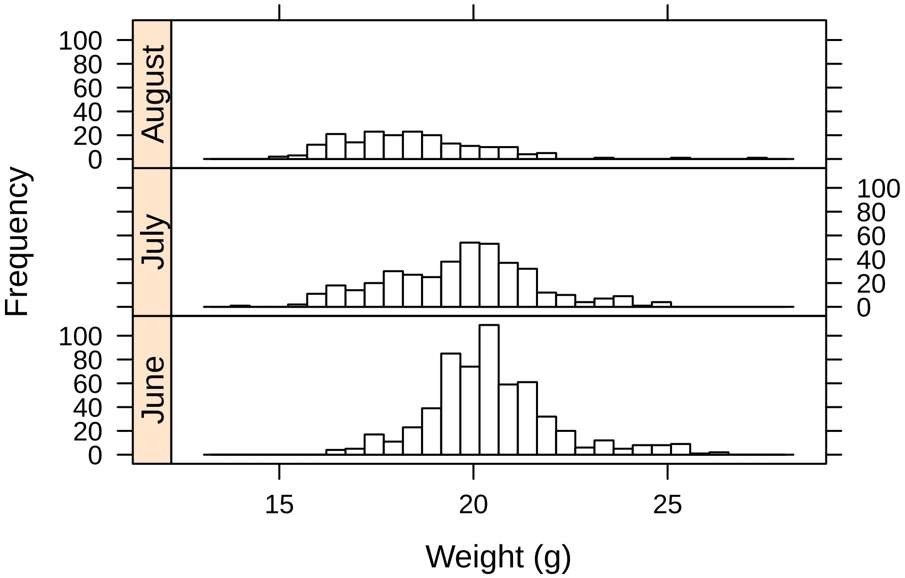

Fig 5b. Histograms for the weight of the sparrows, broken down by month. Note that the centre of the distribution is shifting, and this is causing the skewed distributed for the aggregated data shown in (a)

1. Can you recreate the figures?

2. What issues are you trying to diagnose?
3. What can we learn from these visualizations?

Water birds in rice fields (*Figure 7*)

To reproduce Figure 7, you will need the `BirdData` data again. You may want to create a table, and then plot that…?

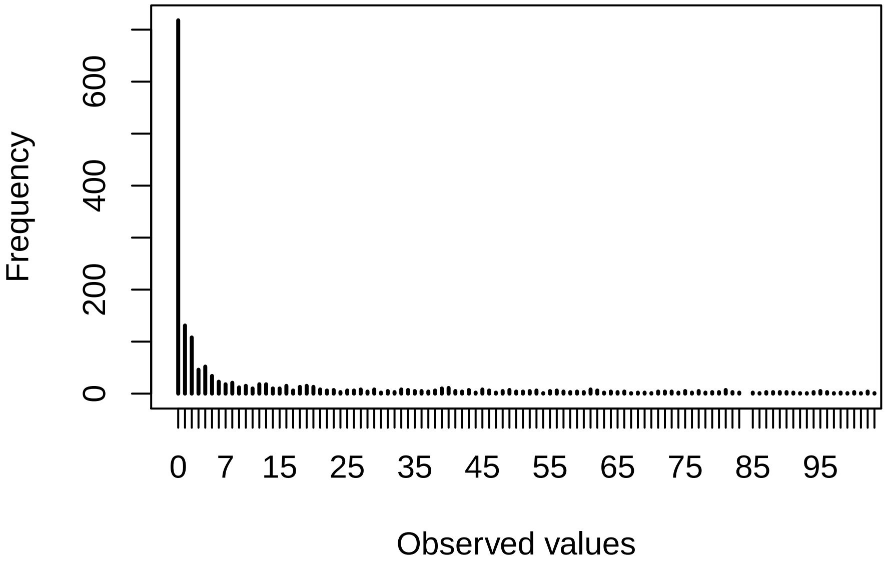

Fig. 7. Frequency plot showing the number of observations with a certain number of waterbirds for the rice field data; 718 of 2035 observations equal zero. Plotting data for individual species would result in even higher frequencies of zeros.

1. Can you recreate the figure?
2. What issues are you trying to diagnose?
3. What can we learn from this visualization?

Sparrow morphometric data (*Figure 10*)

To reproduce Figure 10, you will need the `Sparrows` data again. With this one, though, see if you can change the color, shape, and size of the plotting symbols - the black doesn't help visualize the mass of data. See if you can find some code on the internet to put correlation coefficients in the lower panels.

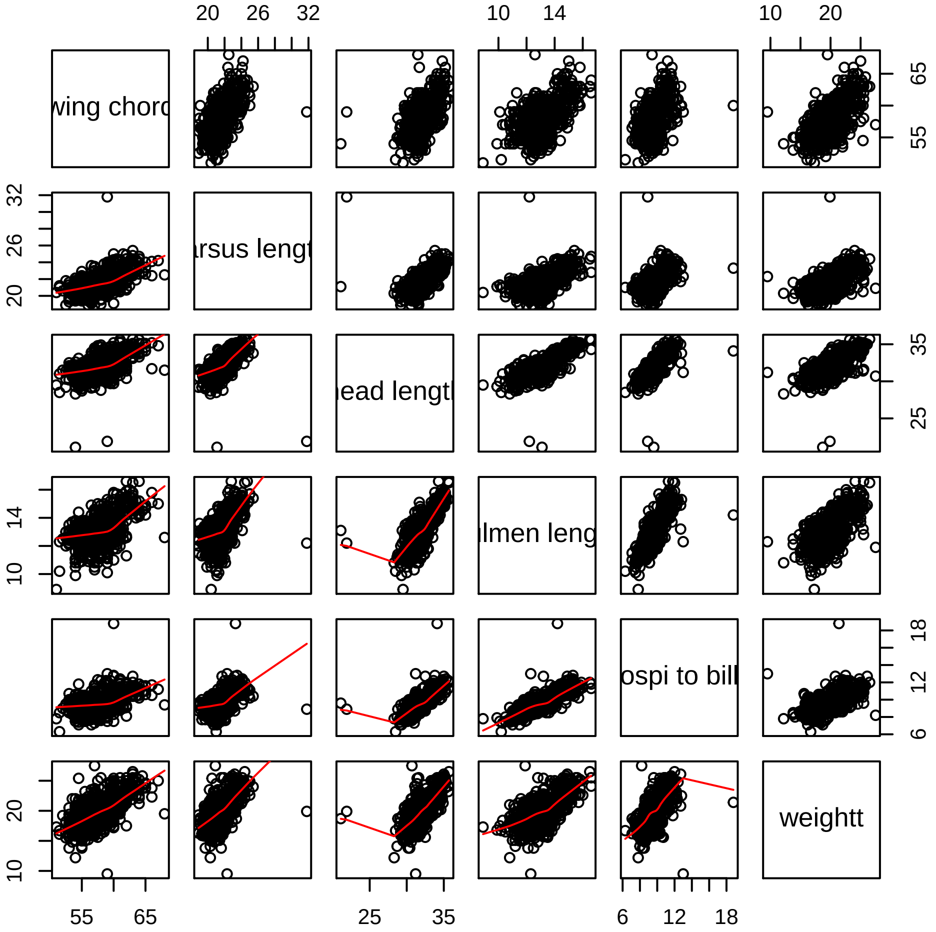

Fig. 10. Multi-panel scatterplot of morphometric data for the 1295 saltmarsh sparrows. The upper right panels show pairwise scatterplots between each variable, and the lower left panels contain Pearson correlationcoefficients. The font size of the correlation coefficient is proportional to its value. Note that there are various outliers.

1. Can you recreate the figure?
2. What issues are you trying to diagnose?
3. What can we learn from this visualization?

Wader time-series and auto-correlation (*Figure 12*)

To reproduce Figure 12, you will need the `wader` data again. The figures on the left are *line* plots, and the figures on the right are auto-correlation function (ACF) plots.

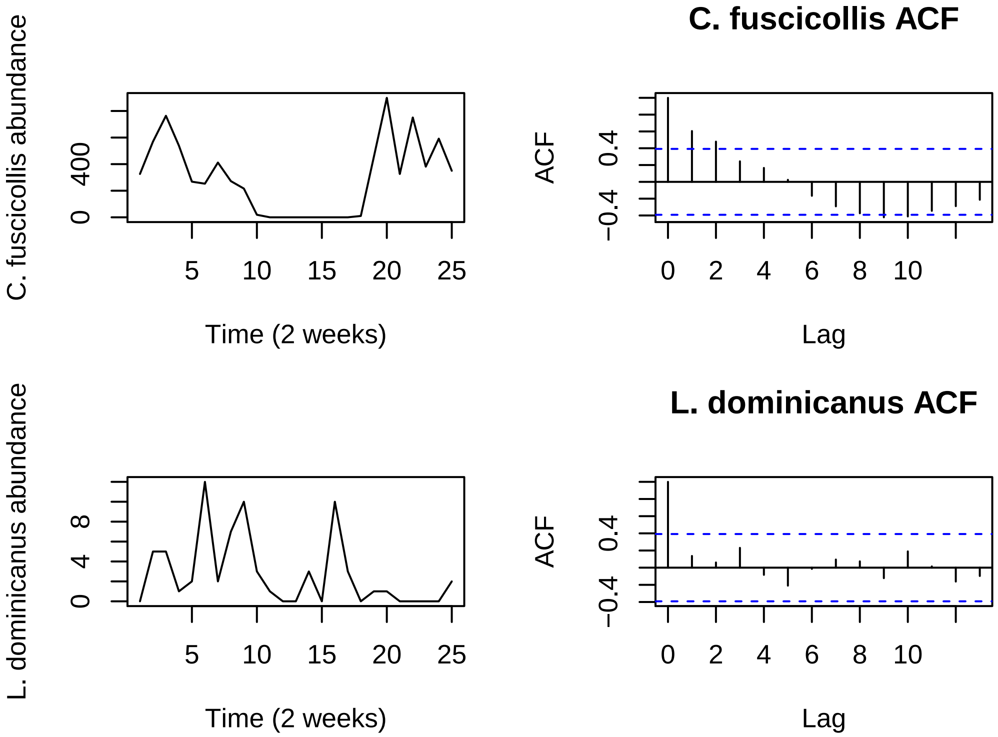

Fig. 12. (a) Number of Calidris fuscicollis plotted vs. time (1 unit = 2 weeks). (b) Auto-correlation function for the C. fuscicollis time series showing a significant correlation at time lags of 2 and 4 weeks (1 time lag = 2 weeks). (c) Number of Larus dominicanus vs. time. (d) Auto-correlation function for L. dominicanus showing no significant correlation. Dotted lines in panels (b) and (d) are c. 95% confidence bands. The auto-correlation with time lag 0 is, by definition, equal to 1.

1. Can you recreate the figure?
2. What issues are you trying to diagnose?
3. What can we learn from this visualization?