Scientific Programming Assignment 1 (SPA1)

MPhil in Computational Biology

October 16, 2022

If there are errors found, I will update the assignment on the moodle.

Due date: 2022-10-31 23:45

Please submit your report to the Moodle website as a single .pdf Name your file `spa1_XXXXXX.Rnw`, where XXXXXX is your six digit coursework number.

Your report must be a maximum of ten pages, excluding the appendix. This course work will consist of 40% towards your overall mark for this module.

The data for this assignment can be found in: `https://github.com/sje30/sp2022/tree/main/assigns/a1`.

1 Cambridge weather files [10 marks]

Download the data file from the github repository. This can be uncompressed using: `tar zxf weather.tar.gz` to create a folder containing daily text, with the most recent entry being `2022_10_16`. This collection has been taken from `http://www.cl.cam.ac.uk/research/dtg/weather/`. Read that page to find out the contents of the data (and any errors).

1. Plot the variation in temperature for one day in the database, 2012-12-25. (2 marks)
2. For each day, compute the mean temperature over the day. What was the hottest day on record, and what was the coldest day on record? (2 marks)
3. Draw a scatterplot of the mean temperature and the total amount of rain in each day. Report the Pearson correlation between the two variables. (2 marks)
4. Draw a graph similar to Figure 1, using our data for Cambridge. How do the graphs compare? (2 marks)
5. Define a "wet day" as one where the total rainfall in a day exceeds a certain threshold that you should clearly define. When was the longest consecutive number of days that it rained according to your method? (2 marks)

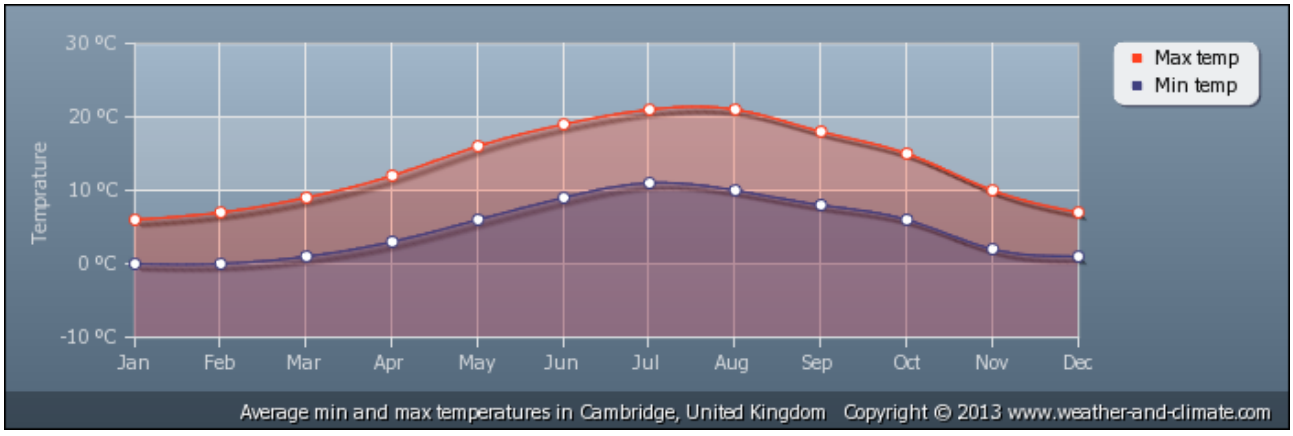

Figure 1: Example min/max mean temperature by month. Graph copied from `http://www.weather-and-climate.com/average-monthly-min-max-Temperature,cambridge,United-Kingdom`.

2 Logistic map [15 marks]

Read the paper by May (1976). Your job is to reproduce the key features of the logistic map, as shown in Figure 4 of the original paper, and shown here. For values of α between 2.8 and 3.6, based on your numerical calculations, can you estimate when period doubling occurs? Show your evidence.

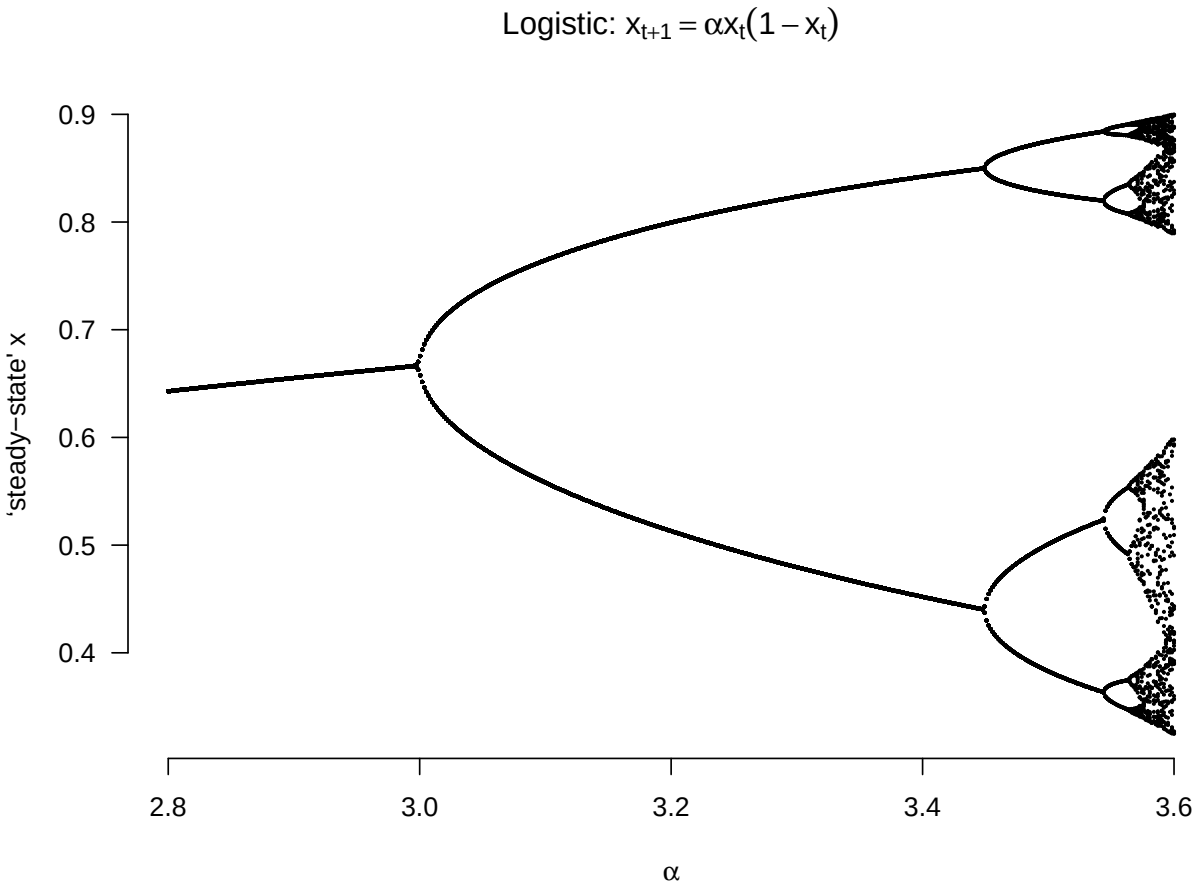

Figure 2: The logistic map, as per figure 4 of May (1976).

May RM (1976) Simple mathematical models with very complicated dynamics. Nature 261:459–467.

`https://paperpile.com/app/p/2d4d413e-df54-0e83-b7fe-8d2b1a26ceb3`

3 RNA sequences [15 marks]

Acknowledgement: source material for this question came from Chapter 5 of the book: *Genomics and Bioinformatics: An introduction to programming tools for life scientists* by Tore Sameulsson. This book is available online, here.

`http://bio.biomedicine.gu.se/cgi-bin/gb/index.cgi?query=about`

Download the data file human.1.rna.fna.gz from the github repository. This file was taken from `ftp://ftp.ncbi.nih.gov/refseq//H_sapiens/mRNA_Prot/human.1.rna.fna.gz`.

This is a relatively large data file (about 500,000 lines). You do not need to uncompress it to read it into R. I also suggest that when starting the assignment, you need only work on a subset of the data. You can do this by e.g. using

```
data <- readLines('human.1.rna.fna.gz',n=5000)
```

to get started. Once you are ready to examine the entire dataset, you can drop the 'n' argument and read in all the data.

This file contains mRNA segments in the FASTA format. Answer the following questions (show your code in the appendix):

1. How many sequences are in the data file? (2 marks)
2. Can you draw a histogram of the length of the sequences? (2 marks)
3. How long is the shortest sequence? How long is the longest sequence? (Give also their SeqID, sequence identifiers). (2 marks)
4. How many sequences contain 8 or more consecutive repeats of `CAG`? Which sequence contains the longest repeat of `CAG`, and how long is it? (4 marks)
5. Write a function that counts the longest run of each four bases, `A`, `C`, `G` and `T` for a sequence. e.g. For the first sequence this should return the following output, a vector of length 4 with appropriate names:

   ```
   > longest_each_base(rnas[1])
   A C G T
   6 6 6 7
   ```

   So, the longest run in the first sequence, `NR_168385.1`, was seven Ts; all other bases only had a run at most of 6 bases. Put the output for the first 10 sequences into a matrix and show that matrix. (2 marks).
6. Can you use the function above to report the mean longest run per sequence for each of the four bases? (1 mark).
7. Is there one base (A, C, G or T) that stands out as having the longest run in each sequence? Show your evidence (2 marks).

3.1 Hints

The FASTA format: `https://www.ncbi.nlm.nih.gov/genbank/fastaformat/`

You might wish to study the following functions and get familiar with their help pages.

```
paste(x, collapse='')
grep(pattern='^>', data)

## regular expressions need some care...
testdata <- c("fake", "1cagcagcag", "2cagcagcagcagcag",
              "2cagcagcagcagcag cagcagcag", "missingcag", "cagcagcagcagcag")
m <- gregexpr( '(c.g){3,}', testdata)
str(m)
regmatches(testdata, m)
attributes(m[[1]])$match.length
```

A sequence ...CAGCAGCAGCAG... would count as 4 consecutive repeats, and would be of length 12 nt.

If you wish to uncompress the data, use `gunzip file.gz` or just to get a subset of lines you can do:

```
zcat human.1.rna.fna.gz | head -5000 > top.txt
```

(mac users may need gzcat rather than zcat)