IMPERIAL COLLEGE LONDON

MSc COURSE IN COMPUTATIONAL METHODS IN ECOLOGY AND EVOLUTION

EXAM 1

For Internal Students of Imperial College of Science, Technology and Medicine

Exam Date: Wednesday, 08th Jan 2018, 14:00 – 17:00

Length of Exam: 3 HOURS

Instructions:

Please note that this exam has three Sections:

- SECTION 1 requires ONE of two questions to be answered
- SECTION 2 requires TWO of three questions to be answered
- SECTION 3 requires ONE of two questions to be answered

THUS, A TOTAL OF FOUR QUESTIONS ARE TO BE ANSWERED, EACH CARRYING EQUAL WEIGHTAGE (25 pts each). So it is a reasonable guideline to spend about 45 minutes on each question.

Read the instructions carefully at the head of each section.

PLEASE PUT ANSWERS TO EACH QUESTION IN A SEPARATE EXAM BOOK.

WE REALLY MEAN IT. THE REASON FOR THIS IS THEN WE CAN PARALLELIZE MARKING AMONG THE DIFFERENT LECTURERS AND YOU GET THE MARKS BACK SOONER.

Continues on next page

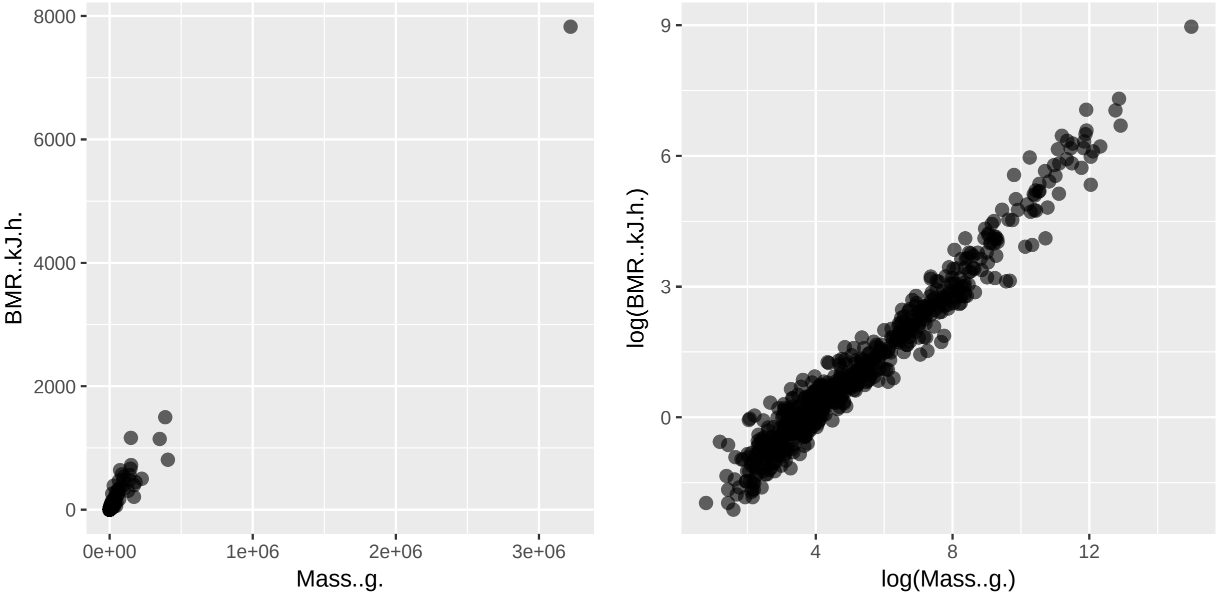
Section 1: Computing, Statistics, Model Fitting

Please select exactly **one question** and answer it. Please indicate clearly in your answer book which question you are answering.

A. You have been given a dataset on basal metabolic rates (rate of energy for maintenance) of mammals across the world. You eagerly load and inspect the data and the metadata files. Two key things the metadata file tells you is that:

- The `Mass..g.` column/field contains the body mass of each mammal in grams
- The `BMR..kJ.h.` field contains the metabolic rate of each mammal in kilojoules/hr.

You would like to makes sense of these data, so decide to examine the relationship between body mass and metabolic rate. Here are two plots that show this relationship:



Now answer the following:

(i) Name and describe two alternative mathematical models you could fit to these data, and how you would determine which model among the two fits better. [40%]

(ii) Explain what biological mechanisms each model could/would capture, and briefly discuss the pros and cons of fitting mechanistic vs. a phenomenological models to these data. [30%]

(iii) *Outline* the appropriate R-/Python-/Pseudo- code that you could write to fit these models to the data. Explain what each command or code-block does with a single-line comment, as you would in the actual script (note that precise syntax is not expected in the answer). [30%]

B. Please read the question carefully. This question requires an essay answer.

You are interested in the effects of climate change on timing of breeding in birds. Your study species is the climate-change sensitive Golden Phoenix (*Phoenix potterus fawkes*). Golden Phoenix eggs burst into flames if they don't hatch by April 20, and also, eggs laid after that date are infertile. You spend the last 4 years collecting data on the date the birds lay their first egg of the first clutch of an individual female in a given year. So, with ongoing climate change, the hope is that more Golden Phoenixes may lay earlier as spring starts earlier every year, and this may aid the species' survival.

You collected data from individual birds attending nests, recording the egg laying data in cumulative days from 1st March. Thus, 14 is March 14, 36 is April 6, and so forth. You collected these data over four years, between 2006 and 2009. You want to analyse whether laying date changed over the years, and in particular, whether it decreased. A decrease (earlier) laying date is what you expect with ongoing global warming. You use two main approaches for data analysis. Find below the R command,

Continues on next page

and the R output. With the given R output, deduce the analysis strategy, and write a methods and results section as you would for a paper or your thesis.

```
> length(PhoenixData$LayingDate) # get total length of data matrix
[1] 108

> var(PhoenixData$LayingDate) # get variance
[1] 539.0041

> summary(PhoenixData$LayingDate)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00   31.75   43.00   48.62   65.25  114.00

> table(PhoenixData$year)

2006 2007 2008 2009
  46   33   10   19

> summary(lm(LayingDate~as.factor(year), data=PhoenixData))

Call:
lm(formula = LayingDate ~ as.factor(year), data = PhoenixData)

Residuals:

    Min      1Q  Median      3Q     Max
-29.804 -16.000  -2.452  12.397  61.697


Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           56.804      3.076  18.467  < 2e-16 ***
as.factor(year)2007   -4.501      4.759  -0.946  0.34643
as.factor(year)2008  -20.704      7.279  -2.844  0.00536 **
as.factor(year)2009  -27.804      5.689  -4.887 3.73e-06 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.86 on 104 degrees of freedom
Multiple R-squared:  0.2152, Adjusted R-squared:  0.1925
F-statistic: 9.505 on 3 and 104 DF,  p-value: 1.325e-05

> summary(lm(LayingDate~year, data=PhoenixData))

Call:
lm(formula = LayingDate ~ year, data = PhoenixData)

Residuals:

    Min      1Q  Median      3Q     Max
-31.247 -15.118  -4.021  11.753  65.205

Coefficients:
             Estimate Std. Error t value Pr(>|t|)

(Intercept) 19017.846   3635.037   5.232 8.52e-07 ***
year           -9.451      1.811  -5.218 9.03e-07 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Continues on next page

```
Residual standard error: 20.81 on 106 degrees of freedom
Multiple R-squared:  0.2044, Adjusted R-squared:  0.1969
F-statistic: 27.23 on 1 and 106 DF,  p-value: 9.026e-07
```

Use all your knowledge from the R week on how to communicate the results of statistical analyses!

Continues on next page

Section 2: GIS, Genomics, C & Data structures

Please select exactly **two questions** and answer them. Please indicate clearly each answer book which question you are answering.

A. You need to generate a habitat suitability map for a species living in Sabah, Borneo that has an optimal habitat defined by steep ($>20°$), mid-altitude slopes (300-600 masl) that have high NDVI values (>0.5).

You are provided with the two data files and one equation:

- A GeoTiff file containing SRTM elevation data (in units of metres above sea level) at 30 metre resolution projected as UTM Zone 50N.
- A GeoTiff containing a scene of MOD09Q1 data: this is an 8-day composite from the MODIS / Terra satellite containing atmospherically corrected reflectance in the red (band 1) and near-infrared band (band 2). The data are continuous ('float') values, at a resolution of 250m and projected as MODIS Sinusoidal.
- The equation for calculating NDVI is:

$$\text{NDVI} = \frac{\text{NIR} - \text{RED}}{\text{NIR+RED}}$$

Describe the series of steps you would go through to integrate the two data sets and develop a simple binary habitat suitability map. Defend your choices of processing options and discuss possible limitations of the data and model.

B. In lectures, we learned that discrete character state data can be 'packed' into bitwise form. For instance, each of the bases of DNA could be represented in a single byte as follows:

```
A: 0001
C: 0010
G: 0100
T: 1000
```

(i) Cytosine and thymine are pyrimidines, while adenine and guanine are purines. Create a pair of 'bit masks' (i.e. field of set bits) for testing if a variable is a purine or a pyrimidine. Use pseudocode to show a statement that performs a check using the bitmask and bitwise operations (30%).

(ii) In the lectures, we performed preliminary ancestral state calculations at a node given left and right descendant nodes. This algorithm is as follows:

I. If both descendant sets have any states in common, construct at their common ancestor the intersection of both descendant sets. Otherwise, go to II.

II. At the common node of both descendants, create the union of descendant state sets.

In C code, a function that performs such an operation would look as follows:

```c
char fitch_downpass(char left, char right)
{
    char result = 0;

    result = left & right;

    if (!result) {
        result = left | right;
    }

    return result;
}
```

Continues on next page

However, the algorithm we examined (Fitch downpass) only provides an initial estimate based on a traversal down the tree (i.e. from the tips towards the roots). To finalise the estimate, a pass from the root of the tree upwards (uppass) should follow the downpass. The Fitch (unordered) parsimony uppass algorithm is as follows:

I. If the preliminary nodal set contains all of the nucleotides present in the final nodal set of ts immediate ancestor, go to II, otherwise go to III.

II. Eliminate all nucleotides from the preliminary nodal set that are not present in the final nodal set of its immediate ancestor and go to VI.

III. If the preliminary nodal set was formed by a union of its descendent sets, go to IV, otherwise go to V.

IV. Add to the preliminary nodal set any nucleotides in the final set of its immediate ancestor that are not present in the preliminary nodal set and go to VI.

V. Add to the preliminary nodal set any nucleotides not already present pro- vided that they are present in both the final set of the immediate ancestor and in at least one of the two immediately descendent preliminary sets and go to VI.

VI. The preliminary nodal set being examined is now final. Descend one node as long as any preliminary nodal sets remain and return to I above.

Translate this algorithm using C-style pseudocode and bitwise operations (following the example above). As with C, remember to declare your variables ahead of time. (40%)

(iii) The Fitch algorithm is unordered and works reasonably well for DNA characters. Unordered parsimony means that transformation between any two states in any direction have the same 'cost'. However, sometimes character information is organized in a cline (i.e. an ordered series) and so we could handle state data in a similar way. Ordered parsimony (Wagner algorithm) lets us to this. In programming, it can be accomplished by counting the number of shifts between two set bits. The following code snippet could be part of a function that accomplishes this:

```
int i = 0;
long bigset; // The bitset determined to be numerically greater
long littleset; // The bitset determined to be numerically smaller

// . . .

while (i < 64 && !(littleset & bigset)) {
        ++i;
        littleset << i;
}
```

Analyse this code snippet, briefly explaining how it works. Explain any potential bugs, issues of safety or portability and how the code could be improved. (30%)

C. A new invasive omnivorous freshwater crab has been found in British rivers. We want to know which native organisms it is preying on. How can we address this question using DNA sequencing data? Please explain the method that you would use, and all the necessary steps involved. Also discuss other methods available and their pros and cons relative to the method you suggest.

Section 3: Neutral theory, HPC

Please select exactly **one question** and answer it. Please indicate clearly in your answer book which question you are answering.

A. Consider a simple individual based neutral model containing a fixed number of individual organisms. In each time step, an individual is chosen at random (according to a uniform distribution) to die and be replaced with the offspring of another 'parent' individual. The parent is also chosen at random according to a uniform distribution. With probability ν the new-born individual is of an entirely new species (speciation) otherwise it is of the same species as it's parent.

You only need to give brief bullet point style answers to the following questions.

(i) For the special case where $\nu = 0$. Given an initial condition where there are many species, what will happen to the number of species in the simulation model over a large number of time steps and why? (20%)

(ii) Considering the same model, what happens in the more general case where $\nu > 0$ starting from a range of different initial conditions. What is the reason for this result? (20%)

(iii) Describe how a pseudo random number generator might need to be used when simulating the neutral model described above. What role does the random number generator seed play in this? (20%)

(iv) Give two advantages or disadvantages to the use of simple models such as the individual based neutral model described here. (20%)

(v) Describe briefly how you could change the way you choose a parent for reproduction in the neutral model described above, in order to build a spatially explicit model of organisms in a flowing stream. (20%)

B. You have written a simulation model, which makes spatially explicit predictions about how a species spreads across a region. The model predictions depend on a key input parameter x. The model produces maps for the species in the form of a large binary matrix in which each entry represents a location; the entry is 1 if the species exists in that location and 0 otherwise.

You only need to give brief bullet point style answers to the following questions.

(i) The simulations take a long time to run, and they are stochastic so repeat simulations are necessary, you also wish to investigate a wide range of different values of x. You decide to use High Performance Computing (HPC) as a way to do this.

a. How might you split up your problem into an array job for running in parallel on HPC? (10%)

b. For running these tasks you would need to write a shell script file. Your shell script contains the code

```
#PBS -l mem=2gb
```

What does this mean and how might you need to change this to perform the required simulations? (10%)

c. Your shell script also contains the code

```
#PBS -l walltime=18:00:00
```

What does this mean and how might you need to change this to perform the required simulations? (10%)

(ii) You wish to find out if the shape occupied by the species exhibits a fractal structure or not, and if it does what the fractal dimension of that shape would be. Describe briefly how you could do this using the box counting method. (40%)

Continues on next page

(iii) You acquire empirical maps showing the distributions in space of three plant species. These maps have already been processed into the same format as the output of your simulation model. You apply your box counting method on the ranges of these species (not their boundaries). In each case, do you expect to see a fractal structure and why?

a. A mammal-dispersed species which follows a Gaussian dispersal kernel (10%)

b. A wind-dispersed species, which follows a fat tailed dispersal kernel (10%)

c. A species which reproduces vegetatively so that offspring are very close in space to their parent (10%)

End of paper