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.

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:
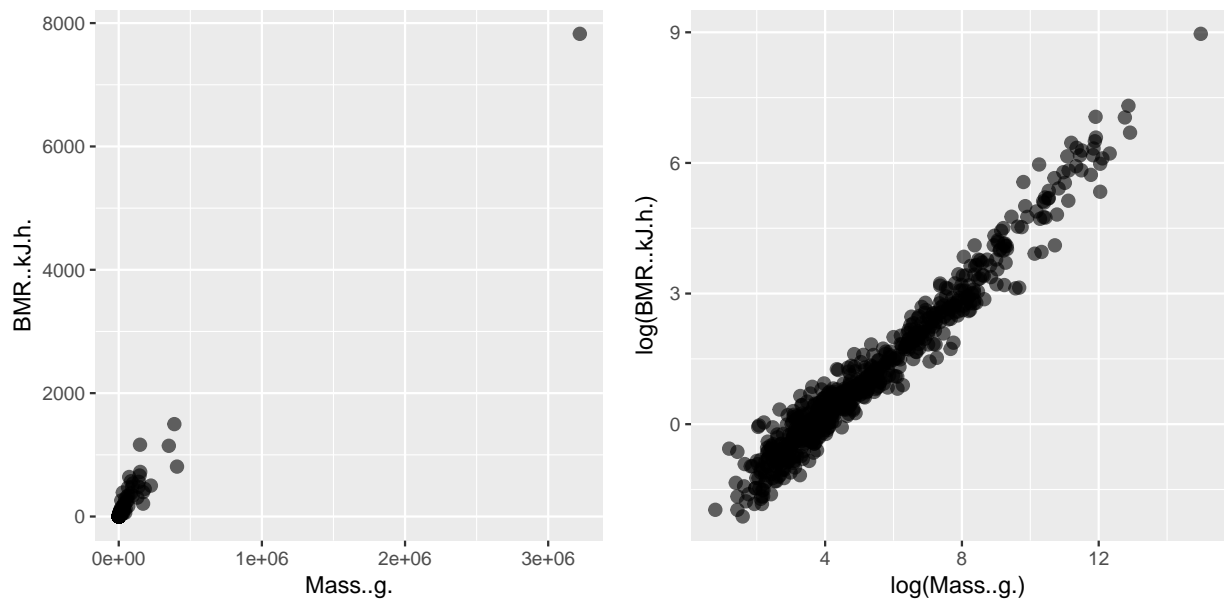
Now answer the following:

(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%]

Model Answer (Markers – Samraat Pawar (first), James Rosindell (second)):

Answers:

(i)
- The ideal answer would have one mechanistic (e.g., Power law; $B = B_0 m^\beta$ and one phenomenological/statistical model (e.g., quadratic). Two statistical models is also acceptable, but would mean a loss of a quarter of this sub-questions mark. Student should state all parameters with their meanings (esp. in case of the mechanistic model). Exceptional answers would give parameter units as well, and also write out the equations for the models.

- Which model among the two fits better: NLLS is an option that was covered in class along with linear models, and model selection using AIC and BIC. Students may pick different methods than these. Exceptional answers would provide some discussion of pros and cons of at least two alternative fitting and model-selection methods methods.

Continues on next page

- Exceptional answers will also comment on and explain why the log-transform is necessary for both choosing the correct model (visualizing the data correctly) and the fitting and model comparison – could be done in log or linear scale, but model comparison would require residuals to be in the same scale (also covered in R visualization and in the NLLS weeks).

(ii)
- Biological mechanisms each model could/would capture should be clearly stated. For example, a power law equation expresses an allometric relationship (this was covered in class pracs in the NLLS example). A statistical model such a straight line or a quadratic would provide insights into gradient of the relationship (the slope coefficient) qualitatively.

- Pros and cons of mechanistic vs. phenomenological modelling: Phenomenological models better for forecasting in many cases, but do not provide much mechanistic insight. Student should weigh that against the problem of over-fitting, and also inference of wrong mechanisms from fitting mechanistic models (Type I error).

(iii) The full outline code should have :

- Loading of packages (e.g., NLLS)

- Data input

- Function specifications of models

- The model fitting perhaps in a loop with `try()` or equivalent

- Parameter extraction from fit object/output (exceptional answers)

- Model selection - AIC or whatever

- Results output

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, 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
```

Continues on next page

```
> 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


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!

Model Answer (Markers – Julia Schroeder (1st), Josh Hodge (2nd)):

Methods: To test whether phoenixes advanced their laying date, I ran a linear model with phoenix laying date as response variable, and year as a continuous variable as explanatory covariate. Laying date was coded such that 1=1st March. I also ran a linear model with year as fixed factor to quantify the differences each year, as the effect may not be completely linear.

Results: I collected data from 108 nests, 46 in 2006, 33 in 2007, 10 in 2008 and 19 in 2009. On average, phoenixes laid their first egg on day 49, that is 18th of April, with a range from 6th March – 114, which is late June. The variance in laying date was high, 539. There was a statistically significant advance in laying date, with on about 9.5 days earlier every year. In 2006, the average laying date was 57 (April

26), in 2007 it was 4.5 days earlier (52.5, April 21), in 2008 it was already 20.7 days earlier (36.3, April 5), and in 2009 it was 27.8 days early (29.2, 29 March).

A distinction answer explains the differing benefits to the two different analyses, has all descriptives, and explains both models correctly including effect sizes. A merit still gives descriptives and explains both models, with little justification, some lack of arguments, or has errors. A pass lacks two (high pass) or three (low pass) of the components (methods, descriptives, model descriptions, model justifications), and may have errors in calculation. A fail describes one component only partially or with errors, or no components. Giving explicit dates can push the mark up a bracket if borderline.

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:

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

 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.

 Model Answer (Markers – Rob Ewers (1st), David Orme (2nd)):

- NDVI needs to be calculated a raster calculator. Note that NDVI could be calculated after the resampling in step ii) – arguably it is more correct to do so on the original data. Any kind of discussion of this point marks good answers.

 In class, the Raster/Map Calculator was used. The standard processing options are to save the new file as a fresh GeoTiff: this kind of attention to detail marks better quality answers, but these aren't vital choices. Good answers would also attempt to describe the calculator syntax. A great answer here would be (modis@2 – modis@1) / (modis@2 + modis@1), showing awareness of the syntax and the dual band file.

- Integrating the data requires that the two rasters be in the same projection and same resolution. The warp function or gdalwarp tool are the correct choices. Answers could use either MODIS sinusoidal or UTM 50N as the shared projection, but it is fairly clearly better to use the local UTM 50N projection and retain the high resolution of the SRTM data. The reprojection requires the selection of a resampling method – better answers will discuss this and choose a method appropriate for continuous values: cubic is appropriate and was used in practicals.

- Raster analysis tools (using the underlying gdaldem tool) are needed to calculate slope from the SRTM data. The default options used on this tool are appropriate, so no discussion is really needed here – detailed answers might again note output options.

- The three data types (elevation, slope and NDVI) now need to be combined to identify suitable areas. In class, a reclassification tool was used to convert continuous layers into integer codes and then a raster calculator was used to find the product of the codes and hence suitable habitat. For example, the following sequence, where good answers are clear about the use of $\leq$ and âĽě to match the specified model and provide a clear explanation of logic of the different final codes.

    – Reclassify:

        ∗ Slope raster: Slope $\leq 20 = 0$, Slope $> 20 = 1$

        ∗ Elevation raster: Elevation $< 300 = 0$; $300 \leq$ Elevation $\leq 600 = 1$; Elevation $> 600 = 2$

            ∗ NDVI raster: NDVI $\leq 0.5 = 0$; NDVI $> 0.5 = 1$

   – Use Raster Calculator or Map Calculator tool to multiply the three values together, giving the following possible outputs:

       ∗ 0 = unsuitable habitat because either the slope, altitude and NDVI is too low for the species

       ∗ 1 = suitable habitat because all three of slope, altitude and NDVI are appropriate for the species

       ∗ 2 = unsuitable habitat because altitude is too high

   – Good answers will then also reclassify these codes to generate a final binary raster.

Alternative methods are possible. Using the raster calculator directly to produce the habitat map would show high technical awareness and is a more flexible method. Any discussion of different approaches marks a good answer. The correct syntax for this would be: srtm@1 $\geq$ 300 AND srtm@1 $\leq$ 600 AND slope@1 $>$ 20 AND ndvi@1 $>$ 0.5

- Good answers might provide detail on generating a map from this data – inclusion of scale bars, north arrows (or a map grid), a title and legend are all details expected from good answers.

- *Data limitations*: SRTM data contains some bad spots: these were not addressed in class so any discussion of this is good. MODIS reflectance data suffers the usual problems of reflectance data: the provided data has been explicitly atmospherically corrected, but that doesn't stop it being cloudy (and having cloud shadow). Good answers might critique the choice of a single 8-day scene and even better ones might discuss seasonality (not so much in the tropics) or El Niño (much more acute).

  Model limitations: Lots of possible improvements could be discussed – relatively low resolution NDVI, simple binary model etc. Good answers might also talk about different variables, using more field data to refine the model or even extending to a species distribution model.

  Answers that present one-quarter to one-third of the expected relevant material (e.g. a sketchy outline of a correct answer), but is marred by major errors or brevity will be awarded a 3rd. Answer gives an account of at least one-half to two-thirds of the expected relevant material, but is marred by defective organisation, omissions or errors that indicate a lack of clear understanding of the concepts will be awarded a 2B. To obtain a 2:1, answer will give a well-organised, mainly accurate account of the relevant concepts and facts, containing at least two-thirds of the expected relevant material and without significant errors of understanding. For Distinction, answers will give an excellent account of virtually all of the expected relevant material. It will show excellent comprehension of the relevant concepts and tools and/or includes some relevant supplementary material.

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:

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

    result = left & right ;

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

    return result ;
}
```

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%)

(i)
```
Purine: 0101
Pyrimidine: 1010

//Check for purine:
Load: DNAvar;
Load: Purine;
IF (DNAvar & Purine) {
print "DNAvar is purine";
}
```

(ii) There are a few ways to do it. This is one possibility:

```
char fitch_upppass
(char left, char right, char n, char ancestor)
{
char result = 0;

result = ancestor & n;

if (n != ancestor) {
if (left & right) {
result = n | ( ancestor & (left | right));
}
else {
result = n | ancestor;
}
}

return result;
}
```

(iii) The code should work because the numerically smaller bitset will have its highest significant bit to the right of the highest significant bit in the larger bitset. First-class answers will note that the code assumes there are no lower bits set in bigset. Otherwise, it could result in errors. Therefore some operations must be performed before the while loop in order to ensure this.

Upper second answers will note the use of a hard-coded 64 for 64 bits. This isn't portable. These answers will note that the machine-specific limit should be set. A first-class answer will know how to do this with C code, showing some evidence of having paid close attention in the demonstrations and reading the code examples. A lower second answer might simply note the issue of portability in hard-coded values but offer little else.

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.

- We would use metabarcoding of the gut contents (or faeces): sequencing of all copies of a plant barcode gene and an animal barcode gene in the sample, to identify the species present. [Extra points if mentioning rbcL as a plant barcode gene and COI as an animal one, or universal ribosomal DNA primers which could be used for both].

- The student could also suggest metagenomics instead of metabarcoding. The pros and cons of each is that metabarcoding is cheaper and easier to analyse, whereas metagenomics provides more certainty of species assignment.

- Steps:
  - Collect the samples (gut contents or faeces of the invasive crab) and store them in a preservative solution.

  - Extract DNA

  - Perform two PCRs to amplify the barcode genes, one PCR with plant barcode [rbcL] primers, one with animal barcode [COI] primers.

  - Sequence the PCR products on a high-throughput sequencer. Extra points for extra reading if they suggest a sequencing platform, e.g. Illumina HiSeq or MiSeq or Solid

  - Check the quality of the sequences and remove those with low quality.

  - Analyse the sequencing data using the program QIIME:
    * Remove chimeras
    * Cluster sequences
    * Compare results with a sequence reference database to identify OTUs (Operational Taxonomic Units).
    * Make an OTU table containing the OTU ID, counts of how many times the OTU appears in each sample, and taxonomy information for each OTU.

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%)

Model Answer (Markers – James Rosindell (1st), Samraat Pawar (2nd)):

(i) It will decay to one species OR monodominance (10%) because there is no speciation to balance extinction so diversity can only decrease to the absorbing state of a single species. (10%)

(ii) Diversity will converge to the same state from any initial condition, or some mention of dynamic equilibrium or similar (10%). The higher the speciation rate the higher the diversity tends to be (10%).

(iii) A pseudo random number generator is the outcome of a deterministic process that produces a set of apparently random numbers, it is needed here to model the stochastic elements of the simulation. (10%)

The random seed is a value used to initialise a pseudo random number generator. Using the same seed again will give you exactly the same sequence of pseudo random numbers. (10%)

(iv) Any two reasonable and well thought out points will do for 1 mark each e.g.

- advantage: can use to study processes of interest in isolation from complicating factors

- disadvantage: not very realistic / ignores important factors

- advantage: computationally / analytically tractable

- advantage: can easily explore parameter space of the model / limited number of parameters

(v) Pick the parent according to a dispersal kernel with some kind of bias in the direction of flow OR pick one of the nearest neighbours but with probability weighted in favour of downstream and against up stream. A solution, which models the stream as a 1D or 2D structure would be equally acceptable (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%)

(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%)

Model Answer (Markers – James Rosindell (1st), Samraat Pawar (2nd)):

(i)   a. Each repeat simulation and each value of x forms an independent job - this is an embarrassingly parallel problem (10%).

    b. This describes how much RAM is needed to be allocated to the simulations, currently 2 gigabytes. If your simulations require more RAM then the number should increase but this will probably incur a penalty when waiting in the queue for jobs to run (10%).

    c. This describes how much CPU time is needed for each simulation currently 18 hours. If your simulations require more time then the number should increase but this will probably incur a penalty when waiting in the queue for jobs to run (10%).

(ii) Place a grid of a certain width over the matrix and count the number of boxes that contain a 1 in the underlying range matrix (10%)

Repeat the process for grids with a variety of different widths (10%)

Plot a graph showing the log of the number of squares needed to cover the range matrix on the y axis against the log of the length of each square on the x axis (10%)

A fractal structure should produce a straight line graph with a gradient equal to $-D$ where D is the fractal dimension. A non fractal graph will produce a line that curves down and converges to a certain value. (10%)

a. This is unlikely to be fractal because the species will spread in fairly compact shapes (10%)

b. This species will probably be fractal because the long distance dispersal events keep spawning new patches of the species from existing patches resulting in a spatially complex structure with many different patches of different sizes. (10%)

c. This is unlikely to be fractal because the species will spread in fairly compact shapes like the mammal dispersed species (10%)