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, 09th 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.

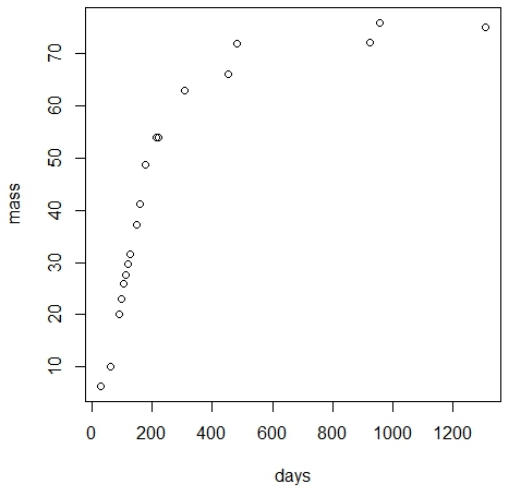
WE REALLY MEAN IT. THE REASON FOR THIS IS THEN WE CAN PARALLELIZE MARKING AMONG THE DIFFERENT LECTURERS AND YOU GET THE MARKS BACK SOONER.

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 obtained data on the growth (increase in biomass per unit area) of a population of Rhododendron after invasion into a new location in the northeast of England. Here is what the population's growth trajectory looks like:



Now answer the following:

(i) Name and describe at least one mechanistic and one purely phenomenological/statistical model you could fit to these data. Explain what biological mechanisms each model could/would capture, how you would determine which model among the two fits better, and the pros and cons of mechanistic vs. phenomenological modelling in this case. [70%]

(ii) Write out the appropriate R-/Python-/pseudo- code that would 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. [30%]

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

Answers:

(i)
- Mechanistic model example: Logistic growth, phenomenological/statistical model example: Cubic polynomial. Other examples are also possible. Student should state all parameters and the execptional answers would give units as well ad descrbe what each parameter does (esp. in case of the mechanistic model). Exceptional answers would also write out the equations for the mechanistic (e.g., logistic growth) and non-mechanistic (e.g., cubic polynomial, and provide more than two models.
- Biological mechanisms each model could/would capture: Logistic growth would capture intrinsic rate of increase and level of intraspecific competition + environmental limits (the carrying capacity). Phenomenological model could also give estimate of carrying capacity, but not intrinsic growth.
- Which model among the two fits better: NLLS fitting with model delection using AIC, BIC, likelihod ratios, etc. Exceptional answers would would provide some dicussion of pros and cons of atleast two alternative methods.
- Pros and cons of mechanistic vs. phenomenological modelling: Student should state the problem with over-fitting and inference of wrong mechanisms from mechanistic models. Phe-

Continues on next page

nomenological models better for forecasting in many cases, but do not provide much mechanistic insight.

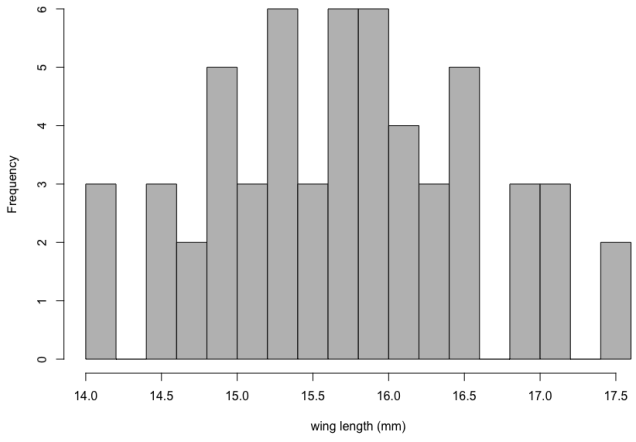
(ii) The code should incorporate:

- Data input statements
- Loading of packages (e.g., NLLS)
- Function specifications of Model(s)
- Function specifications for getting starting values (exceptional answers)
- The NLLS fitting, perhaps in a loop with `try()` or equivalent with different starting valus
- Parameter extraction from fit object/output
- Model selection - AIC or whatever
- Results output

B. You are doing a research project on 57 bird skins that you found in the a natural history museum's collection. These birds are not catalogued so it is unclear what species they belong to. You believe that they are of the species *Parus lundyensis*. The species is clearly sexually dimorphic in plumage, allowing you to identify the sex of the skins. You want to find out whether these 57 bird skins are from the same species. You found a reference to that species in the "Handbook of the mysterious birds of the world" which states "*Parus lundyensis* shows a plumage dimorphism. Both sexes have a wing length of 15.5 – 17.0mm."

You conduct some exploratory data analysis, and then run the main test. Below is the R output from these analyses:

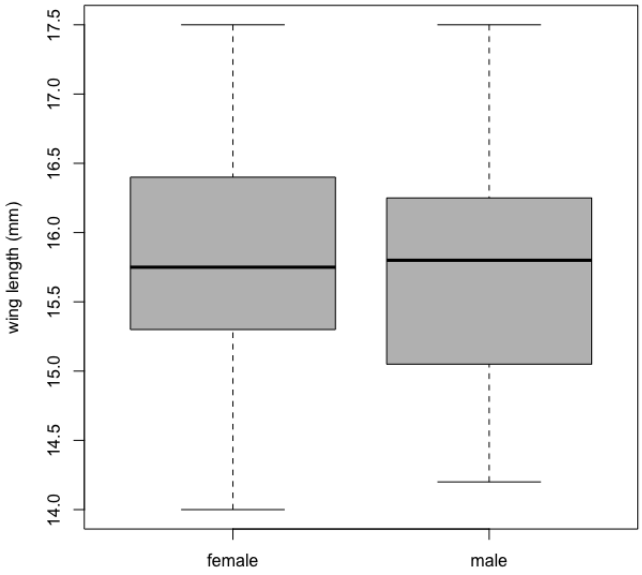
```
'data.frame':   57 obs. of  3 variables:
$ Catalogue_Nr   : int  1 2 3 4 5 6 7 8 9 10 ...
$ wing_length.mm.: num  16.5 15.2 16.1 16.6 14.8 16.6 15.9 15.8 15 15.9 ...
$ sex            : Factor w/ 2 levels "female","male": 2 1 2 1 2 1 1 2 1 1 ...

Catalogue_Nr wing_length.mm.    sex
1            1             16.5   male
2            2             15.2 female
3            3             16.1   male
4            4             16.6 female
5            5             14.8   male
6            6             16.6 female
```




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

Continues on next page

```
14.00   15.20   15.80   15.77   16.30   17.50
```




```
Welch Two Sample t-test

data:  data$wing_length.mm by data$sex
t = 0.074339, df = 55, p-value = 0.941
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
    -0.4422484  0.4763225
sample estimates:
mean in group female   mean in group male
            15.78000             15.76296
```

```
Welch Two Sample t-test

data:  data$wing_length.mm by data$sex
t = 0.074339, df = 55, p-value = 0.941
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
    -0.4422484  0.4763225
sample estimates:
mean in group female   mean in group male
            15.78000             15.76296


One Sample t-test

data:  data$wing_length.mm
t = -4.2154, df = 56, p-value = 9.175e-05
alternative hypothesis: true mean is not equal to 16.25
95 percent confidence interval:
    15.54474 15.99912
sample estimates:
mean of x
    15.77193
```

(i) Write a methods section detailing the aim, methods and statistical analysis you would do to achieve your aim. Fully justify your methods. [30%]

Continues on next page

(ii) Write the R (or pseudo-) code that you need to produce the results starting with a blank R script (including the code to get the above results). Comment each line of code to explain what you do here and why. You can add code that would produce additional results that you think would be good to give. [35%]

(iii) Write a results section using the above output, including descriptive statistics, and effect sizes – quantify the outcome. If you come across issues, discuss them in writing. Write one concluding sentence at the end of this section that links back to the aim. [35%]

Model Answer (Markers – Julia Schroeder (1st), Matteo Fumagalli (2nd)):

We have explicitly done the t-test testing against a mean, and discussed this thoroughly so students should be familiar with it. Also, in all lectures, I have put importance on reporting in methods and results, including the importance of quantifying the results in terms of effect sizes.

(i) Example:

> *We measured the wing length of 57 bird skins found at the museum to the nearest 0.1mm. I first determined the range of the data and plotted the histogram to identify potential outliers. I computed basic descriptive statistics. I also tested for a difference in wing length between the sexes. Then, to find out whether these bird skins had a similar wing length than the existing Parus lundyensis birds, I conducted a one-sided t-test to test against a mean wing length of 16.3mm (Handbook of the mysterious birds of the world).*

To get a distinction on this question, students need to explicitly state the question and aim of the methods before they go on to the t-test. They also should say that they explored the data visually, and that they computed descriptives. A distinction answer also mentions the units that birds were measured in.

A pass (52) just states the t-test and nothing else.

(ii) THE FOLLOWING NEED NOT HAVE THE EXACT SYNTAX (CAN ALSO BE PSEUDOCODE) — PLEASE ASSESS FOR CORRECTNESS OF SEQUENCE AND COMMANDS, NOT EXACT SYNTAX (CMEE STUDENTS ARE NOT EXPECTED TO REPRODUCE CODE SYNTAX FROM MEMORY IN EXAMS).

```
rm(list=ls()) # clear workspace

setwd("~/Box Sync/EEC/Cohort17/Exams17_18") # set working directory; OPTIONAL, ↩
    AS CMEE STUDENTS WERE TOLD NOT TO INCLUDE setwd() IN THEIR R SCRIPT.

data<-read.table("dataStatsQ.txt", header=TRUE) # import data

str(data) # check structure of the data

head(data) # check structure of the data

colnames(data) #check data headers

hist(data$wing_length.mm, breaks=13,col="grey", xlab="wing length
(mm)", main="") # plot histogram

summary(data$wing_length.mm) # explore data

range(data$wing_length.mm) # explore data

var(data$wing_length.mm) # computing the variance

t.test(data$wing_length.mm~data$sex) # check whether there's a difference ↩
    between males and females

boxplot(data$wing_length.mm~data$sex, ylab="wing length (mm)", col="grey") # ↩
    look at sex differences
```

Continues on next page

```
mean(c(15.5,17)) # test for the mean wing length given in the book (assuming ↩
    values in the book refer to a range)

t.test(data$wing_length.mm, mu=16.25) #conduct t-test to see if wing length in ↩
    sample differs from mean wing length given in book (mu)
```

A distinction answer gives all commands to produce the presented output including clear annotation (72). Higher marks can be achieved by adding extra code that the student explains in annotations. A pass (52) is just giving the t-test.

(iii) The 57 skins of birds had a mean wing length of 15.7mm, ranging from 14.0mm to 17.5mm. There was no statistically significant difference between skins of each sex ($t = 0.07$, $df = 55$, $p = 0.94$, two sided t-test). I tested for a statistically significant difference of the wing length of these skins.

This was statistically significantly different from a mean of 16.25, which is the mid-point between the range given in the Handbook of mysterious birds in the world. The measured skins were on average 0.48mm shorter than the bird measurements described in the book. Therefore, one could conclude that the skins are of a different species.

However, the 95% confidence interval of the sampled bird's wing length fully overlaps the range given in the book, therefore, one could conclude that it is indeed the same species. The difference in result here stems from that we did not give the t-test information about the variance of the population sampled in the book, and therefore, we had to assume that the variances are equal. Distinction answers (>72) elaborate on this latter point. A merit needs to quantify the difference between mean and measured skins. A pass would only state that there was a significant difference or similar (52).

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. Using examples, describe the process by which satellite images are converted into data for analysis in a geographic information system [70%].

How does the spectral and spatial resolution of the image affect the accuracy of the data product [30%]?

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

- Image is collected: answers should mention how different bands are used by different satellites and for different data products.
- Image needs georeferencing; orthorectifying (remove perspective and terrain effects); calibration (convert sensor value to reflectance value); atmospheric correction (remove spectral biases from haze, water vapour, etc
- Reflectance profile is generated from field data: albedo varies with surface type
- Reflectance profile observed in each given pixel is compared to best-matching profile of a surface type
- Better data is obtained by having higher spectral resolution (more bands and narrower bands); and higher spatial resolution (smaller pixels to avoid averaging effects).
- Most likely example is land cover mapping using Landsat or ASTER data.

B. In the lectures, we explored the use of binary node structures in C. Answer the following:

(i) Design a structure that could be used to represent a node with an arbitrary number of descendants. [30%]

(ii) Write a C function that uses recursion to traverse a tree constructed of such a node. Partial credit is given for correct pseudocode. [40%]

(iii) What are the safety issues associated with this node structure and traversal method? What are some ways of adjusting the structure or the function to mitigate these risks? [30%]

Model Answer (Markers – Martin Brazeau (1st), Matteo Fumagalli (2nd)):

(i) Example:

```
struct nnode_st {
struct nnode_t **descendants;
struct nnode_t *ancestor;
unsigned ndescendants;
} nnode_st;
```

(ii) Example:

```
void traverse_nnary (struct node *n)
{

if (n->descendants == NULL) { // Optional
return;
}

unsigned c = 0;

do {
traverse_nnary(n->descendants[c]);
++c;
```

Continues on next page

```
} while (c < n->ndescendants);
}
```

The example function must include a call to itself. The example function must naturally terminate, assuming the tree has been correctly constructed.

(iii)
- It's written in C; so don't use C, but some safer language with bounds-checking (for full marks).
- Lack of bounds-checking in C: the entries ndescendants and the number of actual descendent node structs must be kept up to date simultaneous. Otherwise, read errors will occur.
- Assumes that care will be taken to set descendant pointers to NULL when node is terminal.
- Examples that use linked-list structures for descendants (i.e. MrBayes style) would risk failing to set NULL in next member of last member in list.

C. How and why does population structure affect expected genotype frequencies? Explain the relationship between the genotype frequencies and how these change with increasing population structure. How can genotype frequencies be used to test for population structure?

Model Answer (Markers – Jason Hodgson (1st), Matteo Fumagalli (2nd)):

- Expected genotype frequencies in panmictic populations are described by the Hardy Weinberg expansion: $A_1A_1 = p^2$, $A_1A_2 = 2pq$, and $A_2A_2 = q^2$, where Freq(A_1) = p, Freq(A_2) = q, and $p + q = 1$.
- In panmictic populations as Freq(A1) increases the A1A1 homozygotes increase, and A2A2 decrease. The proportion of heterozygotes are maximised when p = q = 0.5. Thus, the proportion of heterozygotes increases as Freq(A1) increases from 0 to 0.5 decreases from 0.5 to 1.
- Panmictic (randomly mating) populations are maximally outbred and maximise heterozygosity for any given polymorphic site for any allele frequency. When population structure develops, people are more likely to mate with individuals more closely related to them than expected given random mating. Closely related individuals are more likely to share genotypes. This increases homozygosity.
- An increase in homozygosity leads to a corresponding decrease in heterozygosity.
- Structured populations have elevated homozygosity and reduced heterozygosity than expected given panmixia. The greater the departure from random mating, the greater the bias in observed genotype frequencies.
- A Chi-Square test of observed versus expected genotype frequencies can be used to test for an excess of homozygosity and a paucity of heterozygosity.

Continues on next page

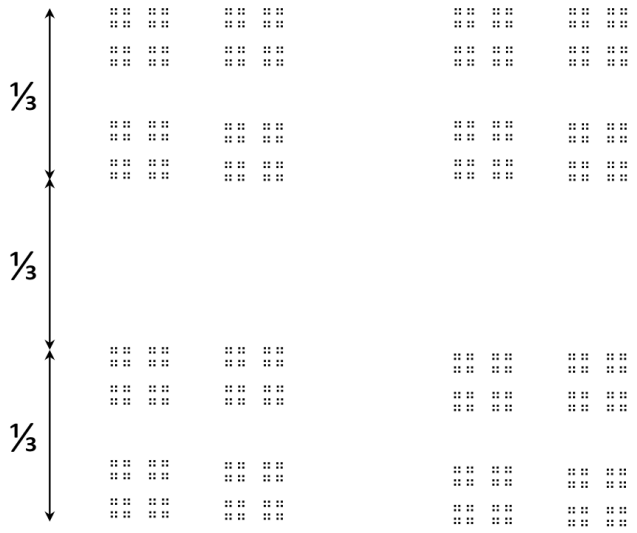
Section 3: Neutral theory HPC

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

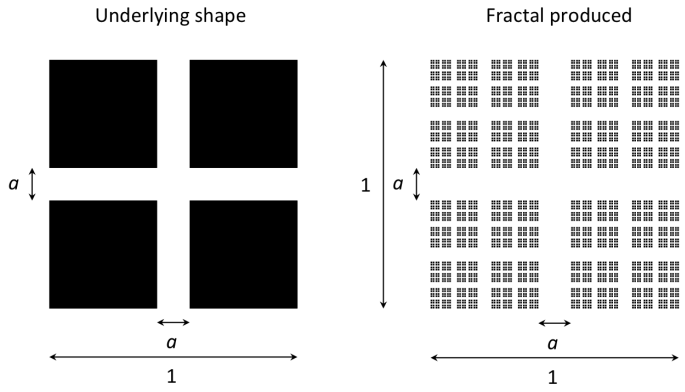
A. Answer the following questions. Please be brief in your answers.

(i) Give three reasons why fractals occur in the natural world giving an example for each. [30%]

(ii) Calculate the dimension of the following fractal, which is known as the Cantor Dust. Please show your workings. [20%]



(iii) The following diagram shows how to construct a range of different fractals depending on variable a where . Write down a formula for the dimension of the fractal in the general case, as a function of the variable a. The Cantor Dust shown in part b) corresponds to the case where $a = \frac{1}{3}$. Check that when you put $a = \frac{1}{3}$ into your formula you do get the same answer as you gave in b) [20%]



(iv) Write a short piece of pseudo code to draw this fractal as a function of a. In pseudo code you can say things like "draw a filled square of width w with the top left corner at `(x , y)`". *Hint*: define a function that calls itself four times and be careful to prevent the resulting loop from continuing infinitely. Your function should have five input parameters: a, w, x, y, threshold [30%]

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

(i) Any three reasonable points [10% each] such as (a) Fitting large surface areas into small volumes (e.g. lungs), (b) Solving transportation problems optimally (e.g. circulatory system), (c) As the result of a simple set of rules operating (e.g. plant growth), (d) Same processes happening at multiple scales (e.g. coastlines). [30%]

Continues on next page

(ii) Requires 4 copies of itself to construct the same shape but 3 times larger. Therefore 4 = 3D where D represents the fractal dimension (10%). [20%]

Log(4) = log(3D) D = log(4)/log(3) = 1.262 [10%]

(iii) Requires 4 copies of itself to construct the same shape but 2/(1-a) times larger [10%]

Therefore 4 = (2/(1-a))D where D represents the fractal dimension.

D = log(4)/log(2/(1-a)) (10%)

Putting in a = 1/3 gives D = log(4)/log(2/(1-(1/3))) = log(4)/log(3) and agrees with answer to ii) (no marks for this last part really because is basically a hint to help them be sure they've got the right answer above)

[20%]

(iv)

```
Draw_fractal(a, width, x, y, threshold)
{
if (width > threshold)
{
// this if statement creates the break to prevent an infinite loop
new_w = (1-a)/2
// easier to define the new width once then use later
Draw_fractal(a, new_w, x, y, threshold)
Draw_fractal(a, new_w, x+a+new_w, y, threshold)
Draw_fractal(a, new_w, x, y+a+new_w, threshold)
Draw_fractal(a, new_w, x+a+new_w, y+a+new_w, threshold)
}
else
{
// draw the fractal points instead of regressing deeper
draw a filled square of width w with the top left corner at (x , y)
}
}
```

[30%]

B. Consider the following spatially explicit neutral model, also known as the voter model. In each time step, an individual is chosen at random to die and be replaced with the offspring of one of its eight immediate neighbours. With probability ν the new-born individual is of an entirely new species in the system (speciation) otherwise it is of the same species as it's parent. This question relates to performing simulations of such a model and observing its species richness.

(i) For the special case where $\nu = 0$, given an initial condition where there are many species in the system, describe what will happen to the diversity of the system as the simulation progresses and why. [20%]

(ii) What is meant by dynamic equilibrium and why is it necessary to have a burn in period for your simulations? [20%]

(iii) You would need to repeat this stochastic simulation many times to get an idea of the overall behaviour of the system and you plan to do this using High Performance Computing (HPC). In the context of this goal describe the meaning of the following and their implications on performing your simulations:

a. The line of R code [20%]:

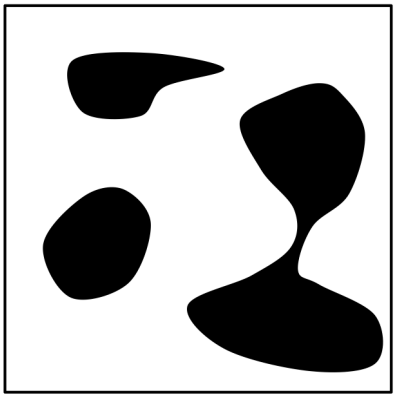
```
xvar <- as.numeric(Sys.getenv("PBS_ARRAY_INDEX"))
```

b. This line of shell script [20%]:

Continues on next page

```
#PBS -l walltime=2:15:00
```

(iv) Suppose now that your simulation model is running on a fragmented landscape where not every position in space can be occupied. The simulation rules remain the same as before except that when a dead individual is replaced, there may not be eight immediate neighbours any more because some of the neighbouring spaces might not be suitable habitat. Consider the following fragmented landscape in which habitat is shown in black and non-habitat is shown in white (the boundary line is in black but does not count as habitat).



Imagine simulating an extremely low (practically zero) speciation rate with a voter model on this landscape, what species richness would you expect to see at dynamic equilibrium? Explain your answer. [20%]

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

(i) (20%) It will decay to one species OR mono-dominance (10%), because there is no speciation to balance extinction so diversity can only decrease to the absorbing state of a single species (10%)

(ii) (20%) Dynamic equilibrium is, this case, a natural balance between speciation and extinction so that species richness is stable over long periods of time (equilibrium) even though the species themselves are turning over (dynamic) (10%).

Burn in is necessary to make sure that simulations have reached dynamic equilibrium and are no longer being affected by arbitrary initial conditions (10%).

(iii) a. (20%) This is an array job, the HPC system will repeat parallel simulations each with different values of `PBS_ARRAY_INDEX`. Your R code needs to read this variable so that your simulations are not identical (10%).

The implications are that you'll want to set the random number seed from `xvar` and also probably use it in naming your output files so that they aren't confused with one another and don't get overwritten or conflicted (10%).

b. (20%) This describes how much CPU time is needed for each simulation currently 2 hours 15 minutes (10%).

The implication is that you probably want to include some kind of timer in your simulation to be sure it doesn't run over, or saves its progress when the time is nearly up, or you'd want to test in advance that 2 hours 15 minutes is definitely enough to complete each simulation. (10%)

(iv) (20%) There will be three species because dispersal is to nearest neighbours only and there is no way to get between the three fragments. Speciation will eventually ensure that each fragment has its own species (10%), but because speciation is practically zero there's no chance of sustaining more than one different species in each fragment. (10%). I'd give a bonus point for realising that one of the fragments is only just connected so as speciation increases a little we'd probably expect a species in the north and another in the south of that 'island'

End of paper