Big Data Exam Details

The final assessment for this unit is a ‘take home exam’, which will be completed **online via timed LMS quizzes.**

You can complete the exam any time during the **24 hour period from 7am on 3rd November**.

The exam will be 'open book' ie you can look at any material you wish during the exam. However, you must do your own work, and the exam does contain ways to test for plagiarism and collaboration, which are considered to be serious offences and will be heavily penalised.

(If you really cannot do the exam on 3rd November eg. you are away in the field, then let Michael know ASAP via email, but also via a post in the LMS forum below.)

The exam has two components:

* a general knowledge / theory component, testing general understanding of big data, modelling and meta-analysis gained from the seminars/lectures and set readings
* a practical component, evaluating your skills in data manipulation and interrogation via specific tasks requiring coding in R

Some parts of the exam will be relatively easy and some parts of the exam will be challenging for many of you. Stay calm and work through the questions one by one. If you don't know how to answer something, then don't panic, just leave it and come back to it later if you have time. Preparation is essential (see below). I suggest you quickly look through the whole exam at the start, to get an overall idea and so you can focus on the easiest questions first.

**You must complete and submit the general knowledge / theory component *before*you can start the practical component.** But you can attempt each of the components whenever you like within the 24 hour period (eg you could do the theory component in the morning and the practical in the evening).

**General knowledge / theory component:**  You will be allowed **2 hours** to complete this component. It will be worth 40% of the total mark. You can start this whenever you want, within the 24 hour period.

Suggested prep for this component is to make sure you can address all the questions listed under 'Course Topics' on the intro page for this unit under the three main headings - you should be able to address these based on the seminars and the readings suggested on the intro page, including the papers and the first few chapters in the Haefner book. You should also probably have notes from the lectures and readings (and/or the readings themselves) ready to refer to. Most questions on the exam will be **multiple choice** or **numerical**, but there will also be **short written answers**. There may also be questions related to lab material and coding, but these will not require you to do anything in R. Some of the questions will be relatively simple, and some will be more challenging. You will definitely want to be sure that you have a good internet connection and a reliable computer.

**Practical component:** You will be allowed 2.5 hours to complete this component. It will be worth 60% of the total mark. You can start this whenever you want, within the 24 hour period, as long as you have finished the other component.

In the practical part of the final assessment, you will be asked specific questions about specific datasets that you will need to answer using R. These data sets could include the species frequency data we used in Lab 2, the temperature button logging data we used in Lab 6, and/or the species abundance data sets used in Lab 7. You are strongly advised to revise the lab scripts for Labs 2, 6 and 7, and have the data sets for those labs ready to be loaded into R (or already loaded) and your modified versions of the lab scripts ready to be used and adapted as required.

* **You will also be asked questions about a new data set, which will be described and provided below a few days before the exam date**.
* You should read this description and familiarise yourself with the data before the exam.
* You are strongly advised to have this data loaded into R before the exam, and to have a pre-prepared script ready to investigate and interrogate the data.
* You are advised to practice applying the scripts from the labs to this new data set as much as possible, before you start your exam.
* Some of the questions will be relatively simple, and some will be more challenging, like in the scripts provided in the labs.
* You will definitely want to be sure that you have a good internet connection, and a reliable computer, and that R is working well on the computer that you are using.

**Example exam questions for the practical component:**

- Which of the 22 temperature buttons from the lab dataset had the highest proportion of time recording a temperature above 22?

- Are there any typo errors in species IDs in the new data set? If so, which species?

- Which species occurred at the least range of latitudes in the first survey in the new data set?

**Important!**I will allow multiple attempts for these quizzes, but unless there are extenuating circumstances it is only your first attempt that will count, and you should thus only have one attempt. Make sure you save your answers as you go, and do not submit your attempt until you are sure it is as good as it can be. Extra attempts are only allowed in case of technical problems. If you do have a temporary technical issue, you should be able to log back in and continue your current attempt. If you do have major technical problems and need to use an extra attempt, then you must start that attempt immediately and the total time for all attempts should be less than the total time allowed, or penalties may apply. If you're not sure about your computer or internet connection then you may wish to do the exam in the UWA computer labs. Make sure you follow all instructions about the correct format of submitting answers, including at least 3 decimal places of accuracy where relevant.

# Exam Data Set – Frequency Dataset

Typographical errors:

|  |  |
| --- | --- |
| **Should be (allfreq /dat1)** | **Incorrectly written (allfreq2/dat2)** |
| Cay.beq | Cay.biq |
| Yiz.vel | Yez.vel |
| Site562: 18.9282758249901 | Site562: 1.892828 |
|  |  |

Summary Stats

|  |  |
| --- | --- |
| No. quadrats | 6 (both years) |
| No. Species | 987 |
| No. Sites | 1056 |
| Latitude Range | 18.01905 – 27.99943 |
| Highest Site/ furthest from equator | site891 |
| Lowest Site/ closest to the equator | site232 |
| Highest Average Freq Year 1 (allfreq) | Peq.mek (3.148674) |
| Lowest Average Freq Year 1 | Bug.kiv (0.2528409) |
| Highest Average Freq Year 2 (allfreq2) | Nuh.yug & Bed.rum (3.113636) |
| Lowest Average Freq Year 2 (allfreq2) | Ved.luv (0) |
| **Species Richness Per Site** | |
| **Highest**: Year 1 | site368 |
| Lowest: Year 1 | site207 |
| **Highest**: Year 2 | Site365 |
| Lowest: Year 2 | Site201 |
| **Year 1** | |
| Most sp. In at least 3 quadrats | Site765 with 217 species |
| Most sp in at least 2 quadrats | Site765 with 269 species |
| Most sp in at least 5 quadrats | Site765 with 147 species |
| Species found at most sites | Hol.huy at 569 sites |
| Species found at greatest geographical range | 46 species |
| Species found at smallest geographical range | Tef.qex (6.435346 degrees) |
| Tef.qex range | 27.9879 – 21.55255 |
| **Year 2** | |
| Most sp. In at least 3 quadrats | Site 272 with 209 species |
| Most sp in at least 2 quadrats | Site 765 with 273 species |
| Most sp in at least 5 quadrats | Sites 272 and 369 with 132 species each |
| Species found at most sites | Gok.tac at 565 sites |
| Species found at greatest geographical range | 49 species |
| Species found at smallest geographical range | (Ved.luv now extinct) next lowest is Gar.kub 7.059504 |
| Gar.kub range | 25.11165 – 18.05214 |

Course Topics

# **Data**

## What is ‘big data’?

* Data that exceeds the analytical capabilities of individuals or disciplines
* Or ‘Four V’s – volume, variety, veracity, and velocity
  + Volume: huge variation
  + Variety: limits the scalability of ecological science
  + Velocity – speed of new data coming in?
  + Veracity: how much the data can be trusted?

## What kind of big data sets might we find that are relevant to biology, conservation, ecology, restoration, and environment.

## Why is big data such a trendy term?

## What are the differences between big data sets and ‘traditional’ data sets derived from a controlled experiment?

## What are some of the concerns or cautions or issues with big data sets?

## How do I decide what kind of data I need to address my scientific question?

## Where and how can we find the data?

## What can I do if I can’t find the data?

## Once I’ve got the data, how do I address the question and obtain results?

# **Meta-analysis**

## What is meta-analysis?

## What is a systematic review?

## How does a systematic review differ to a literature review or a scoping review?

## Why would we do a meta-analysis and/or systematic review?

## How do we get the data for a meta-analysis?

## What kind of data do we need for a meta-analysis?

## What kind of issues should be accounted for in meta-analysis related to confidence and variation?

## What kind of issues should be accounted for in meta-analysis related to independence?

## What kind of issues should be accounted for in meta-analysis related to publication bias?

# **Scientific Modelling**

## What is a model? What does ‘model’ mean?

## Why do modelling? What are the goals or purposes of modelling? How do these depend on our field of research?

## What types of formal models are there? How can models be classified?

## Empirical, mechanistic, process-based, descriptive, ontological, phenomenological, static, dynamic, stochastic, deterministic, space & time: continuous or discrete, DEs/FDEs/PDEs, compartment/transport/particle/individual-based/agent-based, finite state automata/cellular automata/state-transition, Eulerian/ Lagrangian...)

## Model as hypothesis – the classical view (Popper) versus the strong inference view (Platt).

## What is involved in the modelling process?

## What is the relationship between modelling and empirical data? How does modelling fit into empirical data collection and how does data fit into modelling?

## What trade-offs are involved in modelling? (Levins’ triangle 1966, Orzack and Sober 1993, Levins 1993)

## What is ‘individual-based modelling’ or ‘agent-based’ modelling? What advantages/disadvantages do these have compared to other types of modelling?

## How have population dynamics growth models been adapted to take into account age-structure of populations; more than one interacting species with competition, predation etc; density-dependence; stochasticity etc?

## Systems analysis, conceptual modelling, qualitative model formulation

## Turning the concepts into formulas and/or code: quantitative model formulation

## Getting values for parameters – calibration/parameterisation

## Simulation and computational/numerical issues

## How do we evaluate a model? Validation and Testing

## Model analysis: Sensitivity Analysis

## Model analysis: Analysis of Dynamics and Stability

## Models for Management Decision Support

## Optimisation – for calibration, analysis, or... optimisation

## Stochasticity and Uncertainty

# Essential Reading

**Modeling biological systems: principles and applications / James W. Haefner**

Available online through library, hard copy in library as well – this is a key reference for the course. Make sure you read the first few chapters, and other chapters that appear relevant.

**Environmental Modelling: Finding Simplicity in Complexity / John Wainwright, Mark Mulligan**

Scientific modelling from a more environmental perspective (to balance the more biological perspectives in most of the other texts here)

**Introduction to Practical Linear Programming. / David Pannell – Hardcopy BJM Library**

Particularly useful for module on linear programming

**Dynamic Models in Biology / Stephen P. Ellner**

Available online through library – this is a good reference for the course, especially the first, second, second last and last chapters.

[**A Biologist's Guide to Mathematical Modeling in Ecology and Evolution**](http://www.amazon.com/Biologists-Mathematical-Modeling-Ecology-Evolution/dp/0691123446/ref=sr_1_1?s=books&ie=UTF8&qid=1342535430&sr=1-1&keywords=biologists+guide+mathematical+biology)**/**[**Sarah P. Otto**](http://www.amazon.com/Sarah-P.-Otto/e/B001HCYPRE/ref=sr_ntt_srch_lnk_1?qid=1342535430&sr=1-1)**and Troy Day**

Available online through library, hard copy in library as well – a good reference for this course, especially the first two chapters and the epilogue.

**An introduction to ecological modelling : putting practice into theory / Michael Gillman, Rosemary Hails**

Hard copy in library – useful reference for the course.

**A practical guide to ecological modelling : using R as a simulation platform / Karline Soetaert and Peter M.J. Herman**

Available online through library, hard copy in library as well – shows how to use the R software environment and programming language for simulation modelling.

**Ecological models and data in R / Benjamin M. Bolker**

Available online through library, hard copy in library as well – mostly focuses on statistical modelling but has a chapter on dynamic models as well.

**Ecological Modelling: An Introduction /**[**Jorgensen, S.E**](http://catalogue.library.uwa.edu.au/search%7ES1?/aJorgensen%2C+S.E/ajorgensen+s+e/-3,-1,0,B/browse)

Available online through library

**Modelling Complex Ecological Dynamics: An Introduction into Ecological Modelling for Students, Teachers & Scientists / edited by Fred Jopp, Hauke Reuter, Broder Breckling**

Available online through library

**Systems ecology: an introduction to ecological modelling / R.L. Kitching**

Hard copy in library

**Modelling For Field Biologists and Other Interesting People /** Hanna Kokko

Available online through library

**Meta-analysis in ecology** – Gurevitch et al

**TO DO:**

**~~Continue revising Lab 6 & Lab 7 – working through the challenges~~**

**Make notes for papers above & from library**

**Answer each course question**