--- Chapter One ---

Introduction

The aim of this workbook is to introduce biologists to the practical elements of creating graphs and maps using R statistical software. This means it is primarily aimed at undergraduate and postgraduate students who either wish to teach themselves how to create high quality and informative data visualisations in R or who are taking their first courses in how to use R. However, it will be just as useful for more experienced biologists who currently use other data visualisation software packages, but wish to learn how to use R to make effective and informative graphs and maps.

This workbook uses the same task-oriented learning (TOL) approach found in other books in the *PSLS* series, such as *An Introduction to Basic Statistics for Biologists using R*. The TOL Approach helps you learn how to carry out the types of tasks biologists need to be able to do on a regular basis in a practical and meaningful way without getting too tangled up in learning about the underlying theoretical basis for them. This workbook, therefore, does not aim to provide you with information about the principles behind data processing and visualisation. Instead, it focuses on building up practical experience and providing advice about how to create the types of graphs and maps that biologists need to be able to make on a regular basis.

This practical experience and advice comes in the form of a series of exercises which you can work through to learn how to complete specific data visualisation tasks. This might be something simple, such as creating a basic frequency distribution histogram, or something more complicated, like creating a graph of time series data or an animated map from raster data sets. No matter what, for each task, you are provided with all the steps you need to undertake to complete it, starting with getting your data into R and finishing with how to present your data visualisation to others. These exercises use a workflow approach based around flow diagrams to help you understand exactly what you need to do at each step in the process, and where you are using the same basic steps to complete different tasks. This allows you to see how more complicated tasks can be carried out by connecting together simpler individual steps.

The exercises in this workbook are divided into five groups. These are: 1. Creating your first graphs from biological data in R using the GGPlot package (Chapter Three); 2. Creating graphs displaying groups of data with GGPlot (Chapter Four); 3. Creating graphs displaying individual data points with GGPlot (Chapter Five); 4. Creating other types of graphs not covered in Chapters Three to Five (Chapter Six); and 5. Creating maps from biological data in R (Chapter Seven). Together, these represent most of the key tasks that biologists need to be able to do to start creating high quality data visualisations using R in a practical and meaningful way.

R was selected as the basis for the instructions provided in this book because it is free to download and because it is widely used by biologists around the world. While it is a command-driven package, which some people initially find off-putting, it is relatively simple to use, and with the right type of instructions, it is relatively easy for anyone to successfully start using it for data processing and visualisation in a short space of time. However, while the exact instructions will vary for other graphing and data visualisation software, the same basic steps outlined in the boxes on the left hand side of the flow diagrams are required to complete the tasks outlined in this workbook in almost all of them. Thus, it should be relatively easy to adapt the instructions provided here for use with other similar software.

The exercises start with learning how to make a basic graph, such as a frequency distribution histogram, in R using a package called GGPlot (Exercise 1.1), before moving on to customising how your graph looks (Exercise 1.2) and creating multi-series and multi-panel graphs (Exercise 1.4). The basic processes required to make graphs in R using this package learned in these initial exercises are then applied to creating specific types of graphs and other data visualisations. These include as bar graphs (Exercises 2.1 to 2.3), line graphs (Exercises 2.3), scatter plots (Exercise 3.1 to 3.3), graphs of time series data (Exercises 3.4), pie charts (Exercise 4.1), bubble graphs (Exercises 4.2) and maps of biological data (Exercises 5.1 to 5.3). This allows you to build up your data visualisation skills in a logical order as you work through this book one chapter at a time. However, the instructions provided in each chapter are sufficiently complete that you can work through the exercises in each one on their own without having to refer back to the contents of previous chapters. This means if your main aim from reading this book is to work out how to do a specific task, such as making a bar graph or making a map of biological data, you can go straight to the relevant chapter for that task and find out how to complete it. If you do this, however, it

is worth taking the time at a later date to work through the earlier chapters too, as these will help widen your skills base and give you a better understanding of what you are doing in each individual step when creating more advanced data visualisations.

When you start using R you may find it rather frustrating, particularly if you are not already familiar with using command-driven software packages. This is because, rather than picking actions from a menu as you would do with a graphic user interface, you need to enter blocks of code before running them. To make things more complicated, the code in these blocks needs to be entered in a very precise manner, including using exactly the same uppercase and lowercase letters provided in the instructions for each exercise. This means that any typos you make will cause your code not to work, and R will not necessarily provide you with suggestions or indications as to what has gone wrong. Since the aim of this workbook is to help you get up and running with making biological data visualisations using R as quickly as possible, rather than testing your typing skills, you will find a text document called R_CODE_DATA_VISUALISATION_WORKBOOK.DOC in the compressed folder containing all the data required to complete these exercises (instructions for downloading this folder can be found at the start of Chapter Three). This document provides a copy of all the R code used in the flow diagrams that you will find at the start of each exercise. You can, therefore, choose to simply copy and paste the required code into R from this document without having to worry about making mistakes while typing it. Once you are familiar with how to complete a specific task, you can then work out how to modify this code to allow you to do other, related tasks, and you will be given the opportunity to do this as part of each exercise. To help with this, the blocks of code provided in the above document have been colour-coded. This not only makes it easier for you to work out what each different part does, it also makes it easier to work out which bits need be modified to make them do something different. If you wish to learn more about how to create R code to do specific tasks for yourself, we recommend reading *Getting Started with R: An Introduction for Biologists* by Andrew P. Beckerman, Dylan Z. Childs and Owen L. Petchey.

How The Exercises in This Workbook Are Structured:

The exercises in this workbook all follow a standard structure that has been developed to help you to understand what you need to do to complete a specific task in R, to gain

experience in doing it, and to help you work out how to apply it to your own data. First, you are provided with a brief introduction to the task itself and what you will learn by completing it. Next, you will find a flow diagram with all the information you need to work through an example of the task using a specific data set. Once you have worked through this initial example, you will find details of how you can modify the code you used to produce different results, as well as examples of such modifications that you can work through. This will help you gain a deeper understanding of how you can adapt a specific workflow to your own data. This approach of providing detailed step-by-step workflows, along with examples of increasing complexity for you to work through by modifying the commands it contains, means that you can use these exercises to rapidly and efficiently increase your graphing and data visualisation skills. In addition, it provides a resource you can refer back to any time you wish to refresh your knowledge of how to do a specific task.

Why Are Some Instructions And Steps Repeated In Different Exercises?

As you work through the chapters in this workbook, you will quickly notice that there are some instructions and steps that are repeated in many different exercises. If you are not already familiar with the task-oriented learning (TOL) approach used in this book, you may think this repetition is unnecessary. It is not, and it does, in fact, perform a number of important functions which will help you master the use of R for making graphs and maps. First and foremost, it reminds you that there are certain key steps which you need to do each and every time you wish to do a task in R. These include steps like setting your working directory, importing a data set and checking that it has been loaded into R correctly. By repeating them in each individual exercise, it not only helps you to become familiar with these basic, but important, steps, it also serves to reinforce the importance of including them in every workflow that you carry out. Secondly, by including the instructions for the same steps in multiple exercises, it enables you work through a specific task, such as making a bar graph or a map, from start to finish. This means you can concentrate on learning all the steps you need to do to complete that task without becoming distracted by having to refer back to other sections of the book. Finally, by including the same steps in the flow diagrams for different exercises, it helps you see how the same processes are used to create very different types of graphs and maps This makes it much easier to understand how you can create your own custom workflows for types of data visualisations not included in this workbook by using a similar series of steps.

NOTE: As with many things in life, there may be more than one way to do the processes required to complete the exercises outlined in this workbook. The instructions presented here will work for the associated data sets, and this means they should also work in most other circumstances. However, if you find an alternative way to do them which works for you, or if you have someone who can show you how to do them in another way, feel free to do them differently.