National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Preethi Belur Ramesh
Student ID: x20180101

School of Computing
National College of Ireland

Supervisor: Bharat Agarwal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Preethi Belur Ramesh
Student ID:	x20180101
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Bharat Agarwal
Submission Due Date:	15/12/2022
Project Title:	Configuration Manual
Word Count:	630
Page Count:	8

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	1st February 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Preethi Belur Ramesh
x20180101

1 Introduction

The configuration manual contains software and hardware configurations, libraries and highlights the important code snippets of each process carried out in the implementation. This documentation is aimed to help replicating the research project - *A Machine Learning Based Approach to Predict the Species-Habitat Relationship in Australia*

2 System Requirements

2.1 Hardware Requirements

The Hardware requirements is shown in Table 1,

Table 1: Hardware Requirements

Processor	11th Gen Intel(R) Core(TM) i5-11320H @ 3.20GHz 2.50 GHz
RAM	16.0 GB (15.7 GB usable)

2.2 Software Requirements

The R software is mainly used and Table 2, depicts all the R related software used in this research.

Table 2: Software and Versions

Software Type	**Software Name**	**Version**
IDE	RStudio	2021.09.2 Build 382
Programming Language	R	4.1.2
Web Framework	Rshiny	1.7.2
Notebook	Rmarkdown	2.17

2.2.1 Libraries and Packages

Table 3 depicts all the libraries/packages with versions that were installed for the research project.

Table 3: Libraries, Description and Versions

Library	Description	Version
Raster	Used to analyze and model geographic data	3.5-21
sdm	Species Distribution Modelling	1.1-8
sp	Contains classes and methods for spatial data	1.4-6
usdm	Uncertainity analysis for SDM	1.1-18
ggplot2	Data Visualization	3.3.5
dplyr	Data Manipulation	1.0.7

3 Research Implementation

3.1 Data Extraction

The Biodiversity datasets are obtained from GBIF[1]. To access the open source occurrence data, it is necessary to register and login to the user account (shown in Figure 1).

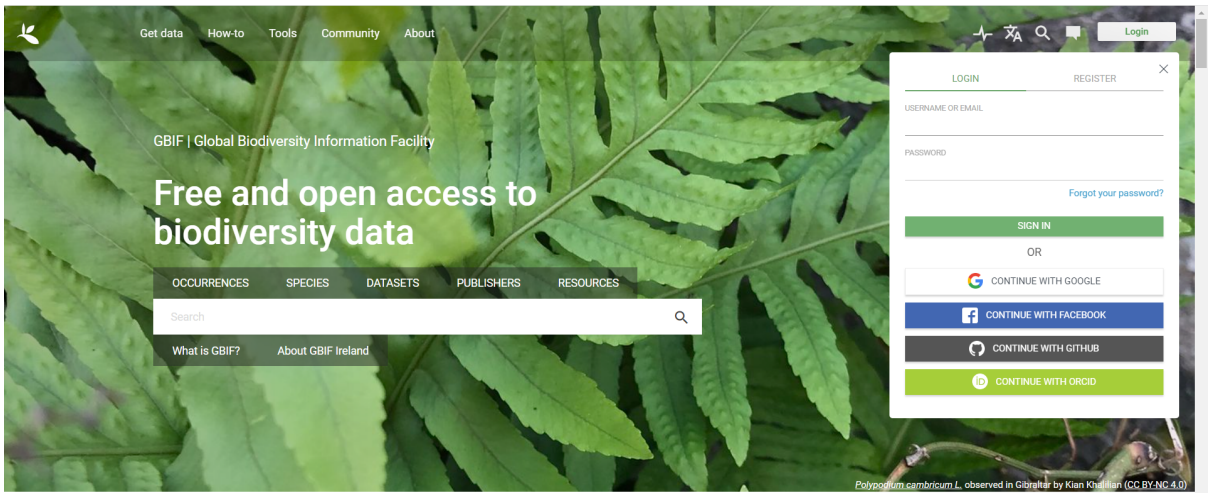
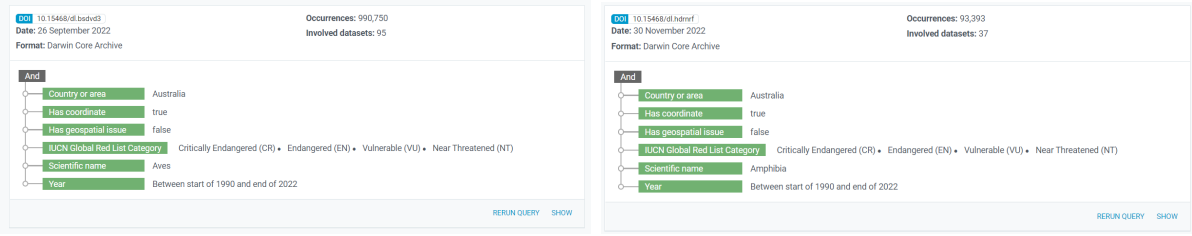


Figure 1: Accessing GBIF Account to Get Biodiversity Data

To access the occurrence data of Australian Birds and Frogs, queries are run with conditions depicted in Figures 2a and 2b.



(a) Query to Obtain Bird Data

(b) Query to Obtain Frog Data

Figure 2: Biodiversity Data Extraction

The queries are run using the Darwin Core Archive Format and are downloaded as a zip folder (Figure 3). The text file named *occurrence.txt* contains the biodiversity data which is opened using the code snippet in Figure 4

[1]https://www.gbif.org/

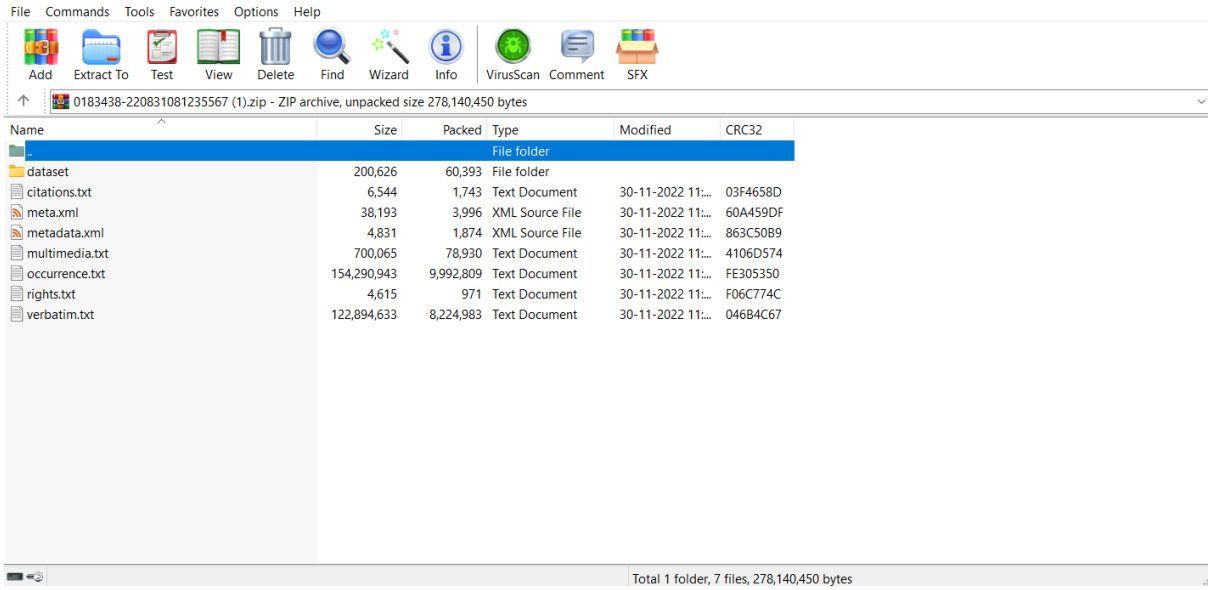


Figure 3: Biodiversity Data : Contents of the Zip Folder

```
## Obtaining the Biodiversity Data
```{r}
occurrence <- read.delim("~/occurrence.txt")
occurrencefrogs <- read.delim("~/occurrencefrogs.txt")
```
```

Figure 4: Code Snippet to Extract Data from the Text File

The present and future climate variables also called as bio-climatic variables are obtained from WorldClim[2]. The data is obtained programmatically using the code in Figure 5. The future climate data is also obtained in the similar fashion represented in Figure 6.

```
## Historic Climate Data
```{r}
myExpl<-raster::getData('worldclim', var='bio', res=5)

names(myExpl) <- c("Annual Mean Temp", "Mean Diurnal Range",
"Isothermality", "Temp Seasonality",
                   "Max Temp of Warmest", "Min Temp of Coldest", "Temp
Annual Range",
                   "Mean Temp of Wettest Quarter", "Mean Temp of Driest
Quarter",
                   "Mean Temp of Warmest Quarter", "Mean Temp of Coldest
Quarter",
                   "Annual Precip", "Precip of Wettest Month", "Precip of
Driest Month",
                   "Precip Seasonality", "Precip of Wettest Quarter",
"Precip of Driest Quarter",
                   "Precip of Warmest Quarter", "Precip of Coldest
Quarter")
```
```

Figure 5: Code Snippet to Obtain Historical Climate Data

[2]https://worldclim.org/

```
# Future Climate Data
```{r}
biof <- raster::getData('CMIP5',var="bio",res="5",rcp=85,model="AC",year=
70)
```
```

Figure 6: Code Snippet to Obtain Future Climate Data

3.2 Data Transformation

The biodiversity data are pre-processed by changing the datatypes of the coordinates column and removing erroneous records shown in Figure 7 and Figure 8.

```
## Transforming Spatial Data (Birds)
```{r}
typeof(occurrence$decimalLatitude)
occurrence$decimalLatitude <- as.double(occurrence$decimalLatitude)
occurrence$occurrenceStatus <- factor(occurrence$occurrenceStatus, levels
= c("PRESENT", "ABSENT"), labels = c(1, 0))
occurrence$occurrenceStatus
```

```{r}
# Check and exclude irrelevant record
occurrence <- subset(occurrence, !is.na(occurrence$decimalLongitude) &
!is.na(occurrence$decimalLatitude)
                     & !is.na(occurrence$individualCount) &
occurrence$individualCount>5)
```
```

Figure 7: Transforming Endotherms Data

```
## Transforming Spatial Data (Frogs)
```{r echo=TRUE}
typeof(occurrencefrogs$decimalLongitude)
occurrencefrogs$occurrenceStatus <-
factor(occurrencefrogs$occurrenceStatus, levels = c("PRESENT", "ABSENT"),
labels = c(1, 0))
occurrencefrogs$occurrenceStatus

occurrencefrogs <- subset(occurrencefrogs,
!is.na(occurrencefrogs$decimalLongitude) &
!is.na(occurrencefrogs$decimalLatitude)
                       & !is.na(occurrencefrogs$individualCount) &
occurrencefrogs$individualCount>5)
```
```

Figure 8: Transforming Ectotherms Data

The dataframes containing the birds and frogs occurrence data are converted into a spatial points dataframe (represented in Figures 9 and 10).

```
## Converting into Spatial Dataframe (Birds)
```{r}
occur <- birdsgeo1[,c('decimalLongitude','decimalLatitude' ,
'species','occurrenceStatus')]
coordinates(occur) <- ~decimalLongitude + decimalLatitude
```
```

Figure 9: Converting Birds Data into a Spatial Points DataFrame

```
## Converting into Spatial Dataframe (frogs)
```{r}
occurfrogs <- occurrencefrogs[,c('decimalLongitude','decimalLatitude' ,
'species','occurrenceStatus')]
coordinates(occurfrogs) <- ~decimalLongitude + decimalLatitude
```
```

Figure 10: Converting Frogs Data into a Spatial Points DataFrame

3.3 Models and Replication

The species distribution models are generated using *sdm* library (Naimi and Araújo; 2016) in R (Team et al.; 2013). Figure 11 shows the code snippet to create SDM object which is created for each species group including parameters such as predictors, dependent variable, and pseudo-absences.

```
## Creation of SDM Object
```{r}
dataobject <- sdmData(occurrenceStatus~., data_sp, predictors = bioc,bg =
list(method='gRandom', n=15))
dataobject
```
```

Figure 11: Creation of SDM Data Object

Figure 12 shows the code snippet used to create models using tree-based ML methods, Random Forest (RF) and Boosted Regression Trees (BRT). The replication of models using 5-fold cross-validation having training and dependent test data maintained at 4:1 ratio.

```
##Models generated using RF and BRT
```{r}
mod1 <- sdm(occurrenceStatus~.,d,
methods=c('brt','rf'),replication=c("cv"),cv.folds = 5,test.p=20,n=3)
mod1
```
```

Figure 12: Models generated using RF and BRT

To obtain the response curves that indicate the climate variables mainly responsible for habitat changes are obtained by generating a model that combines all species. Figure 13 show the code snippet of the combined model and response curves.

```
## Response Curves
```{r}
rcurve(combined_model, id=1:30)
```
```

Figure 13: Code snippet to capture response curves

3.4 Evaluation and Visualization

The Visualization plots are generated to discern the results better. Figure 14 depicts the code snippet to generate lollipop graphs that is used to illustrate the TSS performance across all species group.

```
```{r}
library(ggplot2)

# Create data
data <- data.frame(
  Species=c("Amytornis","Calyptorhynchus","Neophema","Aphelocephala","Dasyornis","Zand
a","Acanthiza","Melithreptus","Oreoscopus","Xanthomyza","Stagonopleura","Sericornis","
Callocephalon","Atrichornis","Lathamus"),
  TSS=c(0.81,0.92,0.9,0.78,1,0.93,1,0.99,1,0.98,0.9,1,0.95,1,0.93)
)

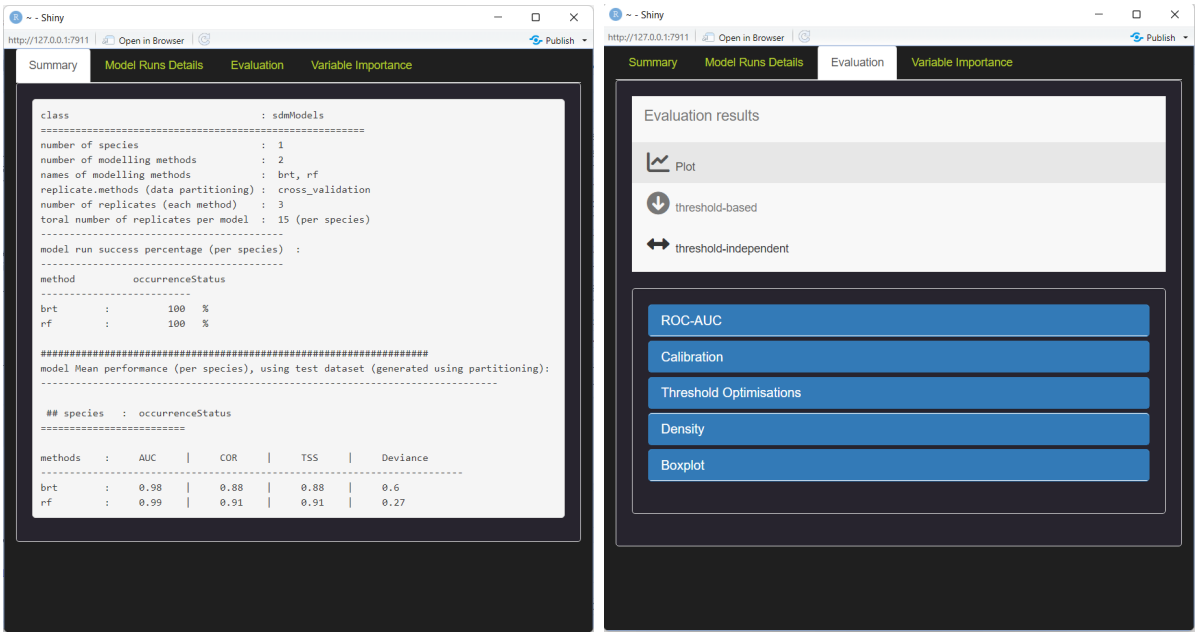
# Horizontal version
p1<- ggplot(data, aes(x=Species, y=TSS)) +
  geom_segment( aes(x=Species, xend=Species, y=0.6, yend=TSS), color="#d15c00") +
  geom_point( color="#d15c00", size=4, alpha=0.6) +
  theme_light() +
  coord_flip() +
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank(),
    panel.background = element_rect(fill="#ebebeb"),
    plot.background = element_rect(fill="#ebebeb")
    #axis.title.x = element_text(color="#d15c00"),
    #axis.title.y = element_text(color="#d15c00"),
    #plot.title = "True Skill Statistics - RF"
  )
p1 + ggtitle("True Skill Statistic - RF") + theme(plot.title = element_text(hjust =
0.5))
```
```

Figure 14: Code Snippet to Generate Lollipop Plot to Show RF Performance

Figures 15a, 15b represent the Rshiny framework used to display different evaluations that are threshold-dependent and independent in a user friendly manner. This is a built-in feature of the *sdm* package.

3.5 Forecast Species Distribution Using Ensemble Technique

Figure 16 represents the code snippet to forecast the species distribution based on present and future climate scenarios. Ensemble modelling techniques are included to project the distributions using weighted evaluation methods. A combined model containing models of all species is used to generate the response curves for both endotherms and ectotherms.



(a) Rshiny: Summary of Model Performance (b) Rshiny: Various Evaluation Results

Figure 15: Graphical User Interface in sdm package

```
## Species Distribution forecast
```{r}
pc <- predict(mc,bioc,filename="pc.img")
plot(pc)
pred1 <- ensemble(mc,pc,filename='enc.img',setting=list(method='weighted',stat='tss
',opt=2))
pred2 <- ensemble(mc,pc,filename='enc1.img',setting=list(method='weighted',stat='au
c'))

plot(pred1)
plot(pred2)
```
```

```
```{r}
biof <- raster::getData('CMIP5',var="bio",res="5",rcp=45,model="AC",year=70)

bioff <- crop(biof,extent(112.9211,159.1092,-55.11694,-9.142176))
plot(bioff)
names(bioff) <- c("bio1", "bio2", "bio3", "bio4",
                  "bio5", "bio6", "bio7",
                  "bio8", "bio9",
                  "bio10", "bio11",
                  "bio12", "bio13", "bio14",
                  "bio15", "bio16", "bio17",
                  "bio18", "bio19")
pf <- predict(mc,bioff,filename="pff.img")
pred3 <- ensemble(mc,pf,'encc2.img',
                  setting=list(method='weighted',stat='tss',opt=2))
plot(pred3)

pred4 <- ensemble(mc,pf,'enc3.img',
                  setting=list(method='weighted',stat='AUC'))
plot(pred4)

```
```

Figure 16: Forecasting the Species Distribution

References

Naimi, B. and Araújo, M. B. (2016). sdm: a reproducible and extensible r platform for species distribution modelling., *Ecography, John Wiley Sons, Ltd* **39**: 368–375.

Team, R. C. et al. (2013). R: A language and environment for statistical computing.