

R data exploration, wrangling, Github
Dr Simon Dedman
2026-01-28
FIU

University of Southampton

Oceanography with
Marine Biology BSc.

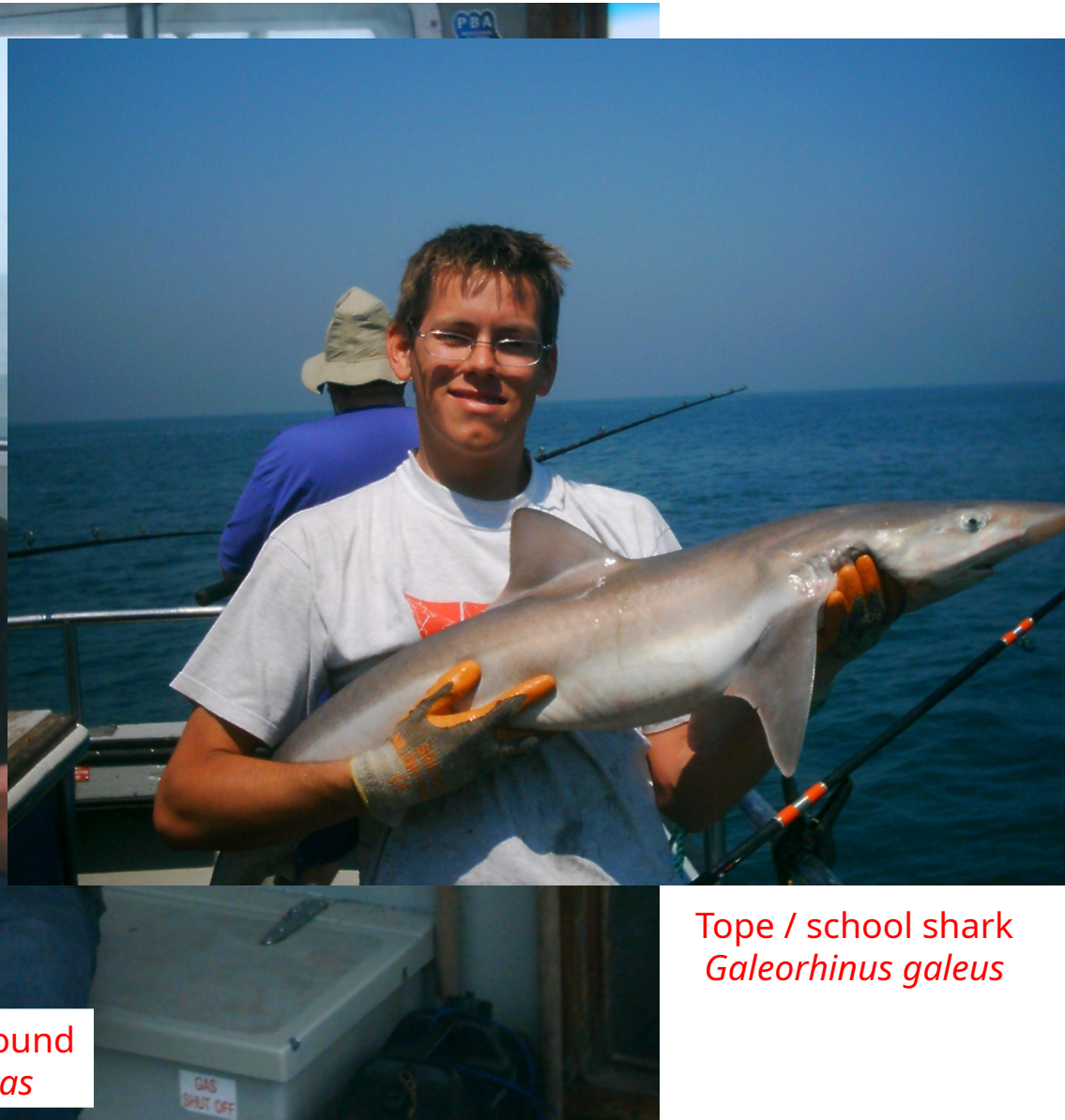
2001-2004



2003 Bachelors thesis

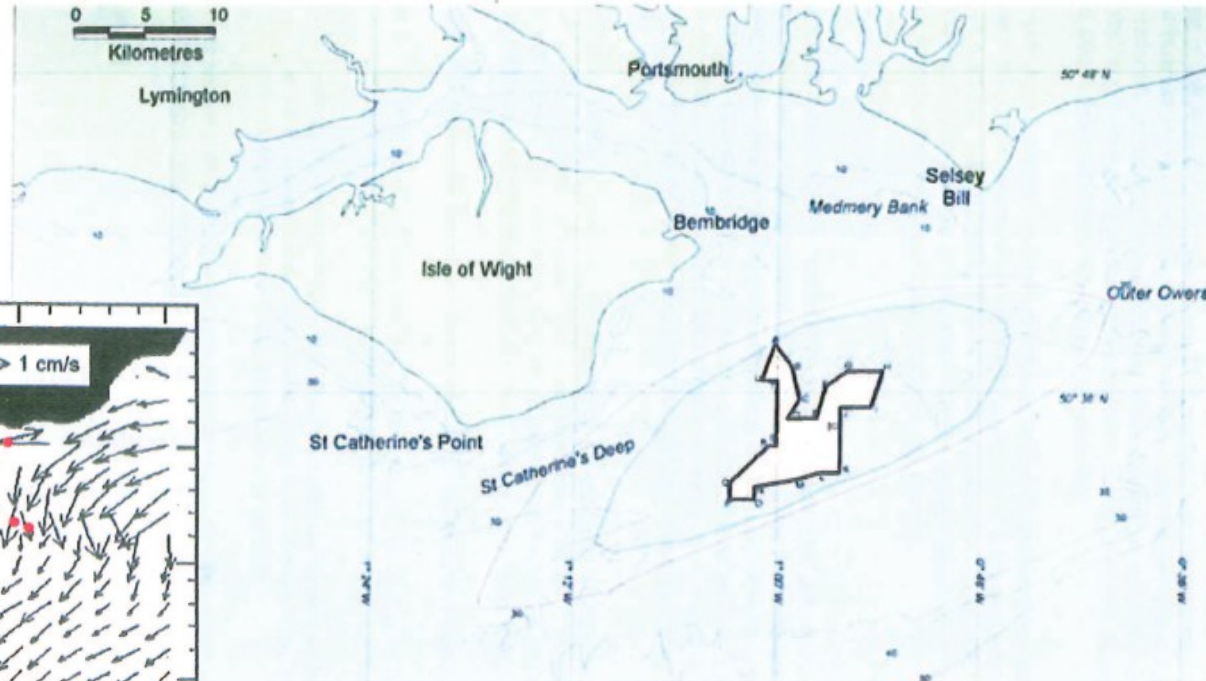
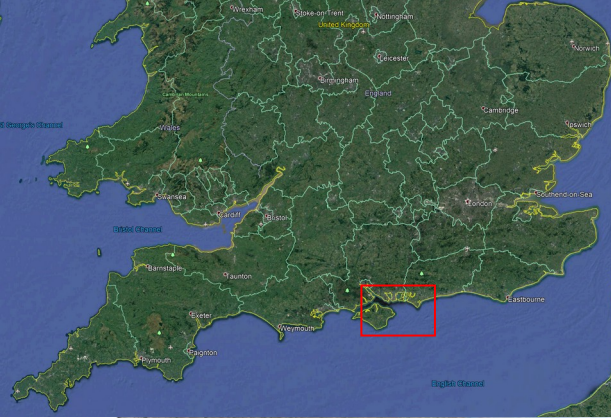


Starry smooth-hound
Mustelus asterias

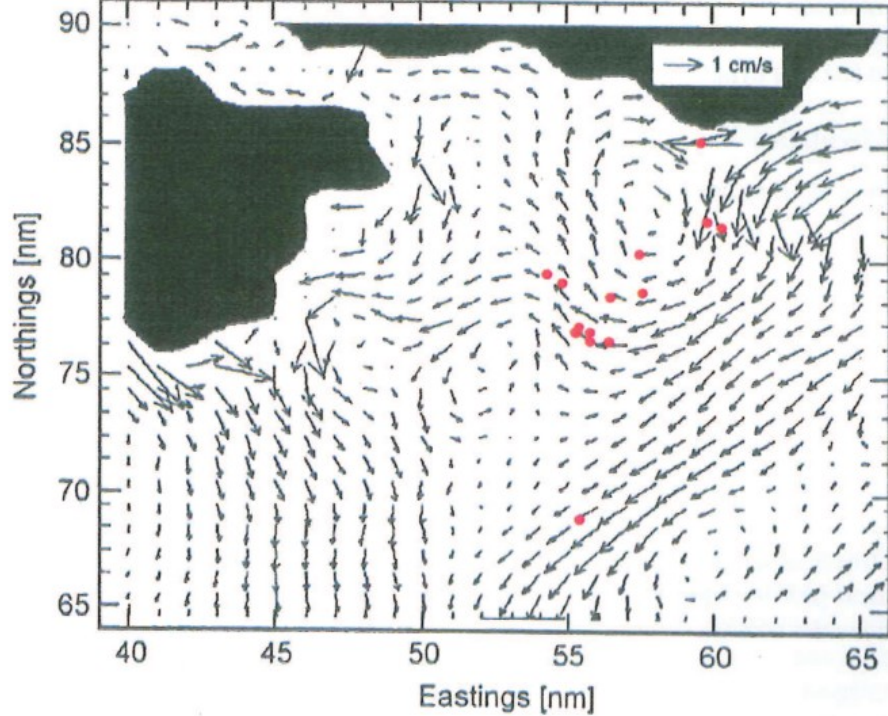


Top / school shark
Galeorhinus galeus

Elasmobranchs of the East Solent



Gravel extraction area



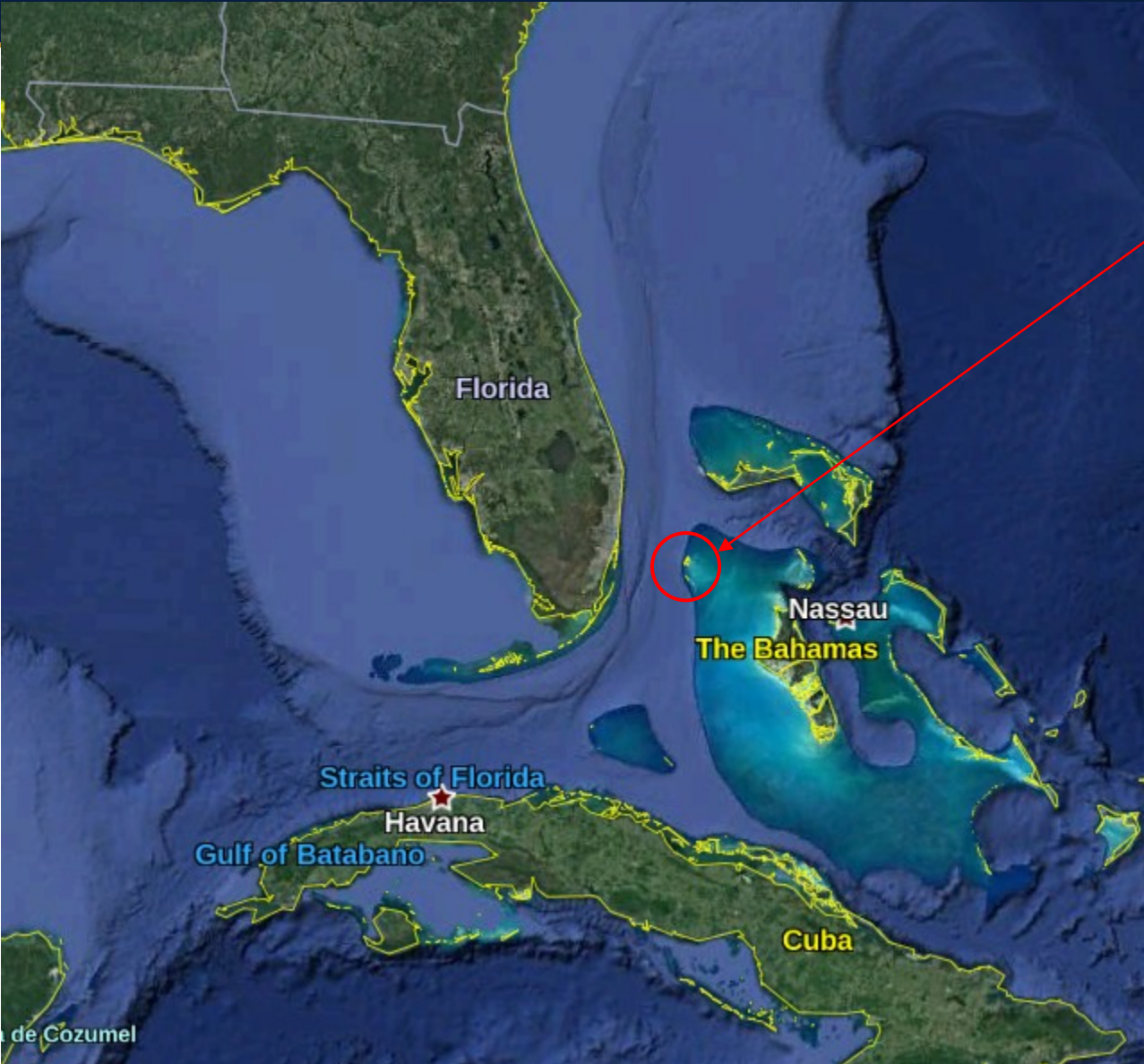
Currents + study sites

University of Aberdeen

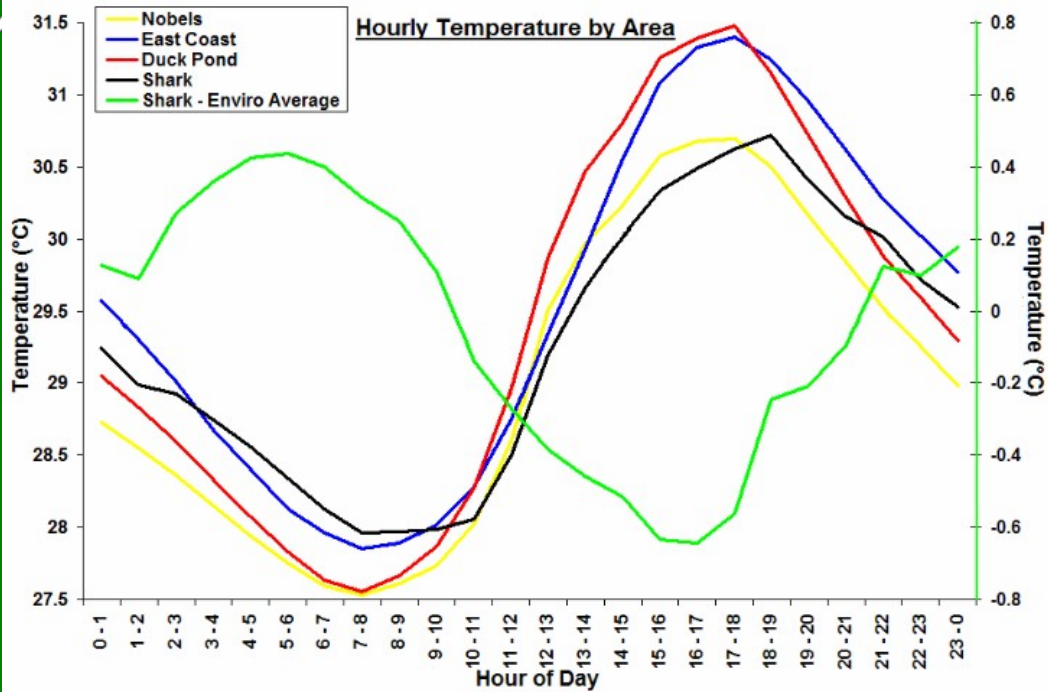
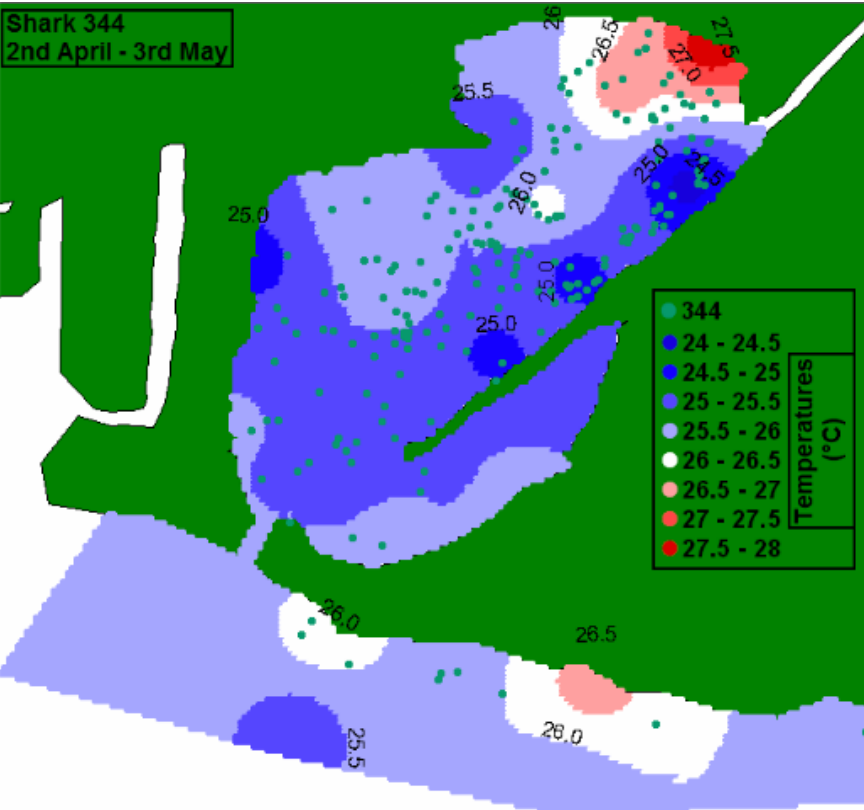
Marine & Fisheries Science MRes.
2005



Bimini, Bahamas, 2006

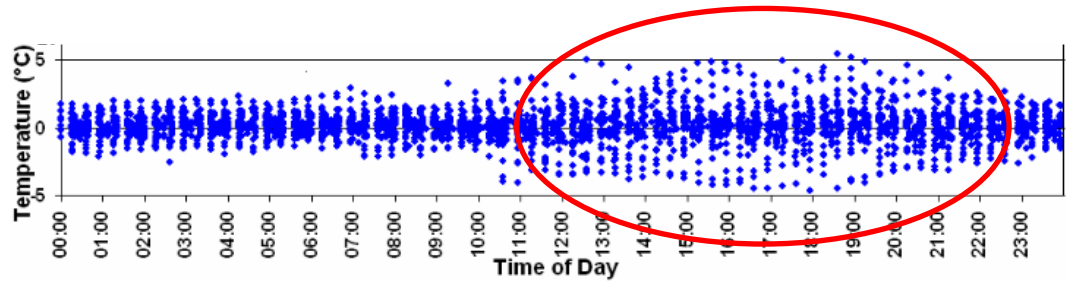


Shark 344
2nd April - 3rd May

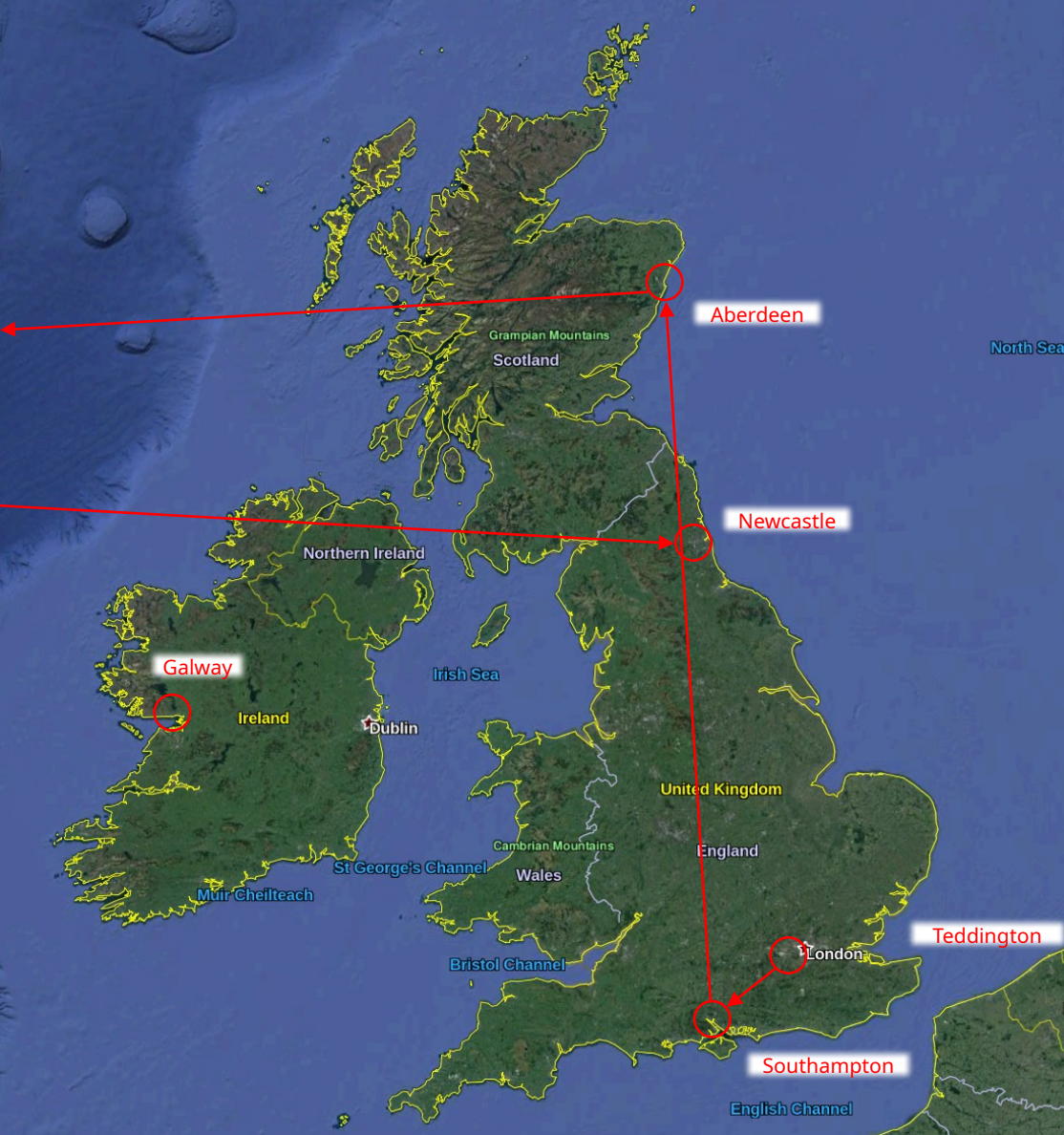


Temperatures: areas, shark, difference

Environmental temperature & shark 344 positions



Shark/environment difference



- Programme Manager, National Institute for Health Research

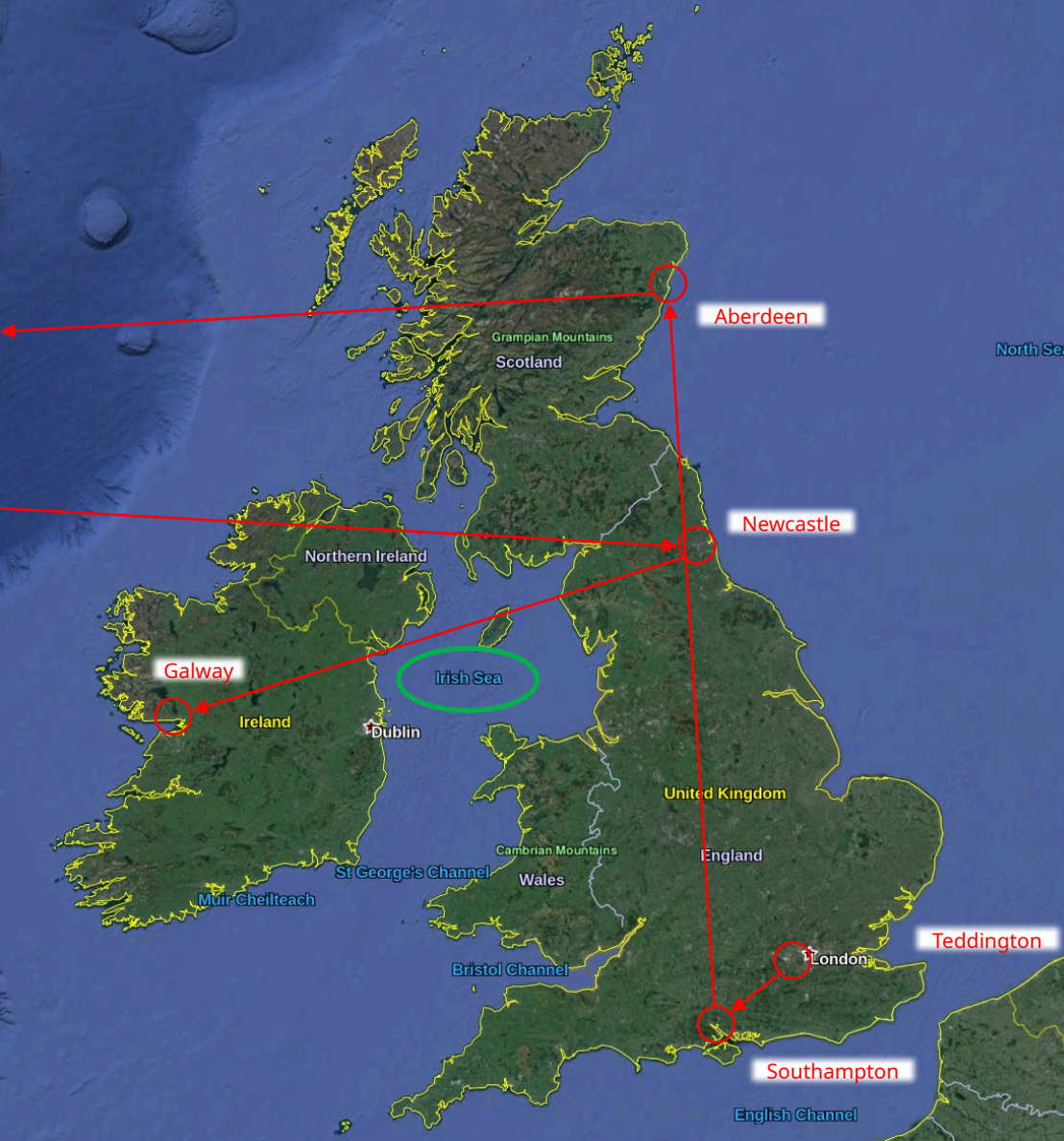
Teddington; 2006-2009

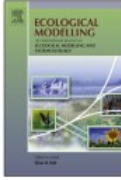
- International Quota Manager, Marine Management Organisation

Newcastle/London; 2009-2012

Galway-Mayo Institute of Technology

2012-15





- 4 Irish sea skate/ray species managed as assemblage – not ideal.
- Differentially vulnerable.
- Spatial management recommended but important habitats unknown.
- Boosted Regression Trees had various advantages over Marxan, MaxEnt, GLMs, GAMs – seemed to be the best approach for data-poor species with presence/absence and abundance data.

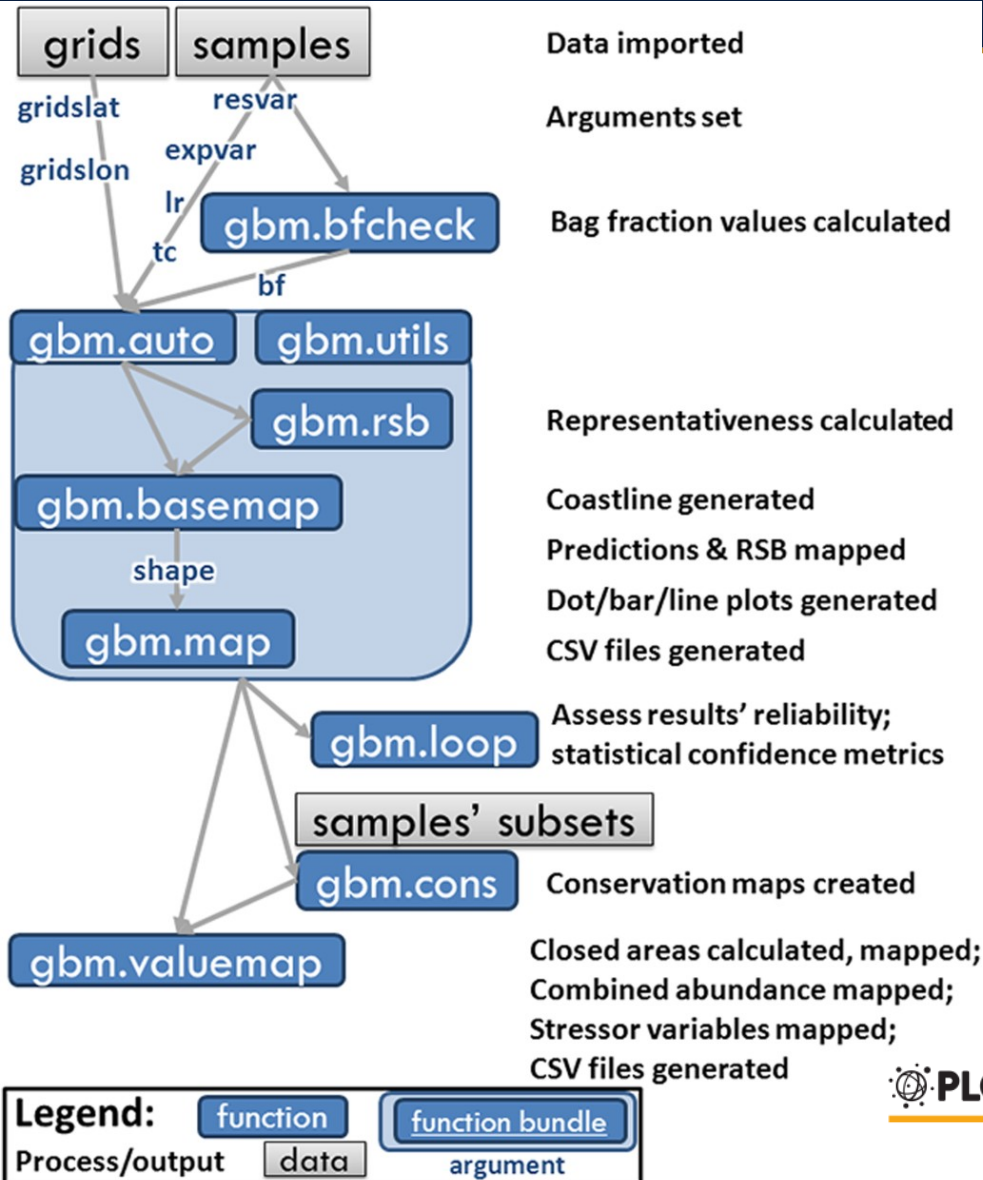
Modelling abundance hotspots for data-poor Irish Sea rays

Simon Dedman^{a,b,*}, Rick Officer^a, Deirdre Brophy^a, Maurice Clarke^b, David G. Reid^b

^a Marine and Freshwater Research Centre, Galway-Mayo Institute of Technology, Galway, Ireland

^b Marine Institute, Rinville, Oranmore, Co., Galway, Ireland





Data imported

Arguments set

Bag fraction values calculated

Representativeness calculated

Coastline generated

Predictions & RSB mapped

Dot/bar/line plots generated

CSV files generated

Assess results' reliability; statistical confidence metrics

Conservation maps created

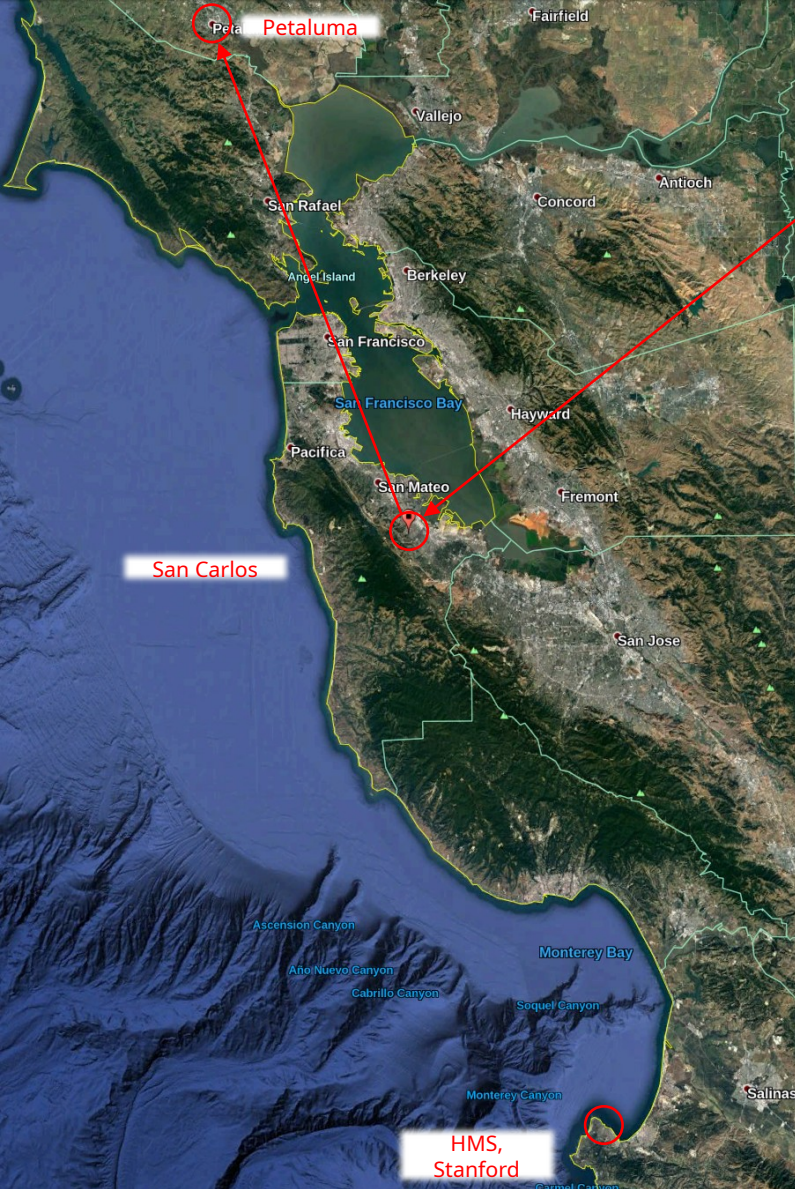
Closed areas calculated, mapped;
Combined abundance mapped;
Stressor variables mapped;
CSV files generated

Gbm.auto: A software tool to simplify spatial modelling and Marine Protected Area planning

Simon Dedman^{1,2*}, Rick Officer¹, Maurice Clarke², David G. Reid², Deirdre Brophy¹

¹ Marine and Freshwater Research Centre, Galway-Mayo Institute of Technology, Galway, Ireland, ² Marine Institute, Rinville, Oranmore, Co. Galway, Ireland






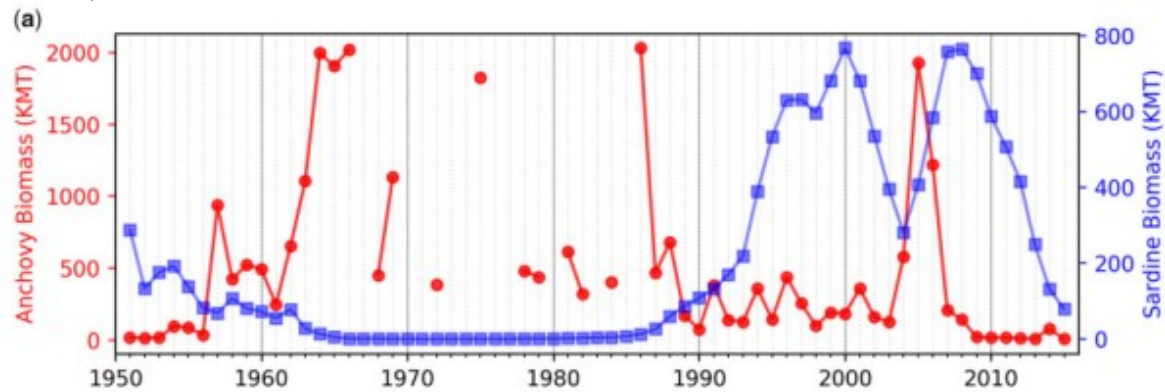
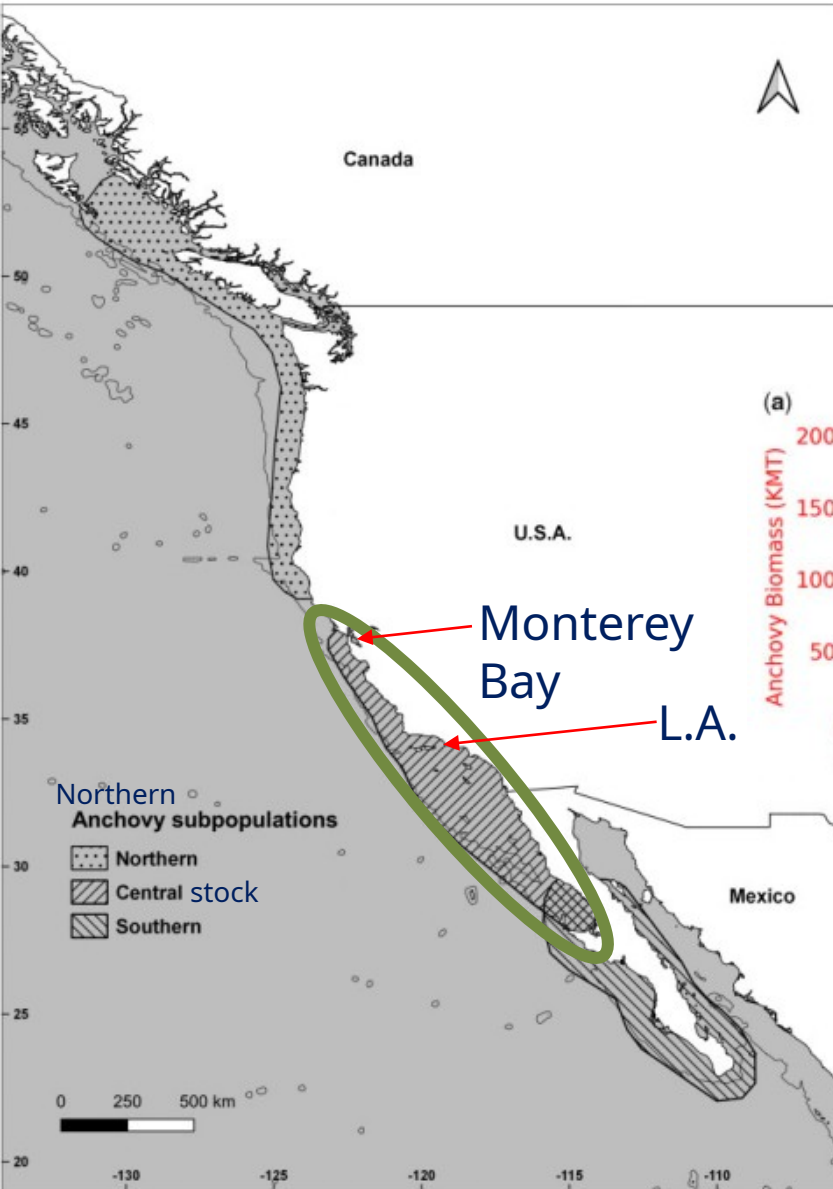


- FishSource (Sustainable Fisheries Partnership, 2017)
- Farallon Institute postdoc (2017-2019)

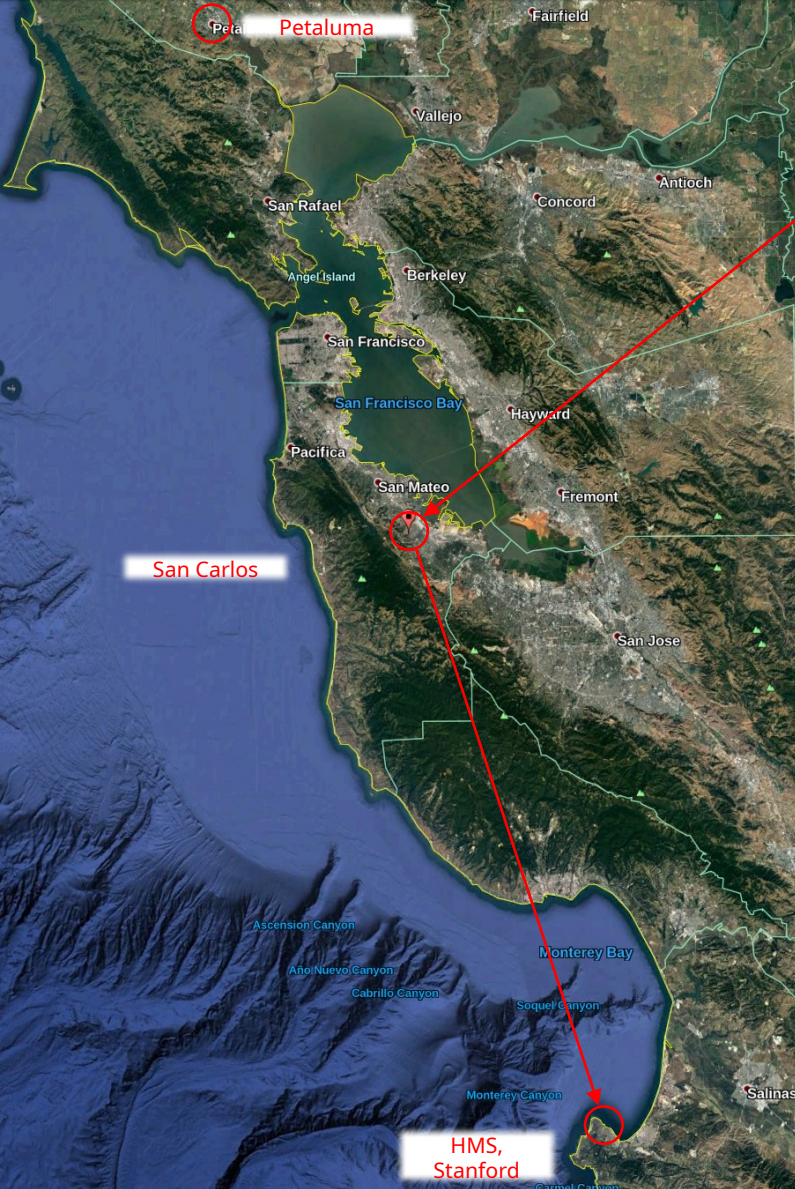
Food for Thought

Sixty-five years of northern anchovy population studies in the southern California Current: a review and suggestion for sensible management

William J. Sydeman *, Simon Dedman , Marisol García-Reyes, Sarah Ann Thompson, Julie A. Thayer, Andrew Bakun, and Alec D. MacCall 

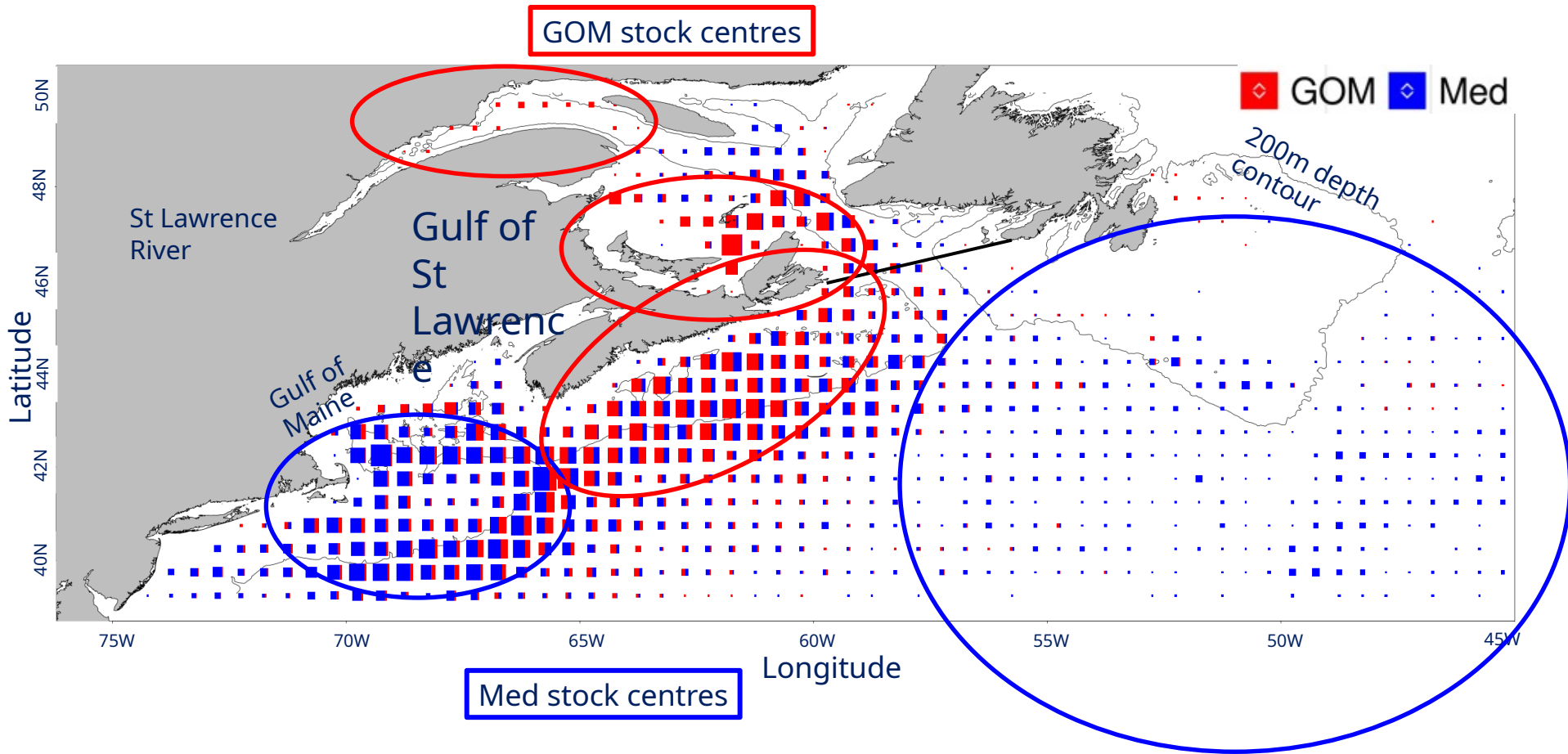


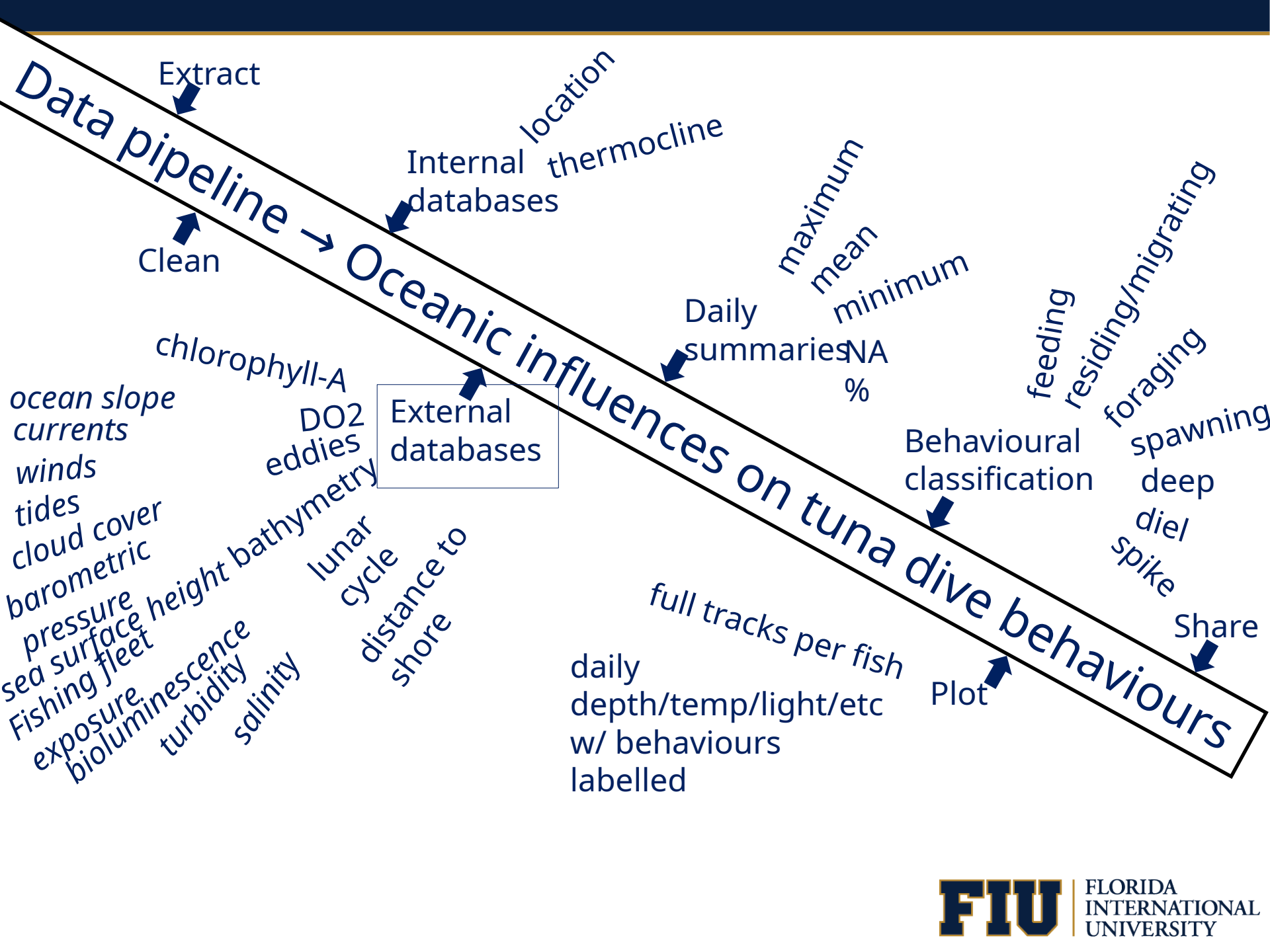
CSNA: central stock of northern anchovy



Blocklab, Hopkins Marine Station, Stanford University

Postdoc (2019-2022)





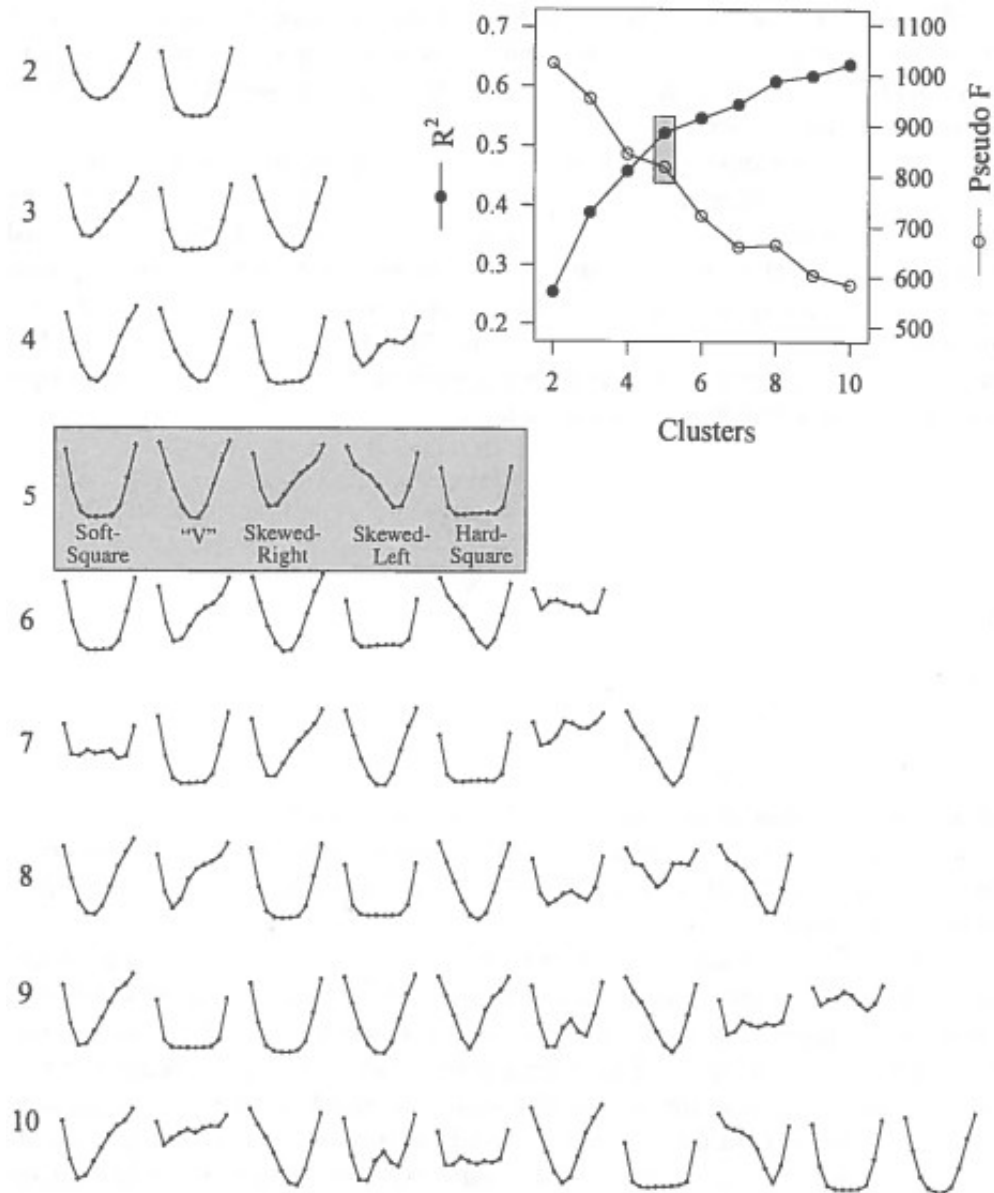
Defining Dive Behaviours

- Manual approach
 - Laborious / slow = small data / 'proof of concept'
 - Arbitrary / contentious / vague
 - Circular

- Algorithmic approach
 - Very quick = big data
 - Slow/complex to build
 - Universal, citable definitions

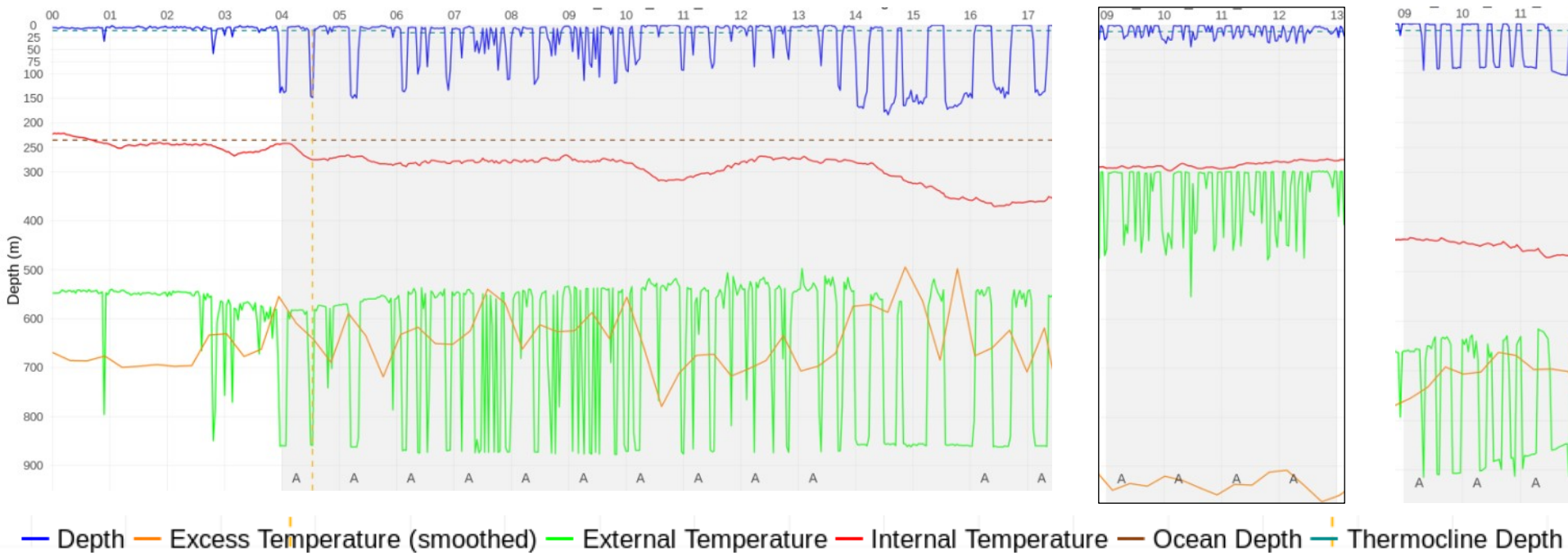
Classification of Dive Profiles: A Comparison of Statistical Clustering Techniques and Unsupervised Artificial Neural Networks

Jason F. SCHREER, R. J. O'HARA HINES, and Kit M. KOVACS

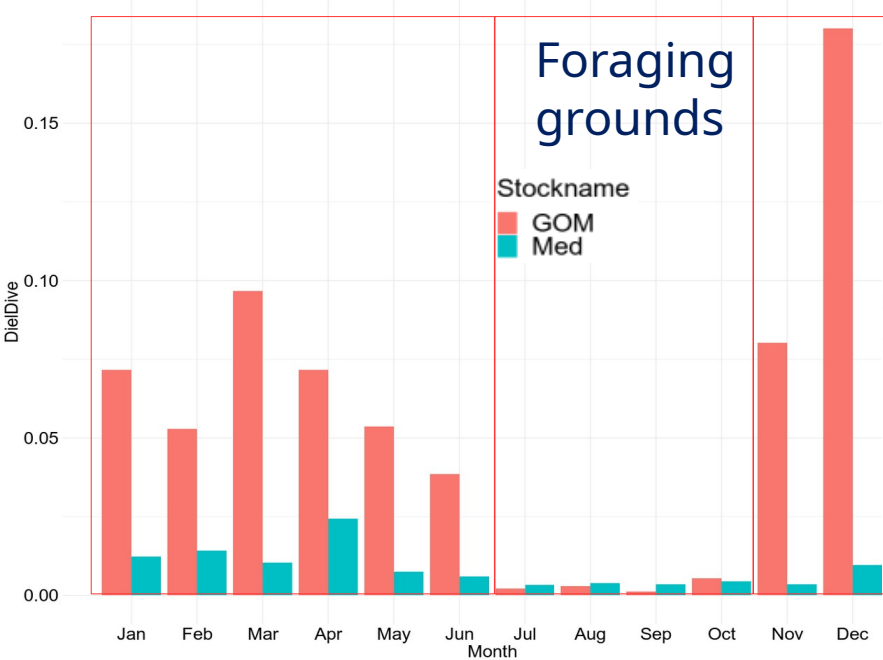


Defining Dive Behaviours: Forage A

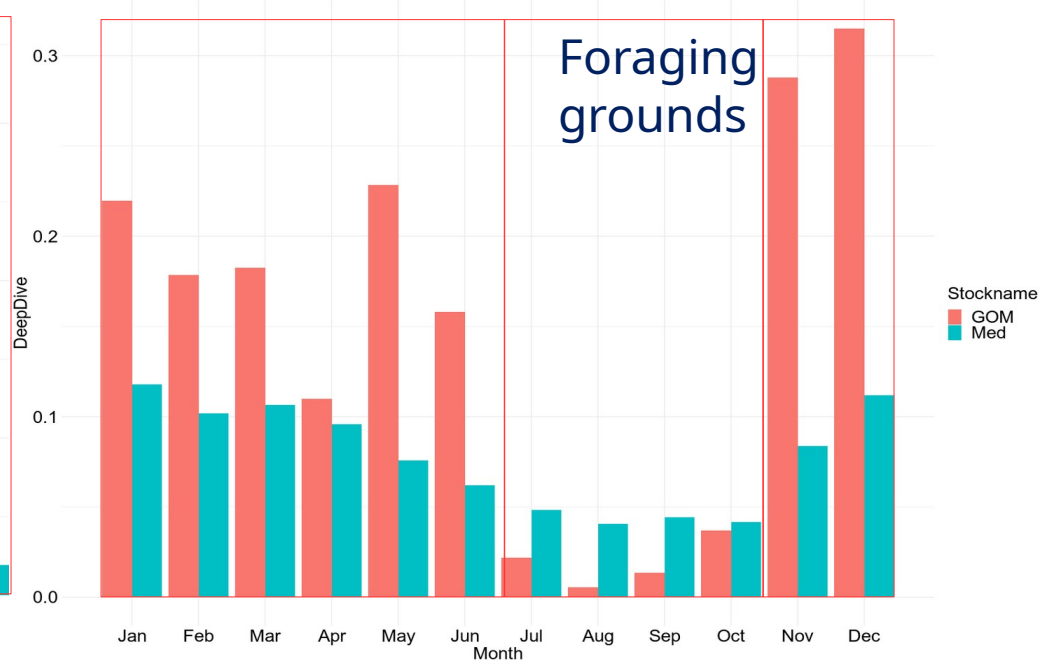
V aka Bounce A: ~100m depth, steep thermocline



Month by DielDive



Month by DeepDive

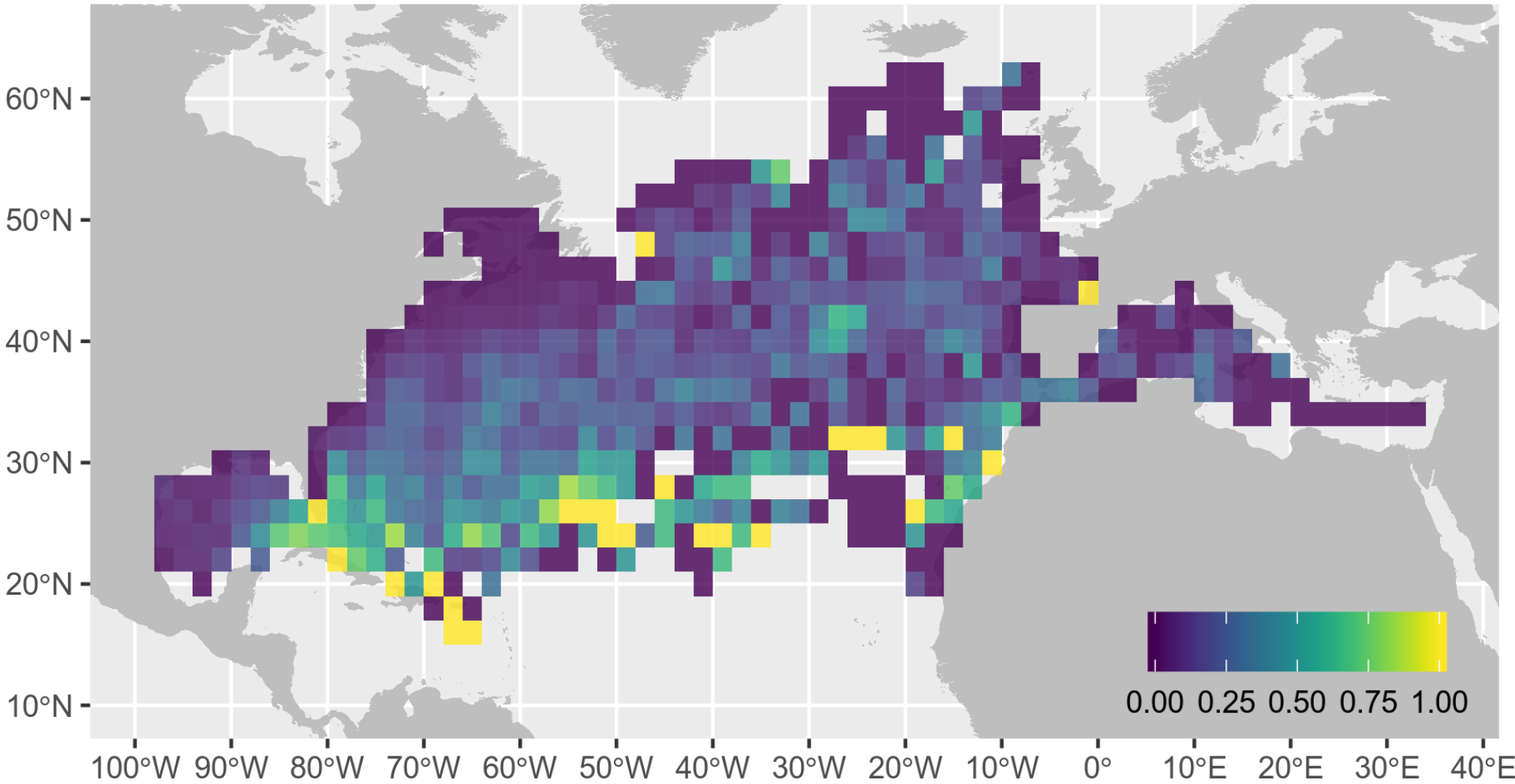


Lots of deep diving.

Hardly any.

DeepDive

Rasterised into 2° cells, aggregation function: mean



Florida International University

Mike Heithaus & Yannis Papastamatiou

Postdoc (2022)

San Carlos

HMS, Stanford

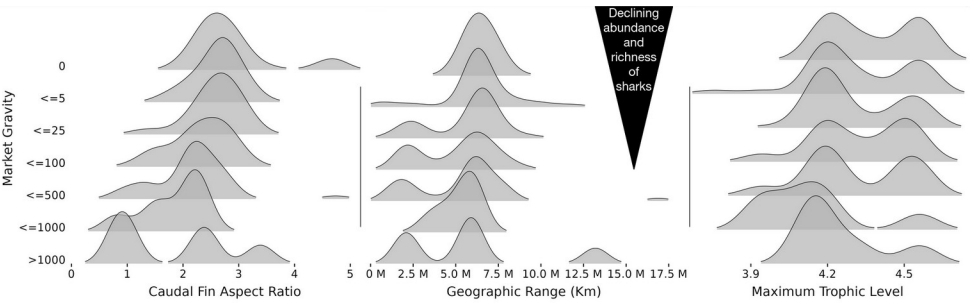
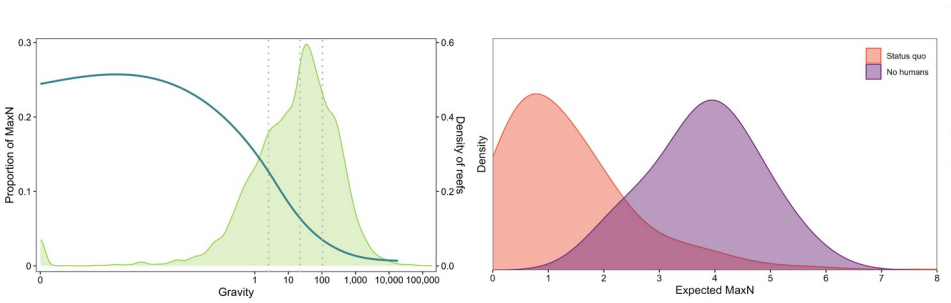
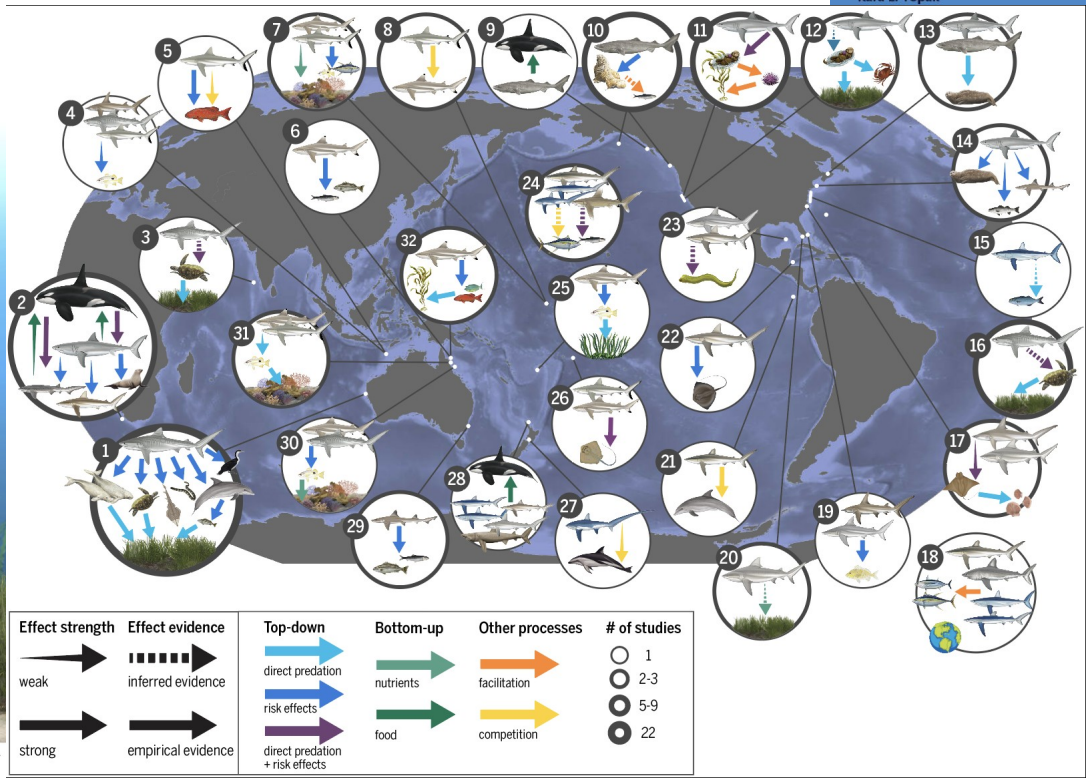
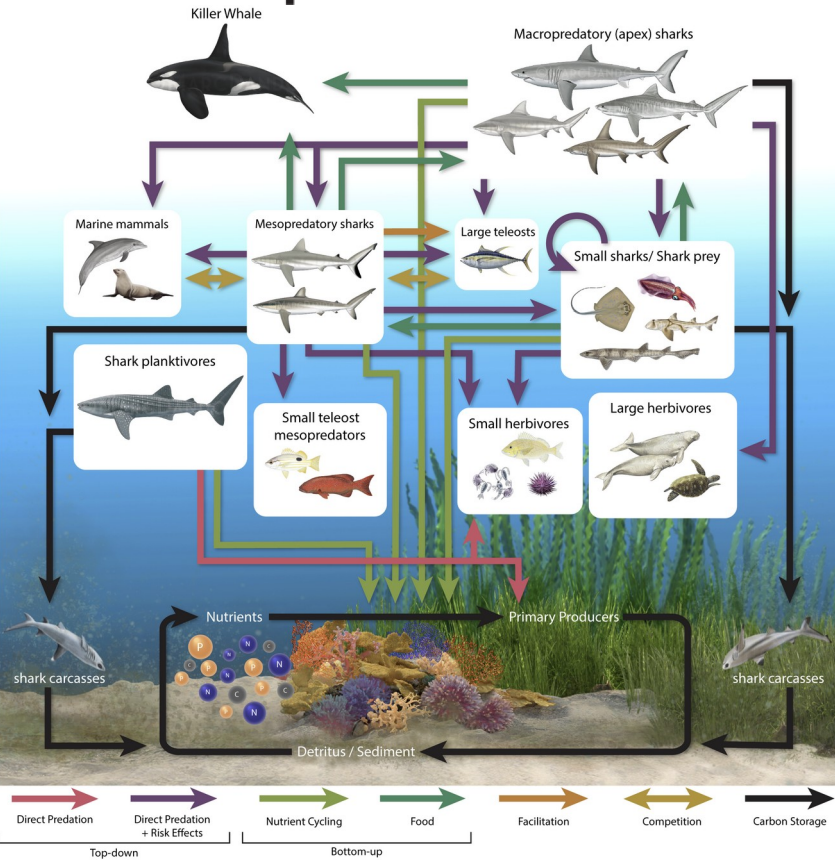
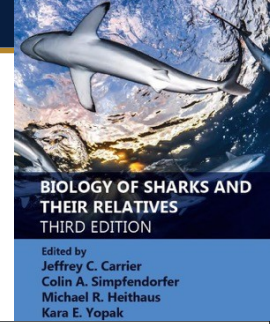
Santa Barbara

Miami

REVIEW

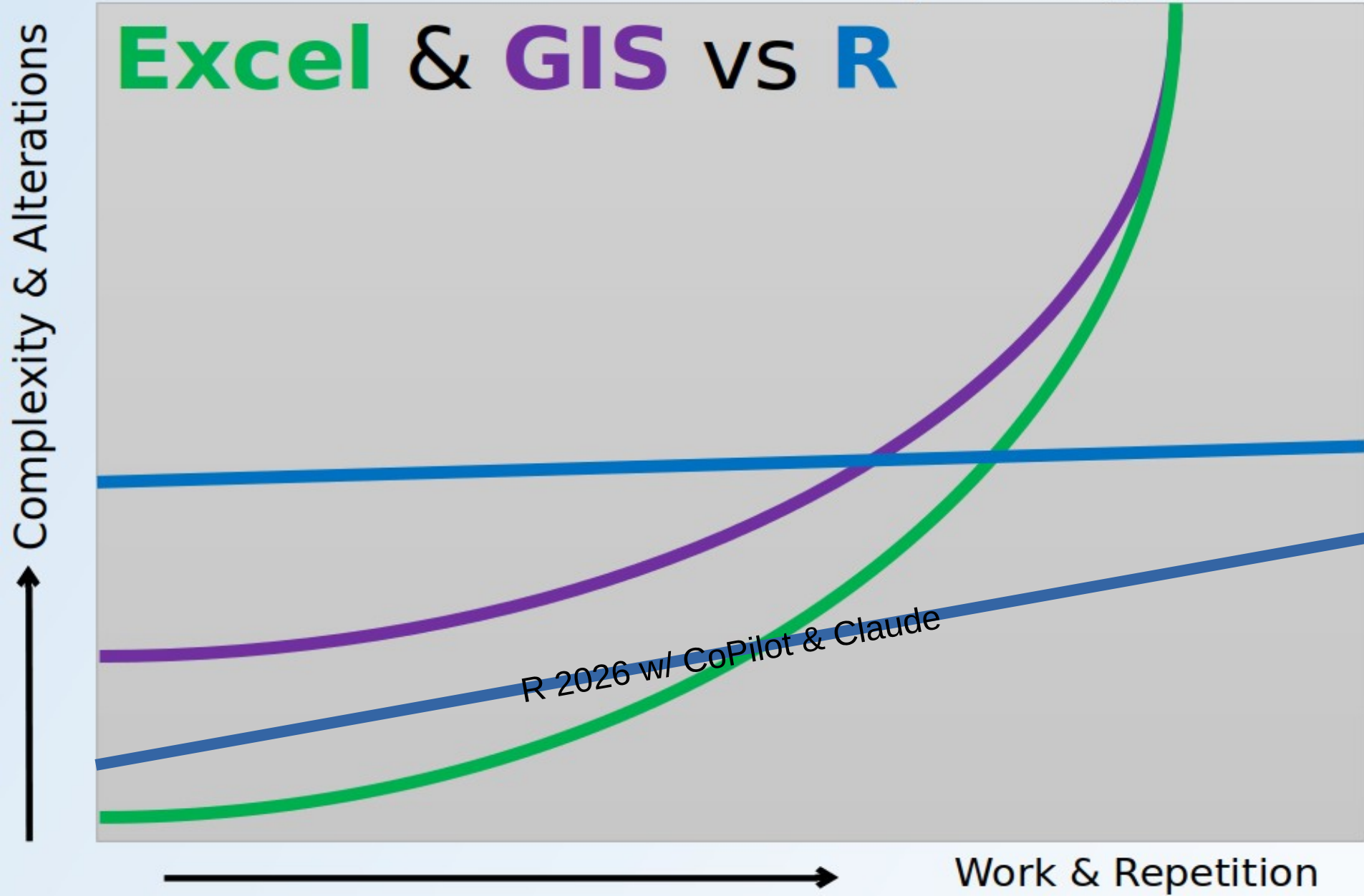
CONSERVATION

Ecological roles and importance of sharks in the Anthropocene Ocean



GUI vs Scripting

Excel & **GIS** vs **R**



System setup: R, Rstudio

Options

- General
- Code**
- Console
- Appearance
- Pane Layout
- Packages
- R Markdown
- Python
- Sweave
- Spelling
- Git/SVN
- Publishing
- Terminal
- Accessibility
- Copilot

- Editing**
- Display
- Formatting
- Saving
- Completion
- Diagnostics

Editing

- Insert spaces for Tab
Tab width:
- Auto-detect code indentation
- Insert matching parens/quotes
- Use native pipe operator, |> (requires R 4.1+)
- Auto-indent code after paste
- Vertically align arguments in auto-indent
- Continue comment when inserting new line
- Allow drag and drop of text
- Enable hyperlink highlighting in editor
Editor scroll speed sensitivity:
- Surround selection on text insertion: ▼
- Keybindings: ▼

Execution

- Focus console after executing from source
- Ctrl+Enter executes: ▼

Snippets

- Enable code snippets

Options

Editing | **Display** | Formatting | Saving | Completion | Diagnostics

General

- Highlight selected word
- Highlight selected line
- Show line numbers
- Relative line numbers
- Show margin
 - Margin column:
- Soft-wrap source files
- Soft-wrap at margin column
- Clamp editor width to margin column
- Show whitespace characters
- Blinking cursor
- Allow scroll past end of document
- Fold style:
- Indentation guides:

Syntax

- Highlight R function calls
- Enable preview of named and hexadecimal colors
- Use rainbow parentheses

General | Code | Console | Appearance | Pane Layout | Packages | R Markdown | Python | Sweave | Spelling | Git/SVN | Publishing | Terminal | Accessibility | Copilot

OK Cancel Apply

Options

Editing Display **Formatting** Saving Completion Diagnostics

Code Formatting

Code formatter:

Format with styler

Use the styler R package to reformat code.

Use strict transformers when formatting code

Reformat documents on save

Also air formatter:
<https://tidyverse.org/blog/2025/02/air/>
<https://posit-dev.github.io/air/editor-rstudio.htm>

General
Code
Console
Appearance
Pane Layout
Packages
R Markdown
Python
Sweave
Spelling
Git/SVN
Publishing
Terminal
Accessibility
Copilot

OK

Cancel

Apply

Options

Editing | Display | Formatting | **Saving** | Completion | Diagnostics

General

- Ensure that source files end with newline
- Strip trailing horizontal whitespace when saving
- Restore last cursor position when opening file
- Reformat documents on save

Serialization

Line ending conversion: Platform Native ▾

Default text encoding: [Ask] Change...

Autosave

- Always save R scripts before sourcing
- Automatically save when editor loses focus

When editor is idle: Backup unsaved changes ▾

Idle period: 1000ms ▾

OK Cancel Apply

General

Code

Console

Appearance

Pane Layout

Packages

R Markdown

Python

Sweave

Spelling

Git/SVN

Publishing

Terminal

Accessibility

Copilot

Options

Editing | Display | Formatting | Saving | **Completion** | Diagnostics

R and C/C++

Show code completions:

- Allow automatic completions in console
- Insert parentheses after function completions
- Show help tooltip after function completions
- Show help tooltip on cursor idle
- Insert spaces around equals for argument completions
- Use tab for autocompletions
- Use tab for multiline autocompletions
- Show data preview in autocompletion help popup
- Include all function arguments in completion list

Other Languages

Show code completions:

Keyword and text-based completions are supported for several other languages including JavaScript, HTML, CSS, Python, and SQL.

Completion Delay

Show completions after characters entered:

Show completions after keyboard idle (ms):

General
Code
Console
Appearance
Pane Layout
Packages
R Markdown
Python
Sweave
Spelling
Git/SVN
Publishing
Terminal
Accessibility
Copilot

OK Cancel Apply

Options

Editing | Display | Formatting | Saving | Completion | **Diagnostics**

R Diagnostics

- Show diagnostics for R
- Enable diagnostics within R function calls
- Check arguments to R function calls
- Check usage of '<-' in function call
- Warn if variable used has no definition in scope
- Warn if variable is defined but not used
- Provide R style diagnostics (e.g. whitespace)
- Prompt to install missing R packages discovered in R source files

Other Languages

- Show diagnostics for C/C++
- Show diagnostics for YAML
- Show diagnostics for JavaScript, HTML, and CSS

Show Diagnostics

- Show diagnostics whenever source files are saved
- Show diagnostics after keyboard is idle for a period of time
Keyboard idle time (ms):

[? Using Code Diagnostics](#)

General | Code | Console | Appearance | Pane Layout | Packages | R Markdown | Python | Sweave | Spelling | Git/SVN | Publishing | Terminal | Accessibility | Copilot

OK | Cancel | Apply

Options

- General
- Code
- Console
- Appearance
- Pane Layout
- Packages
- R Markdown
- Python
- Sweave
- Spelling
- Git/SVN
- Publishing
- Terminal
- Accessibility
- Copilot

RStudio theme: Modern

Zoom: 125%

Text rendering: (Default)

Editor font: Fira Code

Editor font size: 10

Line height (%): 125

Help font size: 10

Editor theme: ChaosSD

Add... Remove

```
# compute five-number summary
fivenum <- function(x) {


  # handle empty input
  n <- length(x)
  if (n == 0)
    return(rep.int(NA, 5))

  # compute quartile indices
  n5 <- 1
  n4 <- ((n + 3) %/% 2) / 2
  n3 <- (n + 1) / 2
  n2 <- n + 1 - n4
  n1 <- n
  i <- c(n5, n4, n3, n2, n1)

  # compute quartile values
  x <- sort(x)
  xf <- x[floor(i)]
  xc <- x[ceiling(i)]
  0.5 * (xf + xc)
}
```

<https://github.com/tonsky/FiraCode>

chaosSD in Canvas

-  General
-  Code
-  Console
-  Appearance
-  Pane Layout
-  Packages
-  R Markdown
-  Python
-  Sweave
-  Spelling
-  Git/SVN
-  Publishing
-  Terminal
-  Accessibility
-  Customization

Package Development

- Use devtools package functions if available
- Use alternate library path when building package
- Save and reload R workspace on build
- Save all files prior to building packages
- Automatically navigate editor to build errors
- Cleanup output after successful R CMD check
- View Rcheck directory after failed R CMD check
- Always use LF line-endings in Unix Makefiles

[? Developing Packages](#)

Go to file/function Addins try these out

bookmarks sidebar

project HERE dropdown here

netcdf_extraction.R x script_order.R x HomeRangeAve.R x StocksPaperFigures.R x SlopeSeaPlots.R x

Environment History Connections Git Tutorial Import Dataset 664 MiB different views Grid

```

2419 cellSummaries <- x %>%
2420   group_by(Cell, !!sym(groupcol)) %>% # group by cell & groupcol
2421   summarise(across(c(X, Y), first), # add back X & Y for xyz fun
2422             Count = n(), # add Count
2423             Proportion = n() / first(groupcount), # total n per group but not cell
2424             !!zCol := match.fun(zFun)(!!sym(zCol), na.rm = T) # output colname zCol create
2425             # https://stackoverflow.com/questions/62906259/r-user-defined-dynamic-summary-
2426             # doesn't work if zCol = "Count" since Count already used in this summarise. B
2427           ) %>%
2428   ungroup
2429
2430 cellSummaries %>% group_by(!!sym(groupcol)) %>% summarise(SumProp = sum(Proportion)) # che
2431 # scaled to maximum fishdays found in any one cell, across all cells, per stock.
2432
2433 xyz <- make.xyz(x = cellSummaries$X,
2434               y = cellSummaries$Y,
2435               z = cellSummaries %>% pull(zCol),
2436               group = cellSummaries %>% pull(groupcol),
2437               FUN = noquote(zFun)) # should do nothing, cellSummaries should be 1:1 size
2438
2439 # reconfigure cell proportions option
2440 if (ScaleToCellMax) xyz$z <-%>% apply(MARGIN = 2, # columns
2441                                     FUN = function(x) (x / max(x, na.rm = TRUE))) # scale
2442
2443 png(filename = paste0(saveloc, plotname, ".png"),
2444     width = pngwidth, #NA default. Manually adjust plot box in Rstudio after ggplot()
2445     height = pngheight, #NA default: Then oasave with defaults, changes from 7x7" to e.g.
2446
2447 barplot2dMap(x, latcol, loncol, origin, cellsize, groupcol, mycolours, legendtitle, zCol, zFun, ScaleToCellMax, baseplot, xlim, ylim, lon1, lon2, lat1, lat2, bathyres,

```

Track lines by month
 Track lines by toppid, facet stock
 2020.08.21 Track lines by toppid, facet...
 2020.09.21 GOM / WTagged Med / ETag...
 2020.09.29 GOM / WTagged Med trackl...
 2020.09.21 W Tagged / Ireland tagged ...
 2020.09.21 E Tagged fish crossing 45W
 2020.08.24 track lines before 1st SG e...
 2020.06.25 Track lines by Stock
 Track lines PAT popups only
 Track lines facet by tagging location
 2021-01-28 Tracklines TOPPID loop spa...
 Tracklines TOPPID loop months/dates/...
 2020.08.24 Fig1 median location track...
 Find offshore tracks off USA Eseaboar...
 Find offshore tracks off USA Eseaboar...
 Plot NearFar Tracks
 2020.07.22 offshore tracks revisited, s...
 Points coloured by stock animate
 2020.05.12 Min Ext Temp Daily points
 Whole area points
 2020.06.30 points + bathy & density c...
 2020.07.09 tracks+bathy contours
 Scotian Shelf zoom points
 Scotian Shelf zoom point month loop
 Scotian Shelf month animate
 Slope Sea ChIA maps
 Negative OceanDepth hunt
 2020.06.09 Track points EddyColour re...
 Bathy & contour map test

Name	Type	Length	Size	Value
AllDailies	tbl_df	143	38.1 MB	30117 obs. of 143 variab...
combofif	matrix	2	344 B	chr [1:2] "GOM" "Med"
datemax	integer	1	56 B	2020L
datemin	integer	1	56 B	1997L
df	data.frame	18	4.1 MB	30117 obs. of 18 variabl...
dflist	list	1	3.4 KB	List of 1
extents	sf	1	5.7 KB	2 obs. of 1 variable
groupcol	character	1	112 B	"Stock"
i	integer	1	56 B	18L
j	integer	1	56 B	1L
k	integer	1	56 B	1L
knames	character	1	112 B	"GOM-Med"
l	integer	1	56 B	2L
loadloc	character	1	168 B	"/home/simon/Documents/Si W...

bookmarks dropdown

R 4.0.4 ~/Dropbox/Blocklab Monterey/Blocklab/

Console Terminal x Render x Jobs x

Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.

[workspace loaded from ~/Dropbox/Blocklab Monterey/Blocklab/.RData]

Registered S3 methods overwritten by 'adehabitatMA':

```

method from
print.SpatialPixelsDataFrame sp
print.SpatialPixels sp












```

Files Plots Packages Help Viewer quick access, pkg install, etc

Home > Dropbox > Blocklab Monterey > Blocklab

Name	Size	Modified
..		
.gitignore	55 B	May 12, 2020, 8:25 AM
.RData	31.1 MB	Oct 5, 2021, 11:57 AM
.Rhistory	22.1 KB	Oct 11, 2021, 10:18 PM
ABFT_StockExtract_MC_SQL.R	2.8 KB	Jul 1, 2019, 3:37 PM
Accelerometry.R	5.6 KB	Jun 1, 2020, 10:19 AM
AgeLonStockPlots.R	5.8 KB	Mar 31, 2020, 12:13 PM
all_data.Rdata	1.5 MB	Aug 18, 2021, 10:29 AM
AllDailies_FishRemove.R	1.3 KB	May 27, 2020, 7:51 PM
AllDailiesAnalysis.R	18.4 KB	May 14, 2020, 1:01 PM
BeforeAfterChangeCheck.R	2.8 KB	Feb 6, 2020, 5:23 PM
BehaviourValidation.R	3.1 KB	Mar 16, 2020, 11:01 AM
Blocklab.Rproj	258 B	Oct 13, 2021, 6:30 PM
ChIApend.R	15.1 KB	Oct 7, 2020, 3:10 PM
DailyDFsRbind.R	3.5 KB	Nov 5, 2019, 2:26 PM

Options

-  General
-  Code
-  Console
-  Appearance
-  Pane Layout
-  Packages
-  R Markdown
-  Python
-  Sweave
-  Spelling
-  **Git/SVN**

Version Control

Enable version control interface for RStudio projects

Sign git commits

Git executable:

[Browse...](#)

SVN executable:

[Browse...](#)




Terminal executable:
















[Browse...](#)

SSH key: [View public key](#)

[Create SSH Key...](#)

[? Using Version Control with RStudio](#)

-  Publishing
-  Terminal
-  Accessibility

-  General
-  Code
-  Console
-  Appearance
-  Pane Layout
-  Packages
-  R Markdown
-  Python
-  Sweave
-  Spelling
-  Git/SVN
-  Publishing
-  Terminal
-  Accessibility
-  Copilot

GitHub Copilot

Enable GitHub Copilot

You are currently signed in as: SimonDedman Sign Out Refresh Diagnostics

Copilot Indexing

Index project files with GitHub Copilot

Copilot Completions

Show Copilot code suggestions:

Show code suggestions after keyboard idle (ms):

Other

Display account and billing messages from GitHub Copilot

Use RStudio project folder as a Copilot workspace

By using GitHub Copilot, you agree to abide by their terms of service.

[? GitHub Copilot: Terms of Service](#)

What IS GitHub?

1. Dropbox/Google Docs but with precise version control
2. Each projects is called a Repository
3. README.md (markdown document) becomes the front page of the repo.
4. 1 person can use it as cloud storage others can see, download, install.
5. 2+ people can collab work on code.
6. Others can fork your repo and work on it themselves.
7. They can propose Pull Requests to fix issues, or go in new directions.
8. Projects gain traction informally here before becoming CRAN packages.
9. You can host HTML on the github.io branch of your repo e.g. <https://github.com/SimonDedman/CarsOrSharks>
10. Your GitHub homepage hosts all your repos and some personal info.

GitHub Setup

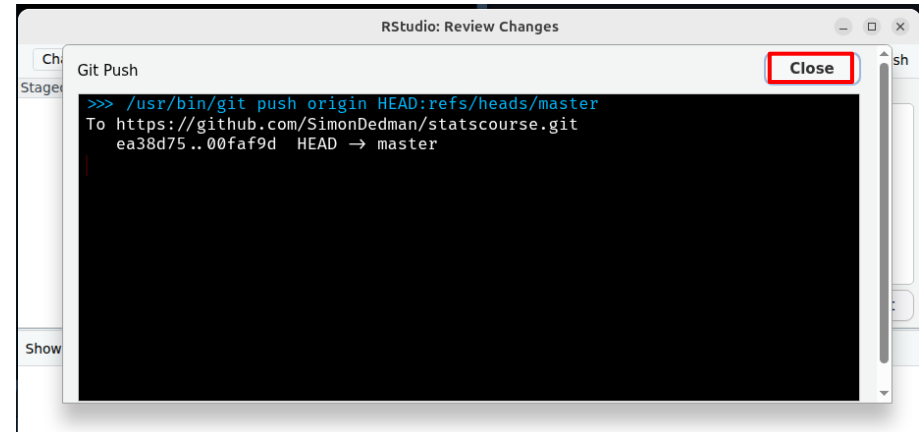
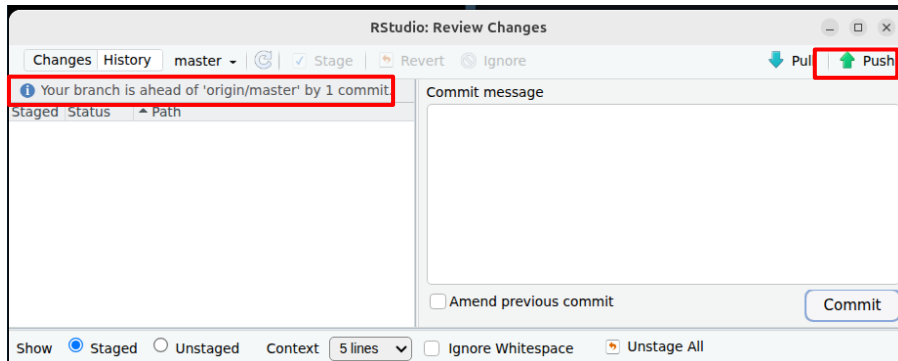
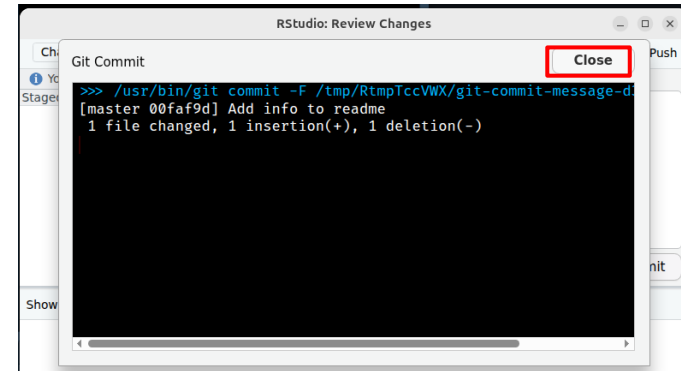
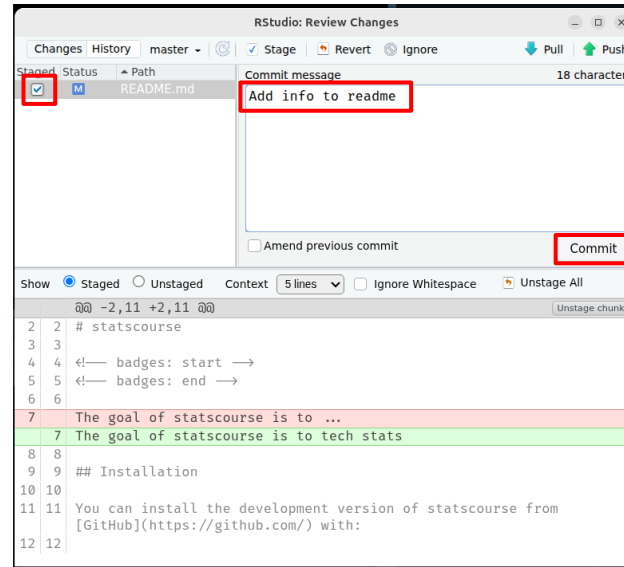
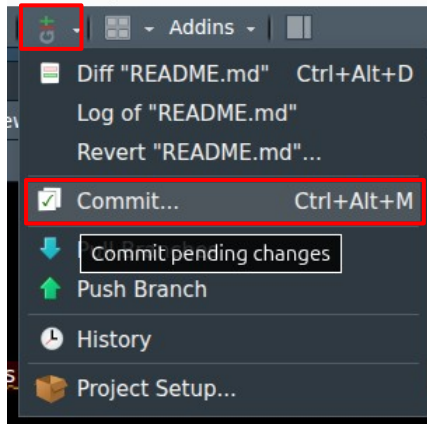
1. Install R - <https://www.r-project.org/>
2. Install RStudio - <https://posit.co/download/rstudio-desktop/>
3. Install Git - <https://happygitwithr.com/install-git>
4. Register a GitHub account - <https://github.com> & <https://happygitwithr.com/github-acct>
5. Setup Personal Access Token for HTTPS (done next) <https://happygitwithr.com/https-pat>
6. Or for SSH - <https://happygitwithr.com/ssh-keys>
7. DON'T Open a new repository on GitHub yet

Bookmark & follow <https://happygitwithr.com>

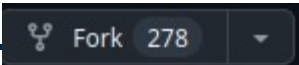
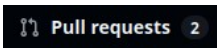
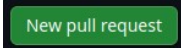

Initialise Project & Connect to GitHub

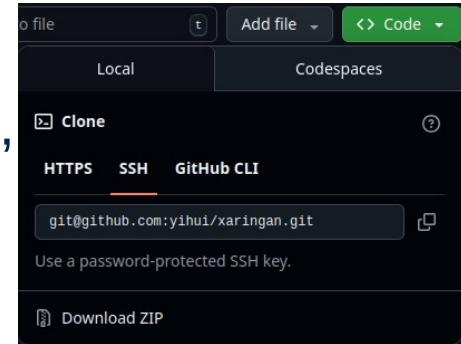
1. Follow Sarah Popov's presentation: <https://popovs.github.io/r-pkg-slides>
2. `install.packages("devtools", "usethis")` `library(devtools)` `library(usethis)`
3. `usethis::create_package("~/Documents/R/fjoRds")`
(replace 'fjoRds' with the name of the project/folder/package you want to create, which doesn't exist)
4. `usethis::use_r("myfunction")`
(give it a real name & add content)
5. Code > Insert Roxygen skeleton
6. Edit DESCRIPTION file, `usethis::use_package_doc()`,
`usethis::use_readme_md()`, `usethis::use_vignette("function_explain")`,
`usethis::use_data_raw()`, `usethis::use_data()`, `usethis::use_test()`,
`use_gpl_license(version = 3, include_future = TRUE)`,
`usethis::use_code_of_conduct("your@email")`, `devtools::document()`,
7. `usethis::use_git()`, `usethis::use_github()`, `usethis::create_github_token()`,
`gitcreds::gitcreds_set()`, replace credentials.
8. Tools > terminal: `git push -u origin master` (*may be needed*)
9. Edit, save

Edit, Save, Stage, Comment, Commit, Push, to Github

















Github Pull Requests: contribute to other packages

1. Fork a project on github (top right). 
2. In your GitHub forked project, click green "code" button, copy git address.
3. In a folder on your filesystem: "git clone [gitaddress]".
4. Go to new subfolder, open Rproj in RStudio.
5. Make changes in your local files in RStudio, test, devtools::document(), 'git commit push' to your own Github fork.  
6. Go to original (other person's) package, pull requests, create, "compare across forks", select your fork as the right hand side. "Create pull request" button. 
7. If it's not automatically named correctly, name it well. See <https://docs.github.com/en/issues/tracking-your-work-with-issues/linking-a-pull-request-to-an-issue>
8. Example: <https://github.com/r-lib/usethis/pull/1898>
9. <https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/proposing-changes-to-your-work-with-pull-requests/creating-a-pull-request?tool=desktop>



System setup:
Theme, Font

Resources:
Guides, Concepts, Code

Name ▲	Created ↕	Last Modified ↕	Modified By ↕	Size ↕
 Others Code Dumps	7:35pm	7:35pm		--
 R Cheat Sheets	6:45pm	6:45pm		--
 2019.06.25_Stanford_Universit...	4:26pm	5:30pm	Simon Dedman	7.1 MB
 2021-10-13_Emil_Aalto_Coding...	4:26pm	8:17pm	Simon Dedman	370 KB
 2021-10-13_Emil_Lab_Meeting...	4:26pm	8:17pm	Simon Dedman	885 KB
 2023-11-16_MyLabMeeting.pptx	8:27pm	8:27pm	Simon Dedman	12.4 MB
 chaosSD.rstheme	4:27pm	4:27pm	Simon Dedman	29 KB
 Data_Organization_SOP.docx	8:27pm	8:27pm	Simon Dedman	203 KB
 Excel-vs-GIS-vs-R.png	4:32pm	4:32pm	Simon Dedman	68 KB
 Filazzola_Lortie.2022.Call_clean...	4:32pm	5:51pm	Simon Dedman	2.1 MB
 JennyBryan.NamingPubsCarpet...	4:32pm	5:51pm	Simon Dedman	5.9 MB
 JennyBryan.NamingThings.pdf	4:32pm	5:51pm	Simon Dedman	2.3 MB
 Project_Organisation_Metadata...	8:27pm	8:27pm	Simon Dedman	89 KB
 template_organised.R	8:04pm	8:04pm	Simon Dedman	50 KB

- Various presentations on coding tips

- Especially my bits in Emil Lab Meeting

- chaosSD.rstheme RStudio custom theme has colours for all elements (some otherwise unhelpfully missing colours)


- template_organised: freshly organised code repository from my years of learning R.

Others code dumps. See also: <https://github.com/SimonDedman/MiscScripts>

Name ▲

 Ally-Jones-spatial-autocorrelation

- Spatial autocorrelation tests

 Blocklab-tuna-movement

- Accelerometer, stocks, age/length, dive behaviour, variables acquire, sensor fixing, depth bins, day/night stats, distance to shore, home range, maps, KS tests, linearity index, plots, scales, etc

 Emil-Aalto

- Graphics & utility functions

 Farallon-Institute


- Lags, logs, pairplots, population modelling

 Gemma-Carroll_ggplot_animate...

- Animated SST fish tracks

 James-Keating

- Pairplots, KS test, XY plots, basic plots

 Matt-Savoca-raincloudplot

- Raincloud plots – sideways violins

 Saving-The-Blue

- Data cleaning, munging, mapping, analyses, reports, track filtering ARGOS & foieGras

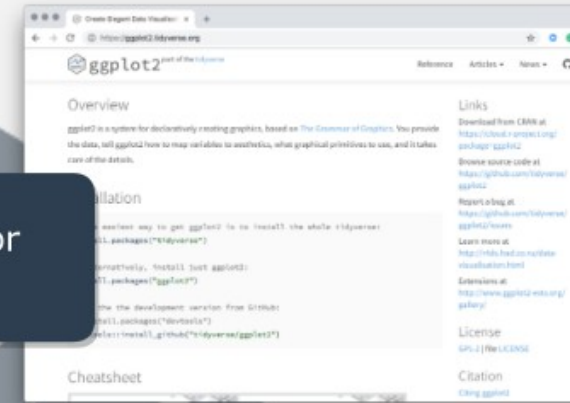
 Simon-Dedman-misc

- Maps, bins, utility functions.

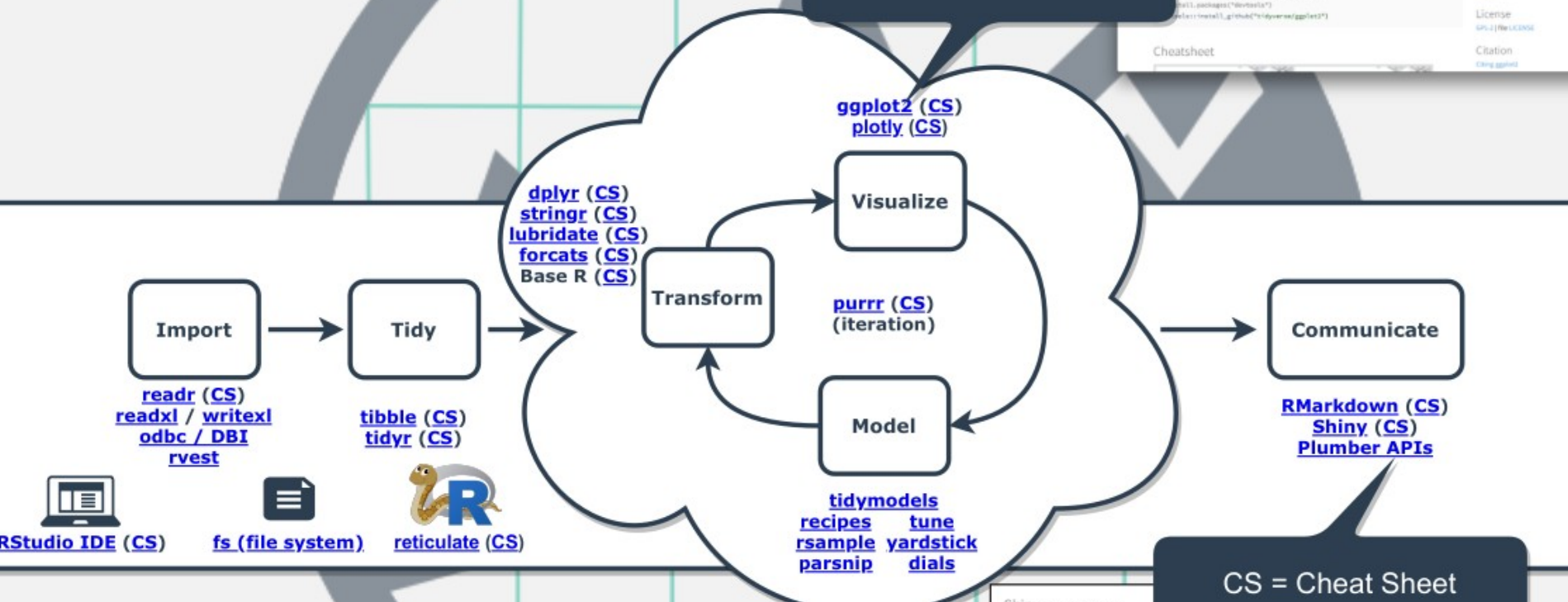
R cheat sheets

Data Science with R Workflow

Learn R following the tidyverse & tidymodels workflow for business analysis in the [R For Business Analysis \(DS4B 101-R\) course](#) through Business Science University.



Click the links for Documentation



42 files: R misc, wrangling, import, data table, factors, ggplot, cloud compute, ML, dates, leaflet/Shiny online plots & maps, colours, markdown, spatial, text strings, survival models

Base R vs TidyVerse (Dplyr/Pipes)



tidyverse.org/blog/

Tidyverse

Packages

Posts



ragnar 0.3.0

package

Tomasz Kalinowski

The new release of ragnar adds faster ingestion, new embedding providers, improved retrieval, and new integrations for using ragnar stores from tools.

2026/01/27



plumber2 0.2.0

package

Thomas Lin Pedersen

The next version of plumber2 has hit CRAN. Read all about the new features such as OpenTelemetry (OTEL) support, authentication, new tags, and performance improvements here.

2026/01/20

```
1 # Base R
2 df <- data.frame(
3   Lat = c(25.1, 25.2, 25.3),
4   Lon = c(121.5, 121.6, 121.7),
5   Depth = c(10, 20, 30)
6 )
7 df[1,3] # 10. [rows,columns] [radio,control]
8 df[3,1] # 25.3
9 df$Lon[3] # 121.7
10 df$Depth[c(1,3)] # 10 30
11 df[1,3] <- 20
12 df[1,3] # 20
13 dfdepth <- df$Depth
14 str(dfdepth) # num [1:3] 10 20 30
15 is.vector(dfdepth) # TRUE
16 dfdepth <- df > select(Depth) # becomes tibble
17
18 # Tidyverse
19 library(tidyverse)
20 df <- tibble(
21   Lat = c(25.1, 25.2, 25.3),
22   Lon = c(121.5, 121.6, 121.7),
23   Depth = c(10, 20, 30)
24 )
25 df > slice(1) > pull(Depth) # 10
26 df > slice(3) > pull(Lat) # 25.3
27 df > pull(Lon) > .[3] # 121.7
28 df > select(Depth) > slice(c(1,3)) > pull() # 10 30
29 length(df$Depth) # 3
30 tbldepth <- df > select(Depth) # tibble
```

Coding, Errors, & Help

- Commenting concepts: help your future self. Script organisation, layout, headings, markdown formatting.
- Errors: AI is nice but the first pass is to actually READ the error messages, run traceback, bring up the ?help file, READ the help file. Loads of people ask me how to fix errors/warnings which are well documented (see gbm.auto help file).

The screenshot displays the RStudio interface with the following components:

- Source Editor:** Contains R code for reading CSV files and saving them as R datasets using `readr::read_csv()` and `usethis::use_data()`.
- Environment:** Shows the current environment with two data objects: `expvardy` (1461 obs. of 8 variables) and `expvarstat` (1461 obs. of 5 variables).
- Console:** Shows the execution of `usethis::use_data(expvardy, overwrite = TRUE)`, resulting in a message: "Saving 'expvardy' to 'data/expvardy.rda'".
- Help Panel:** Displays the documentation for `use_roxygen_md` from the `usethis` package, including a description, usage, and arguments.

```
1 ## code to prepare DATASET dataset goes here
2 sharkdata <- readr::read_csv("data-raw/mydata
3 expvarstat <- readr::read_csv("data-raw/expvar
4 expvardy <- readr::read_csv("data-raw/expvar
5
6 usethis::use_data(sharkdata, overwrite = TRUE)
7 usethis::use_data(expvarstat, overwrite = TRUE
8 usethis::use_data(expvardy, overwrite = TRUE)
9
```

```
> usethis::use_data(expvardy, overwrite = TRUE)
Saving "expvardy" to "data/expvardy.rda".
Document your data (see <https://r-pkgs.org/data.html>).
```

use_roxygen_md {usethis} R Documentation

Use roxygen2 with markdown

Description

If you are already using roxygen2, but not with markdown, you'll need to use [roxygen2md](#) to convert existing Rd expressions to markdown. The conversion is not perfect, so make sure to check the results.

Usage

```
use_roxygen_md(overwrite = FALSE)
```

Arguments

`overwrite` Whether to overwrite an existing Roxygen field in DESCRIPTION with "list(markdown = TRUE)".

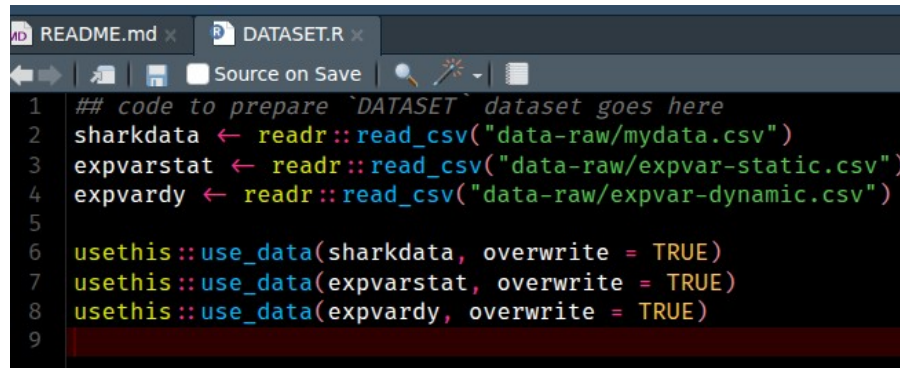
[Package *usethis* version 3.2.1 [Index](#)]

Typical workflow

- Folder setup, Github & R package, here::here
https://github.com/jennybc/here_here
- Import: core & additional data
- Tidy: explore, find errors & outliers, conditional format colour scale, clean
- Transform: join, next slides, plus more, see docx
- Visualise: check autocorrelation, plots
- Model [BRT & hopefully tidymodels 4th March]
- Communicate

Import

- Fork/clone my repo <https://github.com/SimonDedman/statscourse>
- or start your own with the CSV & R files
- `usethis::use_data_raw()` then edit it (DATASET.R)



```
1  ## code to prepare `DATASET` dataset goes here
2  sharkdata <- readr::read_csv("data-raw/mydata.csv")
3  expvarstat <- readr::read_csv("data-raw/expvar-static.csv")
4  expvardy <- readr::read_csv("data-raw/expvar-dynamic.csv")
5
6  usethis::use_data(sharkdata, overwrite = TRUE)
7  usethis::use_data(expvarstat, overwrite = TRUE)
8  usethis::use_data(expvardy, overwrite = TRUE)
9
```

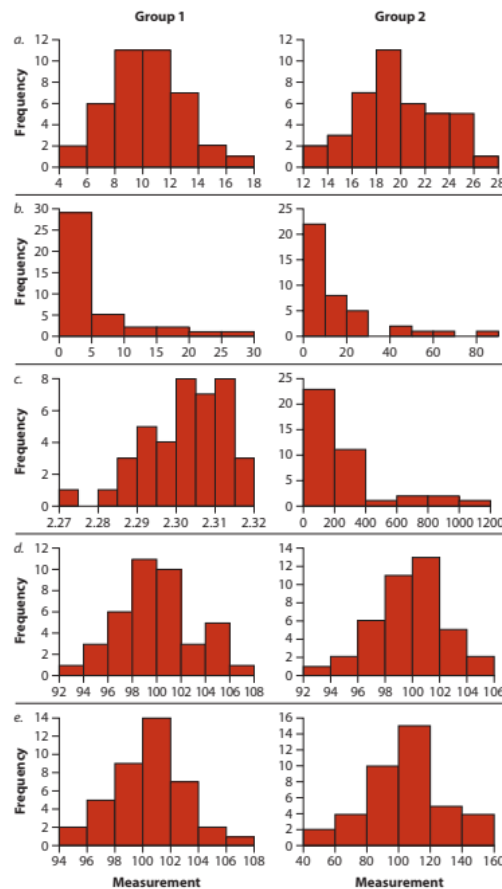
Tidy inc. Plot

- 01_tidy-data.R example
- mydata-bad.csv colour condition example (drumline, no Gibson)

Transform

- Join, assess outliers again
- template_organised.R L1-858 SKIM
- Summarise, pivot, mutate, group, etc

Evaluating Parametric Assumptions & Data Transformations



Common Parametric Test Assumptions

(1) Errors independent & identically distributed

- a. Treatments are applied independently to replicates
- b. A given replicate does not affect other replicates
- c. No autocorrelation among replicates

(2) Homogeneity of variances among treatment groups (homoscedasticity)

(3) Data are normally distributed

Overview: Assumption Checking

- Use tests & graphical examination of residuals to evaluate assumptions
- Minor to moderate violations of assumptions usually not a problem; tests are still reliable
- If assumptions are severely violated, then:
 - 1) try transforming data & retest assumptions
 - 2) rank transform data before analysis
(equivalent to a nonparametric test)
 - 3) redo experiment with appropriate changes in design

Note: Best to test assumptions using residuals for each datum (i.e., group mean – datum value), not the raw data

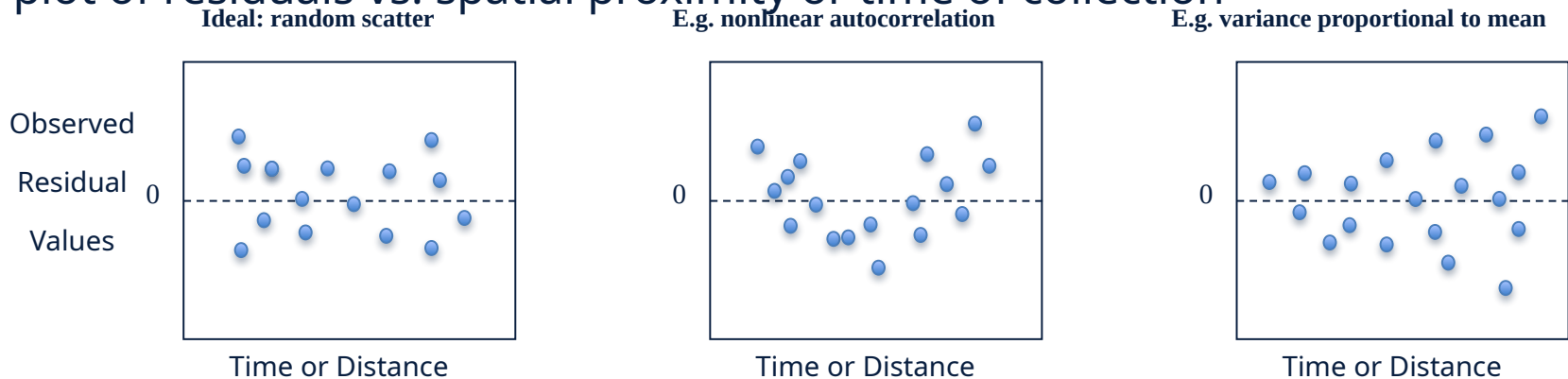
Independence of Error

Effect: ruins the analysis; inflates type I error; lowers precision of variance estimate.

Test: difficult to test, often requires detailed info on experimental protocol (e.g., physical proximity replicates, time order of sampling, etc.) and knowledge of system (e.g., likely sources of non-independence)

If you suspect a problem:

- try plot of residuals vs. spatial proximity or time of collection



Solution: None. If violated, can only redesign and repeat experiment correctly

Homogeneity of Variance

Effect: a problem when you have unequal n and a small # of treatment groups

- often disappears if normality problem corrected first
- general rule: ratio of largest variance to smallest should be <3

a) Bartlett's test – Poor test if data are also non-normal

b) Levene's test – better; doesn't depend on normality of data

```
bartlett.test(response ~ group, data = df)
car::leveneTest(response ~ group, data = df)
```

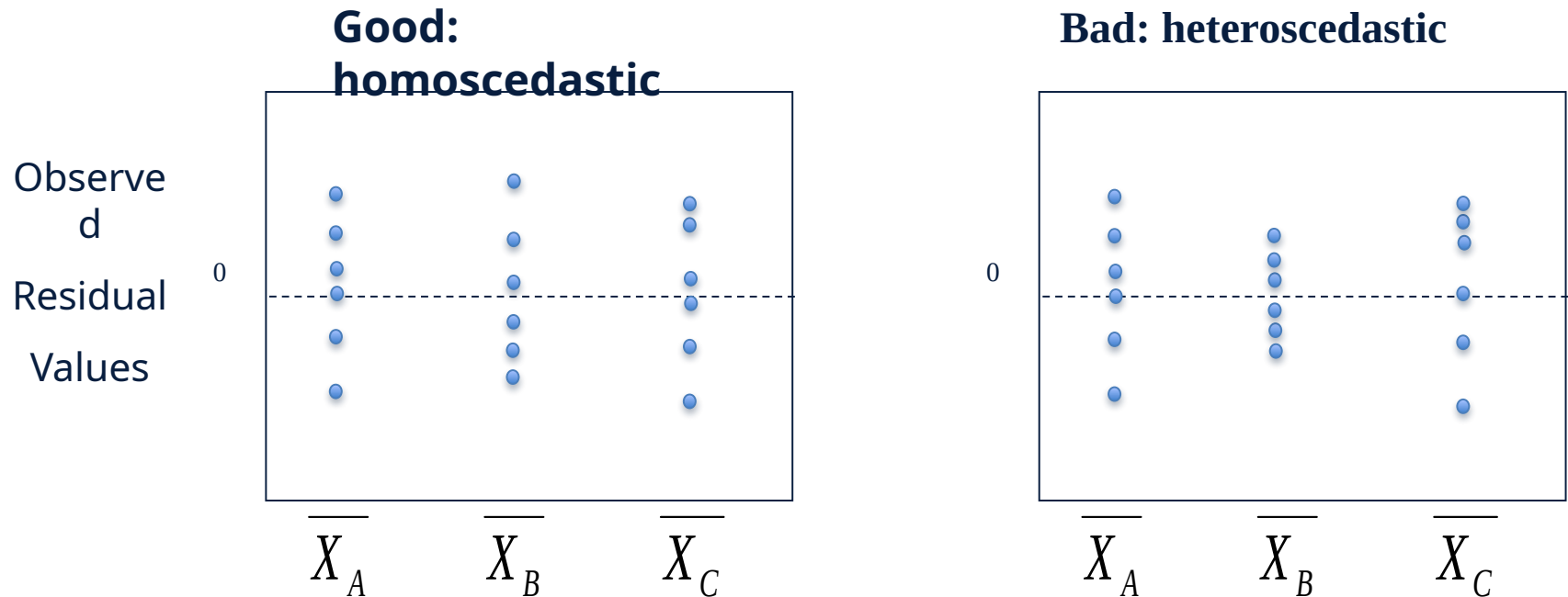
Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
RI	Based on Mean	.291	1	18	.596
	Based on Median	.444	1	18	.514
	Based on Median and with adjusted df	.444	1	16.024	.515
	Based on trimmed mean	.291	1	18	.596

Homogeneity of Variance

Plot Inspection:

Plot of Residuals ($\bar{x}_{ij} - X_j$) vs. Treatment Means



```
ggplot(df, aes(x = fitted, y = residuals)) +  
  geom_point() + geom_hline(yintercept = 0)
```

Normality

Effect:

- not usually a problem, if $n > 10$
- may be problem if $n < 10$ and unequal sample sizes among treatments
- minor to moderate violations of this assumption usually not a problem; hypothesis tests are still reliable

Evaluate: Graphically or with statistical tests:

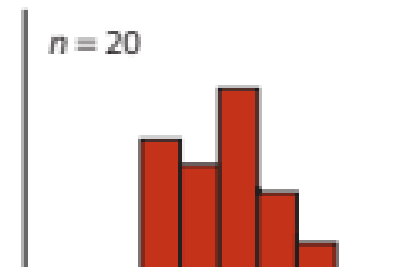
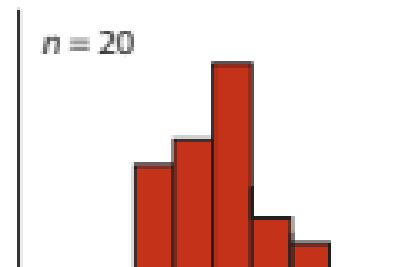
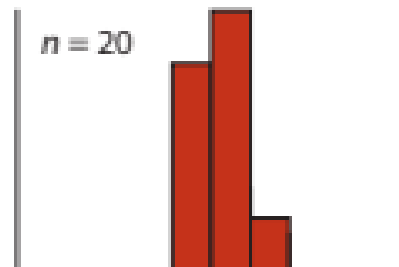
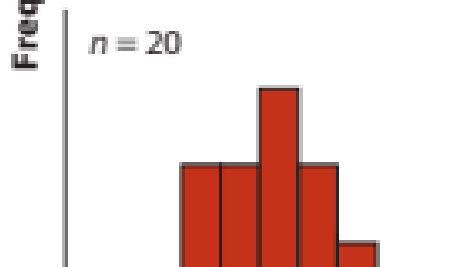
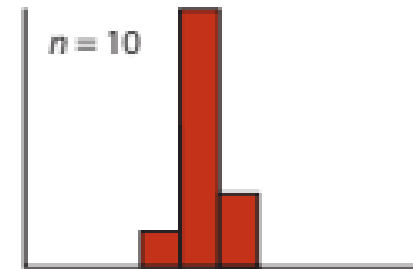
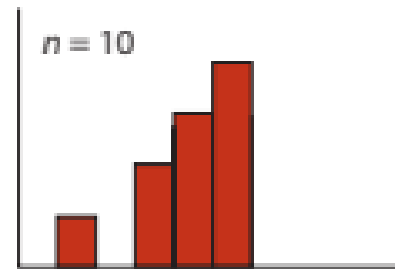
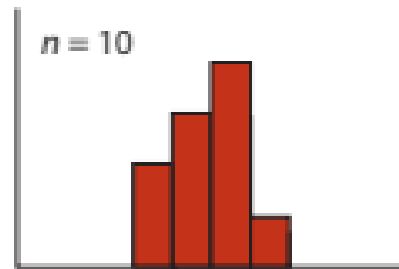
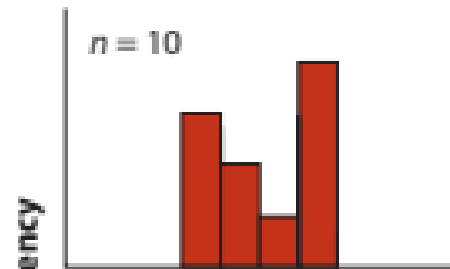
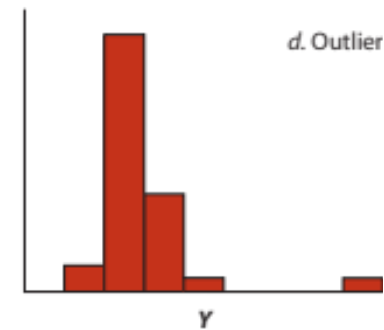
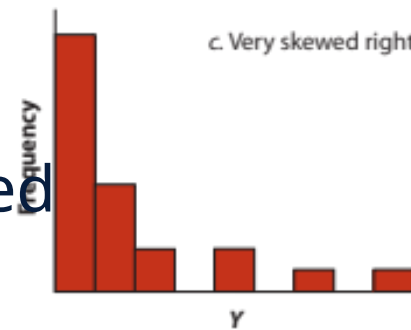
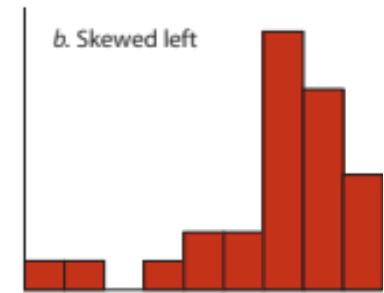
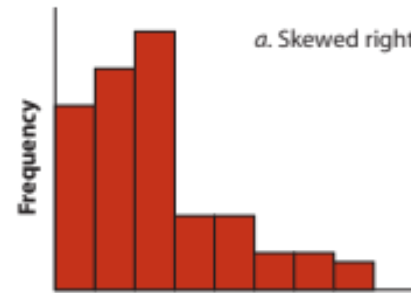
Graphical Approaches:

- a) plot histogram of residuals
- b) normal probability plot (quantile plot)

Inspection of Normality using frequency histograms of data

Problem:

- sample size dependency
- Subjective; not recommended



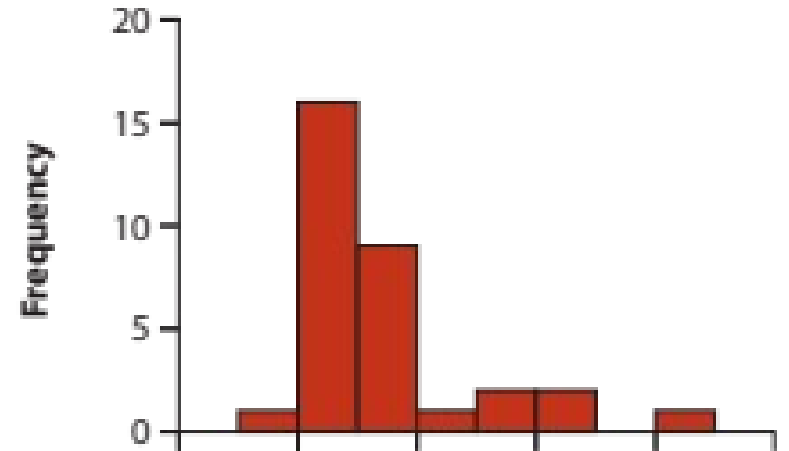
Y

Inspection of Normality

Top:

- frequency histogram

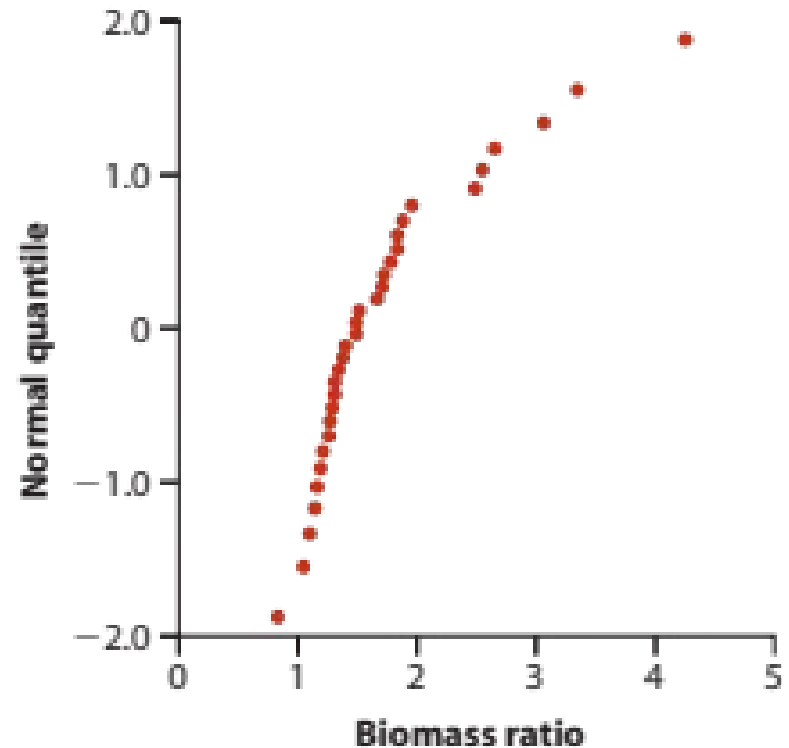
```
ggplot(df, aes(x = residuals)) +  
  geom_histogram(bins = 20)
```



Bottom:

- quantile plot (normal probability plot)

```
ggplot(df, aes(sample = residuals)) +  
  stat_qq() + stat_qq_line()
```

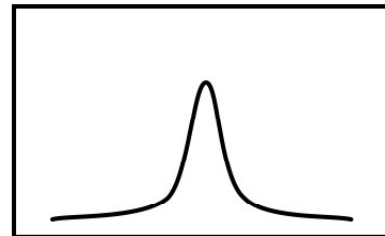
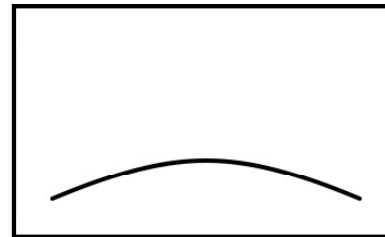
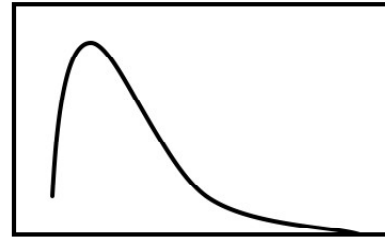
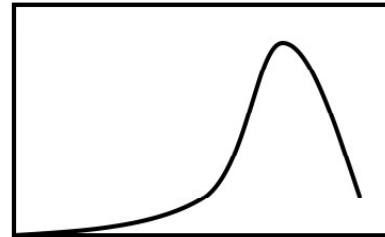


Comparing Normality

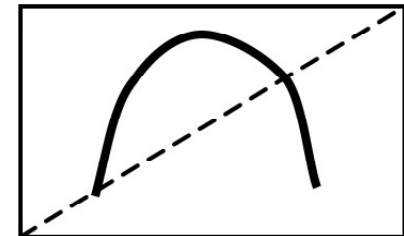
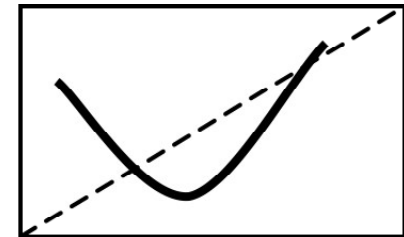
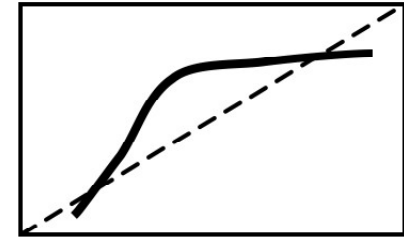
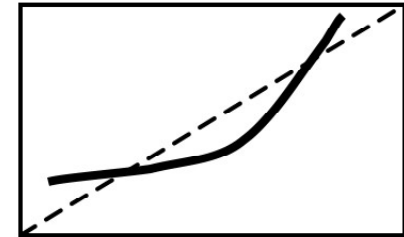
Frequency histograms vs. Quantile Plots

Note: The expected distribution of a normally distributed data set in a quantile plot is a straight line

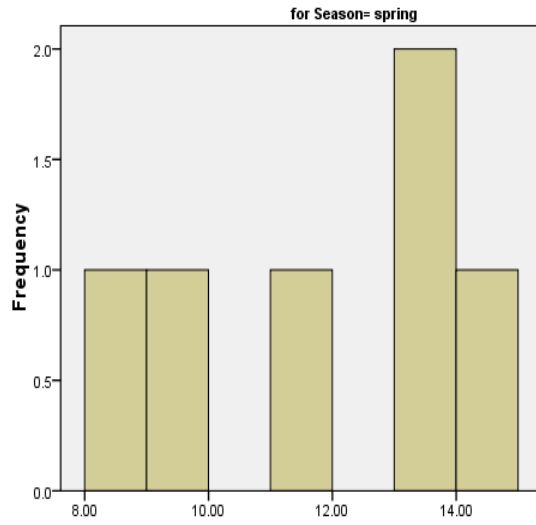
Frequency
Distribution



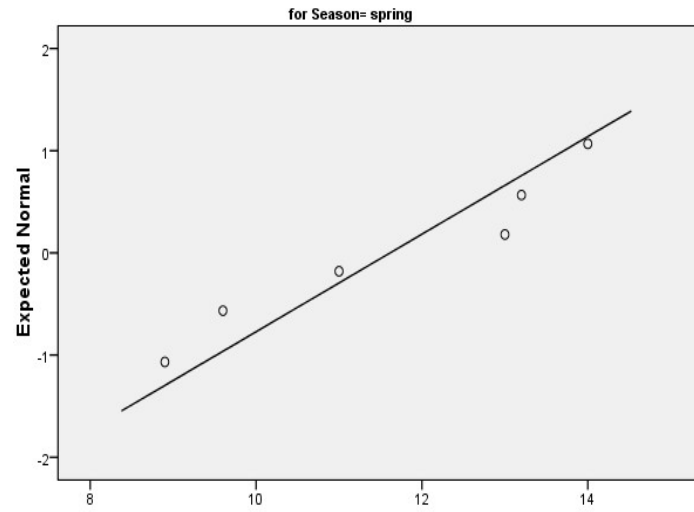
Quantile
Plot



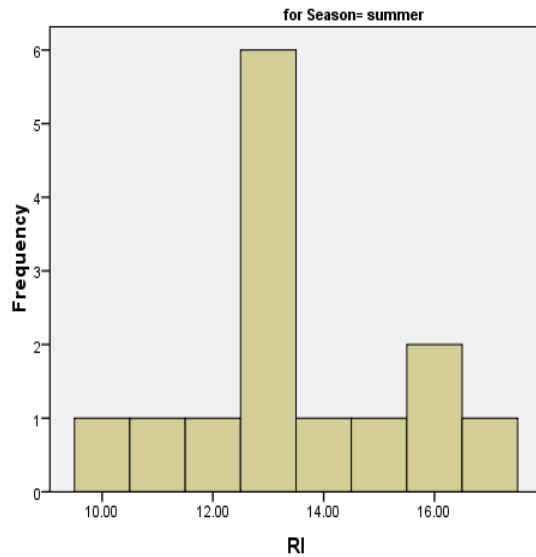
Histogram



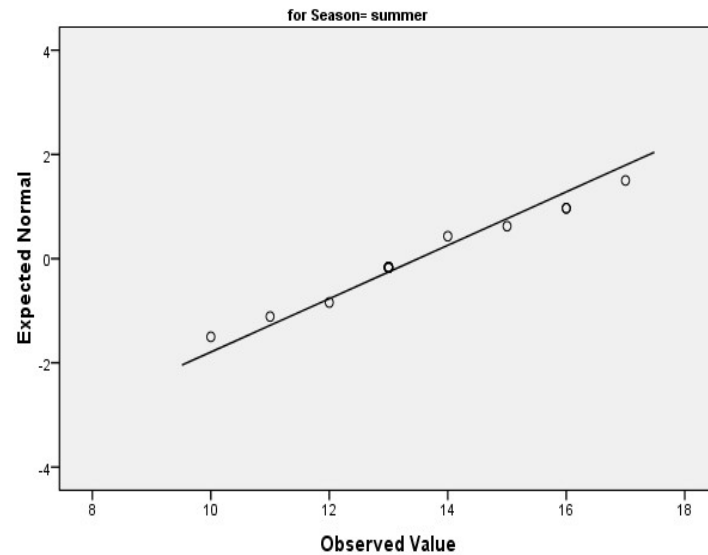
Normal Q-Q Plot of RI



Histogram



Normal Q-Q Plot of RI



Statistical Tests for Normality

- a) **test if skewness & kurtosis are each different from 0**
 - can do this with a 1-sample t-test
- b) **Shapiro-Wilk test (n < 50); or Kolmogorov-Smirnov test (n > 50)**

If Normality is violated:

- **transform data or perform a nonparametric test (e.g., on ranks)**

Tests of Normality

Season		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
RI	spring	.246	6	.200*	.910	6	.434
	summer	.244	14	.023	.935	14	.355

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

```
t.test(skewness, mu = 0) # 1-sample t-test
```

```
shapiro.test(residuals) # n < 50
```

```
ks.test(residuals, "pnorm", mean(residuals), sd(residuals))
```

Outliers

Effect: strong effect (bias) on tests that compare means

How to investigate:

- A) plot residuals vs. treatment means
- B) Histogram

Solution:

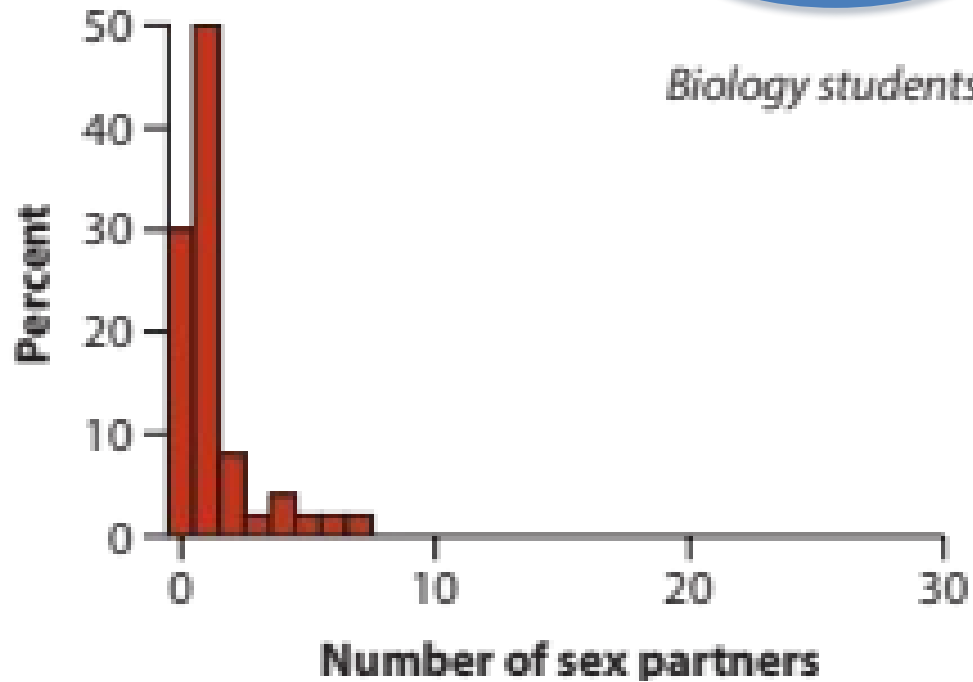
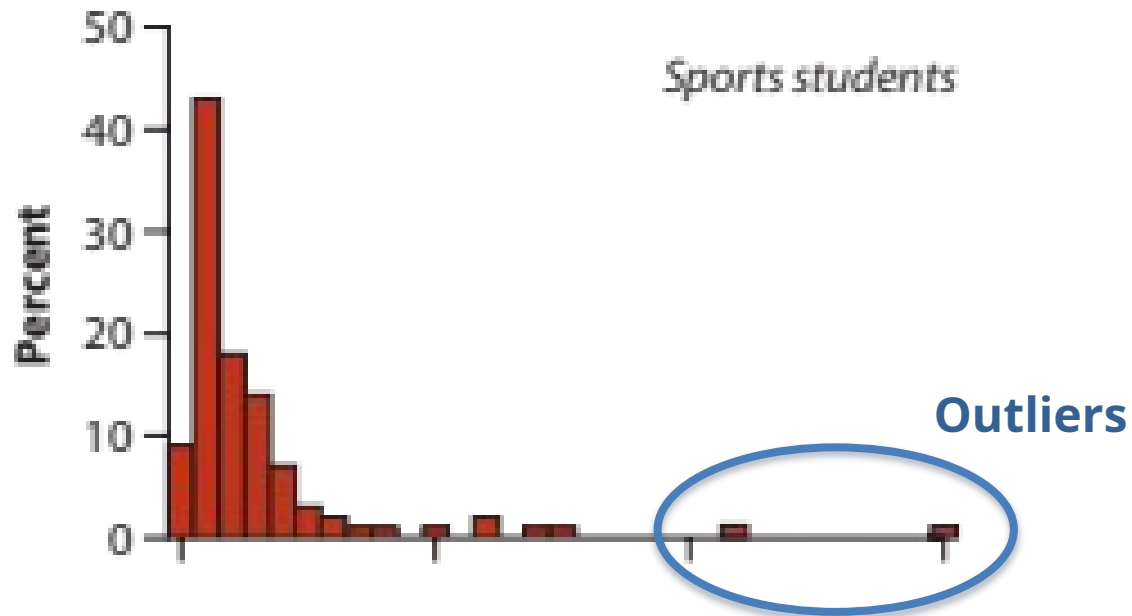
If value is impossible or can attribute it to known cause (error):

- a) omit it
- b) Winsorize: replace with mean value and subtract 1 df from error term in analysis

`DescTools::Winsorize(x, probs = c(0.05, 0.95))`

BUT ALWAYS report how you handled outliers in your document

If value is possible or you can not attribute to cause – leave it in!

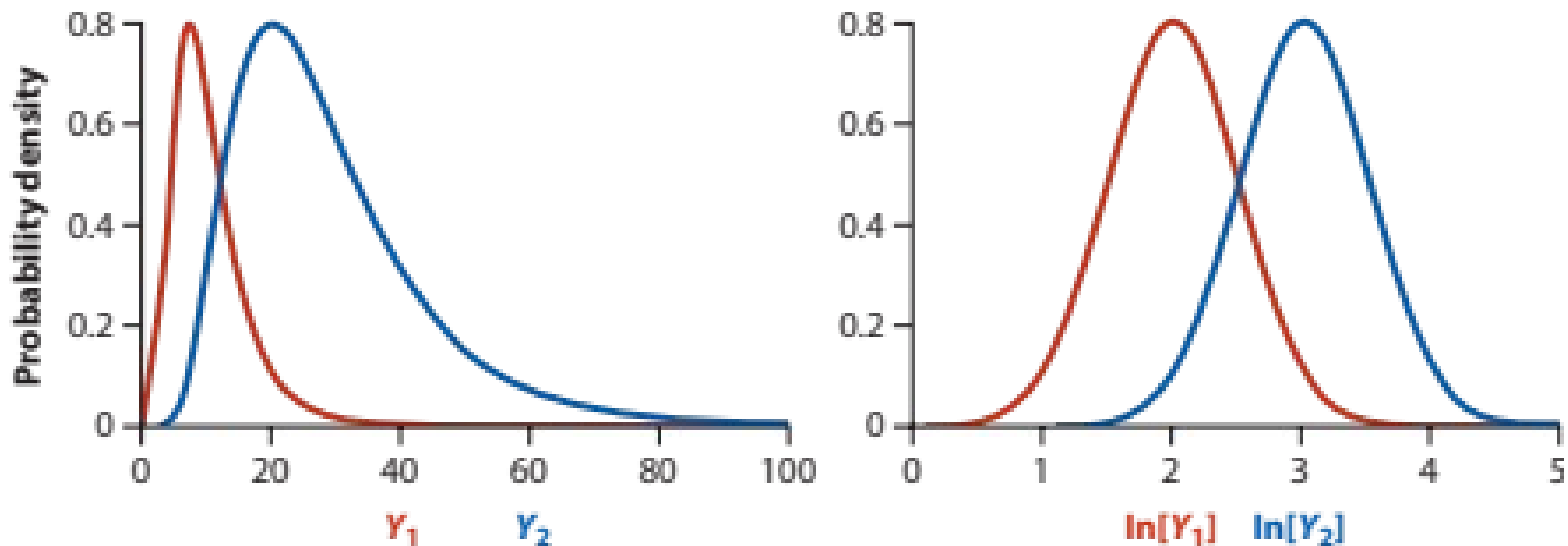


What to do when the assumptions are not met?

1. Obtain more data
2. Try to improve the fit of the data by employing a data transformation
3. Use a procedure that does not depend on parametric assumptions:
 - Rank the data & perform a non-parametric test
 - Resampling methods
 - Generalized Linear Models
 - Etc...

Data Transformations

- A data transformation changes each data point by some simple mathematical formula
- It is not “magic”, nor is it unethical



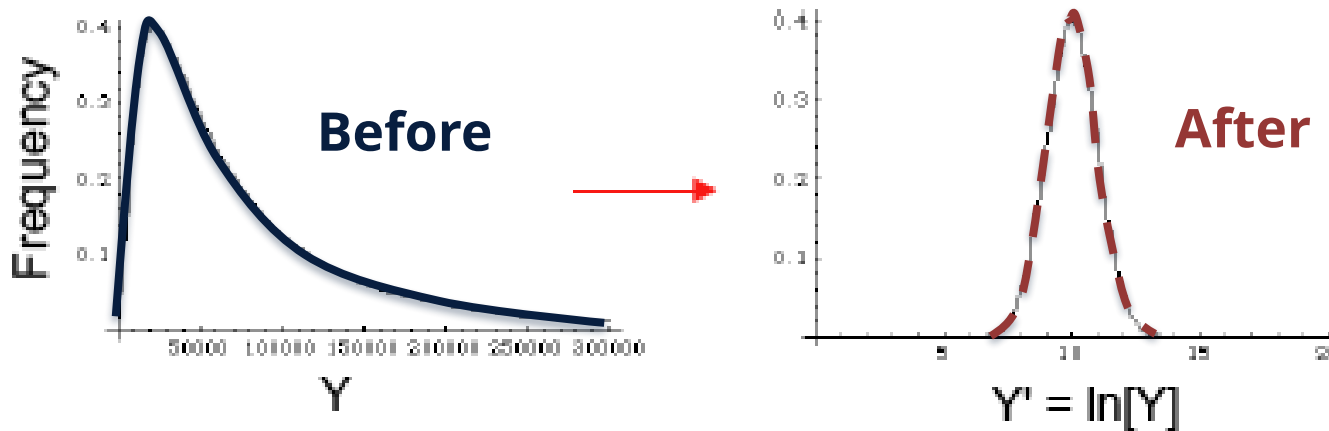
$\log(x)$ # natural log (\ln)
 $\log_{10}(x)$ # log base 10
 $\log_{1p}(x)$ # $\log(x + 1)$, safe for zeros

Valid Transformations

- Require the same transformation be applied to each data point (except for a few types of transformations where “0” and “1” are undefined)
- Must be backwards convertible to the original value without ambiguity
- Must have one-to-one correspondence to original value
- Should not be overly complex
- Theoretically there are infinite number of possible transformations, but a few are common:

Log Transformation

$$Y' = \ln[Y]$$



Various Options:

- \log_{10} or $\ln(X)$
- \log_{10} or $\ln(X + 1)$ - if zeros in data
- \log_{10} or $\ln(X \times 10 \text{ or } 100 \text{ or } 1000, \text{ etc.})$
- if values between 0 and 1

$\log(x)$; $\log_{10}(x)$; $\log_{1p}(x)$

Square Root Transformation

$$\sqrt{x_i} + 0.001 \text{ - if zeros in data}$$

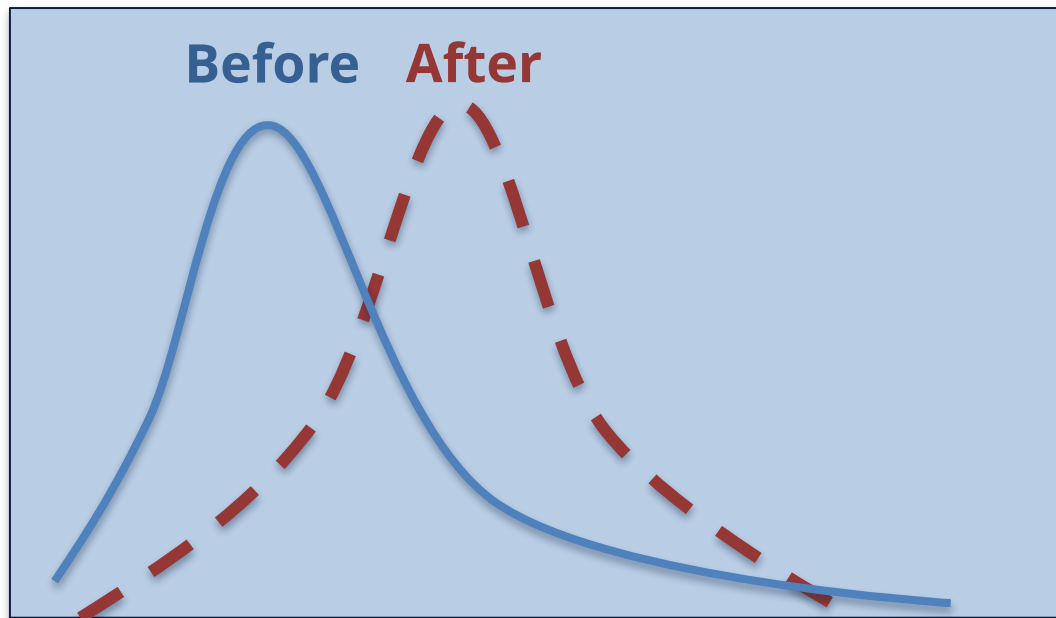
- data are positively skewed ($g_1 > 0$)
- not as “strong” effect on data as log transformation

Reciprocal Transformation

$$1/X$$

$$1/(x + 1) \quad - \text{ use if data contain zeros}$$

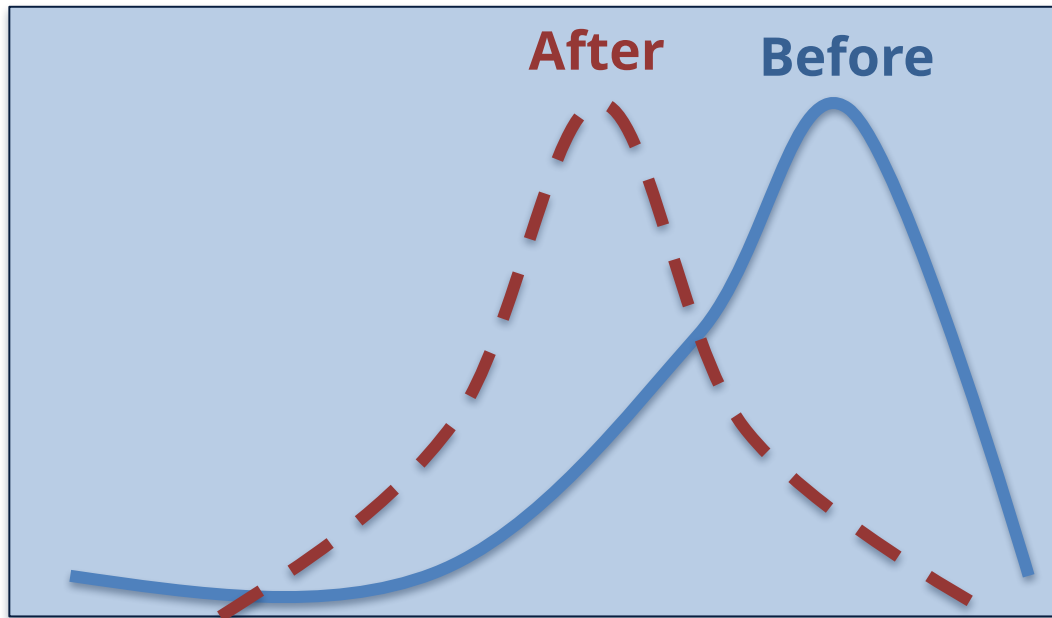
- errors are nonlinear & data positively skewed



Power Transformation

X^y (where $y = 2, 3, \text{ etc.}$)

- data distributions are negatively skewed ($g_1 < 0$)



Angular Transformation

inverse or arcsin

$$\left(\sqrt{x_i}\right)$$

$$\arcsin \left(\sqrt{\frac{1}{4N}}\right) \text{ (for 0 values)}$$

$$\arcsin \left(\sqrt{\frac{1}{1-4N}}\right) \text{ (for 1.00 values); } N$$

= sample #

- data are proportions or percentages or binomial data
- often not needed if values between 0.3 and 0.7
- data distribution often diamond shaped pattern in residual vs. mean plot
- Controversial: new stats papers don't recommend using this

Standardization

- *stdize*
- centre
- scale to SD `df$var_std <- stdize(df$var)`
- rescale 0-1 `df$var_original <- unstdize(df$var_std)`

Range Standardization

- Linear rescale to 0-1, doesn't centre on mean nor scale by SD. Preserves original distribution shape.
- Steinley, D. 2006a. K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology* 59:1–34.
- Steinley, D. 2006b. Profiling local optima in K-means clustering: developing a diagnostic technique. *Psychological Methods* 11:178–192.

```
df$var_range <- range_stdize(df$var)  
df$var_original <- unrange_stdize(df$var_range)
```

One Hot Encoding

- Aka 'binarize wide'
- Convert a single categorical variable to multiple columns of binary variables
- Used by ML models & ML dorksters

```
df |> one_hot_encode(category_column)
```



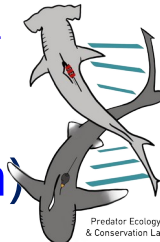



Rank Transformation

(non-parametric transformation)

Method:

- Rank each data point from lowest to highest regardless of treatment group
- DO NOT retest data to see if parametric assumptions are met. It is unnecessary because the resulting test is now considered a non-parametric test in which the assumptions of homogeneity of variances and normality no longer apply.

Thanks. Any questions?

- Projects coded in R & (aim to) require minimal R knowledge 
 - Code / figures / contact / links to all papers & thesis: simondedman.com
 - github.com/SimonDedman/gbm.auto & [/MarSpatAuto](https://github.com/SimonDedman/MarSpatAuto) & [/movegroup](https://github.com/SimonDedman/movegroup)
 - simondedman@gmail.com 
 - AI and the Future of Marine Science (medium.com/@SimonDedman) 
 - Delta sensitivity, or why your results may not reflect true biological preferences (medium.com/@SimonDedman) 
 - Please let me know criticisms/praise/suggestions by email or in person 
- Permission to use ray graphics kindly granted by Marc Dando wildlifeillustrator.com 

M. Castleton, B. Whitlock, A. Carlisle, J. Ganong, A. Boustany, S. Teo, S. Wilson, MJW. Stokesbury.
TAG: Robbie Schallert. TAG/TOPP Data Management. NOAA BTRP. Tag a Giant Fund of The Ocean Foundation, Wildlife Computers & Lotek, Inc. 

