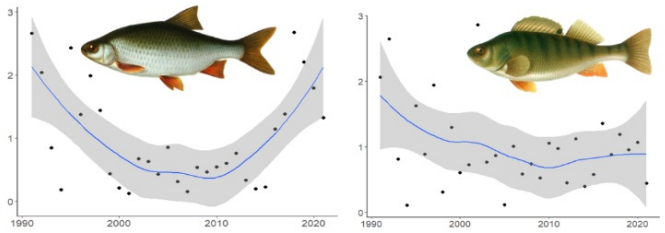
Course 1: Catch per unit effort data standardisation in R for fisheries biologists and practitioners



This short course is aimed at introducing researchers to fisheries data analysis using linear models (LM), generalized linear models (GLM) and generalized linear mixed models (GLMM) in the R working environment. Scientific monitoring and artisanal, commercial or recreational fish catch data is often used to assess population status, but such data are usually complex and require careful standardisation.

By the end of the course, participants should be able to:

- Undertake data exploration to avoid common pitfalls in tackling a data analysis
- Recognise data structures and fit appropriate models to CPUE data
- Understand and apply alternative approaches to model selection
- Interpret and present the results of statistical models

The sessions during November 23-24, 2022 will be a blend of interactive demonstrations, lectures and Q&A time. All materials delivered during the course - including lecture videos, R scripts and resources - will be freely available on this website for future use and independent learning.

The course GitHub page has a discussion group where you can share your challenges and solutions about R and package installations, statistics or other related topics. You will need a GitHub account to post on this discussion group, but creating a GitHub account is easy and useful anyway.

The course is led by Dr Carl Smith (Nature Research Centre, Lithuania and University of Lodz, Poland) with an extensive expertise in statistical analyses and teaching. Additional teaching support is provided by Dr Asta Audzijonyte (Nature Research Centre, Lithuania & University of Tasmania, Australia), Dr Catarina Silva (Nature Research Centre) and Dr Eglė Jakubavičiūtė (Nature Research Centre).

Organisation

- 2 days
- 8 sessions
- Lots of breaks

- (Interrupt)
- (Ask questions)
- Discuss

Outcomes

- Comfortable using R for CPUE standardisation
- Fit models to catch data
- Introduce a 4-step data analysis
- Data exploration and model validation
- This workshop offers a taster only...

I assume:

- Some experience in using R for statistical analysis
- Postgraduate-level understanding of statistical analysis

Why use R?

- It is the “industry standard”
- Transparency
- Integrate data collection, presentation, publication (e.g. R markdown, Shiny, etc.)
- It is free, powerful, flexible, but..
- Not intuitive

CPUE standardisation: What is it and why do we need it?

- Rarely have independent population estimates of fish abundance
- Rely on catch data for population trends to make management decisions
- **Goal of standardisation is to remove variation not due to changes in abundance**

CPUE standardisation: What is it and why do we need it?

- Multiple factors can contribute to variation:
  - Changes in effort (e.g. gear type)
  - Spatial effects (e.g. habitat quality)
  - Temporal effects (e.g. seasonal patterns or cycles)
  - Environmental effects (temperature/rainfall/upwelling, etc.)

Methods for CPUE standardisation

- LM/LMM
- GLM/GLMM
- GAM/GAMM
- Zero-inflated models
- Spatial, temporal and spatial-temporal models
- Bayesian inference
- And many other statistical tools ...
- We will *briefly* look at some of these

Approach

1. Formulate the question
2. Explore the data
3. Model the data
4. Discuss and interpret findings

Approach

1. Formulate the question
   - Ideally before data collection…
   - What data do we need?
   - How will we analyse it?

Approach

1. Formulate the question
2. Explore the data
   - Outliers
   - Collinearity
   - Zeros
   - Independence
   - Type of relationships

Approach

1. Formulate the question
2. Explore the data
3. Model the data
   - Fit model
   - Model selection
   - Model validation
   - Model interpretation

Approach

1. Formulate the question
2. Explore the data
3. Model the data
4. Discuss findings
   - What does it mean?
   - How to present results
   - How to present figures

Model fitting

A GLM comprises three components:

1. **the linear predictor** (linear function of the predictor variables)

2. **the conditional probability distribution** (distribution of the response variable across the regression line for the given set of predictor variables)

3. **the link function** (connects the linear predictor with the mean of the conditional distribution)

Model fitting

Need to choose a **conditional probability distribution** (i.e. Gaussian, binomial, Bernoulli, Poisson, negative binomial, gamma, beta, tweedie, etc.)

- **Not** based on the distribution of the raw response variable
- Instead on variable characteristics (continuous or discrete, bounded or unbounded, presence of zeros)
- Conditional distribution (largely) determines the **link function** (i.e. identity, log, logit, inverse, etc.),
- Link function can be refined as part of the model fitting process

Steps in model fitting for CPUE standardisation

- Select **response variable** (e.g. catch per fishing event)
- Conduct **data exploration**
- Select **probability distribution** for the response variable (Poisson, binomial, gamma, etc.)
- Select **link function** appropriate to the distribution
- Select appropriate **predictor variables** (year, location, season, gear type, depth, etc.)

How to deal with zero catches?

- **Do not ignore zeros** - these are critical data!
- Use an appropriate distribution that can accommodate zero observations
- Simulate from your model to ensure the model accommodates the proportion of zeros in the data
- We will do this (Hilsha analysis)

Which predictor variables to include?

- Those known (or likely) to be important
- Avoid **overparameterising** (data may not contain enough information to estimate model parameters properly)
- Ensure variables have data for all combinations and they are not correlated (e.g. season and ice cover cannot be estimated separately if ice cover is only present in winter)
- Include interactions if warranted (known interactions or highlighted by data exploration)
- Pay attention to possible Year/Season/Region x Gear interactions

North Uist trout

Simple data set (*trout.csv*) comprising:

- **Catch** of trout in kg
- **Effort** as number of rods (anglers)
- Variation due to different **season** (spring, summer, autumn, winter)
- Each sample from a different loch
- R Script in file *Trout.R*

Approach

1. Formulate the question

What variables contribute to variance in trout catch returns?