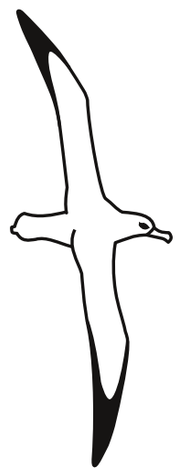
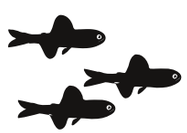
Data analysis portfolio

A selection of completed and ongoing data analysis and software projects



Philipp Hanno Boersch-Supan MRes PhD



http://pboesu.github.io
pboesu

Pelagic scattering layers in the southwest Indian Ocean

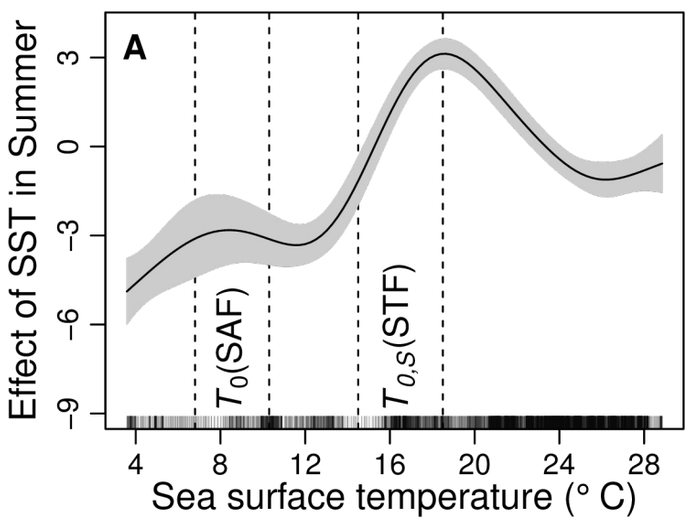
Published as http://dx.doi.org/10.1016/j.dsr2.2015.06.023

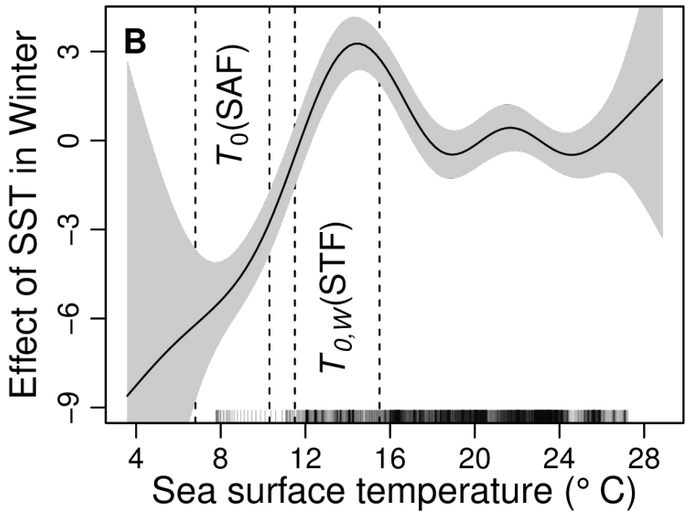
Background:
Much of the animal biomass in the open ocean is concentrated in layers which can be detected using echosounders. These so called scattering layers are often species-rich and include animals like lanternfishes, squids and deep-water prawns. They are an important prey source for predators such as tuna, oceanic sharks and marine mammals. This study investigated the distribution of scattering layers in the southwest Indian Ocean.

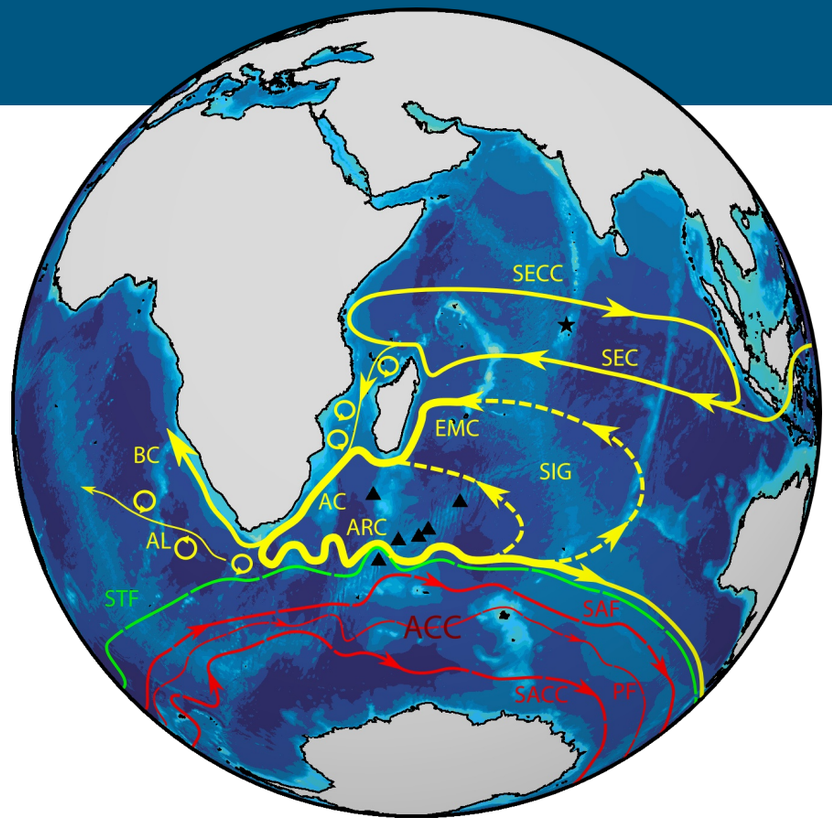
Data sources:
- c. 40,000 echosounder profiles covering over 35,000 km of survey track
  - c. 12,000 collected in the field by me
  - c. 28,000 from the Integrated Marine Observing System database
- gridded satellite sensed environmental data products

Software tools:
- Acoustic data pre-processing with Echoview automized using COM scripting in Windows
- geodata and raster manipulation in GRASS GIS and R
- Data extraction from gridded products using netCDF Operators and raster package in R
- map creation with R, GMT, and Inkscape
- statistical analyses in R

Analysis approach:
- Generalised Additive Mixed Models with autoregressive error structures
- k-medioids clustering







Top: Important surface currents and fronts of the Southwest and Central Indian Ocean. Triangles mark the locations of trawl sampling stations. Dashed lines mark the subsurface return flow of the Southern Indian Subtropical Gyre. I created this map for the introductory chapter of my dissertation.

Left: Smooth terms of generalized additive model showing the effect of various continuous variables on backscattering strength at 38 kHz. The shaded areas are standard errors of the estimated smooths, taking into account the error in the model intercept. Dashed vertical lines in panel A and B indicate the axial temperature ranges T_0 of the subantarctic (SAF) and subtropical fronts (STF), respectively. Subscripts S and W, denote STF Summer and Winter T_0 ranges, respectively.

My contribution:
This study is one of my PhD dissertation chapters, and was conducted together with my advisors Alex Rogers (University of Oxford) and Andrew Brierley (University of St Andrews). I lead all parts of this project including data collection at sea, synthesis of remotely sensed environmental data sets and additional echosounder data from public databases, statistical analyses, figure and manuscript preparation.

Recovery of seabird breeding assemblages after predator eradication

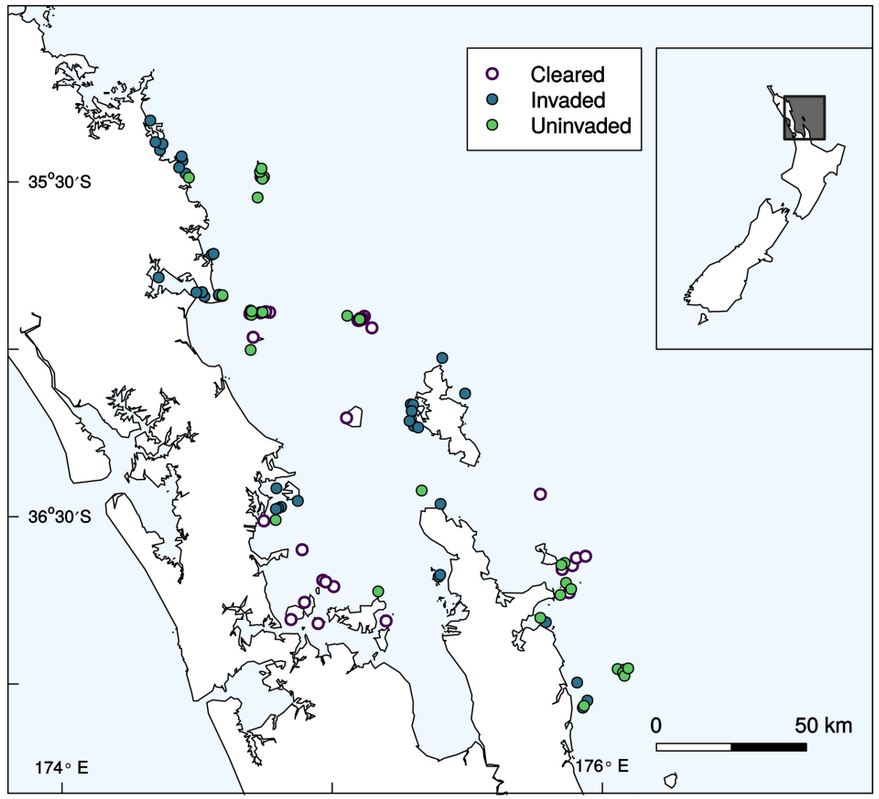
Published as https://doi.org/10.1017/S0030605316000880

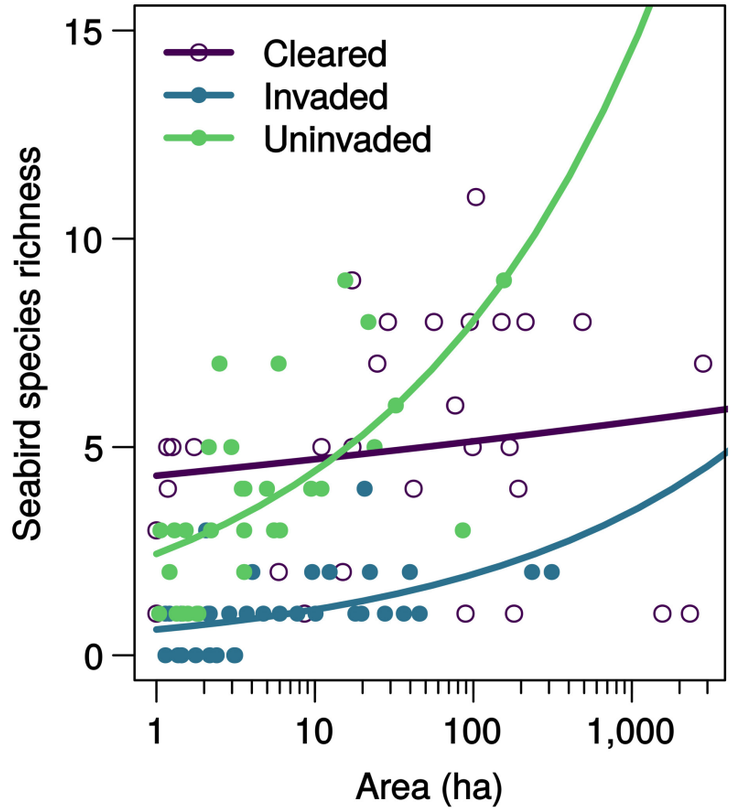
Background:
One of the most acute threats to seabirds are introduced predators, which depredate seabirds at all life stages. Predator eradication is an effective and commonly used approach to seabird conservation. This study investigated the recovery of seabird breeding assemblages on 98 islands cleared of predators in the Hauraki Gulf, New Zealand, which is a hotspot of seabird diversity.

Data sources:
- Breeding bird surveys (Species presence-absence)
- Predator presence, eradication and land-use histories
- Coastline data

Software tools:
- geodetic distance calculations in R
- map creation with R
- statistical analyses in R
- collaborative analysis development on github.com

Analysis approach:
- Generalised Linear Models (Maximum likelihood estimation)
- Chi-squared tests





Left: The Hauraki Gulf, New Zealand, study area and locations and predator eradication status of surveyed seabird islands.

Right: Observed species richness and GLM predictions of species richness as a function of island area for uninvaded islands (intercept = 2.43, slope = 0.26), cleared islands (intercept = 4.31, slope = n.s.), and invaded islands (intercept = 0.62, slope = 0.25).

My contribution:
This was a collaborative study led by Stephanie Borrelle (Auckland University of Technology), and with Chris Gaskins (Northern NZ Seabird Trust), and David Towns (AUT). I led the statistical modelling component, prepared maps and figures, and co-wrote the manuscript.

Biogeography of free living marine bacteria

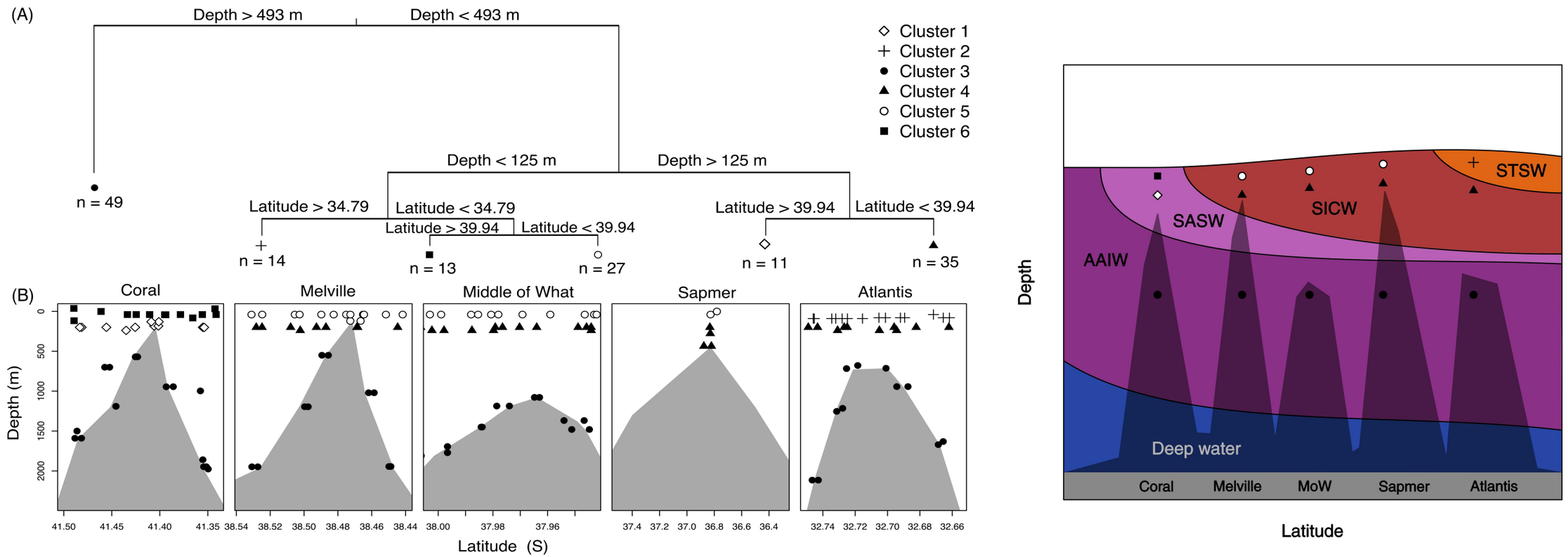
Published as http://dx.doi.org/10.1098/rsos.170033

Background:
This study investigated microbial communities across a dynamic frontal zone in the southwest Indian Ocean and investigate the spatial differences of the microbial community composition with respect to depth and the different water masses separated by oceanic fronts.

Data sources:
- Bacterial taxon abundance from 16S rRNA meta-barcoding data
- Environmental covariates from in-situ oceanographic measurements

Software tools:
- data analysis in R
- scientific illustrations in Inkscape

Analysis approach:
- Multivariate regression trees

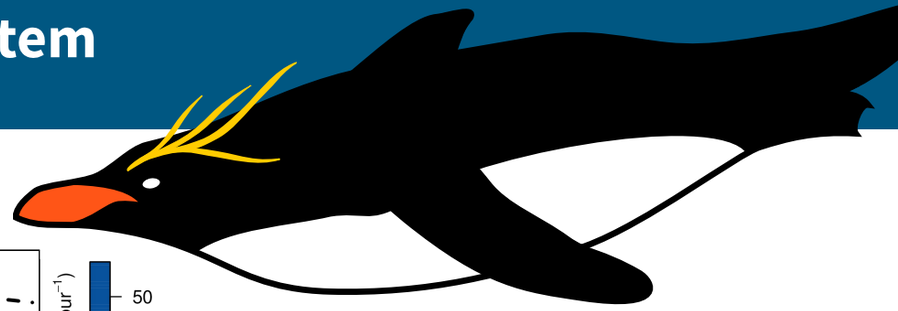


Left: (A) Multivariate regression tree of microbial communities and their structuring by latitude and depth. (B) Locations of MRT clusters across seamounts. **Right:** Schematic of water masses of the Southwest Indian Ridge. Distinct microbial assemblages are illustrated using the same symbols as in (A) and (B).

My contribution:
This was a collaborative study led by Anni Djurhuus (University of Oxford), and with Svein-Ole Mikalsen (University of the Faroe Islands) and Alex Rogers (Oxford). I conducted the regression tree analysis and prepared figures and illustrations. I also helped troubleshooting the bioinformatics pipeline and contributed to the manuscript.

Analysis framework for an automated penguin weighing system

Work in progress

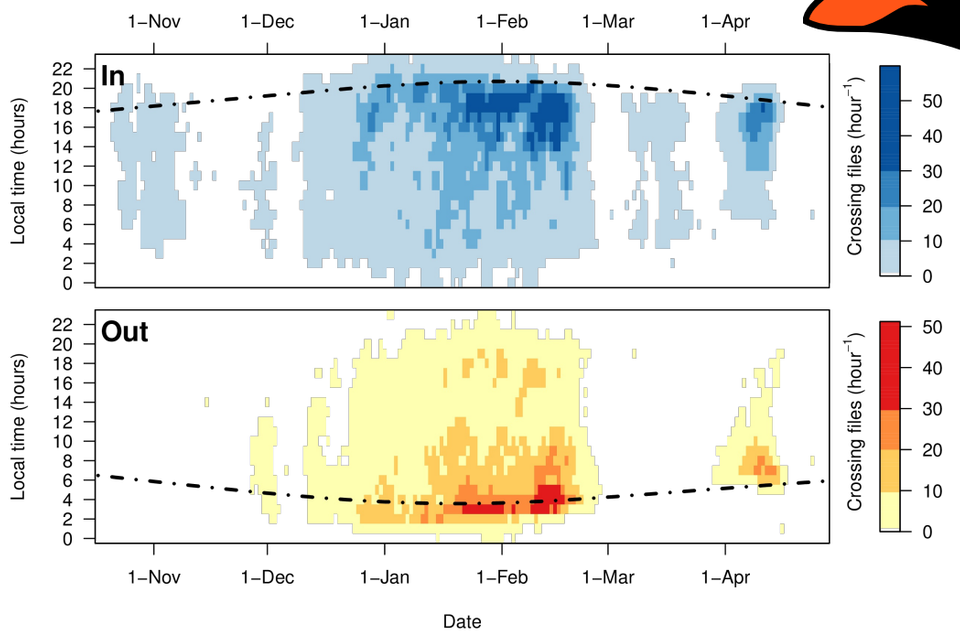


Background:
Climate change is shifting the phenology of animals at all trophic levels and quantifying these changes, and species' potential for phenological plasticity is important for understanding how populations will respond to these changes. This project is aimed at developing an analysis framework for an automated weighing and RFID system to non-invasively monitor a colony of c. 400 pairs of marked and unmarked macaroni penguins.

Data sources:
- over 300,000 high-resolution time series of ground reaction force measurements comprising c. 200,000,000 rows of raw data from multiple sensors.
- over 100,000 RFID tag detections
- diet samples, colony count data, demographic information provided by field technicians
- Remotely sensed gridded environmental covariates

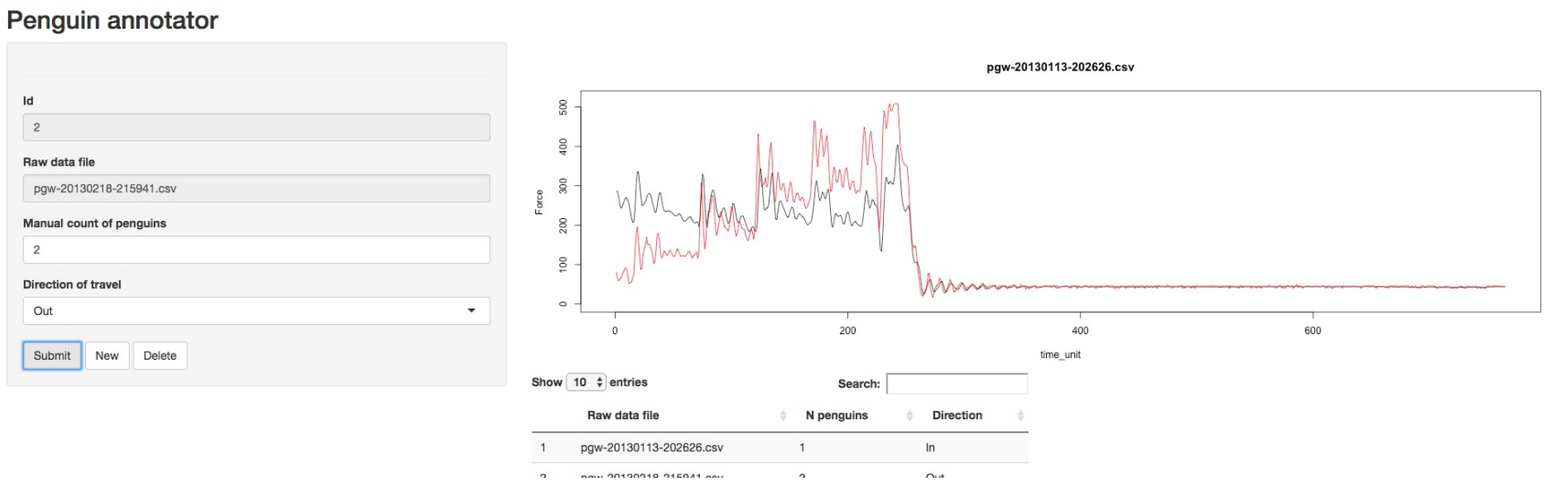
Software tools:
- data is archived in an Oracle database at the British Antarctic Survey
- data processing and analysis with R/Rcpp, using parallelized queries to a local PostgreSQL copy of the archival database
- annotation of test/training data for the classification procedure with a purpose-built R shiny app
- figure preparation with R and Inkscape

Analysis approach:
- Supervised classification using a dynamic time warping sequence similarity measure (see rucrdtw case study)
- Generalised Linear Mixed Models (ML estimation)



Top: Average hourly counts of penguin crossings into and outof the colony reveal a morning and evening "rush hour" on the weighbridge.

Bottom: Screenshot of the interactive R shiny app created to annotate raw weighbridge crossing records. Determining the number of marked and unmarked penguins and their direction of travel in raw data time series is easy for the human eye, but hard for computers. The shiny app simplified and sped up the creation of test and reference data sets for the classification procedure.



My contribution:
This is a collaborative project with Phillip Trathan and Helen Peat at the British Antarctic Survey. I developed the raw data classification procedure and am currently implementing statistical models to quantify and visualise seasonal body mass dynamics, provisioning effort, and changes in phenology. I am furthermore leading the drafting of a corresponding manuscript.

Data analysis portfolio of Philipp Boersch-Supan @pboesu

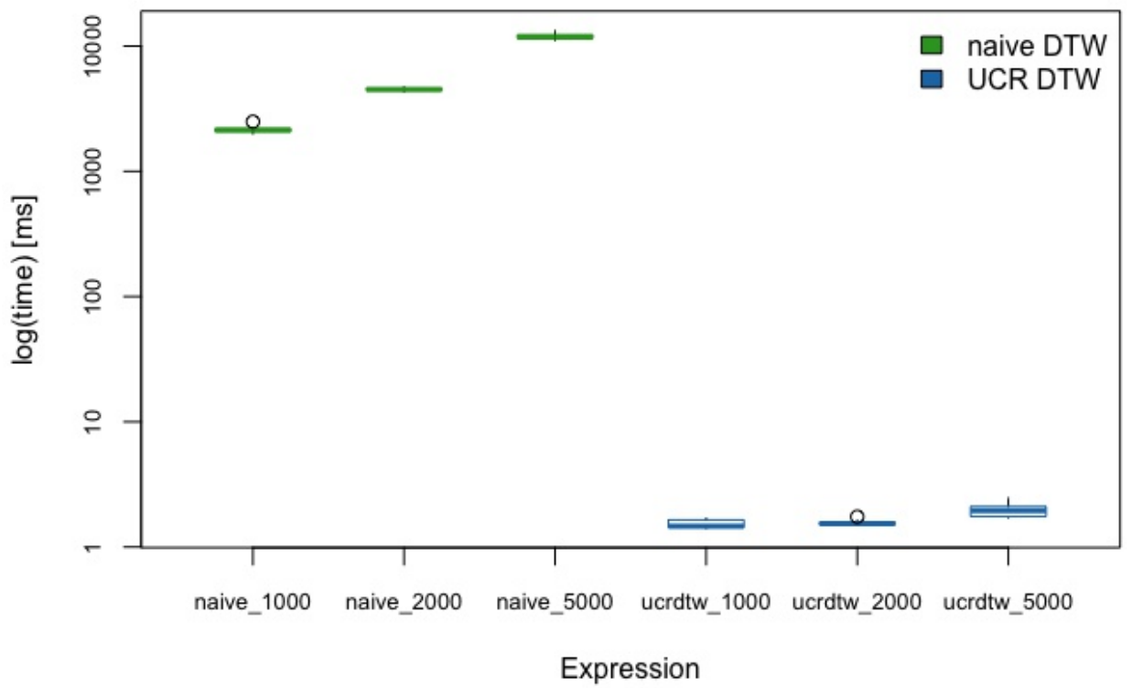
Software tool: rucrdtw - Fast time series subsequence search in R

Published as http://dx.doi.org/10.21105/joss.00100 Software available at https://CRAN.R-project.org/package=rucrdtw

Motivation:
Need for a fast method to classify over 300,000 records of penguins crossing an automated weighing system. Comparing records to manually annotated reference data sets using dynamic time warping (DTW) distances proved to be the most accurate of the trialled approaches, but existing DTW implementations in R were prohibitively slow for the size of the data set.
The highly optimized algorithm for fast multiple DTW comparisons by Rakthanmanon et al. provided a solution but a seamless integration into the existing R and PostgreSQL workflow for the penguin data was lacking.

Implementation:
rucrdtw reimplements previously published C++ code using Rcpp. It provides a fully documented R API to the algorithms and convenience functions for a variety of input data structures.

Code and documentation were peer-reviewed as part of the publication in the Journal of Open Source Software.



Benchmark timings demonstrating that UCR DTW is approximately three orders of magnitude faster than a naive sliding-window search using DTW distance.

My contribution:
I am the creator and CRAN maintainer of the R package, which reimplements the algorithms developed by Rakthanmanon et al. I wrote the documentation, prepared example datasets, and authored the application paper.

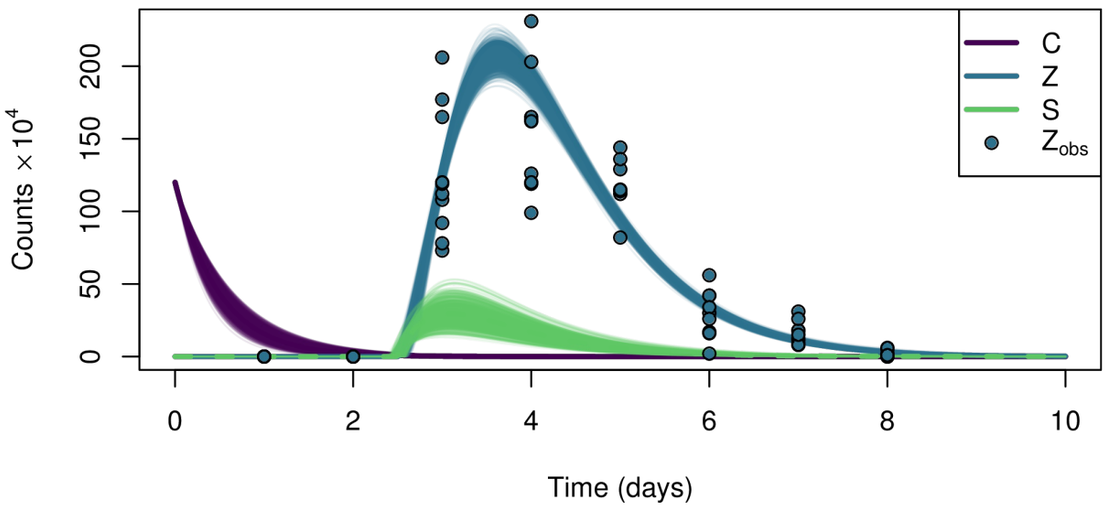
Software tool: deBInfer - Bayesian inference for differential equation models in R

Published as http://doi.org/10.1111/2041-210X.12679 Software available at https://CRAN.R-project.org/package=deBInfer

Motivation:
Need to conduct parameter inference for ordinary and delay differential equation models, such as dynamic energy budget models and disease dynamics models by myself and other members of the Johnson Lab group.

Implementation:
- Metropolis-Hastings sampler implemented in R
- Likelihood evaluation makes use of numerical solvers provided by the deSolve package
- deBInfer provides templates for differential equation models (implemented in R, C, or Fortran) and Bayesian model specification
- deBInfer provides convenience functions for plotting MCMC diagnostics and inference outputs

Code and documentation were peer-reviewed as part of the publication in Methods in Ecology and Evolution.



Posterior model fits for a delay differential equation model of the population dynamics of the fungal pathogen *Batrachochytrium dendrobatidis*. The model was parameterised using observations (Z_{obs}) of a single state variable.

My contribution:
This was a collaborative project with Leah Johnson (Virginia Tech) and Sadie Ryan (University of Florida). I reimplemented a previous version of the software written by L Johnson as an R package, extended the MCMC sampler and convenience functions, wrote most of the documentation and vignettes, prepared figures, and lead the writing of the application paper. I am the package maintainer on CRAN.

Parameter inference for dynamic energy budget models of seabirds

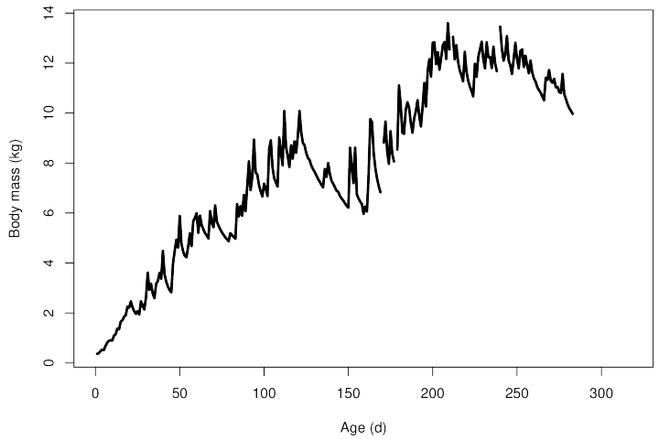
Work in progress Related slide set published at https://doi.org/10.6084/m9.figshare.1591048.v1

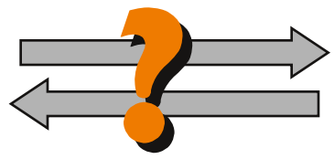
Background:
Chick growth trajectories provide an integrated measure of physiological, behavioural, and environmental processes. This project builds an analytical framework grounded in metabolic theory to examine chick growth in seabirds. Our approach allows us to make a mechanistic link between energy intake and life history processes, such as growth, development, and survival. It provides a generic approach to study the impact of environmental conditions on the life cycle of seabirds.

Data sources:
- Seabird growth curves and trait data obtained from collaborators, digitized from archival records, the literature, and public trait databases.

Software tools:
- digitisation of paper records with ScanTailor and WebPlotDigitizer
- data archaeology with emulated legacy operating systems and/or custom parsers to liberate data from legacy digital files
- Reimplementation and extension of existing MATLAB code for Dynamic Energy Budget models in C and Stan
- Parameter inference with deBInfer and Stan

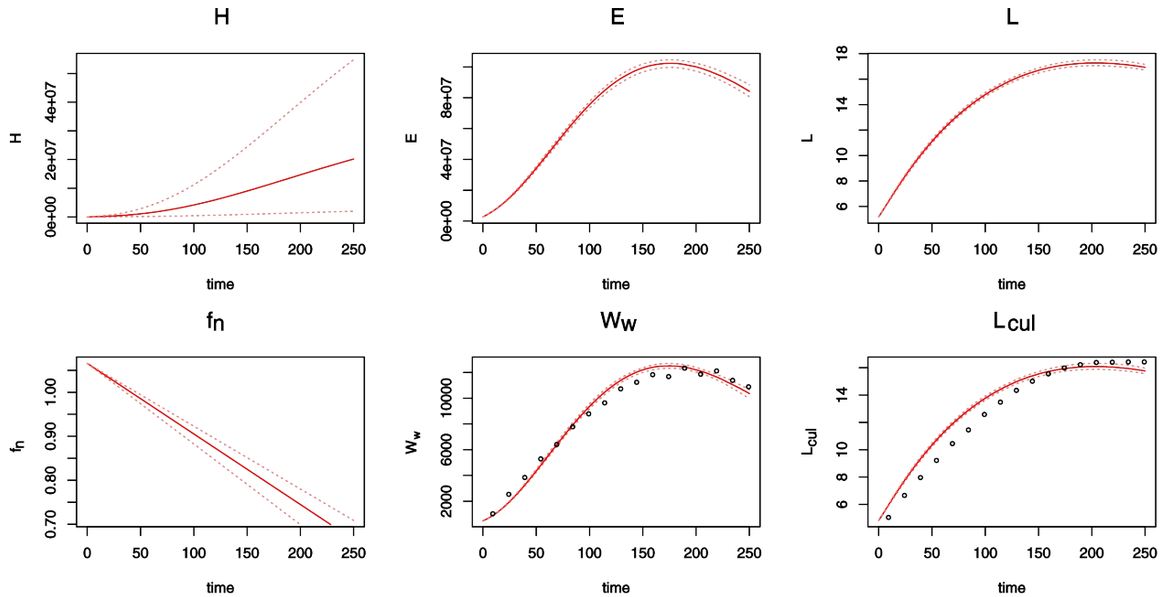
Analysis approach:
- Bayesian parameter inference for differential equations





$$\frac{dL}{dt} = \frac{\dot{r}L}{3}$$
$$\frac{dE}{dt} = \{\dot{p}_{Am}\} f L^2 - \dot{p}_C$$
$$\frac{dE_H}{dt} = (1-\kappa)\dot{p}_C - \dot{k}_J E_H$$
$$\frac{df}{dt} = f_{slope}$$

Growth in seabirds is neither asymptotic, nor continuous (**top right**), preventing the use of commonly used growth models such as the von-Bertalanffy model. Mechanistic models, e.g. from dynamic energy budget theory (**top right**), can capture complex growth trajectories, but linking models to data is statistically challenging.



Left: Fitted DEB model for wandering albatrosses. The fit is based on Bayesian inference that incorporates prior information from physiological and other trait data.

My contribution:
This is a collaborative project with Leah Johnson (Virginia Tech), Sadie Ryan (UF), and Richard Phillips (British Antarctic Survey), and the core research project of my postdoctoral appointment. I adapted mathematical models from the literature, implemented them in R, C, and Stan, collated data from various source for model fitting and implemented the inference procedure. I am in the process of conducting parameter inference for multiple species of albatrosses and petrels, and am drafting a manuscript based on the results.

Environmental drivers of ranavirus prevalence in vernal pools

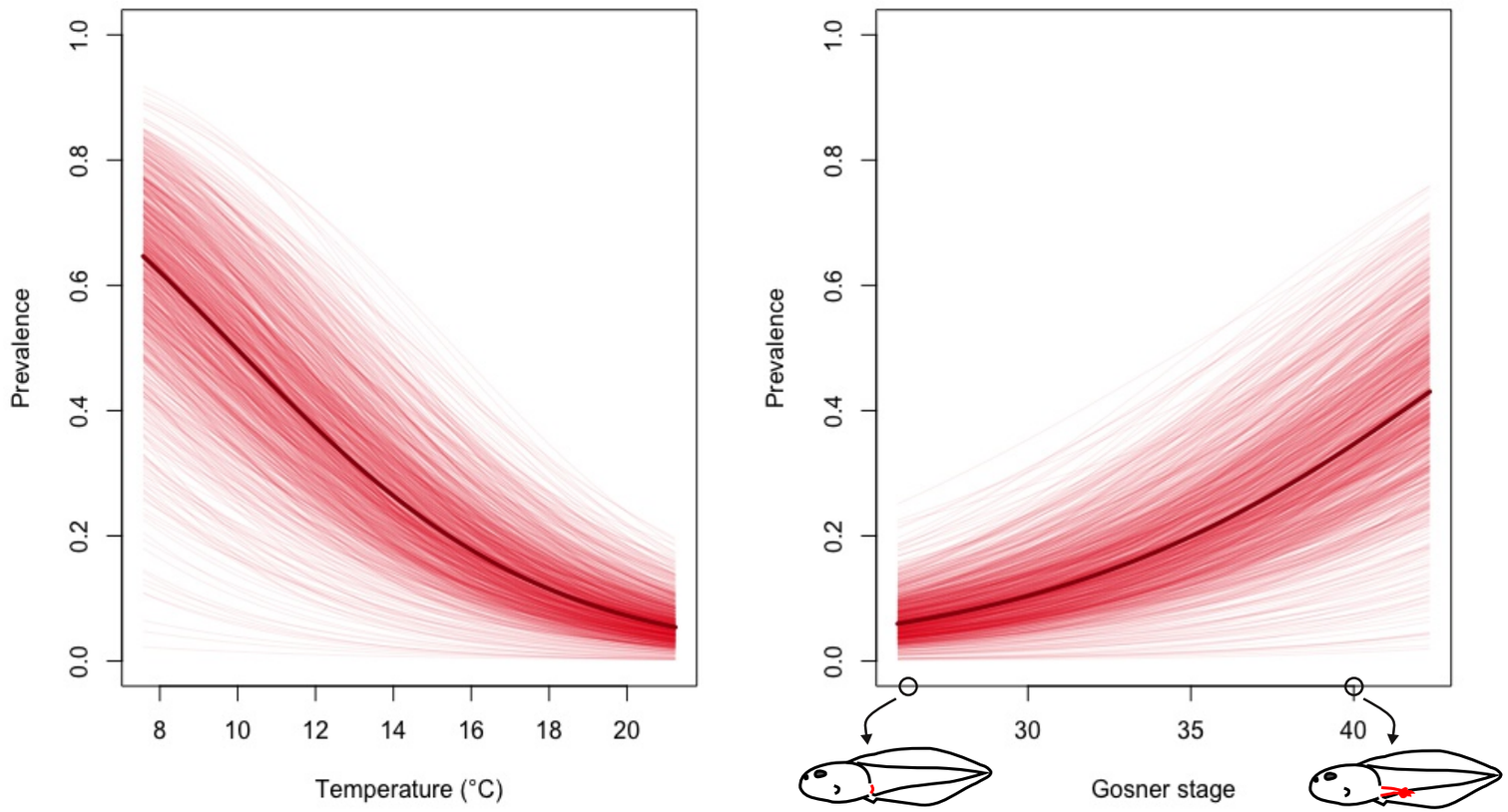
Work in progress

Background:
Mechanisms of emergence for ranaviruses in wild populations are poorly understood, despite the substantial mortality that outbreaks can cause in larval amphibians. This study investigates potential environmental drivers of ranavirus prevalence in wild frog populations in vernal pool habitats.

Data sources:
- Rana virus prevalence data from PCR assays
- Environmental covariates from field surveys
- High degree of missingness in covariates

Software tools:
- Data cleaning in R
- Statistical analyses in R and Stan
- Figure preparation with R and Inkscape

Analysis approach:
- Generalised Linear Mixed Models with joint modelling of missing data (Bayesian estimation)



Plots of marginal effects of temperature (left) and developmental stage (right) on ranavirus prevalence in wild wood frog populations. Bold lines show the posterior mean, thin lines show 1000 posterior realisations of the linear predictor from a multivariate binomial GLMM. Cartoons show tadpole morphologies near the extremes of the observed developmental range.

My contribution:
This is a collaborative study led by Tess Youker-Smith (SUNY-ESF) and Sadie Ryan (UF). I am performing data cleaning, statistical modelling, and figure preparation.