

New statistical methods for eDNA data
Statistical Methods for Post-Genomic Data 2021
25-26 Jan 2021 Virtual Edition (France)

Dr Eleni Matechou,
University of Kent

25/01/21



COLLABORATORS



Dr Alex Diana
Statistical Ecology @ Kent
School of Mathematics, Statistics and Actuarial Science
University of Kent



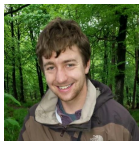
Great crested newt (Triturus cristatus)



Dr Dimitrios Bormpoudakis
Durrell Institute of
Conservation and Ecology
University of Kent



Dr Alex Bush
Lancaster
Environment Centre
University of
Lancaster



Dr Andrew Buxton
Durrell Institute of
Conservation and Ecology
University of Kent



Professor Jim Griffin
Department of Statistical
Science
UCL



Professor Richard Griffiths
Durrell Institute of
Conservation and Ecology
University of Kent



Professor Douglas Yu
School of Biological
Sciences
University of East
Anglia

OUTLINE

MONITORING OF WILDLIFE SPECIES

eDNA

SINGLE SPECIES

MULTIPLE SPECIES

CONCLUSIONS AND FUTURE WORK

MONITORING OF WILDLIFE SPECIES

eDNA

SINGLE SPECIES

MULTIPLE SPECIES

CONCLUSIONS AND FUTURE WORK

- ▶ With many species under threat and biodiversity in decline, the need to assess the state of wildlife populations and intervene as appropriate has never been greater.
- ▶ However, monitoring of wildlife species entails many challenges, one of which is that species absence or presence from a particular site cannot be easily verified.

SPECIES MONITORING

○○●○

eDNA

○○○○○

SINGLE SPECIES

○○○○○○○○○○○○○○○○

MULTIPLE SPECIES

○○○○○○

CONCLUSIONS

○○○○

Reality



Observ



Outcome



- ▶ Large scale monitoring is also expensive and slow and long-term monitoring of a large number of species can be unsustainable.
- ▶ Therefore, there is a need to improve sampling methods by relying less on human expertise and time and more on technological advancements.

MONITORING OF WILDLIFE SPECIES

eDNA

SINGLE SPECIES

MULTIPLE SPECIES

CONCLUSIONS AND FUTURE WORK

- ▶ Environmental DNA (eDNA) is a survey tool with rapidly expanding applications for assessing presence of a wildlife species at surveyed sites.
- ▶ Since the initial proof of concept by Ficetola et al. (2008)¹, the use of eDNA for the assessment of aquatic biodiversity has been rapidly expanding.

¹Ficetola, G. F., Miaud, C., Pompanon, F. and Taberlet, P. (2008) Species detection using environmental DNA from water samples. *Biology Letters*, 4, 423-425.

- ▶ In essence, the eDNA survey method isolates DNA that has become separated from an organism and suspended within the water column, to identify the recent presence of that species within a waterbody.



- ▶ eDNA surveys are now being enshrined within policy and commercial practice.
- ▶ Commercial and political decision-making has started to rely solely on results from eDNA surveys to assess species presence at surveyed sites, whether this be in management decisions around the introduction of invasive species of Asian carp in the USA (Jerder et al. 2011²) or development mitigation decisions by Natural England³ surrounding protected species such as the great crested newt in the UK.

²Jerde, C. L., Mahon, A. R., Chadderton, W. L. and Lodge, D. M. (2011) "Sight-unseen" detection of rare aquatic species using environmental DNA. Conservation Letters, 4, 150-157

³Natural England (2017) Wildlife licensing newsletter March 2017. Tech. rep., Natural England, Peterborough, UK.

However, eDNA methodology is not error-free and both false positive and false negative errors are possible in the two stages of an eDNA survey: the data collection stage (stage 1) and laboratory analysis stage (stage 2).

MONITORING OF WILDLIFE SPECIES

eDNA

SINGLE SPECIES

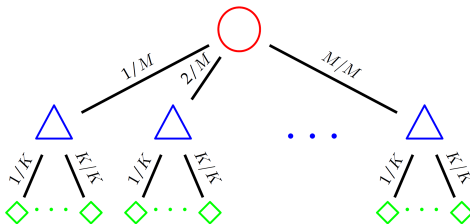
MULTIPLE SPECIES

CONCLUSIONS AND FUTURE WORK

PCR

- ▶ PCR (polymerase chain reaction) is a powerful procedure in which small quantities of DNA are amplified and detected.
- ▶ Quantitative PCR (qPCR) is a method by which the amount of the PCR product can be determined, in real-time.
- ▶ In DNA-based monitoring surveys, a PCR run is positive(negative) if the target DNA is amplified above(below) a certain threshold.

QPCR SAMPLES



Your Pond ID



Test of species presence



Kit ID	Pond ID	Date arrived	GCN Status	eDNA Score	Inhibition	Degradation
GCN1111	Pond 2	22/04/16	Positive	1/12	No	No
GCN1112	Pond 3	22/04/16	Negative	0/12	No	No
GCN1113	Pond 1	22/04/16	Positive	7/12	No	Yes
GCN1114	Pond 5	22/04/16	Inconclusive	0/12	Yes	No
GCN1115	Pond 4	22/04/16	Positive	12/12	No	No



Number of PCR replicates that were positive



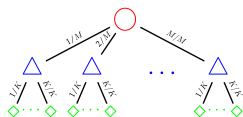
Quality Control

NEW MODEL

- ▶ In Griffin et al. (2019)⁴, we proposed a new Bayesian model for single-species eDNA data that can account for false positive and false negative errors in both stages of eDNA surveys.
- ▶ The model is a multi-scale occupancy model, extending the work by Guillera-Arroita et al. (2017)⁵.
- ▶ In some cases, records that confirm species presence at the site may be available and we showed how such records can be incorporated in the model.

⁴Griffin, J. E., Matechou, E., Buxton, A. S., Bormpoudakis, D., and Griffiths, R. A. (2019). Modelling environmental DNA data; Bayesian variable selection accounting for false positive and false negative errors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

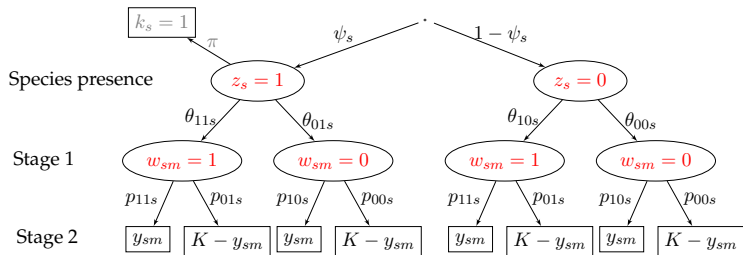
⁵Guillera-Arroita, G., Lahoz-Monfort, J. J., Rooyen, A. R., Weeks, A. R. and Tingley, R. (2017) Dealing with false-positive and false-negative errors about species occurrence at multiple levels. *Methods in Ecology and Evolution*, 8(9), 1081-1091.



Defining $z_s = 1$ if a species is present at site s (and zero otherwise), and $w_{sm} = 1$ if eDNA is present in the m -th sample of the s -th site (and zero otherwise), the model can be written in hierarchical form as

$$\begin{aligned}
 z_s &\sim \text{Bernoulli}(\psi_s), \\
 k_s | z_s = 1 &\sim \text{Bernoulli}(\pi), & \mathbb{P}(k_s = 1 | z_s = 0) &= 0, \\
 w_{sm} | z_s = 1 &\sim \text{Bernoulli}(\theta_{11s}), & w_{sm} | z_s = 0 &\sim \text{Bernoulli}(\theta_{10s}), \\
 y_{sm} | w_{sm} = 1 &\sim \text{Binomial}(K, p_{11s}), & y_{sm} | w_{sm} = 0 &\sim \text{Binomial}(K, p_{10s}),
 \end{aligned}$$

where π is the probability that an occupied site has an associated confirmed species presence.



Schematic representation of our model. Unobservable states are represented by ellipses and data by rectangles.

- ▶ The model suffers from a likelihood symmetry: four solutions in terms of the model parameters give rise to the same likelihood function value.
- ▶ These likelihood symmetries make the model only locally identifiable, since there exist a (countable) number of equally supported solutions (Cole et al. 2010⁶).

⁶Cole, D. J., Morgan, B. J. and Titterton, D. (2010) Determining the parametric structure of models. *Mathematical biosciences*, 228, 16-30

Table of likelihood function symmetries.

Solution	ψ_s	θ_{11}	θ_{10}	p_{11}	p_{10}
1	a	b	c	d	e
2	a	$1 - b$	$1 - c$	e	d
3	$1 - a$	c	b	d	e
4	$1 - a$	$1 - c$	$1 - b$	e	d

- We proposed a set of prior distributions for the regression coefficients that overcomes the identifiability issues of the model, introduced by the likelihood function, and enables us to estimate the probability of species presence at a site without requiring additional sources of information.

- ▶ We exploited the Pólya-Gamma (Polson et al. 2013⁷) data augmentation scheme for logistic regression models, which allows us to efficiently update the model with regression coefficients marginalized out, avoiding the use of trans-dimensional algorithms, such as reversible jump MCMC (Green 1995⁸), that require careful tuning.
- ▶ Our model selection process guarantees that the prior constraints placed on the regression coefficients are always satisfied, regardless of the combination of covariates that are included in the model.

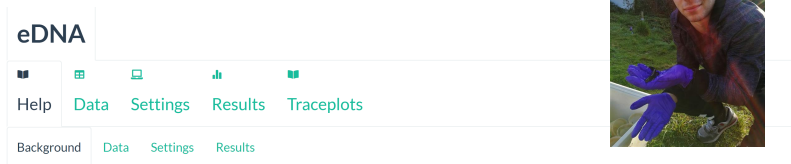
⁷ Polson, N. G., Scott, J. G. and Windle, J. (2013) Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, 108, 1339-1349.

⁸ Green, P. (1995) Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82, 711

R-SHINY APP

The modelling framework has been implemented by Dr Alex Diana into a freely available R-Shiny app <https://www.biorxiv.org/content/10.1101/2020.12.09.417600v1.full>

<https://seak.shinyapps.io/eDNA/>



This app implements the methods developed by Griffin, J. E., Matechou, E. Buxton, A. S., Bormpoudakis, D. and Griffiths, R. A., in *Modelling environmental DNA data: Bayesian variable selection accounting for false positive and false negative errors*, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

The method is appropriate for modelling **eDNA scores** (i.e. the number of positive qPCR replicates) and **accounts for the probabilities of false positive and false negative errors** in the field (data collection - stage 1) and in the lab (data analysis - stage 2) when **estimating the probability of species presence using single-species eDNA data**.

The model is fitted using a **Bayesian approach** and **all model parameters**, that is the probability of species presence and the probabilities of error in both stages, **can be modelled as functions of covariates** with the important covariates for each probability identified using **Bayesian variable selection**.

GREAT CRESTED NEWT DATA

- ▶ Samples were collected as part of a national distribution modelling assessment for great crested newts, commissioned by Natural England.
- ▶ Surveyors were also asked to collect information on additional pond-specific environmental covariates, which we consider as potential predictors for species presence as well as the probabilities of error at the two stages.
- ▶ We have $S = 189$, $M = 1$, $K = 12$ and $q = 0.0794$.

Posterior mean and 95% credible interval for all model parameters at the modal combination of the available covariates.

Parameter	Posterior mean	95% posterior credible interval
ψ	0.14	(0.04, 0.42)
θ_{11}	0.73	(0.45, 0.79)
θ_{10}	0.15	(0.05, 0.27)
p_{11}	0.81	(0.71, 0.90)
p_{10}	0.05	(0.03, 0.07)

Posterior conditional probability of species absence given x positive qPCR replicates, $1 - \psi(x)$, (first row) and posterior conditional probability of x positive qPCR replicates given species presence, $q(x)$, (second row), at the modal combination of the available covariates.

x	0	1	2	3	4	5	6	7	8	9	10	11	12
$1 - \psi(x)$	0.93	0.93	0.93	0.93	0.88	0.58	0.54	0.54	0.53	0.53	0.53	0.53	0.53
$q(x)$	0.159	0.093	0.026	0.005	0.001	0.004	0.014	0.039	0.087	0.151	0.192	0.161	0.069

- ▶ We did not identify any covariates that are linked to the probability of species presence, ψ , or to the probabilities of a stage 1 error, as they all have PIP well below 50%.
- ▶ On the other hand, four covariates with $PIP > 50\%$ have been identified for p_{11} (maximum pond depth, PIP: 1.00, and pond length, PIP: 0.63, presence of macrophytes, PIP: 0.71 and pond density, PIP: 0.91) and one for p_{10} (fish presence, PIP: 0.97).
- ▶ Maximum pond depth and presence of macrophytes have a positive effect on stage 2 true positive probability, while pond length and pond density have a negative effect. Finally, the presence of fish decreases the probability of a stage 2 false positive result.

MONITORING OF WILDLIFE SPECIES

eDNA

SINGLE SPECIES

MULTIPLE SPECIES

CONCLUSIONS AND FUTURE WORK

METABARCODING

- ▶ Metabarcoding is a technique that allows for simultaneous identification of many taxa (species) within the same sample, as opposed to qPCR that only allows the identification single species.
- ▶ The outcome of the metabarcoding procedure is an **OTU table**, which summarises the number of metabarcoding reads for each taxa for each environmental sample.
- ▶ Several bioinformatics procedures are available for generating OTUs from an environmental sample. In our work, we use the BioSoupII workflow, which relies on clustering DNA sequences in several clusters with radius not exceeding a pre-specified threshold.

- ▶ Metabarcoding and the wide availability of remotely sensed environmental covariates that contain biodiversity information (Bush et al. (2017)⁹, Simonson et al. (2014)¹⁰, Bongalov et al. (2019)¹¹) have the potential to relieve the problems of data limitation and analysis with which environmental management has been struggling.
- ▶ This can open the way to near-real-time tracking of state and change in biodiversity and its functions and services over whole landscapes, which will finally allow biodiversity to carry informational weight commensurate with other landscape features in decision-making.

⁹ Bush, A. et al. Nat Ecol Evol 1, 0176–0149 (2017)

¹⁰ Simonson, W. D. et al. Meth. Ecol. Evol. 5, 719–729

¹¹ Bongalov, B. et al. Ecol. Lett. (2019); Davies, A. B. et al. TREE 29, 681–691 (2014)

PROJECT AIM

- ▶ The aim of our project is to realise the potential of these new data collection and analysis methods to inform landscape decision-making, by developing an integrated statistical framework for DNA-based surveys of biodiversity.
- ▶ The framework will allow the estimation of community compositions and the identification of the landscape characteristics that drive them.

jSDMs

- ▶ Joint species distribution models (jSDMs) allow an explicit and flexible explanation of community composition by species' environmental preferences, as well as patterns of co-occurrence and spatial autocorrelation¹².
- ▶ jSDMs models rely on GLMs to jointly model the presence or abundances of the species at the different sites and provide the ideal foundation for this project.
- ▶ We have been focusing on developing alternative ways to model species associations and on efficient Bayesian variable selection for jSDMs.

¹²Ovaskainen, O. et al. Ecol. Lett. 20, 561–576 (2017); Warton, D. I. et al. TREE 30, 766–779 (2015)

- ▶ We have been extending jSDMs in several directions, i.e. species associations, variable selection, to make them appropriate for modelling multi-species eDNA data.
- ▶ The next step is to account for observation error and taxonomic uncertainty.
In fact, many steps in the bioinformatics pipeline used to generate the OTUs are sensitive to subjective choices. For example:
 - ▶ Some metabarcoding reads are discarded based on a user-specified threshold
 - ▶ Species are created using a clustering step, where the size of the cluster is again defined in advance by the user.
 - ▶ and many many more...

MONITORING OF WILDLIFE SPECIES

eDNA

SINGLE SPECIES

MULTIPLE SPECIES

CONCLUSIONS AND FUTURE WORK

- ▶ We developed a new Bayesian model for single-species eDNA data that can be used to infer site-specific probabilities of species presence while accounting for the probabilities of error at both stages of eDNA surveys.
- ▶ Our model overcomes identifiability issues introduced by the likelihood function and our proposed algorithm allows us to perform Bayesian variable-selection efficiently, even in large dimensions.
- ▶ We are currently using the model to explore issues of study design for eDNA surveys.

Currently, we are working on new models for

- ▶ metabarcoding data
- ▶ iDNA data
- ▶ ancient DNA data

Thank you!
Any questions / comments?