

A unifying modelling framework for DNA-based biodiversity monitoring

Alex Diana¹, Eleni Matechou, Jim Griffin, Douglas Yu, Mingjie Luo, Marie Tosa, Alex Bush, Richard Griffiths

¹SMSAS, University of Kent, Sibson Building, Parkwood Rd, Canterbury, CT2 7FS

INTRODUCTION

DNA-based biodiversity surveys have revolutionised the cost-effectiveness, precision and scale at which biodiversity assessment can be performed. These surveys involve collecting physical samples from survey sites and assaying the contents in the laboratory to detect species via their diagnostic DNA sequences, allowing detection of hundreds of species at the same time.

However, the potential of DNA-based surveys comes with many statistical challenges. Firstly, only relative, and not absolute, measures of abundance can be estimated using DNA data alone. Secondly, DNA-based surveys yield data that contain many types of error and noise, some of which are species-specific.

MODELLING APPROACH

We have developed a unifying modelling framework for DNA-based survey data, for the first time simultaneously allowing for all key sources of variation, error and noise in the data-generating process.

Our approach builds on Bayesian hierarchical multivariate models, with inference performed using MCMC and implementing an efficient sampling scheme that allows optimal mixing and scaling even for very large numbers of species.

The model can estimate biomass changes across sites and link those changes to environmental covariates, while accounting for between-species and between-sites correlation. Our novel modelling framework is implemented in the freely available GitHub package *eDNAPlus*.

CASE STUDY

We demonstrate the new framework on a dataset of insects collected using Malaise trap samples in the HJ Andrews Forest, Oregon, in July 2018, for a total of 50 species.

The species-specific effect of elevation on biomass (Fig. 1) suggests low phylogenetic inertia.

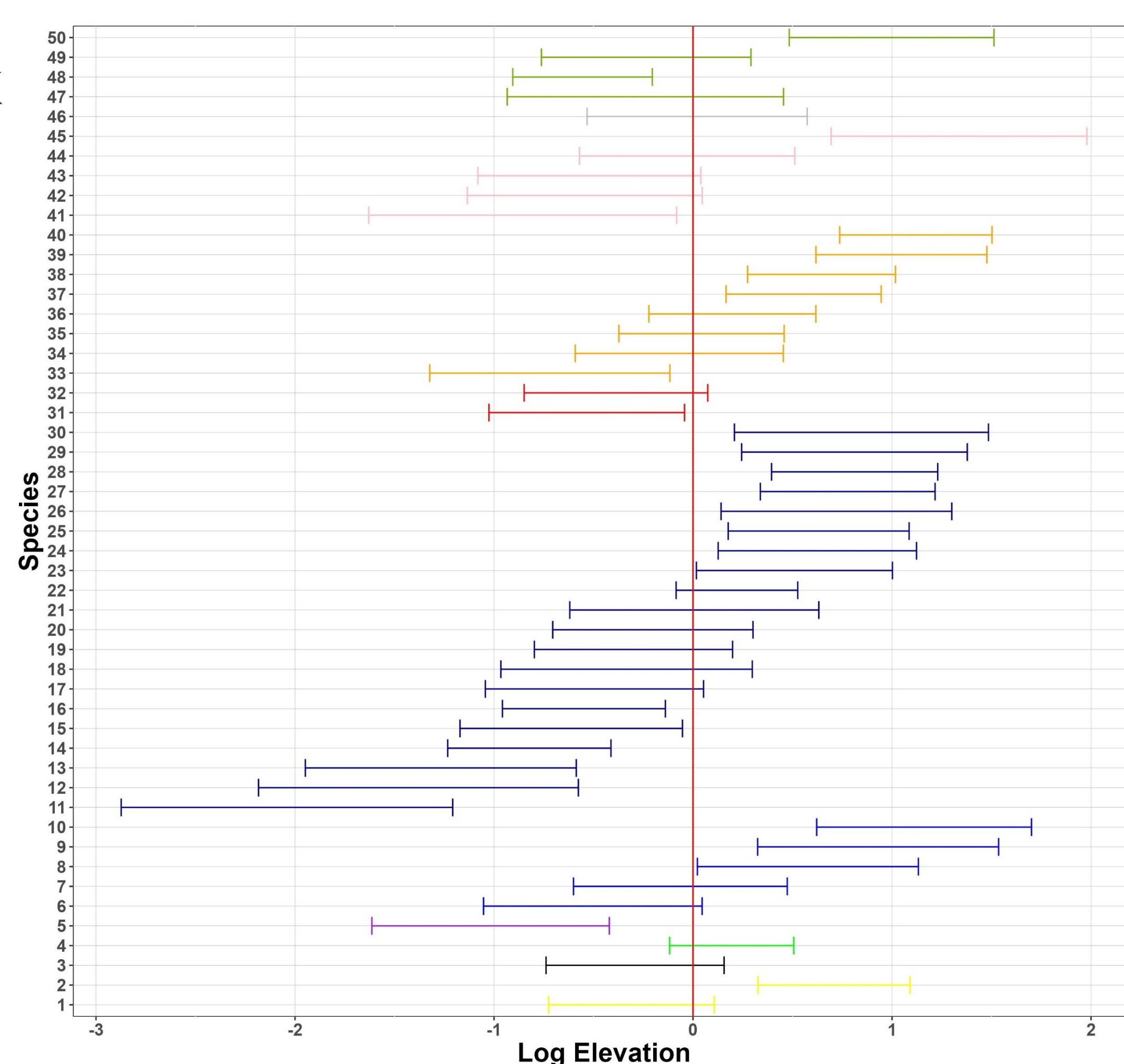


Figure 1

The inferred species correlations (Fig. 2) indicate higher positive correlations within the Diptera species (flies, 14 - 30), as well as with several species in the Hymenoptera (ants, bees, and wasps) and Lepidoptera (butterflies and moths) species.

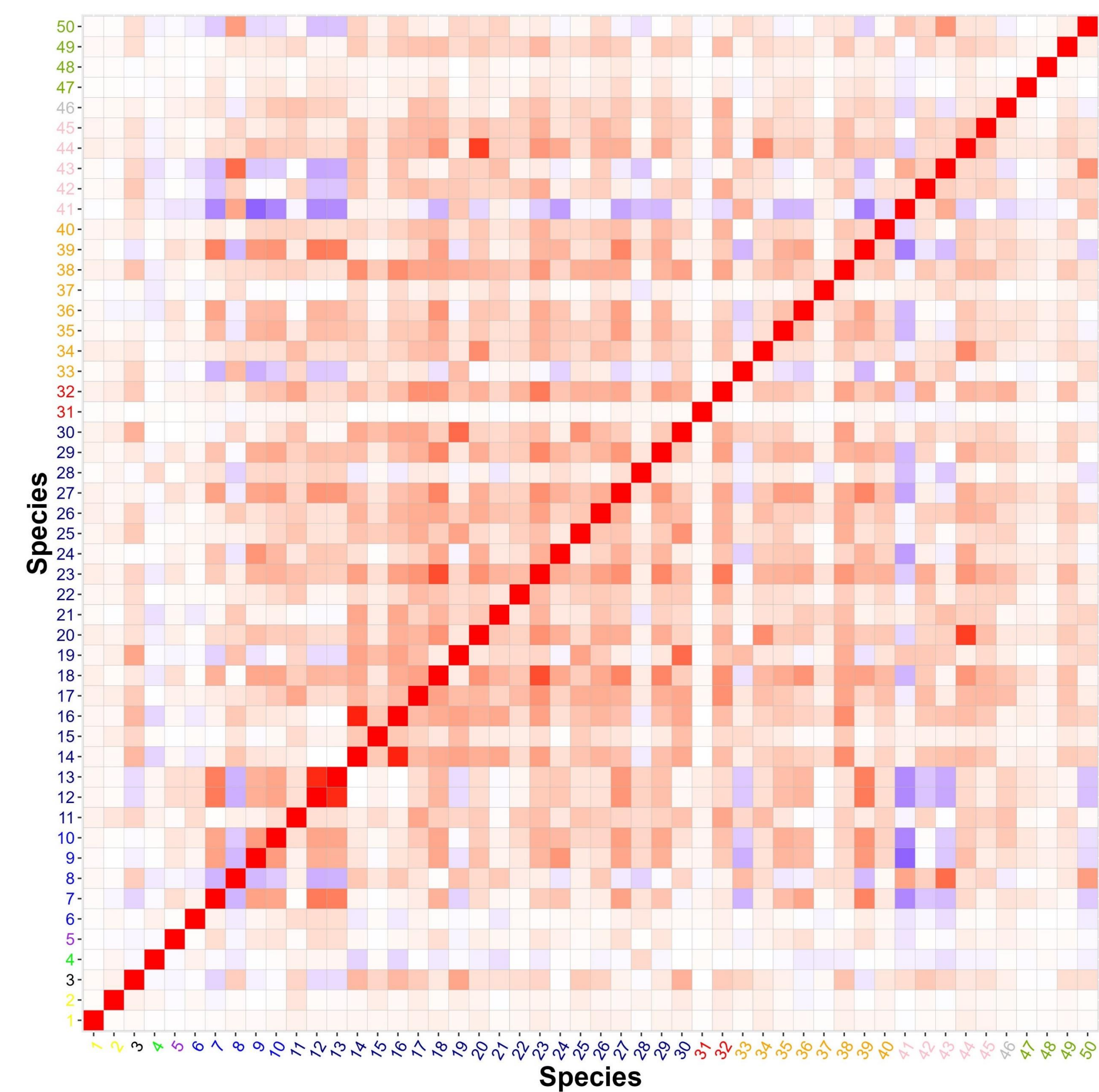


Figure 2

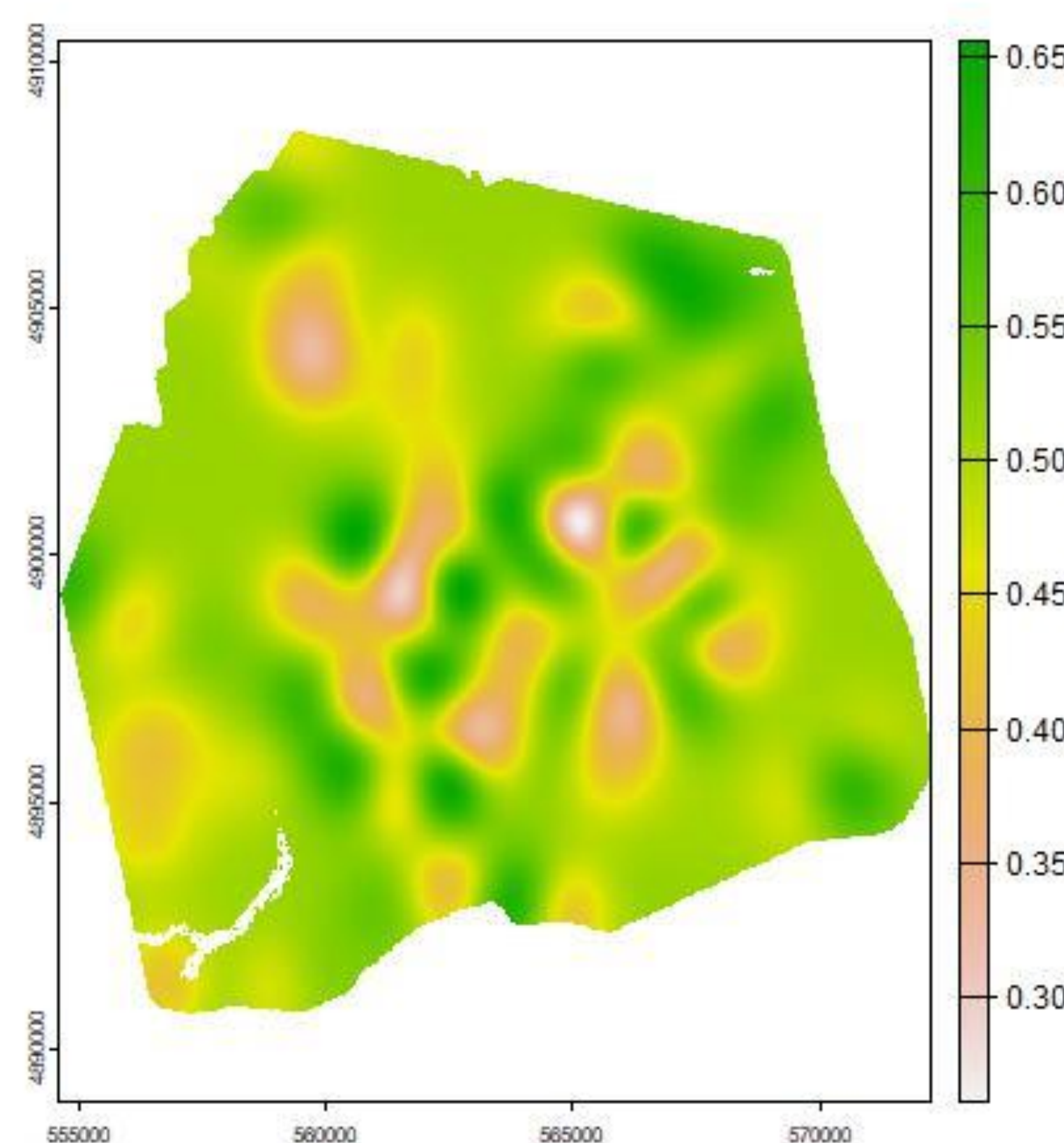


Figure 3

The resulting biodiversity map for the area (Fig. 3), obtained using our model output, is useful for identifying areas of higher species richness and compositional distinctiveness, and hence higher conservation value.

CONCLUSIONS

Over the last decade, DNA-based biodiversity studies, primarily using metabarcoding, have rapidly increased in popularity, and multivariate statistical models are now starting to be deployed to analyse metabarcoding data. By facilitating estimates of within-species changes in biomass and of covariate effects, while accounting for workflow uncertainty and species correlations, our modelling framework provides a step-change improvement in the design and analysis of DNA-based biodiversity monitoring.

REFERENCES

Diana, A., Matechou, E., Griffin, J., Yu, D., Luo, M., Tosa, M., Bush, A. and Griffiths, R., 2022. *eDNAPlus*: A unifying modelling framework for DNA-based biodiversity monitoring. *arXiv preprint arXiv:2211.12213*.