201

Fitting dynamic occupancy models to very large occurrence data sets using hidden Markov models

Dynamic occupancy models are widely used for analysing occurrence data sets (Royle, J. A. and Dorazio, R. M., 2008, Hierarchical Modeling and Inference in Ecology). They are fitted using classical inference in computer packages such as unmarked (Fiske and Chandler, 2011, J. Statistical Software) and Rpresence (https://www.mbr-pwrc.usgs.gov/software/presence.html), and also using Bayesian inference – see for example Clarke and Altwegg, 2019, Ecology and Evolution. Most published applications do not have much more than a thousand sites, and typically far fewer than that. An exception is the Bayesian implementation of a random-walk model (Outhwaite et al, 2018, Ecological Indicators) implemented in Sparta (https://github.com/BiologicalRecordsCentre/sparta. In the United Kingdom, extensive occurrence data are available for many taxonomic groups, and the Biological Recording Centre (BRC) oversees more than 80 recording schemes (brc.ac.uk, Pocock et al., 2015, Biological Journal of the Linnean Society). Sparta has been used to analyse extensive BRC data on invertebrates, bryophytes and lichens (Outhwaite et al., 2019, Scientific Data; 2020, Nature Ecology &amp; Evolution) and pollinators (Powney et al., 2021, Nature Communications). However model-fitting in these cases can be extremely slow, even with a super computer.

We are motivated by the need to fit dynamic occupancy models efficiently to very large data sets, such as those arising from the UK Butterflies of the New Millenium (BNM) database. For example one might have records on some 50,000 sites over 20 years, and there are typically also 70%-80% missing values. In our experience, existing computer packages using classical inference struggle when modelling such data.

We have analysed very large data sets using classical inference and linked hidden Markov models, combined with innovative parallel computing, an approach which is remarkably efficient. The methodology naturally allows inclusion of covariates, model comparisons, and evaluation of goodness-of-fit measures. We illustrate the method using simulation and data on butterfly species from the BNM database, and comparisons are made with unmarked and Rpresence. The ability to model large ecological data sets efficiently provides estimates of key demographic parameters describing survival and colonisation, to enhance understanding, potentially link with integrated population modelling, and ultimately inform and improve management and conservation.