Journal of the Royal Statistical Society Series A, 2024, 1–17
doi: DOI HERE

Efficient statistical inference methods for assessing changes in species' populations using citizen science data

Emily B. Dennis,[1,2] Alex Diana,[3] Eleni Matechou[2] and Byron J.T. Morgan[2,*]

[1] Butterfly Conservation, Manor Yard, East Lulworth, BH20 5QP, Dorset, UK , [2] University of Kent, School of Mathematics, Statistics and Actuarial Science, Canterbury, CT2 7NF, Kent, UK and [3] University of Essex, School of Mathematics, Statistics and Actuarial Science, Wivenhoe Park, Colchester, CO4 3SQ. Essex, UK

[*]Corresponding author. B.J.T.Morgan@kent.ac.uk

Abstract

The global decline of biodiversity, driven by habitat degradation and climate breakdown, is a significant concern. Accurate measures of change are crucial to provide reliable evidence of species' population changes. Meanwhile citizen science data have witnessed a remarkable expansion in both quantity and sources and serve as the foundation for assessing species' status. The growing data reservoir presents opportunities for novel and improved inference but often comes with computational costs: computational efficiency is paramount, especially as regular analysis updates are necessary. Building upon recent research, we present illustrations of computationally efficient methods for fitting new models, applied to three major citizen science data sets for butterflies. We extend a method for modelling abundance changes of seasonal organisms, firstly to accommodate multiple years of count data efficiently, and secondly for application to counts from a snapshot mass-participation survey. We also present a variational inference approach for fitting occupancy models efficiently to opportunistic citizen science data. The continuous growth of citizen science data offers unprecedented opportunities to enhance our understanding of how species respond to anthropogenic pressures. Efficient techniques in fitting new models are vital for accurately assessing species' status, supporting policy-making, setting measurable targets, and enabling effective conservation efforts.

Key words: Biodiversity change; Citizen science; Concentrated likelihood; Generalised abundance index; Occupancy models; Variational Bayes.

[To be read before The Royal Statistical Society at the Discussion Meeting on the 'Analysis of citizen science data' held at the Society's 2024 annual conference in Brighton on Tuesday, 3 September 2024, the President, Dr Andrew Garrett, in the Chair]

Introduction

An existential crisis of our time is the alarming decrease of biodiversity, due to anthropogenic factors such as climate breakdown and loss of habitat. Producing robust measures of change is vital for evaluating species' status, understanding rates of change, and monitoring responses to pressures, as well as progress of conservation actions, such as towards biodiversity targets (Butchart et al., 2010). Both data and appropriate statistical models are critical for measuring biodiversity change. Citizen (or community) science (CS) data, where information is gathered by voluntary participants, are increasingly used for this purpose (Silvertown, 2009; Chandler et al., 2017).

CS data from systematic, designed surveys and monitoring schemes, collected by skilled and committed volunteers, have long been used to produce estimates of species' status, whereas observations of species from less structured, opportunistic or mass-participation sampling, attracting contributions from wider society, are increasingly gathered (Pocock et al., 2017), broadening the scope to measure changes in populations from extensive geographic areas and for a variety of taxa. The many sources of CS data provide vast opportunities for biodiversity monitoring, but require suitable analytical approaches, for example to deal with sources of bias (Isaac and Pocock, 2015). Johnston et al. (2023) outlined a diversity of challenges for biodiversity monitoring using CS data, relating to dealing with observer behaviour, data structures, statistical models, and communication. This paper focuses upon one such key challenge, which is in the computational cost of analysing CS data, particularly with increasing data and complex models, leading to the "necessity to identify and develop suitable modifications to improve computational efficiency and scalability, adapting traditional (and developing new) methods to big data" (McCrea et al., 2023).

In this paper we present illustrations of computationally efficient methods for analysing CS data. Our work is motivated by applications to data for British butterflies, to model changes in both abundance and distribution, but the methods and overall need for efficiency applies to CS data for a range of taxa and locations. The global decline of insects has been widely reported, particularly in western Europe and North America (Wagner et al., 2021), yet there is an ongoing need for robust data and rigorous analysis methods (Thomas et al., 2019; Didham et al., 2020). Many taxa and geographic regions are lacking in sufficient data to appropriately assess trends, but there is a wealth of data gathered on butterflies in the UK. Butterflies are the most comprehensively monitored invertebrate taxon and their population status provides a valuable indicator for changes in biodiversity as they respond rapidly to environmental change.

We demonstrate efficient analysis methods for three sources of CS data for UK butterflies: (i) The UK Butterfly Monitoring Scheme (UKBMS) began in 1976 and is one of the longest-running insect monitoring schemes in the world. Counts of butterflies are made each year by recorders walking transects for six months from the 1^{st} of April, according to a strict protocol (Pollard and Yates, 1993). (ii) The annual Big Butterfly Count (BBC) launched in 2010 and is a UK-wide mass-participation CS project in which members of the public record how many individual butterflies are seen for 15-minute periods during 23-24 days in late July and early August. It is the largest CS project of its kind in the world, with almost 95,000 participants in 2023. (iii) The Butterflies for the New Millenium (BNM) data base collates opportunistic records of where butterflies are seen, and consists of over 14 million records (Fox et al., 2023). These data are unstandardised, and more than 80 similar recording schemes exist for various taxa in the UK (BRC, 2022; Pocock et al., 2015) and are typically used for describing changes in species' distributions.

Analyses of UKBMS and BNM data feed in to regular reporting such as the "State of the UK's Butterflies" (Fox et al., 2023), Red List assessments of extinction risk (Fox et al., 2022), government biodiversity indicators (JNCC, 2022), and multi-taxa outputs such as the State of Nature (Burns et al., 2023), all of which contribute to providing robust evidence for conservation, policy development and the wider state of biodiversity.

During a time of biodiversity change, frequent analysis updates are essential for monitoring species' populations, and for understanding responses to both pressures and conservation action. Efficient computational methods are therefore vital, especially in CS surveys, which typically involve data sets that are large and continuously growing in size.

We are motivated by the need to measure population changes, by modelling abundance and distribution using the data sets mentioned above. In this paper we summarise our recent research, describe new methods and applications, and suggest avenues for future development. We start with models for abundance, where we demonstrate how in certain circumstances it is possible to greatly reduce the dimensionality of the effective model parameter space, considerably reducing computation time. We then consider opportunistic data, where we show how variational inference (VI) can provide an efficient alternative to Markov chain Monte Carlo (MCMC) for Bayesian inference when fitting occupancy models.

Generalised abundance index: GAI

Models for abundance data, such as from the UKBMS, need to account for two key features: seasonality and missing data. Seasonality results in counts that vary throughout the season according to the emergence of one or more generations of adult insects over time in any year. Missing data arise in two ways. Firstly, some visits are missed by volunteers. UKBMS transects are sampled by committed and skilled volunteers, walking transects weekly across six months of the year, but inevitably some weeks are missed: Dennis et al. (2013) estimated that this occurs for roughly 8 of 26 weeks of the transect season. Secondly, there is turnover in transects sampled each year; for example transects may cease to be monitored, and new transects continue to be introduced to the scheme.

A variety of methods have produced analyses of these data. Dennis et al. (2013) presented a method based on using generalised additive models (GAMs) to model seasonality, followed by a generalised linear model (GLM) analysis to account for the annual variation in the transects sampled. This approach produced annual indices of abundance for each species, which can be used to estimate time trends in abundance. Alternatively, Matechou et al. (2014) proposed a stopover model approach, which enables estimation of within-season survival probabilities. Dennis et al. (2016) then proposed the generalised abundance index (GAI) approach, which provides a framework for both of these types of models, as well as a third alternative, which describes seasonality parametrically using an appropriate mixture of distributions, such as Normal (see Dennis et al., 2022, for examples).

In outline, for a given species, in any year we suppose that counts are obtained at S sites, each visited on at most V occasions. Each count, $y_{s,v}$, for site s and visit v is regarded as the realisation of a random variable, such as Poisson, with expectation $\lambda_{s,v} = N_s a_{s,v}$, where the likelihood then takes the form

$$L(\boldsymbol{N}, \boldsymbol{\theta}; \mathbf{y}) \propto \prod_{s=1}^{S} \prod_{v=1}^{V} \exp(-N_s a_{s,v})(N_s a_{s,v})^{y_{s,v}}. \tag{1}$$

Here the N_s are parameters describing site abundance, and $a_{s,v}$ denotes a function determined by parameters, $\boldsymbol{\theta}$, describing seasonal variation. Here and later the product over visits only includes terms corresponding to when visits are made. Background is provided by Dennis et al. (2016), who derive a concentrated-likelihood approach that substantially reduces the number of parameters to estimate via maximimum likelihood. Briefly, using maximum likelihood, the site parameters can be estimated by

$$N_s = \frac{y_{s,.}}{a_{s,.}}, \tag{2}$$

where we use the dot notation to indicate summation corresponding to visited sites. The total observed count for each site is therefore re-scaled to account for incomplete sampling within the season. Substitution of the estimates of Equation (2) into Equation (1) results in a Poisson likelihood which can be maximised with respect to only the parameters $\boldsymbol{\theta}$. Counts can be expected to be over-dispersed relative to the Poisson and/or contain additional zeros, for example due to small counts at the ends of the season. Alternative discrete distributions such as negative binomial and zero-inflated Poisson, respectively, may then be appropriate, as described in Dennis et al. (2016).

Two adaptations of the GAI

We now provide two new adaptations of the GAI. Firstly, the model is extended to include appropriate site and year effects using an *annual model* integrated within the GAI, combined with the use of concentrated likelihood. Secondly, we adapt the GAI to improve the analysis of BBC data, a "snapshot", mass-participation CS scheme.

An extended GAI with site and time effects

The basic GAI, described above, is a static model, where data for each year are analysed separately. However, to deal with the turnover of sites surveyed in the UKBMS each year, the Poisson GLM stage of Dennis et al. (2013), based on the *annual model* of ter Braak et al. (1994), is typically used. Here the abundance estimates $\{\hat{N}_{s,r}\}$, for sites s and now year r, are the dependent values for an additive model with year and site effects and the estimated year effects are then used to form abundance indices. Use of log-linear regression in this way is widely used in ecology, for example using TRIM (Trends and Indices for Monitoring Data, van Strien et al., 2004; Bogaart et al., 2020). This two-stage GAI approach is now a valuable tool for analysing seasonal count data in the UK (UKBMS, 2023; JNCC, 2022; Fox et al., 2021, 2023) and beyond (Schmucki et al., 2016; Van Swaay et al., 2020). In practice, the proportion of the species' flight period surveyed for a given site and year is typically included as a weighting in the GLM stage, such that better sampled sites have a higher contribution to the estimated species' abundance indices (Brereton et al., 2018).

A drawback of the two-stage GAI approach is the need to bootstrap to account for variance propagation between the two model stages. Bootstrapping is time consuming for large data sets, and can also be problematic when re-sampling small data sets, for example for rare species. We now describe an alternative approach which effectively incorporates the annual model within the GAI, extending the model to consider all years at once, and thus solving the issue of variance propagation.

In brief we expand the Poisson likelihood of Equation (1) to counts $y_{s,v,r}$ where r denotes one of Y successive years, with a corresponding expansion to $a_{s,v,r}$. We then incorporate the expression for the annual model $N_{s,r} = e^{\alpha_s+\beta_r}$ - see ter Braak et al. (1994) - which results in the following expression for the log-likelihood, ignoring an additive constant.

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}) = \sum_r \sum_s \sum_v \{-e^{(\alpha_s+\beta_r)} a_{s,v,r} + y_{s,v,r}(\alpha_s + \beta_r) + y_{s,v,r}\log(a_{s,v,r})\}. \tag{3}$$

Here $\{\alpha_s\}$ and $\{\beta_r\}$ are respectively site and year effects to be estimated. As above, we use a concentrated likelihood approach to form maximum-likelihood parameter estimates efficiently, by concentrating out the parameters $\boldsymbol{\alpha}$, resulting in the log-likelihood

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}) = \sum_r \sum_s \sum_v \left[-\frac{y_{s,.,.} e^{\beta_r} a_{s,v,r}}{\sum_j e^{\beta_j} a_{s,.,j}} + y_{s,v,r}\{\beta_r - \log(\sum_j e^{\beta_j} a_{s,.,j})\} \right], \tag{4}$$

which we can maximise efficiently with respect to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. Full details are given in Section S1 of the Supplementary Material.

The aim here is to produce a new model which can be fitted efficiently and produce estimates of uncertainty. From the above modelling, estimates of error result from inverting the estimated Hessian at the maximum-likelihood estimates, which is far more efficient than bootstrapping. Profile confidence intervals are also easily obtained.

We illustrate this extended GAI approach with application to UKBMS data spanning 1976-2022 for two species: Chalk Hill Blue *Polyommatus coridon* and Gatekeeper *Pyronia tithonus*. Chalk Hill Blue is a species confined to calcareous grassland in southern England, with UKBMS counts of around 900,000 individuals from $\sim$ 460 sites, whereas Gatekeeper is a widespread butterfly species across southern Britain, with UKBMS counts from more than 4,300 sites, counting over 3.3 million individuals.

Options for describing seasonal variation (parameters associated with $\boldsymbol{\theta}$) are flexible, as in the original GAI, but here we use a Normal distribution, since both species are univoltine (one generation per year), and thus parameters $\{\mu_r\}$ and $\{\sigma_r\}$ are estimated for each year, which describe the mean flight dates and length of the flight period, respectively (see Figure S1 where we see the tendency for earlier emergence over time). Flight periods are thus assumed to be fixed over sites within each year.

Figure 1 shows the estimates of $\{\beta_r\}$ from the extended GAI, with 95% confidence intervals (CI) produced from the estimated Hessian at the maximum-likelihood estimates. Error estimates were scaled to account for overdispersion in the Poisson GAI formulation.

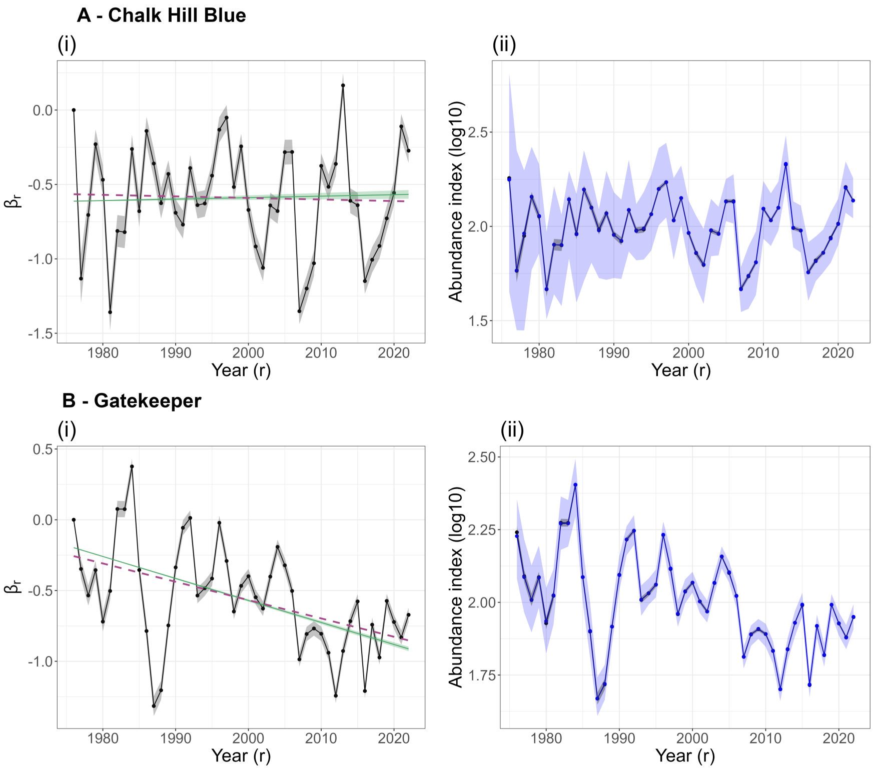

Fig. 1: Results from applying the extended GAI to two species: Chalk Hill Blue (A) and Gatekeeper (B). Plots (i) show the estimates of $\{\beta_r\}$ (black) with 95% confidence intervals (CI). Linear trend estimates for $\{\beta_r\}$ from the extended GAI are shown in green (with 95% CI). The trend line from a posthoc linear regression through the estimates of $\{\beta_r\}$ is also shown (purple dashed line). Plots (ii) compare abundance indices from the extended GAI (black) and two-stage GAI (blue), with 95% CI produced from the estimated Hessian at the maximum-likelihood estimates and from non-parametric bootstrapping respectively. Abundance indices correspond to $\{\beta_r\}$ converted to the $\log_{10}$ scale with a mean of 2, as is standard practice for UKBMS indices.

In order to estimate a linear trend over time, the parameters $\boldsymbol{\beta}$ may also be expressed by a linear form, where we set $\beta_r = \gamma r, \forall r$, in Equation (4) and maximise the resulting log likelihood with respect to the parameters $\boldsymbol{\theta}$ and γ. The intercept parameter for a linear regression on year cannot be estimated due to confounding. This is also the case for the annual model (McCullagh and Nelder, 1989, p.63, van Strien et al., 2004), where we therefore set $\beta_1 = 0$. Trend lines based on the extended GAI with the linear form are shown in Figures 1(i), and due to the lack of intercept, for ease of comparison we equate the ordinate at the middle year to that from a posthoc linear regression (also shown).

Abundance indices produced from the extended GAI are virtually indistinguishable from indices produced from the two-stage GAI (Figure 1(ii)). The two-stage GAI was based on the implementation used for annual analyses of UKBMS data, where a GAM is used for the first stage. Despite the similarity in the indices, the 95% CI produced from bootstrapping the two-stage GAM are much wider than those estimated from the Hessian for the extended GAI. This difference is due to differences between the models, rather than due to differences in the methods of error estimation. For example the extended GAI does not make a distributional assumption (Poisson) about the site abundance estimates, $\mathbf{N}$, as is the case in the GLM of two-stage GAI. The standard errors from the extended GAI may also be underestimated if the data are overdispersed relative to the Poisson distribution assumed, which could be explored in future work using alternative discrete distributions.

The implementation of the two-stage GAI includes a weighting in the GLM, accounting for the proportion of the flight period sampled, whereas the extended GAI does not. Interestingly, the close resemblance of the two indices in Figure 1(ii) implies that the new extended GAI effectively has this weighting built-in, by modelling the counts directly so that better-sampled sites have a greater contribution to the likelihood.

The extended model took 42 minutes to run for the Chalk Hill Blue, whereas the non-parametric bootstrap (with 1000 replicates) for the two-stage GAI took 98 minutes. For the more widespread Gatekeeper, the extended GAI took 1.3 hours, whereas the two-stage GAI bootstrap took approximately 7 hours, but based on only 200 replicates. Analyses were parallelised across four CPUs on a computer equipped with Intel Core i7-8700@3.2GHz with 32GB of RAM. Clearly there are substantial time savings to be had for the 59 UK butterfly species, and potentially far more for more numerous taxa. We have seen that the advantages of the extended GAI are not confined to efficiency, but may also involve greater parameter precision provided the model and distributional assumptions are appropriate. In the Discussion we outline the potential contributions of future work on the extended GAI.

Analysis of Big Butterfly Count data

The Big Butterfly Count (BBC; `https://bigbutterflycount.org`) is an annual survey of common butterfly species which encourages wide participation particularly from members of the general public. The BBC has a high media profile and attracts a large number of participants, many of whom may have limited or no prior experience of biodiversity monitoring. The sampling protocol is minimal: participants simply count numbers of individuals seen of widespread butterfly (and day-flying moth) species for 15 minutes from any location.

The BBC generates a large amount of data for the UK's widespread butterfly species - more than 11 million butterflies have been counted since 2011 - but to date few analyses have been undertaken and annual reporting of BBC results is based on simple comparisons with respect to the previous year only. Using data for 2011-2014, Dennis et al. (2017a) demonstrated that estimates of change in abundance from BBC were comparable to those estimated from standardised monitoring (UKBMS data), but that the short snapshot sampling period of three weeks results in bias caused by the inter-annual variation in species' flight periods.

This is demonstrated in Figure 2A which shows the estimated flight period for Marbled White *Melanargia galathea* for four example years. This species is a univoltine (single-generation), summer-flying species, which is therefore likely to be particularly susceptible to phenological bias (in the estimation of timing) with respect to the BBC sampling period. Flight periods were estimated from UKBMS data using the GAI with a spline formulation for $\{a_v\}$, which is fixed across sites within a given year. The timing of peak emergence varies year-to-year, and thus the proportion of the flight period sampled by the BBC varies annually (see Figure 2B). For example in 2011 the BBC only captures the tail end of the flight period, whereas in 2012 more than half of the flight period is captured.

Here we describe a modification of the GAI for producing abundance indices and trends from BBC data. The GAI was developed for standardised monitoring data, typically collected along transects, where sites are clearly defined and revisited many times within and across years. Mass-participation CS data such as from the BBC does not have such a structure; it consists of many locations, which often only have one count undertaken. Hence we define a BBC "site" to be a 1km x 1km square and pool BBC data to this spatial scale. BBC sites may then have counts across multiple days within the BBC sampling period, and potentially multiple entries (15 minute counts) on a given day.

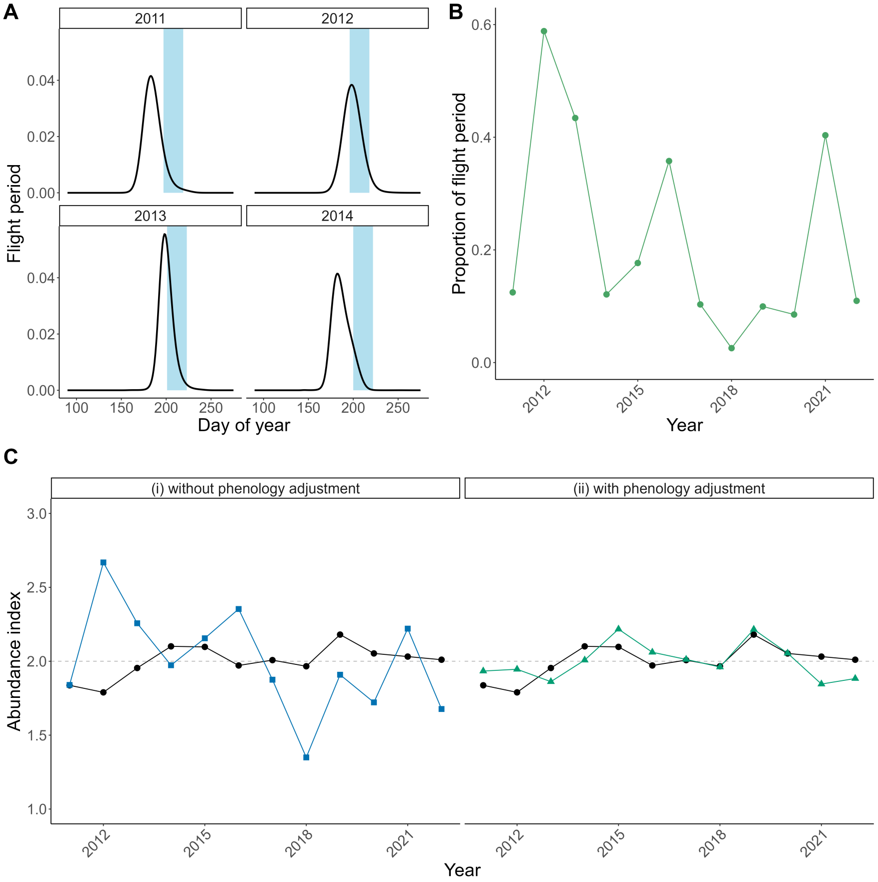

Fig. 2: Demonstration of the phenology adjustment approach for the Marbled White butterfly: A) flight period curves estimated from UKBMS data for four years. The blue shaded areas represents the BBC sampling period each year. B) the proportion of the Marbled White flight period covered by the BBC sampling period each year. C) relative abundance indices produced from the GAI applied to UKBMS data (black), from BBC data without phenology adjustment (i, blue squares), and from BBC data with phenology adjustment (ii, green triangles). Indices are on the $\log_{10}$ scale with a mean value of 2 (indicated by the horizontal dashed lines).

To accommodate these multiple entries the GAI likelihood for a given year now takes the form

$$L \propto \prod_{m=1}^{M} \prod_{v=1}^{V} \exp(-N_m a_{m,v} \kappa_{m,v})(N_m a_{m,v})^{y_{m,v,\cdot}}, \tag{5}$$

where $\kappa_{m,v}$ is the number of entries for site m and visit v, and $y_{m,v,\cdot} = \sum_\kappa y_{m,v,\kappa}$ is the sum of the counts over those entries.

The number of sites in a mass-participation CS dataset such as BBC is high, and we adopt a concentrated likelihood approach to reduce the functional size of the parameter space, as in the previous section, where Equation (2) becomes

$$N_m = \frac{y_{m,\cdot,\cdot}}{\sum_{v=1}^{V} a_{m,v} \kappa_{m,v}} \tag{6}$$

The BBC sampling period represents just a snapshot of most species' flight periods, therefore accurate flight period estimation from applying the GAI to BBC data is not possible for all species. Instead, for each species and year, we produce estimates of $\{N_m\}$ for each BBC site (1km x 1km square) from Equation (6) using daily flight period estimates of $\{a_v\}$ from the GAI applied to UKBMS data, where flight periods are assumed to be the same across sites. The average of $\{N_m\}$ provides an overall measure of BBC abundance per year and species (Dennis et al., 2016).

Figure 2C demonstrates the benefit of the phenology adjustment approach for the Marbled White butterfly. Fluctuations in the BBC abundance index produced without phenology adjustment (Figure 2C(i)) largely reflect year-to-year variation in the proportion of the flight period captured by the BBC each year, whereas adjusting for phenology produces an index that shows a pattern more similar to the UKBMS abundance index (Figure 2C(ii))). Applying the phenology adjustment approach is less influential for a multivoltine species such as the Comma *Polygonia c-album* (Figure S2), for which the BBC sampling period covers a smaller, and less variable, proportion of the overall flight period each year.

Dennis et al. (2024) describe the new phenology adjustment approach for snapshot citizen science data in full, including the use of bootstrapping to estimate uncertainty. The approach is applied to BBC data for 17 species and explored further via simulation. The method enables data from snapshot CS schemes such as BBC to contribute to monitoring biodiversity. BBC results receive high-profile media coverage and are already beginning to reflect the advantages of the new analyses outlined here.

Efficient occupancy model fitting

Occupancy models (MacKenzie et al., 2018; Altwegg and Nichols, 2019) are widely used for inferring species distributions from presence/absence data at multiple sites, across a single or multiple seasons. They have been employed to extract meaningful species distribution trends from opportunistic citizen science data, addressing challenges associated with non-systematic sampling and variable observation effort (Kéry et al., 2010; Isaac et al., 2014). However, as the size of the corresponding presence/absence data increases, occupancy models can be computationally demanding, especially in a Bayesian framework.

We start by defining the standard occupancy model that we fit in this section. We assume that there are n sampling units, where each sampling unit corresponds to a site in a specific year. In each sampling unit, we have a set of observations y_i, equal to 1/0 if the species was detected/not detected. The sampling unit to which observation i belongs is denoted by k_i. The hierarchical model representation is

$$\begin{cases} \text{logit}(\psi_j) = X_j^{\psi}\beta_{\psi} \\ z_j \sim \text{Be}(\psi_j) \\ \text{logit}(p_i) = X_i^{p}\beta_p \\ y_i \sim \begin{cases} \text{Be}(p_i) & \text{if } z_{k_i} = 1 \\ 0 & \text{if } z_{k_i} = 0 \end{cases} \end{cases} \tag{7}$$

where:

- z_j is the occupancy state of sampling unit j
- β_{ψ} are the covariate coefficients of the occupancy probability ψ
- β_p are the covariate coefficients of the detection probability p
- X_j^{ψ} and X_i^{p} are the available covariates for sampling unit j and observation i, respectively.

Similar versions of this model have been considered in Diana et al. (2023) and in Doser et al. (2023), who employed Gaussian processes to account for spatio-temporal autocorrelation in the occupancy probability. Bayesian inference for these occupancy models can be easily performed using Markov chain Monte Carlo (MCMC) (see Diana et al., 2023; Doser et al., 2023, who employed a Pólya-Gamma sampling scheme for logistic regression models (Polson et al., 2013)). In this MCMC framework, the z terms from Equation (7), indicating species presence/absence at each site, are treated as latent variables and hence inferred and updated, typically at each MCMC iteration. Inferring the latent variables z allows us to easily write the complete data likelihood - see for example King (2014) and Newman et al. (2023) - for the observations in y, as shown in Equation (7). However, when the number of sites is large, this leads to a computationally intensive MCMC, even when efficient model-fitting approaches, such as the Pólya-Gamma scheme, are employed for updating the model parameters. In addition when the complete data likelihood is used there may be correlations between estimators which can also slow down MCMC - see Newman et al. (2023) and Borowska and King (2022), where a subset of latent states are not treated as auxiliary variables but are integrated out numerically to reduce the correlation between latent variables. Therefore, MCMC-based inference can be prohibitively slow, which limits its application to large data sets such as CS data. For example, for the data sets considered in Diana et al. (2023), obtaining acceptable effective sample sizes from the posterior distributions of all parameters requires running times of around 19 hours (on an Intel Core i7-10610U@1.8GHz). An obvious alternative is to use classical inference to fit occupancy models since it can be much faster (see for example the approach of Dennis et al., 2017b). In this case, the likelihood function is written by marginalising over the z variables and hence the observed data likelihood is used for inference. Expressions for the complete and observed data likelihood for occupancy models are given in Section S2 of the Supplementary Material. However, quantifying uncertainty around functions of parameters can sometimes be computationally intensive, relying on bootstrap methods when closed form expressions of variances are not available. Bayesian inference also offers the potential to more readily account for spatio-temporal autocorrelation in the occupancy probabilities.

Variational inference (VI) has been proposed as an alternative tool to overcome the computational issues of MCMC (Jordan et al., 1999). VI is traditionally faster than MCMC-based approaches because it transforms the problem from sampling (from a posterior distribution) to optimization. Therefore, VI combines the speed of classical inference, with the interpretability of Bayesian inference. However, while MCMC-based inference always recovers the true posterior (given enough MCMC iterations), in VI the true posterior distribution $p(\theta|y)$ is approximated using an appropriate flexible family of distributions, which is called the variational family. We denote the variational distributions by $q_\lambda(\theta)$, where λ is the set of variational parameters, the observed data likelihood by $p(y|\theta)$ and the prior distribution by $p(y|\theta)$. The parameters λ corresponding to the optimal variational distribution can be found by minimizing the Kullback-Leibler (KL) divergence between the true posterior distribution $p(\theta|y)$ and the variational distribution $q_\lambda(\theta)$. It can be proved that this is equivalent to finding the $q_\lambda(\theta)$ that minimizes the quantity

$$\mathbf{E}_{\theta \sim q_\lambda(\theta)} \left[\log\, p(y,\theta) - \log\, q_\lambda(\theta)\right] \tag{8}$$

which is known as the Evidence Lower BOund (ELBO), and forms the basis of VI inference. We note that $\log\ \ p(y,\theta) = \log\ \ \{p(y|\theta)p(\theta)\}$ and more details on VI can be found in Blei et al. (2017).

In VI, it is common to assume that parameters are a-posteriori independent, which is equivalent to assuming a variational family of the form $q(\theta) = \prod_j q_j(\theta_j)$ (the *mean-field assumption*) since this assumption considerably simplifies the inference. However, if the assumption is not valid, then posterior variance is underestimated (Wang and Titterington, 2005). In the case of occupancy models, occupancy and detection probability are independent conditionally on z, but not independent of z. Hence, assuming (according to the mean field assumption) that they are independent of z implies that ψ and p are independent of each other a-posteriori, which clearly does not hold. If the dependence structure of the model is ignored, then it leads to underestimation of the posterior variance of the occupancy and detection probability parameters, (β^ψ, β^p) (Clark et al., 2016).

However, if the observed data likelihood is used for inference, instead of the complete data likelihood, with the latter being common practice in a Bayesian framework, then we do not need to assume that (β^ψ, β^p) are independent of z, or of each other. In the observed data likelihood case, we assume, as is standard in VI (Titsias and Lázaro-Gredilla, 2014), that $q_\lambda(\theta)$ is a multivariate normal distribution, that is, if $\theta = (\beta^\psi, \beta^p)$ are the model parameters, we assume $\theta \sim q_\lambda(\theta) = N(\mu, \Sigma)$, where the variational parameters are $\lambda = (\mu, C)$, with μ the mean of the variational distribution and C the Cholesky factor of the covariance matrix Σ.

We perform inference using stochastic gradient descent. Computing the gradient of Equation (8) with respect to λ is complicated by the fact that λ itself appears in the expectation. To overcome this issue, we use the reparameterisation trick (Kingma and Welling, 2013), which is a cornerstone of variational inference as it allows us to easily obtain these types of gradients. More details on the inference are presented in Section S2 of the Supplementary Material.

To investigate the efficacy of our novel VB approach, we have performed a small preliminary simulation study to assess the coverage of Bayesian credible intervals generated using our procedure. Across the simulations, we varied the number of sites and the average occupancy and detection probability. We have chosen n, the number of sites, to be 500, 1000 and 2000 and ψ and p to be 0.25, 0.5 and 0.75. For each simulation, we ran 500 replications. The coverage was computed across 4 covariate coefficients for occupancy and 4 covariate coefficients for detection, with the coefficients randomly set to be either -1 or 1. Results are reported in Table 1 where it is seen that the 95% posterior credible intervals have the nominal coverage.

Table 1. Coverage of the 95% Bayesian credible intervals generated using variational Bayes, for varying n, the number of sites, and values of ψ and p.

n / ψ=p	.25	.5	.75
500	0.950	0.949	0.961
1000	0.961	0.953	0.952
2000	0.951	0.948	0.951

We also analyzed the dataset of Ringlet butterflies collated through the Butterflies for the New Millennium (BNM) recording scheme run by Butterfly Conservation, using records collected between 1970 and 2014, which is also used in Diana et al. (2023). The data set consists of $>$ 2 million records from $\sim$ 140,000 unique 1 km^2 (defined as sites), of which $>$ 218,000 detections of Ringlet have been made, and non-detections were produced using observations of other butterfly species (Kéry et al., 2010). In this case, we do not account for spatial autocorrelation, but we model year as a factor variable and assume independence between sites, a point which we discuss in the next section. We also use relative list length, obtained by dividing the list length, which is the number of species recorded for a given site/date, divided by the maximum recorded list length in a neighboring area of 50 km, as a covariate for detection probability and model the effect of the day of the year using a second polynomial of Julian date.

In this case, convergence was determined by assessing when parameter updates were smaller than a pre-specified tolerance, which required approximately 1 hour of computation time (results were again obtained on an Intel Core i7-10610U@1.80GHz). In Figure 3, we present the posterior distribution of the occupancy index, which is the year-specific occupancy probability, here assumed the same for all sites. The increase in precision over time reflects the growth of the underlying opportunistic CS data set, which has shown ongoing expansion, particularly since the mid 1990s (Fox et al., 2023).

The occupancy probabilities show similarities with results presented in Diana et al. (2023), in particular an increasing trend in recent years reflecting an expansion in the range of Ringlet in the UK and similar drops in the species' prevalence in mid 70s and early and late 90s are identified. However, in this case, the pattern is less smooth, as expected, since the effect of year is not constrained in any way. Furthermore, a direct comparison cannot be made, since Diana et al. (2023) fit a more complex model accounting for spatio-temporal correlation, and plot an occupancy index rather than the annual occupancy probabilities.

In Figure 4, we present the posterior distribution of the detection probability throughout the year which can be interpreted as an estimate of species' flying time. Unlike in Diana et al. (2023), in this example detection probability does not vary with year, and thus represents an average for 1970-2014. We also note here that we do not account for spatial effects on the probability of detection. However, the species' flight patterns are likely to vary in space, and hence the curve in Figure 4 represents a mixture of spatially-varying curves. Additionally, we note that the peak of detection probability in this case is lower than that obtained by Diana et al. (2023), which is due to the different model structure for detection and occupancy probabilities, as discussed above. The 95% posterior credible interval of the coefficient of relative list length is $(0.881, 0.883)$. We have shown that using a VI framework to fit occupancy models efficiently to CS data

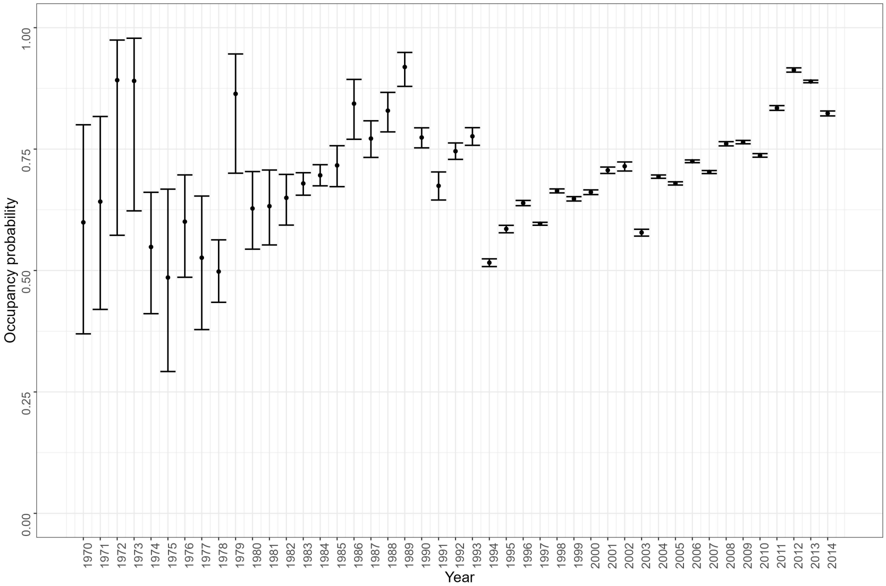

Fig. 3: 95% posterior credible intervals of the occupancy probabilities of each year for Ringlet. The dots represent the posterior medians.

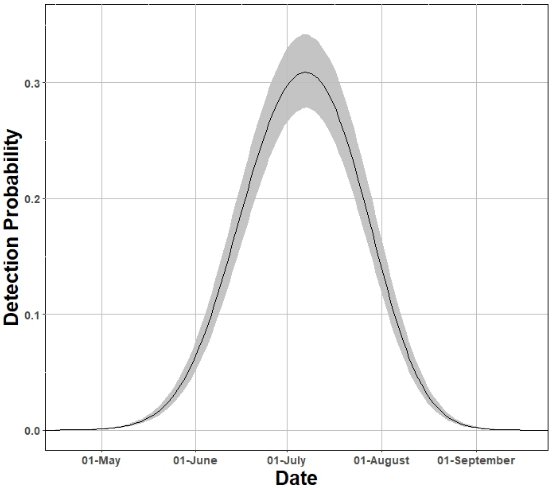

Fig. 4: Posterior median and 95% posterior credible intervals of the detection probability in each week for Ringlet.

shows promising results, with various avenues for future development - some of which we discuss in the next section.

Discussion

CS data have an important rôle in the future of biodiversity monitoring, but the ongoing growth of such data requires novel statistical models along with efficient inference methods and available computer code. In this paper, we have built upon recent work and demonstrated that concentrated likelihoods can greatly reduce the dimensionality of the effective model parameter space and that variational inference can substantially reduce computational time for large CS data sets. The approaches proposed in this paper address computational problems, but also lead to model developments for CS data.

The GAI is efficient due to the use of concentrated likelihood (Dennis et al., 2016). There is a broad analogy here with the efficiency that can result from adopting a hidden Markov model, when appropriate, and forming a forward algorithm for likelihood construction; see for example Cowen et al. (2017). We have shown concentrated likelihood to be useful in the new extension of the GAI to incorporate the annual model which conveniently allows for formal variance propagation, negating the need for time-consuming bootstrapping. A similar approach is given in Bravington et al. (2021), where the two stages for density surface models, one involving a detection probability and the other a GAM, are combined into one.

The extended GAI has been introduced in this paper, with application to two univoltine species, but future work will test wider application to more species, in particular with varying quantities of data and life histories, including species with two or more generations per year, requiring more complex functions to describe seasonal variation (Dennis et al., 2016). As mentioned, alternative distributions to the Poisson may also be explored, in particular to assess for effects of overdispersion and the associated impact on estimates of uncertainty.

In both adaptations of the GAI presented, species' flight periods were assumed to be the same across sites within each year, as is typical in the production of abundance trends for UK butterflies (UKBMS, 2023). However, greater flexibility in the seasonal variation function $\mathbf{a}$ can be readily accounted for through the inclusion of appropriate spatial covariates. For example this is done by Schmucki et al. (2016) and Dennis et al. (2022). This is similarly a future direction to explore for occupancy models, accounting for spatial variation in detection probabilities, for example due to spatial variation in species' phenologies, via spatial covariates or spatial effects. See also for example Clark and Altwegg (2019) and Dennis et al. (2019) for models with spatial variation in occupancy.

The extended GAI also provides a basis for further extensions within this computationally efficient approach, such as formal data integration by maximising a joint likelihood; see Besbeas et al. (2002) and Schaub and Kéry (2021). For example, UKBMS (count data from standardised monitoring) and BBC (mass-participation CS data) could be used to produce new urban butterfly indicators. Integrated modelling approaches can optimise the use of available CS data sets, but present outstanding questions and challenges (Isaac et al., 2020; Zipkin et al., 2021; Johnston et al., 2023), including the need for computationally efficient approaches for data integration.

We have introduced the basic occupancy model within a VI framework and have discussed the use of the observed, instead of the complete, data likelihood for avoiding the standard issue of underestimated posterior variances when parameters are assumed to be independent. In this case, the dependence structure in the model parameters and latent variables was such that the use of the observed data likelihood allowed us to use VI without having to assume that parameters are a-posteriori independent. Intuitively, the same approach could be employed in other ecological models within a VI framework, although at the moment this is only an intuition and future work would need to explore the quality of inference for different types of data and corresponding models.

We believe that VI provides a powerful and versatile framework for efficiently fitting a wide range of ecological models. The advantage of VI is that it combines the speed of classical inference with the interpretability of Bayesian inference. VI relies on the ability to obtain the gradient of the likelihood function, which might seem like an obstacle. However, it is possible to take advantage of recent developments of deep learning methodologies such as the use of automatic differentiation, which can automatically compute gradients of this type (as long as the likelihood function is tractable), such as for example using the package TMB (Kristensen et al., 2016). We envisage that analyses currently based on hidden Markov models for likelihood computation - see for example Cowen et al. (2017), Besbeas and

Morgan (2019), Besbeas and Morgan (2020) and McClintock et al. (2020) - can be a fruitful research avenue, since automatic differentiation engines would allow us to compute gradients by differentiating through the forward recursion used to compute the likelihood in this case. The use of automatic differentiation also enables us to consider highly non-linear extensions, such as neural networks, to be introduced in the model. However, as discussed, particular attention needs to be paid to the chosen variational family, since that will determine the accuracy of the approximation.

We have considered a simple occupancy model, in comparison to models that have been employed previously for data of this type, as a means of discussing VI and the use of observed data likelihood within the VI framework. The flexible spatio-temporal models, for example based on Gaussian processes (Doser et al., 2022; Diana et al., 2023), which have been developed within an MCMC framework could be considered within this VI framework in future work. However, since this leads to a regression with as many covariates as the number of support of points, assuming a variational approximation with a full covariance matrix is computationally prohibitive, since the number of parameters of the Cholesky factor C scales quadratically with the number of covariates. One option to overcome this problem is to induce sparsity on the inverse of the covariance matrix by zeroing elements of the Cholesky factor C (Tan and Nott, 2018). For example, it is possible to assume a variational approximation where the covariate coefficients for the spatial approximation are independent in the posterior. Although this step can potentially reintroduce bias in the model, since it assumes a-posteriori independence of parameters, it leads to a substantial reduction in the number of parameters, making the model feasible to estimate. Finally, in cases where the observed data likelihood cannot be obtained, for example in complex data-generating processes or models with individual random effects, then the complete data likelihood and efficient MCMC inference (see for example King et al., 2023, who devise an importance sampling approach for ecological models with random effects) may provide the only viable alternative, at least at the moment.

To some extent formal design considerations do not arise with the data that we have considered, however there is an issue of non-random sampling (Boyd et al., 2023b; Johnston et al., 2023) and the need to account for issues such as preferential sampling and spatial and temporal biases (Altwegg and Nichols, 2019; Boersch-Supan et al., 2019; Conn et al., 2017; Pati et al., 2011). See for example papers by King et al. (2023) and by Lahoz-Monfort et al. (2014), respectively on sampling the data, and on how to design studies when resources are limited.

We have demonstrated efficient analysis methods for sources of CS data for UK butterflies, but efficient statistical inference methods are needed for understanding population changes from CS data for a wide range of taxa and locations. For example the GAI approach, or related models, has been applied to moths, bees and beetles (Fox et al., 2021; Matechou et al., 2018; Dennis et al., 2021). Development of efficient inference for occupancy models is also vital given their application to various taxa (for example Burns et al., 2023; Outhwaite et al., 2019; Boyd et al., 2023a). The need for efficiency will continue to increase with the growth of data sets such as the Global Biodiversity Information Facility (GBIF), which has amassed more than 2.5 billions occurrences of more than one million species (GBIF.org, 2023).

CS data are increasingly "big", not only in terms of volume, but also involving characteristics such as variety. Farley et al. (2018) and McCrea et al. (2023) discuss the "Four Vs Framework" in which data may be characterised as "big". Analysing CS data for biodiversity monitoring presents various challenges (Johnston et al., 2023), but methods also need to be suitably scalable for these increasingly large data sets.

The examples in this paper are based upon analyses of data featuring observations of species submitted by citizen scientists, but computational challenges also arise from other data types, for example with the growth of data from technological advances such as automated interpretation of images submitted by citizen scientists (Terry et al., 2020; van Klink et al., 2022).

It may be argued that high performance computing (HPC) and cloud computing can be used to address the challenge of fitting computationally demanding models to CS data (Farley et al., 2018). For example, a supercomputer has enabled the production of occupancy trend estimates for thousands of UK species using Bayesian occupancy models (Outhwaite et al., 2019; Boyd et al., 2023a). However ultimately, as data sets continue to grow, and models become increasingly complex, we argue that a trade off is needed, and that using more and more computing resources is not a simple solution. Fitting computationally demanding methods can make appropriate model validation and inference difficult: for example variable selection may become impractical (Johnston et al., 2023), as well as suitable goodness-of-fit assessment. Achieving model convergence for all parameters in MCMC may also become difficult (Outhwaite et al., 2019; Boyd et al., 2023a).

Furthermore, to maximise the use of CS data in biodiversity monitoring, there is a need for statistical approaches that are appropriately disseminated and accessible for use in practice. Johnston et al. (2023) suggest that "accessible communication of novel methods could democratise analysis of these data and thus enable CS data to reach their broadest potential". Statistical methods that depend upon HPC may be a barrier for analysis to those without easy or affordable access to such resources, thus hindering the potential of biodiversity monitoring with CS data globally (Pocock et al., 2018).

Efficient methods are also crucial for producing frequent analysis updates for reporting on the status of species, particularly as time lags in data availability continue to reduce. The need for accurate reporting is ever necessary in monitoring species' status, measuring against biodiversity targets, supporting policy-making and guiding effective conservation effort. This paper has presented examples for fitting computationally efficient models to CS data, but, as also suggested by Johnston et al. (2023), with the growth of such data and its importance for biodiversity monitoring (Pocock et al., 2018; Callaghan et al., 2021), there is an ongoing need to develop efficient statistical inference methods, with the potential to learn from developments in mainstream statistics.

Competing interests

No competing interest is declared.

Author contributions statement

All authors devised the paper. ED curated the data sets. ED did the computing for the new GAI modelling and AD did the computing for the VI analyses. All authors collaborated in the writing of the paper, and ED and AD produced the Supplementary material.

Acknowledgments

We thank the Associate Editor and two referees for their positive and helpful reviews. The UKBMS is organised and funded by Butterfly Conservation, the British Trust for Ornithology (BTO), and the Joint Nature Conservation Committee (JNCC). The UKBMS is indebted to all volunteers who contribute data to the scheme. We are also very grateful to all of the volunteers who have contributed to the Butterflies for the New Millennium project, which is run by Butterfly Conservation with support from Natural England, and also the citizen scientists who submitted Big Butterfly Count data. Byron Morgan was supported by a Leverhulme Emeritus Fellowship. This work was partly funded by NERC grant NE/T010045/1.

Data availability

Data and code associated with the analyses in this paper are available as follows: (i) example of the extended GAI approach applied to UKBMS data `https://github.com/EBDennis/Extended_GAI_UKBMS_example` (ii) example of the GAI applied to BBC data `https://github.com/EBDennis/GAI_BBC_example` (iii) example of VI applied to BNM data for Ringlet `https://github.com/AlexDiana/VBOccupancy`.

References

Altwegg, R. and Nichols, J. D. (2019) Occupancy models for citizen-science data. *Methods in Ecology and Evolution*, **10**, 8–21.

Besbeas, P., Freeman, S. N., Morgan, B. J. T. and Catchpole, E. A. (2002) Integrating mark–recapture–recovery and census data to estimate animal abundance and demographic parameters. *Biometrics*, **58**, 540–547.

Besbeas, P. and Morgan, B. J. T. (2019) Exact inference for integrated population modelling. *Biometrics*, **75**, 475–484.

— (2020) A general framework for modelling population abundance data. *Biometrics*, **76**, 281–292.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017) Variational inference: A review for statisticians. *Journal of the American Statistical Association*, **112**, 859–877.

Boersch-Supan, P. H., Trask, A. E. and Baillie, S. R. (2019) Robustness of simple avian population trend models for semi-structured citizen science data is species-dependent. *Biological Conservation*, **240**, 108286.

Bogaart, P., van der Loo, M. and Pannekoek, J. (2020) *rtrim: Trends and Indices for Monitoring Data*. URL: `https://CRAN.R-project.org/package=rtrim`. R package version 2.1.1.

Borowska, A. and King, R. (2022) Semi-complete data augmentation for efficient state space model fitting. *Journal of Computational and Graphical Statistics*, **32**, 19–35.

Boyd, R. J., August, T. A., Cooke, R., Logie, M., Mancini, F., Powney, G. D., Roy, D. B., Turvey, K. and Isaac, N. J. (2023a) An operational workflow for producing periodic estimates of species occupancy at national scales. *Biological Reviews*, **98**, 1492–1508.

Boyd, R. J., Powney, G. D. and Pescott, O. L. (2023b) We need to talk about nonprobability samples. *Trends in Ecology & Evolution*, **38**, 521–531.

ter Braak, C. J. F., van Strien, A. J., Meijer, R. and Verstrael, T. J. (1994) Analysis of monitoring data with many missing values; which method? In *Bird Numbers 1992 Distribution, Monitoring and Ecological Aspects* (eds. E. J. M. Hagemeijer and T. J. Verstrael).

Bravington, M. V., Miller, D. L. and Hedley, S. L. (2021) Variance propagation for density surface models. *Journal of Agricultural, Biological, and Environmental Statistics*, **26**, 306–323.

BRC (2022) Biological Records Centre Home Page. Available from: www.brc.ac.uk (accessed November 13, 2023).

Brereton, T. M., Botham, M. S., Middlebrook, I., Randle, Z., Noble, D. and Harris, S. e. a. (2018) United Kingdom Butterfly Monitoring Scheme report for 2017. *Tech. rep.*, Centre for Ecology & Hydrology, Butterfly Conservation, British Trust for Ornithology and Joint Nature Conservation Committee.

Burns, F., Mordue, S., al Fulaij, N., Boersch-Supan, P. H., Boswell, J., Boyd, R. J., Bradfer-Lawrence, T., de Ornellas, P. et al. (2023) State of Nature 2023, the State of Nature partnership, available at www.stateofnature.org.uk.

Butchart, S. H. M., Walpole, M., Collen, B., van Strien, A., Scharlemann, J. P. W., Almond, R. E. A., Baillie, J. E. M., Bomhard, B., Brown, C., Bruno, J. et al. (2010) Global biodiversity: indicators of recent declines. *Science*, **328**, 1164–1168.

Callaghan, C. T., Poore, A. G., Mesaglio, T., Moles, A. T., Nakagawa, S., Roberts, C., Rowley, J. J., VergÉs, A., Wilshire, J. H. and Cornwell, W. K. (2021) Three frontiers for the future of biodiversity research using citizen science data. *BioScience*, **71**, 55–63.

Chandler, M., See, L., Copas, K., Bonde, A. M., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G. et al. (2017) Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, **213**, 280–294.

Clark, A. E. and Altwegg, R. (2019) Efficient Bayesian analysis of occupancy models with logit link functions. *Ecology and Evolution*, **9**, 756–768.

Clark, A. E., Altwegg, R. and Ormerod, J. T. (2016) A variational Bayes approach to the analysis of occupancy models. *PloS one*, **11**, e0148966.

Conn, P. B., Thorson, J. T. and Johnson, D. S. (2017) Confronting preferential sampling in wildlife surveys: diagnosis and model-based triage. *Methods in Ecology and Evolution*, **8**, 1535–1546.

Cowen, L., Besbeas, P. T., Morgan, B. J. T. and Schwarz, C. (2017) Hidden Markov models for extended batch data. *Biometrics*, **73**, 1321–1331.

Dennis, E. B., Brereton, T. M., Morgan, B. J. T., Fox, R., Shortall, C. R., Prescott, T. and Foster, S. (2019) Trends and indicators for quantifying moth abundance and occupancy in Scotland. *Journal of Insect Conservation*, **23**, 369–380.

Dennis, E. B., Fagard-Jenkin, C. and Morgan, B. J. T. (2022) rGAI: An R package for fitting the generalized abundance index to seasonal count data. *Ecology and Evolution*, **12**, e9200.

Dennis, E. B., Freeman, S. N., Brereton, T. and Roy, D. B. (2013) Indexing butterfly abundance whilst accounting for missing counts and variability in seasonal pattern. *Methods in Ecology and Evolution*, **4**, 637–645.

Dennis, E. B., Kéry, M., Morgan, B. J. T., Coray, A., Schaub, M. and Baur, B. (2021) Integrated modelling of insect population dynamics at two temporal scales. *Ecological Modelling*, **441**, 109408.

Dennis, E. B., Morgan, B. J. T., Brereton, T. M., Roy, D. B. and Fox, R. (2017a) Using citizen science butterfly counts to predict species population trends. *Conservation Biology*, **31**, 1350–1361.

Dennis, E. B., Morgan, B. J. T., Freeman, S. N., Brereton, T. M. and Roy, D. B. (2016) A generalized abundance index for seasonal invertebrates. *Biometrics*, **72**, 1305–1314.

Dennis, E. B., Morgan, B. J. T., Freeman, S. N., Ridout, M. S., Brereton, T. M., Fox, R., Powney, G. D. and Roy, D. B. (2017b) Efficient occupancy model-fitting for extensive citizen-science data. *PLoS ONE*, **12**, e0174433.

Dennis, E. B., Morgan, B. J. T., Harrower, C. A., Bourn, N. A. D. and Fox, R. (2024) Incorporating phenology to estimate species' population trends from snapshot citizen-science data. *In revision.*

Diana, A., Dennis, E. B., Matechou, E. and Morgan, B. J. T. (2023) Fast Bayesian inference for large occupancy data sets, using the Pólya-Gamma scheme. *Biometrics*, **79**, 2503–2515.

Didham, R. K., Basset, Y., Collins, C. M., Leather, S. R., Littlewood, N. A., Menz, M. H., Müller, J., Packer, L., Saunders, M. E., Schönrogge, K. et al. (2020) Interpreting insect declines: seven challenges and a way forward. *Insect Conservation and Diversity*, **13**, 103–114.

Doser, J. W., Finley, A. O. and Banerjee, S. (2023) Joint species distribution models with imperfect detection for high-dimensional spatial data. *Ecology*, **104**, e4137.

Doser, J. W., Finley, A. O., Kéry, M. and Zipkin, E. F. (2022) spoccupancy: An R package for single-species, multi-species, and integrated spatial occupancy models. *Methods in Ecology and Evolution*, **13**, 1670–1678.

Farley, S. S., Dawson, A., Goring, S. J. and Williams, J. W. (2018) Situating ecology as a big-data science: Current advances, challenges, and solutions. *BioScience*, **68**, 563–576.

Fox, R., Dennis, E. B., Brown, A. F. and Curson, J. (2022) A revised Red List of British butterflies. *Insect Conservation and Diversity*, **15**, 485–495.

Fox, R., Dennis, E. B., Harrower, C. A., Blumgart, D., Bell, J. R., Cook, P., Davis, A. M., Evans-Hill, L. J., Haynes, F., Hill, D. et al. (2021) The State of Britain's Larger Moths 2021. *Butterfly Conservation, Rothamsted Research and UK Centre for Ecology & Hydrology, Wareham, Dorset, UK.*

Fox, R., Dennis, E. B., Purdy, K. M., Middlebrook, I., Roy, D. B., Noble, D. G., Botham, M. S. and Bourn, N. A. D. (2023) The State of the UK's Butterflies 2022. *Butterfly Conservation, Wareham, UK.*

GBIF.org (2023) GBIF Home Page. Available from: https://www.gbif.org (accessed November 13, 2023).

Isaac, N. J. B., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A. et al. (2020) Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, **35**, 56–67.

Isaac, N. J. B. and Pocock, M. J. O. (2015) Bias and information in biological records. *Biological Journal of the Linnean Society*, **115**, 522–531.

Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P. and Roy, D. B. (2014) Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, **5**, 1052–1060.

JNCC (2022) UK Biodiversity Indicators 2022 - C6 Insects of the wider countryside (butterflies). Available from: https://jncc.gov.uk/our-work/ukbi-c6-insects-of-the-wider-countryside/ (accessed November 8, 2023).

Johnston, A., Matechou, E. and Dennis, E. B. (2023) Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, **14**, 103–116.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999) An introduction to variational methods for graphical models. *Machine learning*, **37**, 183–233.

Kéry, M., Gardner, B. and Monnerat, C. (2010) Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, **37**, 1851–1862.

King, R. (2014) Statistical ecology. *Annual Review of Statistics and Its Application*, **1**, 401–426.

King, R., Sarzo, B. and Elvira, V. (2023) When ecological individual heterogeneity models and large data collide: An importance sampling approach. *The Annals of Applied Statistics*, **17**, 3112–3132.

Kingma, D. P. and Welling, M. (2013) Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114.*

van Klink, R., August, T., Bas, Y., Bodesheim, P., Bonn, A., Fossøy, F., Høye, T. T., Jongejans, E., Menz, M. H. M., Miraldo, A. et al. (2022) Emerging technologies revolutionise insect ecology and monitoring. *Trends in Ecology & Evolution*, **37**, 872–885.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. and Bell, B. M. (2016) TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software*, **70**, 1–21. URL: `https://www.jstatsoft.org/index.php/jss/article/view/v070i05`.

Lahoz-Monfort, J. J., Harris, M. P., Morgan, B. J. T., Freeman, S. N. and Wanless, S. (2014) Exploring the consequences of reducing survey effort for detecting individual and temporal variability in survival. *Journal of Applied Ecology*, **51**, 534–543.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L. and Hines, J. E. (2018) *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence, Second Edition.* Academic Press, New York.

Matechou, E., Dennis, E. B., Freeman, S. N. and Brereton, T. (2014) Monitoring abundance and phenology in (multivoltine) butterfly species: a novel mixture model. *Journal of Applied Ecology*, **51**, 766–775.

Matechou, E., Freeman, S. N. and Comont, R. (2018) Caste-specific demography and phenology in bumblebees: Modelling beewalk data. *Journal of Agricultural, Biological and Environmental Statistics*, **23**, 427–445.

McClintock, B. T., Langrock, R., Gimenez, O., Cam, E., Borchers, D. L., Glennie, R. and Patterson, T. A. (2020) Uncovering ecological state dynamics with hidden Markov models. *Ecology Letters*, **23**, 1878–1903.

McCrea, R., King, R., Graham, L. and Börger, L. (2023) Realising the promise of large data and complex models. *Methods in Ecology and Evolution*, **14**, 4–11.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models (2nd edition)*. Chapman & Hall, London.

Newman, K., King, R., Elvira, V., de Valpine, P., McCrea, R. S. and Morgan, B. J. T. (2023) State-space models for ecological time series: practical model-fitting. *Methods in Ecology and Evolution*, **14**, 26–42.

Outhwaite, C. L., Powney, G. D., August, T. A., Chandler, R. E., Rorke, S., Pescott, O. L., Harvey, M., Roy, H. E., Fox, R., Roy, D. B. et al. (2019) Annual estimates of occupancy for bryophytes, lichens and invertebrates in the UK, 1970–2015. *Scientific Data*, **6**, 1–12.

Pati, D., Reich, B. J. and Dunson, D. B. (2011) Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, **98**, 35–48.

Pocock, M. J. O., Chandler, M., Bonney, R., Thornhill, I., Albin, A., August, T., Bachman, S., Brown, P. M. J., Cunha, D. G. F., Grez, A. et al. (2018) A vision for global biodiversity monitoring with citizen science. In *Advances in Ecological Research*, vol. 59, 169–223. Elsevier.

Pocock, M. J. O., Roy, H. E., Preston, C. D. and Roy, D. B. (2015) The Biological Records Centre: a pioneer of citizen science. *Biological Journal of the Linnean Society*, **115**, 475–493.

Pocock, M. J. O., Tweddle, J. C., Savage, J., Robinson, L. D. and Roy, H. E. (2017) The diversity and evolution of ecological and environmental citizen science. *PloS one*, **12**, e0172579.

Pollard, E. and Yates, T. J. (1993) *Monitoring butterflies for ecology and conservation: the British Butterfly Monitoring Scheme*. Chapman & Hall, London.

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.

Schaub, M. and Kéry, M. (2021) *Integrated Population Models:Theory and Ecological Applications with R and JAGS*. Elsevier.

Schmucki, R., Pe'er, G., Roy, D. B., Stefanescu, C., Van Swaay, C. A. M., Oliver, T. H., Kuussaari, M., Van Strien, A. J., Ries, L., Settele, J. et al. (2016) A regionally informed abundance index for supporting integrative analyses across butterfly monitoring schemes. *Journal of Applied Ecology*, **53**, 501–510.

Silvertown, J. (2009) A new dawn for citizen science. *Trends in Ecology & Evolution*, **24**, 467–471.

van Strien, A., Pannekoek, J., Hagemeijer, W. and Verstrael, T. (2004) A loglinear Poisson regression method to analyse bird monitoring data. *Bird Census News*, **13**, 33–39.

Tan, L. S. L. and Nott, D. J. (2018) Gaussian variational approximation with sparse precision matrices. *Statistics and Computing*, **28**, 259–275.

Terry, J. C. D., Roy, H. E. and August, T. A. (2020) Thinking like a naturalist: Enhancing computer vision of citizen science images by harnessing contextual data. *Methods in Ecology and Evolution*, **11**, 303–315.

Thomas, C., Jones, T. H. and Hartley, S. E. (2019) "Insectageddon": A call for more robust data and rigorous analyses. *Global Change Biology*, **25**, 1891–1892.

Titsias, M. and Lázaro-Gredilla, M. (2014) Doubly stochastic variational bayes for non-conjugate inference. In *International conference on machine learning*, 1971–1979. PMLR.

UKBMS (2023) UKBMS official statistics. Available from: https://ukbms.org/official-statistics (accessed November 8, 2023).

Van Swaay, C. A. M., Dennis, E. B., Schmucki, R., Sevillega, C. G., Aghababyan, K., Åström, S., Balalaikins, M., Bonelli, S., Botham, M., Bourn, N. et al. (2020) Assessing Butterflies in Europe - Butterfly Indicators 1990-2018 Technical report. Butterfly Conservation Europe & ABLE/eBMS (www.butterfly-monitoring.net).

Wagner, D. L., Grames, E. M., Forister, M. L., Berenbaum, M. R. and Stopak, D. (2021) Insect decline in the anthropocene: Death by a thousand cuts. *Proceedings of the National Academy of Sciences*, **118**, e2023989118.

Wang, B. and Titterington, D. M. (2005) Inadequacy of interval estimates corresponding to variational bayesian approximations. In *International Workshop on Artificial Intelligence and Statistics*, 373–380. PMLR.

Zipkin, E. F., Zylstra, E. R., Wright, A. D., Saunders, S. P., Finley, A. O., Dietze, M. C., Itter, M. S. and Tingley, M. W. (2021) Addressing data integration challenges to link ecological processes across scales. *Frontiers in Ecology and the Environment*, **19**, 30–38.