 DOI: 10.1111/biom.12506

A Generalized Abundance Index for Seasonal Invertebrates

Emily B. Dennis,[1,3,*] Byron J. T. Morgan,[1] Stephen N. Freeman,[2] Tom M. Brereton,[3] and David B. Roy[2]

[1]Centre for Ecology & Hydrology, Benson Lane, Crowmarsh Gifford, Wallingford, Oxfordshire, U.K.
[3]Butterfly Conservation, Manor Yard, East Lulworth, Wareham, Dorset, U.K.
[*]*email:* E.B.Dennis@kent.ac.uk

Summary. At a time of climate change and major loss of biodiversity, it is important to have efficient tools for monitoring populations. In this context, animal abundance indices play an important rôle. In producing indices for invertebrates, it is important to account for variation in counts within seasons. Two new methods for describing seasonal variation in invertebrate counts have recently been proposed; one is nonparametric, using generalized additive models, and the other is parametric, based on stopover models. We present a novel generalized abundance index which encompasses both parametric and nonparametric approaches. It is extremely efficient to compute this index due to the use of concentrated likelihood techniques. This has particular relevance for the analysis of data from long-term extensive monitoring schemes with records for many species and sites, for which existing modeling techniques can be prohibitively time consuming. Performance of the index is demonstrated by several applications to UK Butterfly Monitoring Scheme data. We demonstrate the potential for new insights into both phenology and spatial variation in seasonal patterns from parametric modeling and the incorporation of covariate dependence, which is relevant for both monitoring and conservation. Associated `R` code is available on the journal website.

Key words: Butterflies; Citizen science; Concentrated likelihood; Normal mixtures; Phenology; UKBMS.

1. Introduction

We shall illustrate the work of the article with UK butterfly monitoring data, though the approach may be applied to other similar insect data, possibly with modification.

1.1. *Butterfly Monitoring*

Indices of abundance are vital for monitoring the population status of a species and measuring responses to changes in climate and land-use. Indices play an important rôle in assessing progress toward targets to reduce biodiversity loss at both national (Defra, 2013) and global scales (Convention on Biological Diversity, 2006; Butchart et al., 2010).

Insects are an important component of our ecosystems and account for a major proportion of the world's biodiversity (Gaston, 1991), but most groups are not well monitored. Butterflies are the most comprehensively monitored invertebrate taxon; their population status provides a valuable indicator for changes in biodiversity and phenology as they respond sensitively and rapidly to changes in climate and habitat (Thomas, 2005; van Swaay et al., 2008). Butterfly monitoring schemes that collect count data exist in many countries and continue to be established (van Swaay et al., 2008; Dennis et al., 2013). Similar schemes for monitoring abundance of other insect taxa also exist, for example, for moths, dragonflies, and bees (Dennis et al., 2013). In the United Kingdom, abundance indices for butterflies form one of 25 indicators employed by the UK government for the assessment of general trends in biodiversity (Defra, 2013). Novel methods for accurately and efficiently deriving indices are continually sought.

A key modeling problem for monitoring-scheme data for insects such as butterflies is the seasonal nature of the data. Butterflies have multi-stage life cycles and counts are usually only made of the most visible adult stage. Hence, count data fluctuate within each year in response to the emergence of butterflies as adults and additionally many species are multivoltine, with more than one brood of adults per year, a feature which may itself be increasing with climate change (Altermatt, 2010).

Count data for UK butterflies are primarily gathered through the UK Butterfly Monitoring Scheme (UKBMS). The scheme consists of a national system of transects on which recorders make counts of butterflies on a weekly basis under favorable conditions during the main butterfly flight period between early April and the end of September (Pollard and Yates, 1993). Roughly 30% of potential counts in the 26 week season are missed (Dennis et al., 2013), for example due to unsuitable weather conditions or recorder unavailability. The scheme began with 34 sites, and has steadily grown to over 1200 sites sampled in 2013, from which long-term and 10-year trends are reported annually for 56 of the 59 species occurring regularly in the United Kingdom (Brereton et al., 2014).

1.2. *Current Methods of Analysis: Notation, Models, and Assumptions*

Suppose that counts are recorded at S sites, each visited on at most T occasions within a single year; $T = 26$ weeks for the UKBMS. Each count, $y_{i,j}$, for the ith site and jth visit, at

occasion $t_{i,j}$, is regarded as the realization of a Poisson random variable, with expectation $\lambda_{i,j}$.

1.2.1. Use of GAMs. Currently, the main approach to account for missing values and seasonal variation in UKBMS counts involves nonparametric curve fitting using generalized additive models (GAMs, Wood, 2006). The original approach by Rothery and Roy (2001) has been extended by Dennis et al. (2013). The approach involves three stages, only the first of which actually involves fitting a GAM. The following approach is applied for each species.

For any year, we write

$$\lambda_{i,j} = \exp\left\{\eta_i + s^f_{i,j}\right\}.$$

The seasonal effect is described through the function $s^f_{i,j} = \alpha_0 + \sum_{d=1}^{f} \alpha_d B_d(t_{i,j})$, where $B_d(t_{i,j})$ are the basis functions for cubic splines (Chambers and Hastie, 1991) and f is the degrees of freedom, which is estimated within the `mgcv` package in `R` (Wood, 2006; R Core Team, 2015), which is employed. GAMs are fitted to the data from all sites, for each year separately. A limitation of the GAM is its restriction to the same seasonal shape, $s^f_{i,j}$, for all sites in any year, though the shapes are scaled differently for different sites, through the $\{\eta_i\}$.

Once the GAMs have been fitted, then the estimated values for the curve describing the seasonal patterns for the kth year are included in a GLM as offsets:

$$\mathbb{E}[y_{i,j,k}] = \exp\{\alpha_i + \beta_k + \hat{s}^f_{i,j}\},$$

where we extend the specification of $y_{i,j}$ to include the year, which we denote by the subscript k. Here, $\{\alpha_i\}$ and $\{\beta_k\}$ denote site and year effects, respectively. The objective of this fit is to impute missing weekly counts.

An index value for any year and site i is then obtained from the estimated area under the seasonal curve, given from the trapezoidal rule as follows:

$$\text{Index}_i = \sum_{j=2}^{T} \frac{(y_{i,j} + y_{i,j-1})(t_{i,j} - t_{i,j-1})}{2}, \qquad (1)$$

where $y_{i,j}$ are real or imputed counts.

Once this is done, then a further GLM is carried out for all the site/year values of the indices, using a Poisson distribution and log link with site and year as additive terms, and the year terms are used as annual indices of abundance. In our experience, this last stage is not necessary if the number of sites is large.

This is a time-consuming process, due mainly to the many GLMs fitted when the offsets are used. As the approach involves multiple stages and the nonparametric estimation of offsets, error estimation is via the bootstrap, which is particularly time consuming. Application of the GAM approach has been limited to assuming a Poisson distribution and a log link.

An emphasis in the GAM approach lies in the interpolation of missing values. This is because site-specific indices in full weekly detail could be of interest to site managers. The option to have maximum resolution on spatio-temporal changes is appealing. This will assist in the estimation of timings of peak count, for example, which can be obtained without the distributional assumptions made in the other models we discuss.

1.2.2. Stopover model. Matechou et al. (2014) have proposed a stopover model incorporating a mixture of terms to parameterize the seasonal pattern. Unlike the GAM, it incorporates specific parameters that relate to the butterfly lifecycle. In this case,

$$\lambda_{i,j} = N_i p_{i,j} \sum_{d=1}^{j} \beta_{i,d-1} \left(\prod_{k=d}^{j-1} \phi_{k,c} \right),$$

for $j = 1, \ldots, T$ and $c = k - d + 1$, where N_i is defined as the size of a super-population of butterflies, which provides an index of abundance, for site i.

It is assumed that the total number of adults at site i is distributed over visits through the parameters $\{\beta_{i,d-1}\}$. The seasonal effect is achieved through these parameters, and is modeled using a mixture of B normal distributions, corresponding to B broods, so that

$$\beta_{i,d-1} = \sum_{b=1}^{B} \omega_{i,b}\{F_{i,b}(t_{i,d}) - F_{i,b}(t_{i,d} - 1)\},$$

where $F_{i,b}(t_{i,d}) = \Pr(X \le t_{i,d}$, for $X \sim N(\mu_{i,b}, \sigma^2_{i,b})$, where $\mu_{i,b}$ is the mean date of emergence for the bth brood at site i and $\sigma_{i,b}$ denotes the corresponding standard deviation. We define $\phi_{j,c}$ to be the probability that an individual which has been at a site for c occasions and is present at visit j will remain until visit $j+1$. Additionally, the $\{p_{i,j}\}$ are appropriate detection probabilities of a surviving individual being detected.

This approach was illustrated on data from 50 sites from 1 year of a single numerous British butterfly, the Common Blue, using a Poisson distribution, though others could also be employed. In this application, the approach was not used for constructing an index of relative abundance.

Surprisingly little is known about adult butterfly survival, and so a model which explicitly includes survival parameters has great potential. It is shown by Matechou et al. (2014) that the detection probabilities may be estimated if they are regressed on suitable covariates, such as the temperature when a visit takes place; however, imputation may be necessary, as temperature is not always recorded by observers. Additionally, the model may be near singular (Catchpole, Kgosi, and Morgan, 2001) if either there are insufficient data to allow precise estimation of the effect of covariates on the detection probabilities, or if suitable covariates vary little. In these cases stable, separate estimation of N_i and $p_{i,j}$ may not be possible. An important study of detection probabilities is provided by Isaac et al. (2011). They conclude that detectability varies appreciably between species and sites, that the variation in detection is small compared to the variation in true abundance, and that estimates of relative abundance from the transect sampling correlate highly with estimates obtained from the alternative approach of distance sampling (which is not a practical alternative for the collection of national invertebrate data).

As a result, in the work of this article we shall not use detection probabilities, though except for the GAM approach they can be included straightforwardly, if there were suitably detailed and abundant data to warrant this. The assumption of constant detection probability is typically made when forming indices of relative abundance (van Swaay et al., 2008).

1.3. *Motivation and Structure of the Article*

The current modeling approaches applied to the UKBMS data require optimization of a likelihood with potentially many parameters corresponding to relative abundance for each site. Given the often large amount of data available from monitoring schemes such as the UKBMS, fitting these models for hundreds or thousands of sites over many years, for multiple species, is computer-intensive. Long-term monitoring schemes require annual updates, and time-consuming methods, particularly the nonparametric bootstrapping required for GAM error estimation, lead to appreciable lags in data processing. The need for more efficient data analysis methods motivates the work of this article, which contains several new developments. In addition, we respond to the need for a flexible approach, as different species exhibit different characteristics, which cannot be described by a single model.

We present a general abundance index, GAI, which incorporates the stopover model and also a nonparametric alternative, corresponding to the first stage of the GAM, as special cases. A new mixture model, which is intermediate between the two existing methods, is presented as a further special case of the GAI. In all cases, a range of alternative distributions is possible. Due to the adoption of a concentrated likelihood approach, the members of the GAI are all fitted extremely efficiently. The GAI thus provides a range of useful tools for practical use, and different components can be applied according to the requirements of different applications. If very detailed data are available on a small number of species, then the stopover model may be appropriate, whereas for routine use on national data describing many different species the new mixture model may be appropriate, though for species with more than two broods the nonparametric approach may be best.

The GAI is described in Section 2: Section 2.1 gives three alternative expressions for the modeling of seasonal counts; the concentrated likelihood approach is outlined for the Poisson case in Section 2.2, and extended by an iterated concentrated likelihood method for the negative binomial and zero-inflated Poisson distributions in Sections 2.3 and 2.4. Efficiency is discussed in Section 2.5. The derivation of the GAI is provided in Section 2.6. Section 3 presents a series of examples of the GAI applied to UKBMS data, chosen to illustrate the flexibility of the approach. The last example indicates how a multi-year model can be formed, using covariates. The article ends with a discussion and further new approaches are given in the Supplementary Material.

2. Generalized Abundance Index

Here, each count $y_{i,j}$ in a given year is treated as the realization of an appropriate discrete random variable, which may be Poisson, negative binomial, or zero-inflated Poisson. 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. We shall not impute missing data as our focus is not primarily site specific.

Also, in this article the expectation of the distribution, $\lambda_{i,j}$, will be modeled as a product of the site parameter, N_i, which represents the relative abundance for the ith site, and general $a_{i,j} \equiv a_i(t_{i,j}, \boldsymbol{\theta})$, which denotes a function describing the seasonal variation in counts in terms of a small set of parameters $\boldsymbol{\theta}$:

$$\lambda_{i,j} = N_i a_{i,j}.$$

We specify a particular GAI using the x/z notation, with x denoting the distribution and z the choice for $a_{i,j}$. In this article, we consider x as P, ZIP, and NB for the Poisson, zero-inflated Poisson, and negative-binomial distributions, respectively. In the following sections possible options for z are described.

2.1. *Functions for $a_{i,j}$*

The function $a_{i,j}$ may be any general function which describes the seasonal variation in counts over the monitoring period, and we present both nonparametric and parametric options.

2.1.1. Splines. For illustration, we adopt simple cubic B-splines as in the GAM, such that

$$a_{i,j} = \exp\left\{\alpha_0 + \sum_{d=1}^{f} \alpha_d B_d(t_{i,j})\right\},$$

where $B_d(t_{i,j})$ are the basis functions and f is the degrees of freedom, defined as the sum of the degree of the spline (in this case 3 for cubic splines) and the number of knots minus one. Six knots were used in the example in this article (Section 3.1), but other choices had minimal effect on the results. To formulate the B-spline basis matrix in $\{a_{i,j}\}$ within the concentrated likelihood framework, we use the `splines` package in `R` (R Core Team, 2015), rather than the `mgcv` package used in the GAM approach as the latter is too complex to use in this context. The optimal number of knots could be selected automatically, for example, using cross-validation, as in the `mgcv` package. Model notation is x/C. Apart from the choice of f, the P/C GAI corresponds to fitting the first stage of the GAM approach and the seasonal pattern is the same across sites, as for the GAM approach.

2.1.2. Mixture model. In this case, $a_{i,j}$ is taken as a mixture of B Normal probability density functions so that

$$a_{i,j} = \sum_{b=1}^{B} w_{i,b} \frac{1}{\sigma_{i,b}\sqrt{2\pi}} \exp\left\{-\frac{(t_{i,j} - \mu_{i,b})^2}{2\sigma_{i,b}^2}\right\}, \qquad (2)$$

where $w_{i,b}$, $\mu_{i,b}$, and $\sigma_{i,b}$ correspond to the weight, mean, and standard deviation, respectively, for the ith site and bth brood, and $\sum_{b=1}^{B} w_{i,b} = 1$. For a univoltine species, where $B = 1$, $a_{i,j}$ would be the single Normal probability density function

$$a_{i,j} = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{-\frac{(t_{i,j} - \mu_i)^2}{2\sigma_i^2}\right\}. \qquad (3)$$

This model has the potential flexibility of having different mean and scaling parameters for different sites, which is an advantage compared to the GAM and spline. Model notation is x/N_B.

2.1.3. Stopover model. The stopover model has been described in Section 1.2.2. It was originally proposed to describe counts of migrating birds breaking their journey to rest and feed. An additional attractive feature of this model is that it can account for individuals being seen in multiple weeks. In Web Appendix A, we provide links between stopover and mixture models. The stopover model valuably estimates survival probabilities, but it will also do this when applied to data simulated from the mixture model, without survival parameters. When we have examined this feature, we have found that the resulting estimates of survival are small, in line with the predictions of equations (1) and (2) in Web Appendix A. This finding could be potentially misleading if the model is used uncritically. For more discussion, including a simulation study of performance and discussion of parameter redundancy, see Matechou et al. (2014). Model notation is x/SO_B.

2.2. *Concentrated Likelihood for the Poisson Case*

The Poisson distribution with expectation $\lambda_{i,j} = N_i a_{i,j}$ gives the likelihood

$$\mathrm{L}(\mathbf{N}, \boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^{S}\prod_{j=1}^{T} \frac{\exp(-N_i a_{i,j})(N_i a_{i,j})^{y_{i,j}}}{y_{i,j}!}. \tag{4}$$

Maximization of this likelihood is straightforward but cumbersome when data arise from many sites. However, the number of parameters to estimate can be reduced appreciably by optimizing a concentrated (or profile) likelihood as follows. Using the notation $a_{i,\cdot} = \sum_{j=1}^{T} a_{i,j}$,

$$\begin{aligned}\ell = \mathrm{Log}(\mathrm{L}) = &-\sum_{i=1}^{S} N_i a_{i,\cdot} + \sum_{i=1}^{S} y_{i,\cdot}\log(N_i) \\ &+ \sum_{i=1}^{S}\sum_{j=1}^{T} y_{i,j}\log(a_{i,j}) - \sum_{i=1}^{S}\sum_{j=1}^{T}\log(y_{i,j}!).\end{aligned}$$

Then,

$$\frac{\partial \ell}{\partial N_i} = -a_{i,\cdot} + \frac{y_{i,\cdot}}{N_i},$$

and equating to zero we obtain

$$N_i = \frac{y_{i,\cdot}}{a_{i,\cdot}}, \tag{5}$$

which estimates $\{N_i\}$ by scaled site totals. It is interesting to compare this expression with that of equation (1). In that case, if the sampling times are roughly equidistant then we can see that the index in (1) is also, approximately, proportional to the sum of the site-specific annual totals, as $y_{i,1} \approx y_{i,T} \approx 0$, for all i. Substituting the expression for $\{N_i\}$ in (4) results in a Poisson likelihood with expectation $\lambda_{i,j} = \frac{y_{i,\cdot}}{a_{i,\cdot}} a_{i,j}$, which we refer to as the concentrated likelihood, which is maximized with respect to only the parameters, $\boldsymbol{\theta}$, associated with $\{a_{i,j}\}$. Estimation of $\{N_i\}$ is then straightforward, by deriving $\hat{a}_{i,\cdot}$, and substituting into (5). An alternative approach for reducing the number of parameters, by treating the site parameters as random effects, is shown to generalize (5) (Web Appendix B).

2.3. *Negative-Binomial Case*

For the negative-binomial case (using the NB-2 form, Hilbe, 2011), the likelihood is given by

$$\mathrm{L}(\mathbf{N}, \boldsymbol{\theta}, r; \mathbf{y}) = \prod_{i=1}^{S}\prod_{j=1}^{T} \frac{\Gamma(y_{i,j}+r)}{\Gamma(r)y_{i,j}!}\left(\frac{N_i a_{i,j}}{r+N_i a_{i,j}}\right)^{y_{i,j}}\left(\frac{r}{r+N_i a_{i,j}}\right)^{r},$$

where r is the dispersion parameter and the expectation of $y_{i,j}$ is again $N_i a_{i,j}$. Hence,

$$\begin{aligned}\ell = \mathrm{Log}(\mathrm{L}) = \sum_{i=1}^{S}\sum_{j=1}^{T}\Bigg[&\log\left\{\frac{\Gamma(y_{i,j}+r)}{\Gamma(r)y_{i,j}!}\right\} + y_{i,j}\log(N_i a_{i,j}) \\ &- (r+y_{i,j})\log(r+N_i a_{i,j}) + r\log r\Bigg],\end{aligned} \tag{6}$$

leading to

$$\frac{\partial \ell}{\partial N_i} = \sum_{j=1}^{T}\left\{\frac{y_{i,j}}{N_i} - \frac{(r+y_{i,j})a_{i,j}}{r+N_i a_{i,j}}\right\}. \tag{7}$$

An exact solution for N_i does not result in this case from equating to zero. However, given that $\mathbb{E}(y_{i,\cdot}) = N_i a_{i,\cdot}$, if we make the approximation $y_{i,j} \approx N_i a_{i,j}$, then (7) reduces to

$$N_i = \frac{y_{i,\cdot}}{a_{i,\cdot}},$$

as in (5), which provides an approximation for a concentrated likelihood, which can be fitted as for the Poisson case. Exact maximum-likelihood parameter estimates can then be obtained as follows:

(i) Maximize the approximate concentrated likelihood from (6) with $N_{i,} = \frac{y_{i,\cdot}}{a_{i,\cdot}}$ to give parameter estimates for $\hat{a}_{i,j}$.
(ii) Based on $\hat{a}_{i,j}$, solve (7) numerically for N_i.
(iii) Insert the N_i from (ii) into (6) and optimize for the parameters for $\hat{a}_{i,j}$.
(iv) Iterate steps (ii)–(iii) until convergence.

2.4. *Zero-Inflated Poisson Case*

The approach for the negative-binomial applies also for the zero-inflated Poisson. The likelihood has the form

$$\begin{aligned}\mathrm{L}(\mathbf{N}, \boldsymbol{\theta}, \psi; \mathbf{y}) = &\prod_{i=1}^{S}\prod_{j=1}^{T}\left\{1-\psi+\psi e^{-N_i a_{i,j}}\right\}^{1-\delta_{i,j}} \\ &\times\left\{\frac{\psi e^{-N_i a_{i,j}}\left(N_i a_{i,j}\right)^{y_{i,j}}}{y_{i,j}!}\right\}^{\delta_{i,j}},\end{aligned}$$

where $1-\psi$ accounts for additional zeros, and

$$\delta_{i,j} = \begin{cases} 1 & \text{if } y_{i,j} > 0 \\ 0 & \text{if } y_{i,j} = 0. \end{cases}$$

Then,

$$\ell = \text{Log(L)} = \sum_{i=1}^{S}\sum_{j=1}^{T}\Big\{(1-\delta_{i,j})\log\big(1-\psi+\psi e^{-N_i a_{i,j}}\big) + \delta_{i,j}\log\left(\frac{\psi}{y_{i,j}!}\right) - \delta_{i,j}N_i a_{i,j} + \delta_{i,j}y_{i,j}\log\big(N_i a_{i,j}\big)\Big\}, \tag{8}$$

and differentiating with respect to N_i gives

$$\frac{\partial \ell}{\partial N_i} = \sum_{j=1}^{T}\left\{\frac{-\psi a_{i,j}(1-\delta_{i,j})e^{-N_i a_{i,j}}}{1-\psi+\psi e^{-N_i a_{i,j}}} - \delta_{i,j}a_{i,j} + \frac{\delta_{i,j}y_{i,j}}{N_i}\right\}. \tag{9}$$

Steps (i)–(iv) in Section 2.3 can then be applied to obtain maximum-likelihood parameter estimates, but replacing (6) and (7) with (8) and (9), respectively.

2.5. *Increased Efficiency*

Step (ii) in Section 2.3 is easily achieved using the `uniroot` function in `R` (R Core Team, 2015) and only a few iterations of steps (ii)–(iii) are generally needed. The concentrated likelihoods are functions of S fewer parameters than the original likelihoods. Substantial reductions in computation time are then made, which we demonstrate via simulation in Web Appendix C.

2.6. *Generalized Abundance Index*

For each year for any particular model, we use the average of the estimated site parameters, $\{\hat{N}_i\}$, as a measure of abundance, given by

$$G = \frac{1}{S}\sum_{i=1}^{S}\hat{N}_i. \tag{10}$$

If desired, the GLM final stage of the GAM-based approach could be employed to account for different sites being sampled in different years. Although the resulting additional computation is fast, it is unlikely to be necessary for large data sets, including those for most species in the UKBMS.

The model is fitted separately for each year, G is calculated in each case and the results are plotted against time to provide an index of abundance. Errors may be derived by nonparametric bootstrapping, where for each replicate the GAI is fitted to data for a random sample of sites, drawn with replacement, or by standard inversion of the estimated Hessian at the likelihood maximum, followed by use of the multivariate delta method.

3. Examples

In work not reported here, we have checked the accuracy of the GAI by application to simulation data (Dennis, 2015, Tables 5.3 and 5.4). We now apply the GAI for a series of examples of butterfly transect counts from the UKBMS, to illustrate the range of modeling alternatives. The species selected cover univoltine, bivoltine, and a multivoltine species, where adults have one, two, and more than two flight periods per year, respectively. Supplementary tables and figures are provided in Web Appendix D. Latin names for the species studied in this article are provided in Web Table 4.

3.1. *Splines*

A spline is advised for species with complex seasonal flight patterns, which may not be easily modeled parametrically. We demonstrate the P/C GAI for Speckled Wood, a multivoltine species whose flight pattern tends to exhibit three overlapping broods per year. The flight period is further complicated since the Speckled Wood overwinters as both caterpillar and pupa, which may emerge at different times. We apply the GAI and GAM to data from 1980 to 2011 for a subset of 100 sites.

Comparable seasonal pattern curves are predicted from the GAM and P/C GAI (Figure 1), as well as similar indices of relative abundance (Web Figure 3), despite the simplicity and greater speed of fitting the GAI, compared to the GAM approach.

3.2. *Mixture Model*

We examine the performance of the x/N_2 GAI for a selection of bivoltine UK butterfly species. For demonstration, we fit the model where x is Poisson, zero-inflated Poisson, and negative-binomial for five species and make comparisons with the GAM approach. For each species, models were fitted to data for each year from 1978 to 2011 separately, and an index of abundance then formed as defined in Section 2.6. Confidence intervals were derived via bootstrapping.

In order to compare the two methods, each index was standardized to have zero mean and unit variance. Where a species has been observed in more than 100 sites within a given year (true of all species but Small Blue), each model was fitted to a common random sample of 100 sites. For illustration we consider the homoscedastic case, where $\sigma_1 = \sigma_2$, and also no site variation in distribution location and scale parameters. We let $\mu_2 = \mu_1 + \mu_d$, where $\mu_1, \mu_d \geq 0$ to ensure $\mu_2 \geq \mu_1$. For $B = 2$, we denote $w_1 = w$, and $w_2 = 1 - w$.

There were minimal differences in the indices derived from the P, ZIP, and NB GAIs, but NB performed best in terms of AIC and dispersion (Web Figures 4 and 5). The latter is unavailable for ZIP. The negative-binomial GAI was also found to perform best when compared to a Poisson GAI and an alternative hierarchical model in Web Appendix B, although the Poisson GAI produces similar results and is more computationally efficient. The indices of abundance from the GAM and GAI show similar patterns (Figure 2). The differences seen are in part due to the use of just 100 sites in the analyses of the article. The differences diminish appreciably when either the last GLM stage of the GAM is employed for the GAI, to account for different sites being sampled in different years, or when much larger numbers of sites are analyzed. There is a greater difference for the Small Blue,

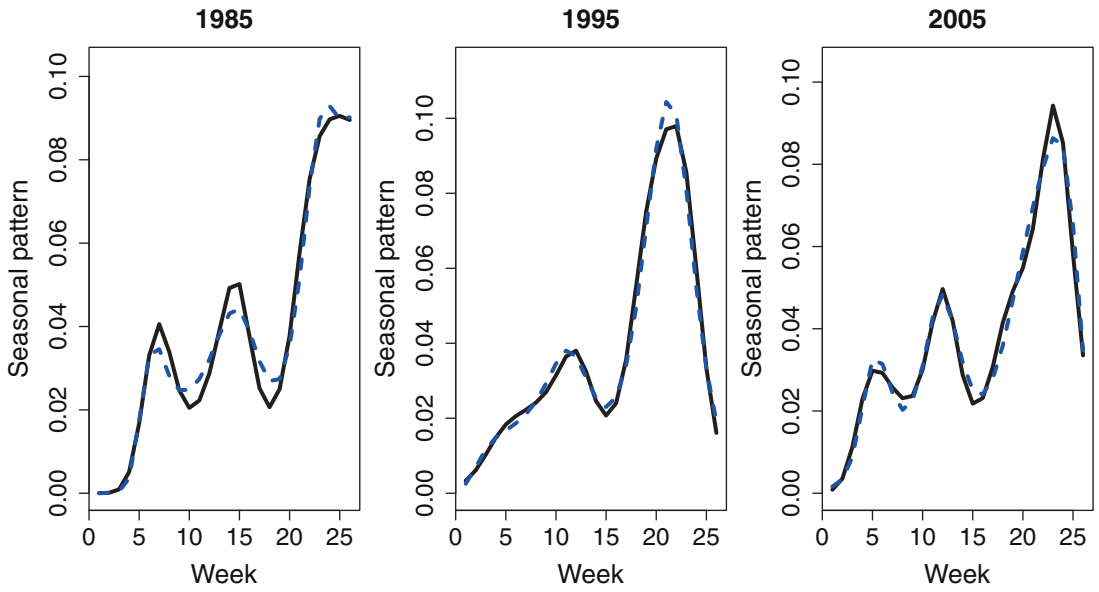

Figure 1. Predicted seasonal pattern for each week since the start of the season for the GAM approach (solid) and P/C GAI (dashed) for Speckled Wood. This figure appears in color in the electronic version of this article.

particularly for earlier years in the index, which may be due to the lack of sites available for this species which is a habitat specialist.

The confidence intervals for the GAI are narrower than those for the GAM for three of the five species, and are never greatly wider (Table 1 and Web Figure 6). The GAI is substantially quicker than the GAM (Table 1). This is of vital importance when data on multiple species are analyzed each year, and bootstrap confidence intervals are also required.

3.3. *Stopover Model*

For illustration, we apply the P/SO_1 GAI to data for two univoltine species to assess changes in survival probability ϕ over time (Figure 3a). Matechou et al. (2014) considered data for one species in a single year but explored allowing ϕ to vary with time or age. Here, only constant ϕ (within each year) is considered.

Analysis was restricted to start from the first year in which the species was recorded at at least 30 sites. Higher survival probabilities are correlated with earlier emergence in the season (Figure 3b), which generates an hypothesis for further investigation; for example, earlier emergence may expose individuals to cooler temperatures leading to increased longevity.

3.4. *Goodness of Fit*

As in Matechou et al. (2014), we may use dispersion (residual deviance/degrees of freedom) as an overall measure of goodness of fit, and a check of whether there is overdispersion present which needs to be taken account of. An illustration is provided in Web Table 1. Web Figure 5 shows the overall improvement in fit of moving from model P/N_2 to model NB/N_2. Goodness of fit may also be examined graphically, to check for outliers and model deficiencies, and Matechou et al. (2014) do this on a site basis for one species and 1 year, plotting observed values against expected. Another possibility, not shown here, is to combine the data from all sites for individuals years, to reduce the number of plots involved.

3.5. *Regressing Parameters on Year and Northing*

In this section, we demonstrate the flexibility of the GAI for the inclusion of covariates, which was not possible with the GAM approach. Rather than fitting the model separately to data for each year, a single concentrated likelihood can be maximized over multiple years. The number of parameters can be reduced by restricting appropriate parameters over time, for example, to be constant or linearly time-varying.

For demonstration, we apply models to data for Wall Brown. We use the P/N_2 GAI, but fit a single multi-year model. The parameters w, the mixing probability for the two broods, μ_1 and μ_d, the mean flight date of the first brood and the separation of the two broods, could vary linearly with year, northing, or an additive or multiplicative combination of both. We allowed the standard deviation σ to be constant or linearly varying with year but consider only the homoscedastic case where $\sigma_1 = \sigma_2$.

The most complex model, which had 14 parameters and included an interaction between northing and year for w, μ_1, and μ_d, was favored in terms of AIC and has a dispersion value of 1.8 suggesting moderate overdispersion. The estimated seasonal pattern is provided for 3 years in Figure 4, each for a sample of northing values. The positive value of the slope for year for w at all but the most extreme latitudes suggests an overall trend for an increase in size of the first brood relative to the second brood over time (Web Table 5). The timing of the first brood is later further north, but has become earlier over time, and the difference in the timing of the two broods has increased over time. The standard deviation has changed minimally with time.

A different approach to analyzing data from multiple years is presented in Dennis et al. (2016); by making additional

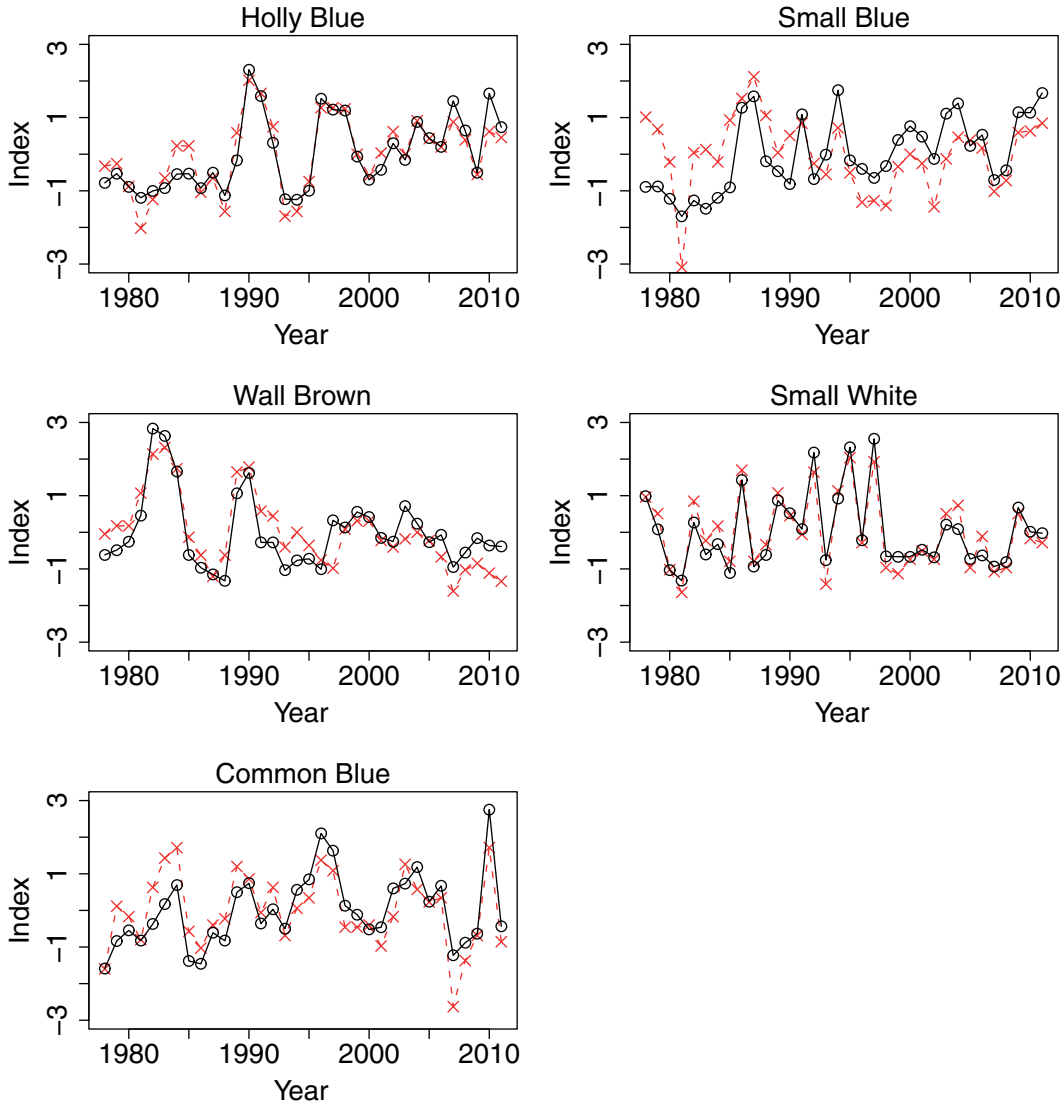

Figure 2. Relative abundance indices from the NB/N_2 GAI (solid line, circles) and GAM approach (dashed line, crosses). This figure appears in color in the electronic version of this article.

Table 1

Comparison of efficiency and accuracy for the GAM and P/N_2, ZIP/N_2, and NB/N_2 GAI, where m and s denote minutes and seconds, respectively

	Time for a single run				Mean CI width			
		GAI				GAI		
Species	GAM	P	ZIP	NB	GAM	P	ZIP	NB
Holly Blue	9 m	20 s	3 m	1 m	0.862	0.664	0.703	0.627
Small Blue	32 m	13 s	2 m	1 m	3.091	1.892	1.949	1.871
Wall Brown	39 m	23 s	3 m	2 m	0.860	1.089	1.147	1.096
Small White	23 m	28 s	3 m	3 m	0.998	0.954	0.954	0.938
Common Blue	22 m	26 s	3 m	2 m	1.066	1.305	1.328	1.338

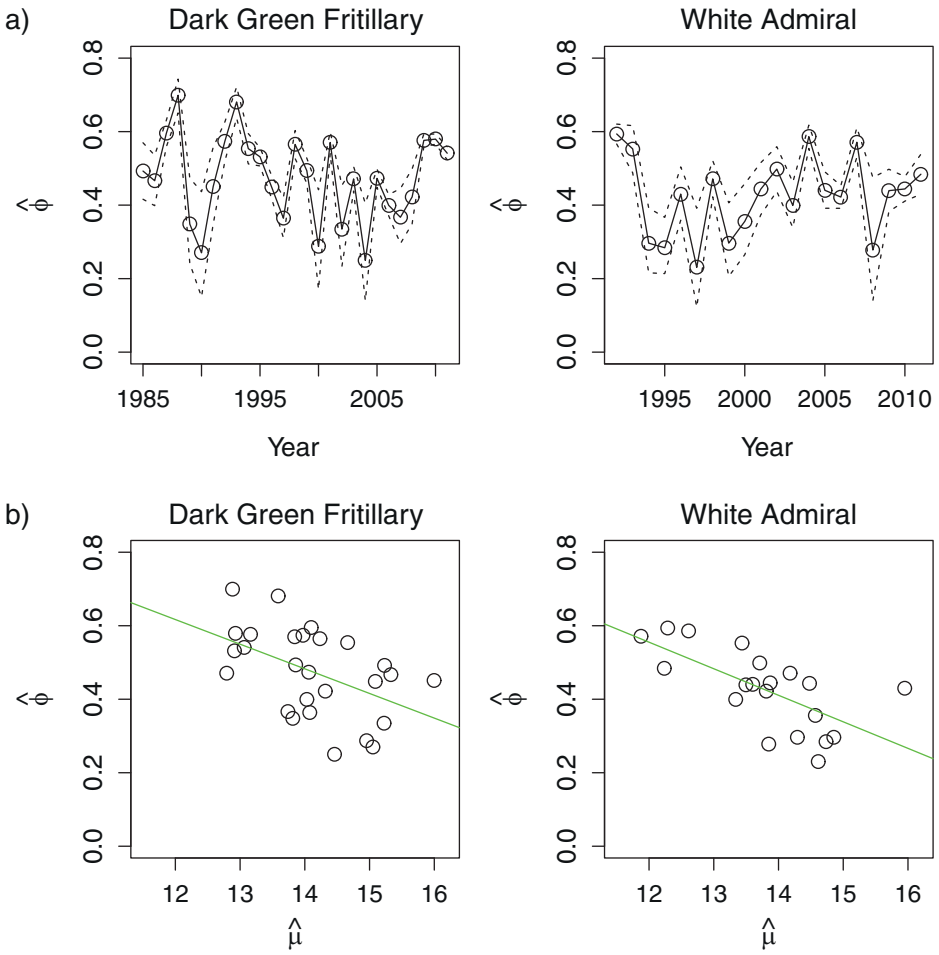

Figure 3. (a) Predicted survival probability (weekly) for each year and (b) average week of emergence (μ) versus predicted survival probability (weekly). A P/SO_1 GAI was fitted to data for the two univoltine species.

modeling assumptions population sizes are linked deterministically between years and broods via appropriate productivity parameters.

4. Discussion

We have presented a generalized abundance index which unifies and extends existing methods for estimating abundance of seasonal invertebrates. It is fitted efficiently using maximum-likelihood estimation and a concentrated likelihood. The GAI is suitably general for parametric or nonparametric functions to be chosen specific to the study species and scenario. Splines may be preferable for species with complex flight periods, such as migrants. The new mixture model is a simplification of the stopover model. The stopover model provides additional insights via the estimates of survival. However, for wider scale analysis, the mixture model is more efficient and akin to the methods currently used for deriving abundance indices. The mixture model may also be more suitable in cases with limited data, since the stopover model has greater demands on data in order to estimate survival. When spatio-temporal models are fitted to long-term data for many species and sites, an important consideration is the computational effort required. Model fitting is very time-consuming for the GAM approach. When there are many sites, bootstrapping can take weeks for a single UKBMS species. The GAI shows substantial improvements in computation time which will reduce the time and resources required for data processing, leading to faster outputs and feedback of results to recorders and policy makers. The provision of such feedback is essential for the motivation and retention of participants in citizen science projects such as the UKBMS.

The GAM approach assumes the seasonal pattern to be static across sites within each year. In principle, geographic variation could be incorporated in the smoothing component but in practice that does not appear to be straightforward and robust. The parametric approaches within the GAI can readily incorporate available covariates, such as northing, land cover, weather, or growing degree days (Hodgson et al., 2011). Novel description of spatial and temporal variation in seasonal pattern will benefit phenological studies, which for butterfly data have involved measures such as mean first appearance, mean peak appearance, and mean length of the flight period (Roy and Sparks, 2000; Karlsson, 2014; Roy et al., 2015). Hodgson et al. (2011) utilized GAMs for studying spatio-temporal variation in phenology, but changes in phenology and voltinism can be studied more flexibly through the GAI,

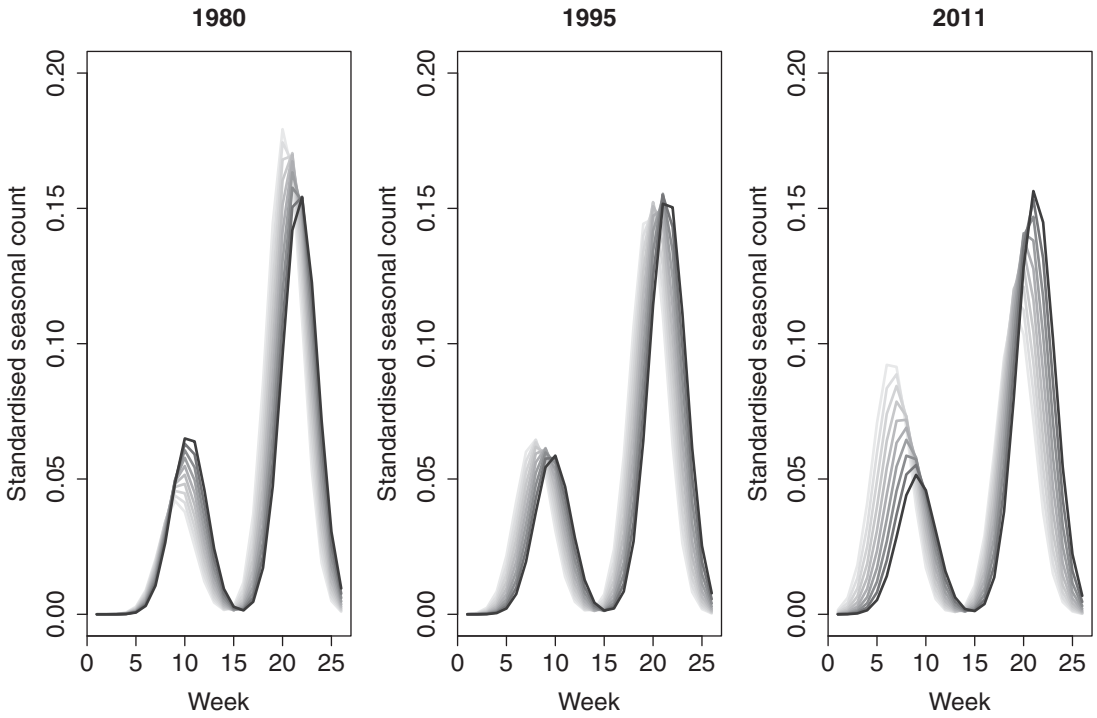

Figure 4. Predicted seasonal pattern (standardized seasonal count) for each week since the start of the season for the multi-year P/N_2 GAI (1978-2011) for Wall Brown for 3 years. Each line represents one of 10 equally spaced Northing values between 17 km (light gray) and 667 km (dark gray).

extending the capacity to study the nonuniform effects of climate change.

The GAM approach accounts for turnover in sites between years. This is not done in the GAI, but comparable results to the GAM are produced despite the simplicity of the model. If necessary, an additional GLM stage can be added to the GAI. Time variation in sites sampled may need to be accounted for when there is a limited number of sites. Trends in relative abundance for individual sites can be estimated by the GAI, which may be of interest for conservation and monitoring of certain locations. For the GAM approach trends in abundance are assumed to be spatially constant, which may be an unrealistic assumption.

The presented examples demonstrate the generality of the GAI framework, and application to multiple years and species for the first time outside the GAM context. In practice, wider model selection would be required in any application of the GAI. Alternatives to the Normal distribution in the parametric approaches, such as asymmetric distributions to account for skewness in emergence are also possible (Calabrese, 2012). Clearly the "best" model choice will be dependent on both the purpose of the study and the species of interest.

The gains in efficiency achieved by the GAI arise from maximizing a concentrated likelihood. The proposed iterative concentrated likelihood approach for negative-binomial and zero-inflated Poisson obtains the correct result and is still considerably quicker than previous methods. The Poisson distribution may be sufficient if an index is the required output of a study, since the resulting GAIs are quick with minimal differences in accuracy. Using random effects to describe $\{N_i\}$ is slower and less straightforward than the concentrated likelihood method (Web Appendix B).

The GAI is a robust and flexible framework that can produce new insights relevant to the monitoring and conservation of invertebrates with both efficiency and accuracy. An `R` program for the GAI is available in the Supplementary Material.

5. Supplementary Materials

The Web Appendices referenced in Sections 2, 3, and 4, together with `R` code, are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

This work was part-funded by EPSRC grants EP/1000917/1 and EP/P505577/1. We thank an associate editor, two referees, and Martin Ridout for their useful comments. The UKBMS is operated by the Centre for Ecology & Hydrology and Butterfly Conservation and funded by a multi-agency consortium including the Countryside Council for Wales, Defra, the Joint Nature Conservation Committee, Forestry Commission, Natural England, the Natural Environment Research Council, the Northern Ireland Environment Agency, and Scottish Natural Heritage. The UKBMS is indebted to all volunteers who contribute data to the scheme.

References

Altermatt, F. (2010). Climatic warming increases voltinism in European butterflies and moths. *Proceedings of the Royal Society B, Biological Sciences* **277**, 1281–1287.

Brereton, T. M., Botham, M. S., Middlebrook, I., Randle, Z., Noble, D. G., and Roy, D. B. (2014). United Kingdom Butterfly Monitoring Scheme report for 2013. Technical Report, Centre for Ecology and Hydrology and Butterfly Conservation.

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

Calabrese, J. M. (2012). How emergence and death assumptions affect count-based estimates of butterfly abundance and lifespan. *Population Ecology* **54**, 431–442.

Catchpole, E. A., Kgosi, P. M., and Morgan, B. J. T. (2001). On the near-singularity of models for animal recovery data. *Biometrics* **57**, 720–726.

Chambers, J. M. and Hastie, T. J. (1991). *Statistical Models in S*. Boca Raton: Chapman & Hall/CRC.

Convention on Biological Diversity (2006). Framework for monitoring implementation of the achievement of the 2010 target and integration of targets into the thematic programmes of work, COP 8 Decision VIII/15. *www.cbd.int/decisions*

Defra (2013). UK Biodiversity indicators in your pocket 2013. *Published by Defra on Behalf of the UK Biodiversity Partnership, Defra, London*.

Dennis, E. B. (2015). *Development of statistical methods for monitoring insect abundance*. PhD thesis, University of Kent.

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., Morgan, B. J. T., Freeman, S. N., Roy, D. B., and Brereton, T. (2016). Dynamic models for longitudinal butterfly data. *Journal of Agricultural, Biological, and Environmental Statistics* **21**, 1–21.

Gaston, K. J. (1991). The magnitude of global insect species richness. *Conservation Biology* **5**, 283–296.

Hilbe, J. M. (2011). *Negative Binomial Regression*. New York: Cambridge University Press.

Hodgson, J. A., Thomas, C. D., Oliver, T. H., Anderson, B. J., Brereton, T. M., and Crone, E. E. (2011). Predicting insect phenology across space and time. *Global Change Biology* **17**, 1289–1300.

Isaac, N., Cruickshanks, K., Weddle, A., Marcus Rowcliffe, J., Brereton, T., Dennis, R., et al. (2011). Distance sampling and the challenge of monitoring butterfly populations. *Methods in Ecology and Evolution* **2**, 585–594.

Karlsson, B. (2014). Extended season for northern butterflies. *International Journal of Biometeorology* **58**, 691–701.

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.

Pollard, E. and Yates, T. J. (1993). *Monitoring Butterflies for Ecology and Conservation: The British Butterfly Monitoring Scheme*. London: Chapman & Hall.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rothery, P. and Roy, D. B. (2001). Application of generalized additive models to butterfly transect count data. *Journal of Applied Statistics* **28**, 897–909.

Roy, D. B., Oliver, T. H., Botham, M. S., Beckmann, B., Brereton, T., Dennis, R. L. H., et al. (2015). Similarities in butterfly emergence dates among populations suggest local adaptation to climate. *Global Change Biology* **21**, 3313–3322.

Roy, D. B. and Sparks, T. H. (2000). Phenology of British butterflies and climate change. *Global Change Biology* **6**, 407–416.

Thomas, J. A. (2005). Monitoring change in the abundance and distribution of insects using butterflies and other indicator groups. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 339–357.

van Swaay, C. A. M., Nowicki, P., Settele, J., and van Strien, A. J. (2008). Butterfly monitoring in europe: methods, applications and perspectives. *Biodiversity and Conservation* **17**, 3455–3469.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton: Chapman & Hall/CRC.

Received June 2015. *Revised November* 2015.
Accepted January 2016.