Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014, **5,** 1373–1382 doi: 10.1111/2041-210X.12279

SPECIAL ISSUE. MODELLING DEMOGRAPHIC PROCESSES IN MARKED POPULATIONS: PROCEEDINGS OF THE EURING 2013 ANALYTICAL MEETING

Goodness-of-fit of integrated population models using calibrated simulation

Panagiotis Besbeas[1,2*] and Byron J.T. Morgan[2]

[1]*Department of Statistics, Athens University of Economics and Business, 76 Patission Str, Athens 10434, Greece; and* [2]*National Centre for Statistical Ecology, School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury CT2 7NF, UK*

Summary

1. Integrated population modelling is proving to be an important and useful technique in statistical ecology. However, there is currently no simple formal method for judging how well models fit data, when potentially several different data sets described by different structured models are being analysed in combination.

2. We propose and evaluate a new approach, of calibrated simulation. Here, comparative data sets are obtained from simulating data when model parameter values are obtained from the assumed asymptotic normal distribution of the maximum-likelihood estimators from the real data. The approach is motivated and justified by Bayesian *P*-values. Calibration of the resulting statistics is achieved as repeated data sets are easily simulated from the fitted model. The method requires the specification of model discrepancy measures, and we show how different measures can highlight different aspects of fit.

3. Calibration is only strictly necessary if the statistics proposed may appear to be extreme.

4. The approach of using calibrated simulation to check the goodness-of-fit of integrated population models is demonstrated by application to data sets on lapwings and herons. In each case, there are two data sets involved in the integrated analysis, and for each component data set, discrepancy measures of goodness-of-fit are obtained. For the lapwing application, as replication is efficient, it is possible to calibrate the procedure simply by using additional simulations. The heron application is shown to be feasible, but is substantially harder to calibrate, due to the presence of productivity thresholds that need to be estimated using profile likelihood methods. We demonstrate the importance of taking more than one discrepancy measure for time-series data. Avenues for future research are outlined. This article has supplementary materials on line.

Key-words: asymptotic normality, discrepancy measure, goodness-of-fit, herons, integrated population modelling, kernel density estimation, lapwings

Introduction

When independent data sets are obtained from observations on wild animals then after appropriate stochastic models are constructed, likelihoods can be formed for each data set and multiplied together to give a single joint likelihood. This was done by Besbeas *et al.* (2002) (hereafter BFMC), for census and demographic data on birds. The advantages of maximizing the joint likelihood were that common parameters were estimated with greater precision and in addition it was possible to provide a coherent estimate of a productivity parameter, together with its standard error, that would not otherwise have been possible. The approach generalizes naturally to when there are additional data and corresponding likelihoods, for example on productivity, and is called integrated population modelling; see for example, Tavecchia *et al.* (2009). The importance of the independence assumption is examined by simulation in Besbeas, Borysiewicz & Morgan (2009) and Abadi *et al.* (2010). More flexible approaches, based on hierarchical modelling and Bayesian analysis, are provided by Chandler & Clark (2014) and Mazzetta, Morgan & Coulson (2010), in which different observations can be obtained on the same animals, which are described by an underlying stochastic, possibly spatial, process. A useful bibliography of recent work on integrated population modelling has been provided by Schaub & Abadi (2011).

In BFMC, the demographic data were ring recoveries, providing information on survival which is typically expressed in terms of annual probabilities, and the corresponding likelihood was product multinomial; see for example, Freeman & Morgan (1992). The census data were described using a state-space model incorporating survival and productivity parameters; see for example, Durbin & Koopman (2001). In that case, an approximate likelihood resulted from using the Kalman fil-

*Correspondence author. E-mail: p.t.besbeas@kent.ac.uk

ter. Goodness-of-fit was considered graphically, separately for the two individual component data sets, using plots of observed vs. expected values for the demographic data and Q-Q plots for the prediction errors from the Kalman filter used to produce the likelihood for the census data. In other applications, goodness-of-fit of time-series components of integrated population modelling has been checked visually by superimposing observed and fitted trajectories, as done for instance in Besbeas & Morgan (2012). Thus, up to now, goodness-of-fit checking for integrated population models is either not done, or done in an *ad hoc* fashion, and a new approach is required. In this paper, we follow the same strategy as BFMC, considering the fit of the models to component data sets separately. However, we adopt a common procedure for all components of integrated population analysis and also present a simulation-based method for formally evaluating the resulting measures. Brooks, King & Morgan (2004) introduced Bayesian inference for integrated population modelling, and an advantage of the Bayesian approach is the availability of Bayesian *P*-values for judging goodness-of-fit. However, Bayesian *P*-values are implicitly dependent on the priors (see e.g. King *et al.* 2009, p138). In addition, more recently, King (2011) has identified problems with MCMC mixing, as did also Besbeas & Morgan (2012). Therefore, as in BFMC, here we use classical inference for model fitting. Our particular focus in this paper is on checking the goodness-of-fit of models that have already been chosen for data sets. Model selection in integrated population modelling is the topic of a companion paper by P.T. Besbeas, R.S. McCrea and B.J.T. Morgan, in prep.

Methods

STATE-SPACE MODELS

State-space models involve two linked stochastic equations, a transition equation and an observation equation; see for example, Durbin & Koopman (2001). As an illustration, we give below the transition equation for the lapwing *Vanellus vanellus* census data, taken from BFMC. We assume no sex effect on survival and that breeding starts at age 2:

$$\begin{pmatrix} N_{1,t} \\ N_{a,t} \end{pmatrix} = \begin{pmatrix} 0 & p_t\phi_{1,t} \\ \phi_{a,t} & \phi_{a,t} \end{pmatrix} \begin{pmatrix} N_{1,t-1} \\ N_{a,t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{a,t} \end{pmatrix} \qquad \text{eqn 1}$$

Here $N_{1,t}$ and $N_{a,t}$ denote the numbers of one-year-old female birds and (adult) female birds aged ≥2 years, respectively, at time t, and $\phi_{1,t}$ and $\phi_{a,t}$ are, respectively, the annual survival probabilities of birds in their first year of life and of birds aged 1 year and older at time t. The parameter p_t denotes the annual productivity of females per female, and the $\varepsilon_{1,t}$ and $\varepsilon_{a,t}$ terms are errors, which are taken as normally distributed with variances given by suitable Poisson and binomial expressions; see BFMC for details.

We are not able to observe both $N_{1,t}$ and $N_{a,t}$, as information is available only on the number breeding, $N_{a,t}$, and this feature is described by the observation equation:

$$y_t = (0,1) \times (N_{1,t}, N_{a,t})' + \eta_t \qquad \text{eqn 2}$$

which includes a term to describe measurement error. A simple possibility, which we adopt, is to assume that $\eta_t \sim N(0,\sigma^2)$.

PARAMETER REGRESSIONS

For lapwings, following model selection, the survival probabilities, which are common to both likelihoods in the integrated population model, are logistically regressed on a measure of winter severity, ω_t, the number of days when the temperature was below freezing in year t, as measured in central England. Specifically, in an obvious notation,

$$\text{logit}(\phi_{1,t}) = \beta_0 + \beta_1\omega_t \qquad \text{eqn 3}$$

$$\text{logit}(\phi_{a,t}) = \xi_0 + \xi_1\omega_t. \qquad \text{eqn 4}$$

The reporting probability of dead birds, λ_t, which is a component of the ring–recovery likelihood, is logistically regressed on time, this reflecting the decline over time in reporting probability of dead birds in England. In addition, productivity, p_t, which only appears in the state-space model, is logarithmically regressed on time, as it has been inferred to be decreasing, and as such is responsible for the declining population size of this species in England; see BFMC. We may denote the integrated population model as $\phi_1(\omega_t),\phi_a(\omega_t)/\lambda(\text{year})/p(\text{year})$.

USE OF SIMULATION TO CHECK GOODNESS-OF-FIT

Constructing diagnostics for judging the goodness-of-fit of state-space models to data is complex, as explained in Newman *et al.* (2014, p117). A particular use of simulation was suggested by Brooks, Catchpole & Morgan (2000) in the context of the analysis of mark–recovery and recapture data from wild birds, and it is also proposed by Johnson (2004). We extend the approach to integrated population modelling. The work is motivated by Bayesian *P*-values; see for example, Brooks, Catchpole & Morgan (2000), where multiple simulations are obtained from the posterior distribution for the parameters of the model being considered. Once the integrated model is fitted to all of the data, then s sets of simulated data sets, of dimensions matched to those of the real data sets, are obtained repeatedly from the component models. Each set is simulated with parameter values drawn from the assumed asymptotic multivariate normal distribution of the maximum-likelihood parameter estimates from fitting the real data.

In detail, suppose that $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}$ are, respectively, the maximum-likelihood estimates from fitting the real data, and associated dispersion matrix obtained from inverting the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$. For each simulated parameter value, $\boldsymbol{\theta}_i \sim \text{N}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}})$, we might calculate a measure of the discrepancy between the data, $\mathbf{x}$ and the corresponding model, $D(\mathbf{x}; \boldsymbol{\theta}_i)$, and we also simulate a new data set $\mathbf{x}_i$ from the model. For each new data set, we then calculate $D(\mathbf{x}_i; \boldsymbol{\theta}_i)$, and a scatter plot is obtained of $D(\mathbf{x}_i; \boldsymbol{\theta}_i)$ vs. $D(\mathbf{x}; \boldsymbol{\theta}_i)$. If the model fits the data well, then one would expect approximately half of the

points in the scatter to be above the line of unit slope through the origin. We denote the proportion of points above the line by $p_c = n_c/s$, where n_c is the corresponding number of points above the line. For integrated population modelling, we can obtain such plots and proportions separately for each of the data sets in the analysis. An attraction of this approach is that there is complete freedom in the choice of the measures of discrepancy that may be used, and furthermore more than one might be used for each data set, as recommended by Gelman, Meng & Stern (1996). For example, Millar & Meyer (2000) used four different measures when assessing the fit of a surplus-production model for fisheries data: one was a standard chi-square, while the other three were specific to the problem. They obtained *P*-values of 0·69, 0·27, 0·50 and 0·42 which they judged indicated that the model fitted the data sufficiently well. However, we note the variation in the values obtained, which indicates the importance of taking several measures. As observed by Johnson (2004), the distribution of *P*-values is also unknown, and they cannot be easily calibrated due to the computation time required. By running simulations for replicated versions of the real data, we provide such a calibration for the methods in this paper, without the need for multiple Markov chain Monte Carlo simulations required to calibrate Bayesian *P*-values.

If uninformative prior distributions are assumed for the model parameters, and if the assumption of asymptotic normality for the distribution of maximum-likelihood estimators is justified, then simulating as we do from the multivariate normal distribution will be similar to simulating from a posterior distribution for the parameters, producing Bayesian *P*-values. It is therefore important to check the assumption of multivariate normality for the problems that we consider, and we do that in the next section. Should the assumption of multivariate normality not hold then a possible approach, which we do not consider here would be to sample from a kernel density estimate from additional bootstrap sampling. An alternative would be to employ an appropriate reparameterization.

Simulation check of multivariate normality

The papers of Besbeas, Lebreton & Morgan (2003) and McCrea *et al.* (2010) have both made the multivariate normal assumption for the distribution of maximum-likelihood estimators in integrated population modelling when data arise from single-site and multi-site mark–recovery data, respectively. In each case, the assumption was found to perform well for the data considered. They checked profile likelihoods for selected parameters, and as a simpler alternative here we focus on univariate marginal distributions and examine all model parameters. We have conducted further simulations, of single-site mark–recovery data, based on the lapwing data set (see Results) and for several modifications of that set in which there are successive reductions in sample sizes. This has also been done for integrated modelling, where independent census data were also available. Illustrations are given in Fig. 1. In order to obtain replicate data sets matched to the real data, we have used nonparametric bootstrap sampling for the ring–recovery data and bootstrapping in the state-space framework (Stoffer & Wall 1991) for the time series. We believe that both of these approaches are novel for integrated population modelling. Even when mark–recovery data are very sparse indeed, the assumption of approximate univariate normality is satisfied for the parameter estimators. We attribute this to the fact that mark–recovery likelihoods will be approximately products of binomial distributions, corresponding to animals not recovered, and the good approximation of those binomial distributions by normals. Additional results are in the Appendix S1. The bootstrapping also provides a useful check of the Hessian-based standard errors which we have used. The agreement, shown in the Appendix S1, is excellent.

Choice of discrepancy measure

MARK–RECOVERY DATA

For mark–recovery data, there are different discrepancy measures that may be used. Brooks, Catchpole & Morgan (2000) use the Freeman–Tukey statistic (Freeman & Tukey 1950) in which, for expected values $\{e_i\}$, we define the following discrepancy measure:

$$D_{\mathrm{FT}}(\mathbf{x};\theta) = \sum_i (\sqrt{x_i} - \sqrt{e_i})^2, \qquad \text{eqn 5}$$

and an alternative is the Pearson chi-square statistic, incorporating an amalgamation level m to accommodate small values. Details of these two measures and their asymptotic equivalence when the model is correct are provided by Bishop, Fienberg & Holland (1975, p513). The difficulty with using the chi-square measure when data are sparse is the need for pooling cells with small expected values, which is not only arbitrary but results in differential weighting of the cells. We demonstrate this in Fig. 2 for model $\phi_1(\omega_t),\phi_a(\omega_t)/\lambda$(year) fitted to the lapwing data. We can see here that chi-square provides not a single discrepancy measure, but rather an infinite family of such measures, indexed by the amalgamation level used. Small amalgamation can give rise to acceptable p_c values, but for larger amalgamation, extreme non-zero counts, which may occur in the real data, are moved towards the main diagonal of the recovery table, resulting in increased discrepancies compared with simulated data, for which such extreme values would be relatively rare, and resulting small p_c values. If matching such extreme values is seen to be important, then the chi-square discrepancy measure will indicate poor fit of the model. This explains how different discrepancy measures can lead to different values and indeed different conclusions. This is further explored for these data in the Appendix S2, using simulation, when extreme values are less likely to occur. Also shown in Fig. 2 is a Monte Carlo investigation of the residual deviance, $2(\ell_{\max} - \hat{\ell})$, where $\ell_{\max}$ is the log likelihood under the maximal model, which fits a parameter for each observation, and $\hat{\ell}$ is the maximized log likelihood for model $\phi_1(\omega_t),\phi_a(\omega_t)/\lambda$

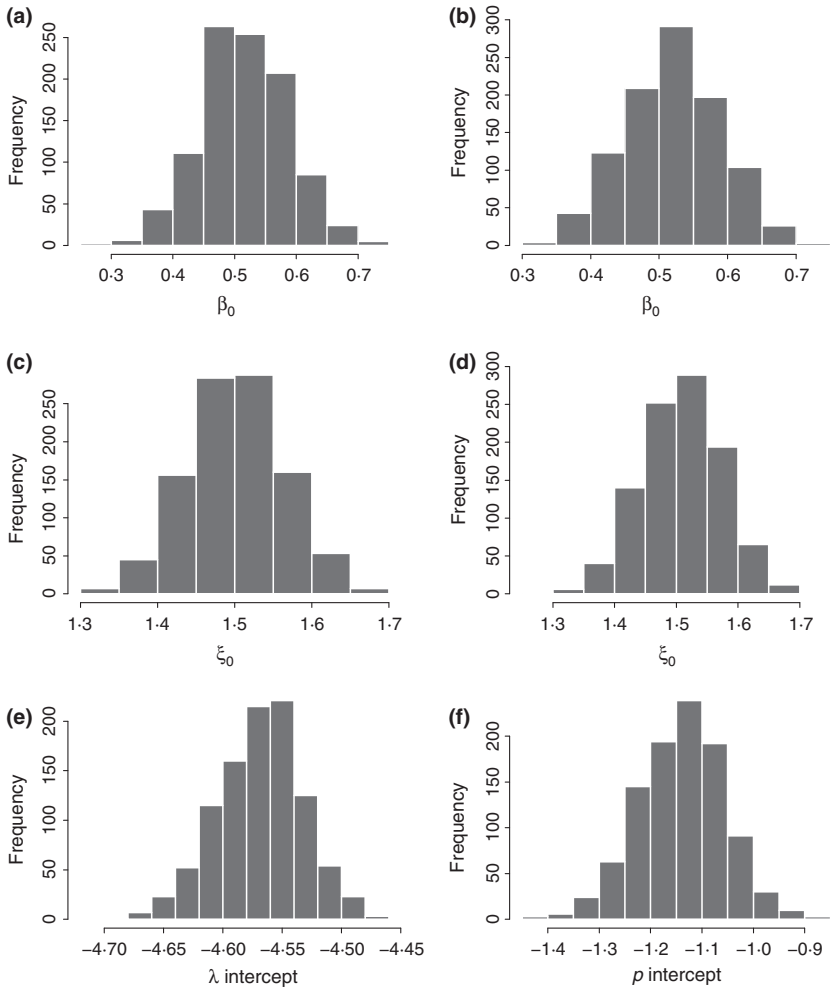

Fig. 1. Illustration of sampling distributions for estimated parameters from ring–recovery data alone, left column, and integrated population modelling, right column, derived by bootstrapping for the lapwing example. The parameters are (a,b) ϕ_1 intercept, β_0, (c,d) ϕ_a intercept, ξ_0, (e) λ intercept, (f) p intercept. 1000 bootstrap replicates were used in each case.

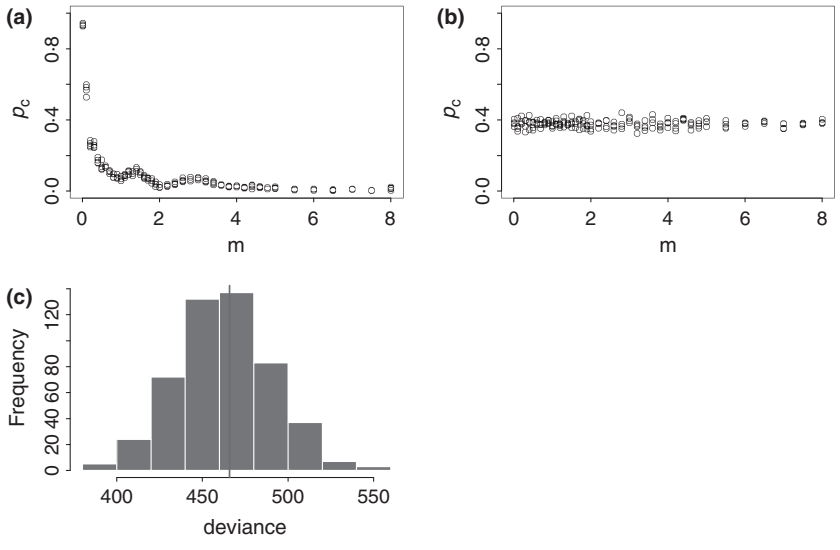

Fig. 2. The amalgamation level m determines the value at which small cells are pooled when forming a Pearson chi-square discrepancy measure. Panel (a) shows the effect of m on the chi-square discrepancy, for model $\phi_1(\omega_t),\phi_a(\omega_t)/\lambda$(year) fitted to the lapwing ring–recovery data. For comparison, panel (b) trivially shows the stability of the Freeman–Tukey discrepancy, as it does not involve m. There are four circles per amalgamation level, corresponding to four replicate runs. Panel (c) provides a histogram summarizing the distribution of the observed deviance, indicated by the vertical line, from 500 Monte Carlo simulations.

(year); see for example, Catchpole (1995). This is in agreement with the Freeman–Tukey results in Fig. 2b, and we therefore select the Freeman–Tukey measure for use in the work of this paper. This perspective applies also to mark–recapture data in general.

CENSUS DATA

For any time series $\{y_t\}$, there are many alternative discrepancy measures that can be used, based on the prediction errors, $\{y_t - \hat{y}_t\}$, where $\hat{y}_t$ are fitted values. For illustration, we use

two simple measures in the paper; these are the mean absolute percentage error (MAPE),

$$D_{\mathrm{MAPE}}(\mathbf{y};\theta) = \frac{100}{n}\sum_{t=1}^{n}|(y_t - \hat{y}_t)/y_t|, \qquad \text{eqn 6}$$

where n is the number of (non-missing) prediction errors, and the maximum percentage error (MPE),

$$D_{\mathrm{MPE}}(\mathbf{y};\theta) = 100\max\{(y_t - \hat{y}_t)/y_t\}. \qquad \text{eqn 7}$$

In both cases, the observations where $y_t = 0$ are ignored. As we shall see, in practice careful thought needs to be given to the selection of an appropriate measure(s) for particular ecological time series, and there is a wide range of alternatives that may be appropriate in different applications.

Results

We illustrate the performance of calibrated simulation by application to the real data sets on birds analysed by BFMC, and by simulation. There are two of these real data sets, on northern lapwings, described earlier, and on grey herons, *Ardea cinerea*. In each case, there are both national ring–recovery data on birds ringed as young, and national count data. For the herons the count data can be taken as a census, while for lapwings, the count data provide an index of abundance. The two examples contrast in interesting ways: lapwings are in decline, whereas grey herons, following a large fall in numbers as a result of the severe winter of 1962 in Britain, have been increasing since then. The count data for both species are plotted in BFMC. In each case, the data have been obtained from the British Trust for Ornithology.

LAPWINGS

We present in Fig. 3 results for the lapwing data. Based on 500 simulations, the resulting values for p_c (n_c) are 0·39 (195) for the ring–recovery data, 0·72 (362) for the MAPE and 0·52 (259) for MPE. We judge therefore that the selected model $\phi_1(\omega_t),\phi_a(\omega_t)/\lambda(\mathrm{year})/p(\mathrm{year})$ fits the data well, in agreement with the conclusions of BFMC.

SIMULATION RESULTS: LAPWINGS

In order to provide a calibration of these values, we now simulate 100 data sets matched to the real data, using the maximum-likelihood parameter estimates from fitting the selected model to the real data. Parameter values used in the simulation are given in Table 1(a). Note the qualitative difference between this simulation and those needed to produce the original p_c values. The simulation produces replica data sets, and we then obtain p_c values for each of these replicated data sets. Then for each of these replications, we fit the selected model to the simulated data and form p_c values, in each case based upon 500 simulations from the appropriate assumed multivariate normal distribution. The resulting samples of 100 values for p_c are summarized in Fig. 4d. We can see that the values of p_c obtained for the real data are in agreement with the box plots, in accordance with the findings of BFMC. The distribution of p_c values obtained for the recovery data indicates that when the model is correct a wide range of p_c values may be obtained. A uniform distribution for p_c values is desirable, as it would remove the need for calibration. The location of the box plot for the MAPE measure of discrepancy in Fig. 4d is interesting, as one might expect it to be centred at 0·5, as is true of the other two measures considered in that figure panel. This is partly a consequence of the behaviour of the MAPE discrepancy for the relatively short length of ecological time series that we consider, and we return to this point later in the paper. Sensitivity studies not reported here also show that the MAPE discrepancy is affected by the size of σ used. Also in Fig. 4, for comparison, we provide box plots of p_c values from the three following wrong models for the data: (a) a model where all parameters are constant, $\phi_1,\phi_a/\lambda/p$; (b) a model in which ϕ_a

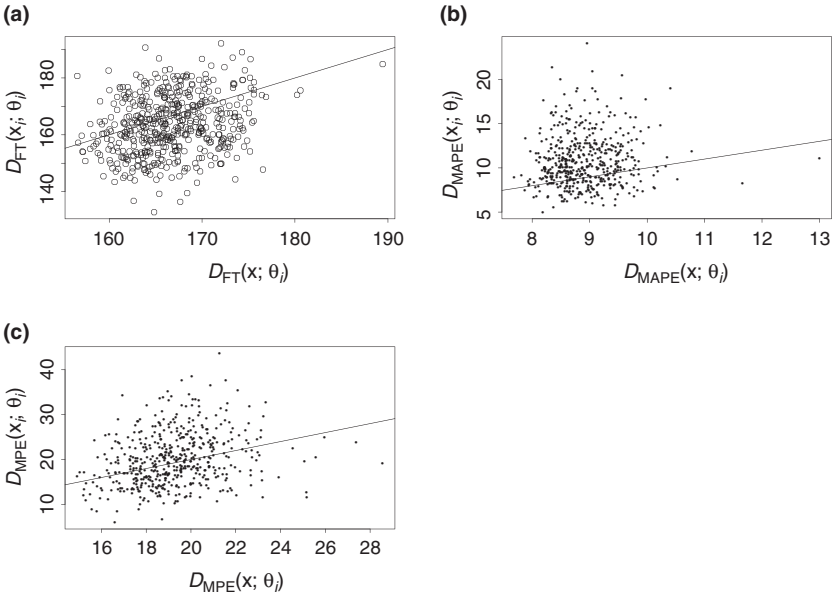

Fig. 3. Simulation results to construct p_c values for model $\phi_1(\omega_t),\phi_a(\omega_t)/\lambda(\mathrm{year})/p(\mathrm{year})$ of demographic (a), and count data on lapwings, (b) using mean absolute percentage error and (c) using maximum percentage error. 500 simulations are used. Here and later, circles are used for the Freeman–Tukey discrepancy and dots for the discrepancy measures for the time-series data.

Table 1. Parameter values used in simulations to calibrate p_c estimates: (a) for lapwings, (b) for herons with one productivity threshold and (c) for herons with two productivity thresholds. Here the parameters v_0,v_1, v_2 specify the thresholds measured on a logarithmic scale; see Besbeas & Morgan (2012). The values are maximum-likelihood estimates of models fitted to the real data. Data-based initial populations (not shown) were used to generate the abundance data for both species

Parameters	(a)	(b)	(c)
ϕ_1 intercept (β_0)	0·523	−0·187	−0·185
ϕ_1 slope (β_1)	−0·023	−0·023	−0·022
ϕ_2 intercept	–	0·388	0·391
ϕ_2 slope	–	−0·019	−0·017
ϕ_3 intercept	–	0·906	0·925
ϕ_3 slope	–	−0·020	−0·019
ϕ_a intercept (ξ_0)	1·519	1·355	1·357
ϕ_a slope (ξ_1)	−0·028	−0·018	−0·020
λ intercept	−4·564	−2·027	−2·027
λ slope	−0·584	−0·831	−0·831
p intercept	−1·175	–	–
p slope	−0·425	–	–
v_0	–	−0·112	−0·041
v_1	–	0·280	0·528
v_2	–	–	−0·142
σ	160·01	381·74	316·64

and p are both constant, $\phi_1(\omega_t),\phi_a/\lambda(\text{year})/p$ and (c) a model with constant productivity, $\phi_1(\omega_t),\phi_a(\omega_t)/\lambda(\text{year})/p$. Thus, in order, the three wrong models (a–c) are of decreasing misspecification. In Fig. 4, as we move from (d) to (c), we see that there is no major change to the box plot for the recovery data, this being because the only change to the model is to productivity, which does not feature in the likelihood for the recovery data. However, as we then move in order to the plots of (b) and (a), we progressively make the modelling of the recovery data worse, and as a result the recovery box plots shift in the direction of smaller values. In contrast, as we move to (c) and then to (b) and (a), the time-series box plots move in the direction of larger values. This is because the models for the time-series data are becoming steadily less realistic, which results in larger estimates of observation error variance in the state-space model for the census data, accommodating the lack of fit. Thus, in this case $D(\mathbf{x}_i; \boldsymbol{\theta}_i)$, values will tend to be greater than $D(\mathbf{x}; \boldsymbol{\theta}_i)$ values.

The comparison between the performance of the MAPE and MPE measures is interesting, demonstrating the need for ecological time series to use more than one discrepancy measure. In separate work, we have investigated the behaviour of a range of alternatives, and a preliminary conclusion is that there may be advantages to using symmetric MAPE (SMAPE) measures; see the Appendix S3.

HERONS

We can examine the fit to the heron data for the selected model of BFMC in the same way as we have for the lapwings. This model is more complex, involving four age classes for survival, corresponding to birds aged 0, 1, 2 and more than 2 years, and has a more sophisticated model for productivity, incorporating density dependence. As with the lapwing application, each survival probability is regressed on ω_t and the reporting probability is regressed on time, but the productivity parameter, instead of being regressed on time, is now related to population size through a threshold dependence; see Besbeas & Morgan (2012). This approach is motivated by the rapid population growth of the population following the population crash of 1962, so that the model has productivity that is driven by population size, relative to fixed threshold sizes. The final model selected by Besbeas & Morgan (2012) was one with three thresholds, with productivity determined by the position of the population size relative to those thresholds. Likelihoods do not change when threshold parameters vary over intervals that do not contain any time-series data, and as a consequence thresholds are estimated from profile likelihoods.

We can see this from the histograms of Fig. 5, which are the result of multiple simulations of matched data from a fitted

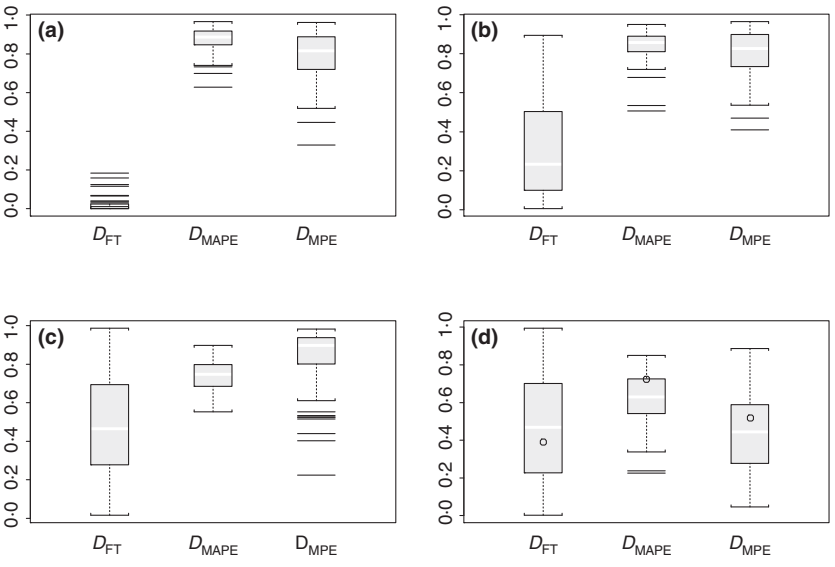

Fig. 4. Box plots of p_c values for four models for simulated lapwing data under model $\phi_1(\omega_t),\phi_a(\omega_t)/\lambda(\text{year})/p(\text{year})$: (a) model with all parameters constant; (b) model with constant adult survival and constant productivity; (c) model with constant productivity; (d) correct model. In (d), the position of the p_c values for the real data is indicated by circles.

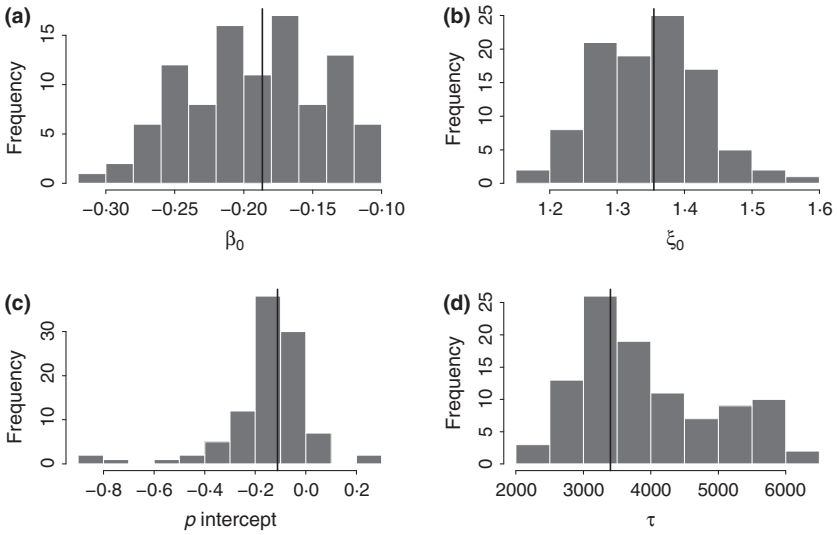

Fig. 5. Histograms of maximum-likelihood estimates for four of the parameters in the one threshold model for heron data, resulting from fitting 100 simulated data sets. The parameters are (a) ϕ_1 intercept, β_0, (b) ϕ_a intercept, ξ_0, (c) intercept p, (d) the single threshold value τ. The values used in the simulation are indicated by vertical lines, and we note the non-normal appearance for parameters β_0 and τ. See Table 1(b) for the parameter values used in the simulation.

heron model with one threshold parameter τ, each simulation fitted by maximum likelihood. Consequently, for this application, we conduct calibrated simulation conditional upon the maximum-likelihood estimates of threshold parameters obtained from the real data.

Resulting scatter plots of $D(\mathbf{x}_i; \boldsymbol{\theta}_i)$ vs. $D(\mathbf{x}; \boldsymbol{\theta}_i)$ for the heron census data are given in Fig. 6. We can see that increasing the number of thresholds improves the fit to the census data. The plots of Fig. 7 indicate good description of the census data when there are three thresholds for productivity. The p_c values

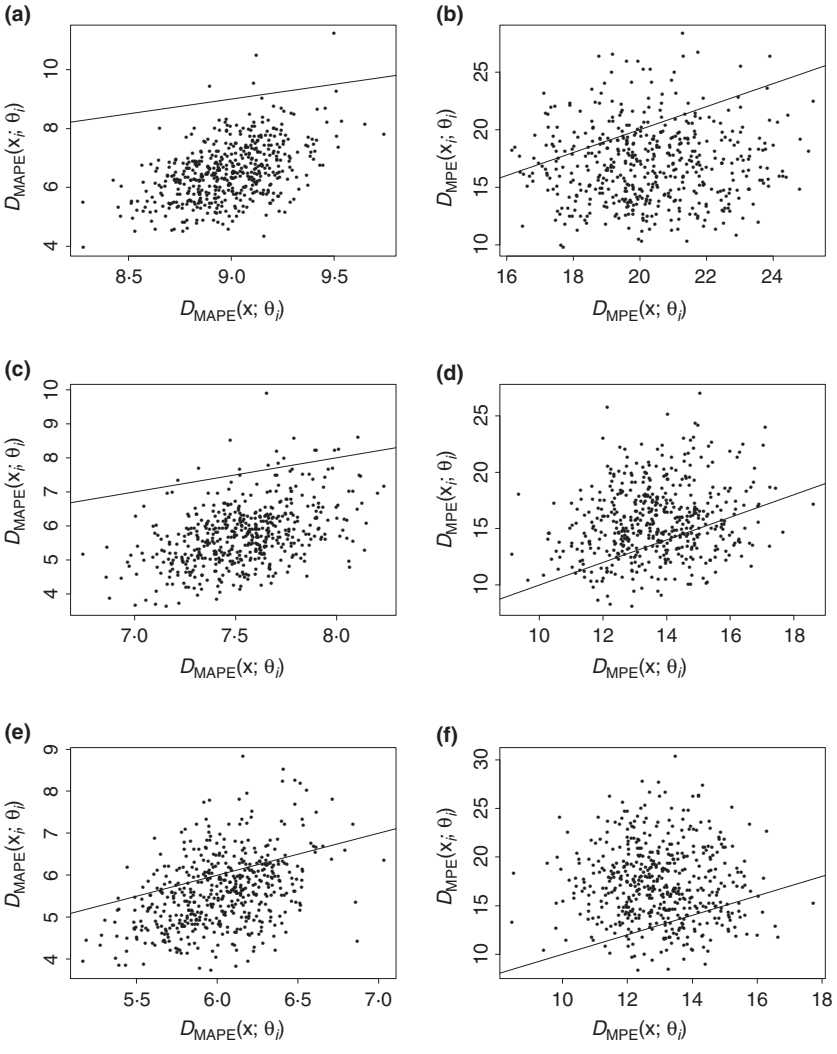

Fig. 6. Scatter plots to obtain p_c values for determining the fit of integrated population models of demographic and count data on grey herons. Plots (a) and (b) correspond to no threshold for productivity, taken as constant, plots (c) and (d) correspond to a single productivity threshold, and plots (e) and (f) correspond to two thresholds for productivity. Mean absolute percentage error (MAPE) discrepancy measures appear on the left and those for maximum percentage error (MPE) are on the right.

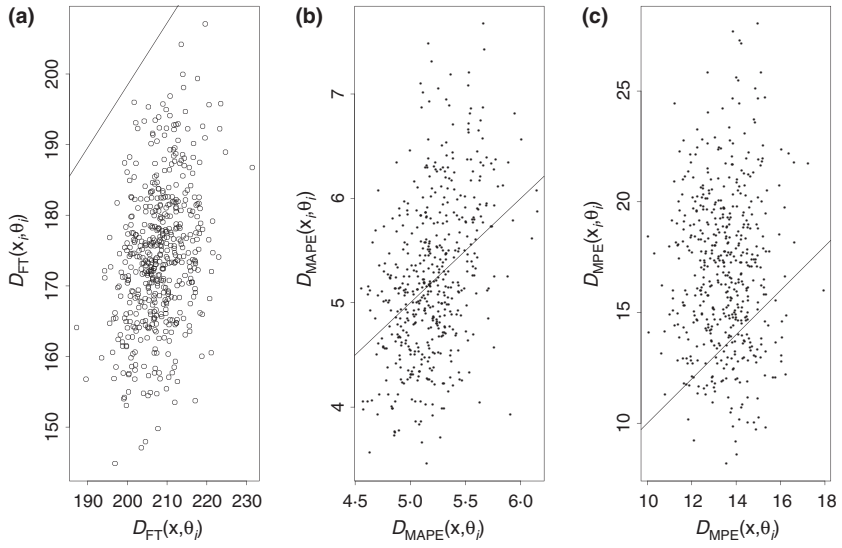

Fig. 7. Scatter plots to obtain p_c values for determining the fit of integrated population models of demographic and count data on grey herons when three thresholds for productivity are used. The three panels correspond to the following three discrepancy measures described in the text: (a) Freeman–Tukey; (b) mean absolute percentage error; (c) maximum percentage error.

for MAPE and MPE are 0·49 and 0·81, respectively. It appears that the recovery data are not well described, and this may well be due to overdispersion. We can therefore see the advantage of checking goodness-of-fit separately for the different components of integrated population modelling. We do not discuss here the incorporation of overdispersion for the recovery data, but this can be done in different ways; see for example Barry *et al*. (2003), Besbeas, Borysiewicz & Morgan (2009), Burnham & Rexstad (1993) and Pledger & Schwarz (2002). However, overdispersion is not revealed as such a problem for the census data. BFMC do not comment on overdispersion at all, and for goodness-of-fit to the census data rely solely on superposition of observed and fitted time series.

SIMULATION RESULTS: HERONS

We illustrate the method with the use of single and two threshold models, where productivity takes different values, depending on a threshold population size(s). Thus, this model for simulation is structurally quite different from that for lapwings. The parameter values that were used in the simulations were the maximum-likelihood estimates from fitting the corresponding model to the heron data, and are given in Table 1(b) and (c). In each case, recovery data were simulated by using the observed ringing totals and winter severity measures ω_t. These were combined for joint analysis with independent sets of simulated abundance data, the values p_c for a range of models were calculated based upon 500 simulations, and the process was repeated 100 times. The sample sizes were selected to match the heron data sets. The census data were generated by simulating a population from a state-space model, with the threshold determined by the corresponding true values of the population size. For the case of a single threshold, and each pair of simulated ring–recovery and abundance data, the parameters were estimated two ways, once employing constant productivity, and once with the threshold determined by the corresponding observed values y_t; see Besbeas & Morgan (2012). For two thresholds, the same approach is adopted. The threshold model is fitted in the manner of Besbeas & Morgan (2012), by searching over a grid of population values for the threshold, as direct optimization with respect to parameters and τ is not feasible.

We can see from Fig. 8 that the box plots are better centred on 0·5 when the model is correct, compared to when there is no assumed threshold. However, the differences are small and will inevitably depend upon the value of the threshold adopted. The results of Fig. 9 are more interesting, and we can see that the incorrect model with no thresholds results in box plots for the time-series discrepancy measures that are clearly incorrectly centred, signalling an incorrect model. The structural difference between the single threshold and two threshold models is not as great as between threshold models and a model with no threshold, and we can see this from panels (b) and (c) of Fig. 9. As with the lapwing simulations, there are differences in behaviour of the two time-series discrepancy measures, and this is the subject of further research.

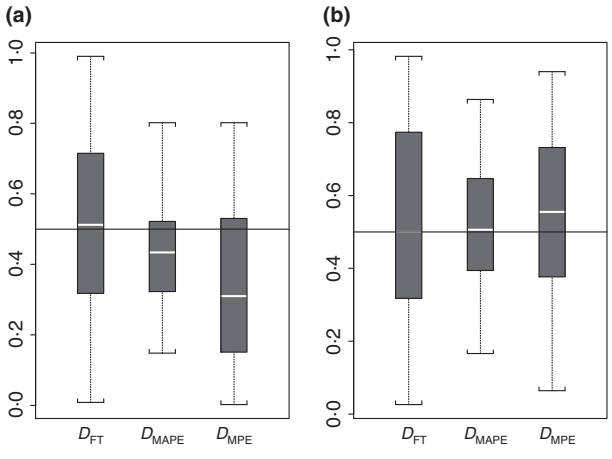

Fig. 8. Box plots of p_c values for two models for simulated heron data with a single productivity threshold: (a) the wrong model is fitted, with constant productivity; (b) the correct model is fitted, with one productivity threshold.

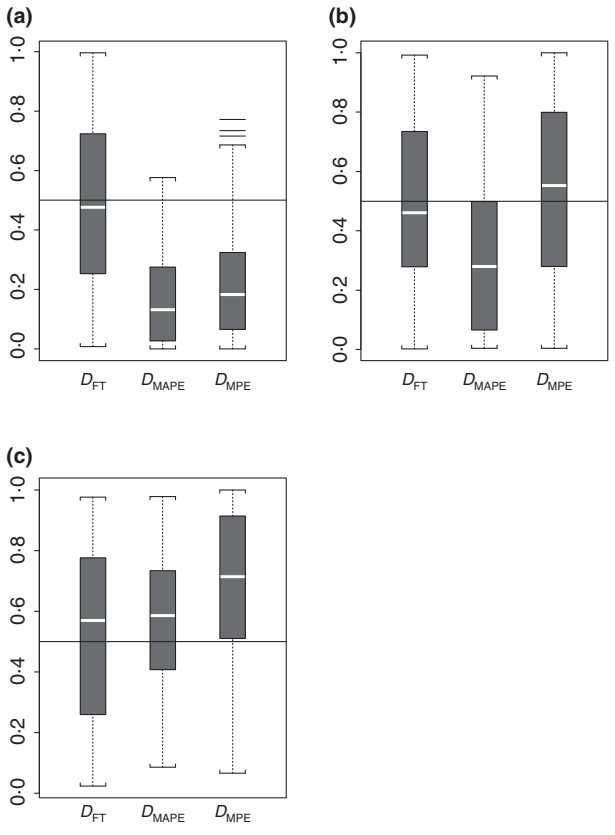

Fig. 9. Box plots of p_c values for three models for simulated heron data with two productivity thresholds: (a) the wrong model is fitted, with constant productivity; (b) the wrong model is fitted with one productivity threshold; (c) the correct model is fitted with two productivity thresholds.

Discussion

By means of two examples, we have demonstrated the potential of calibrated simulation for gauging the fit of integrated population models. In most cases, applications will be straightforward, corresponding to the lapwing example. If p_c values are obtained which are judged to be acceptable, the calibration would not be necessary, and it only needs to be used to check values considered to be too high or too low. We believe that this approach will prove to be useful in future applications of integrated population modelling. The value of the heron example is in demonstrating that in certain cases modifications to the standard procedure are needed. Ongoing research will evaluate the use of calibrated simulation for assessing the fit of mark–recapture data in general. For the ring–recovery data analysed in this paper, the calibration of the Freeman–Tukey discrepancy measures shows that the distribution of p_c values is approximately uniform when the model fits the data well. However, this is a consequence of the data and measure employed, and less uniform results are obtained for the time-series data and measures. Further research on Bayesian *P*-values is described by Zhang (2014).

Constructing discrepancy measures for different types of data and models produces interesting new challenges. We have seen that for time-series and mark–recovery data different measures need to be used in tandem and that sensitivity analyses also need to be carried out. This is the topic of current research.

Acknowledgement

We thank David Fletcher for his helpful discussions, the British Trust for Ornithology for the use of the two real data sets, and an Associate Editor and two reviewers for their comments. This work was part funded by EPSRC grant EP/100917/1.

Data accessibility

The lapwing and heron mark–recovery data that were used in this study are available online as a supplementary material of BFMC. The computer programs used in the work of the article are available on request from the authors.

References

Abadi, F., Gimenez, O., Arlettaz, R. & Schaub, M. (2010) An assessment of integrated population models: bias, accuracy, and the violation of the assumption of independence. *Ecology*, **91**, 7–14.

Barry, S.C., Brooks, S.P., Catchpole, E.A., Morgan, B.J.T. (2003) The analysis of ring-recovery data using random effects. *Biometrics*, **59**, 54–65.

Besbeas, P. & Morgan, B.J.T. (2012) A threshold model for heron productivity. *JABES*, **17**, 128–141.

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

Besbeas, P., Lebreton, J.-D. & Morgan, B.J.T. (2003) The efficient integration of abundance and demographic data. *Applied Statistics*, **52**, 95–102.

Besbeas, P., Borysiewicz, R.S. & Morgan, B.J.T. (2009) Completing the ecological jigsaw. *Modelling Demographic Processes in Marked Populations* (eds D.L. Thomson, E.G. Cooch, & M.J. Conroy) Springer series: Environmental and Ecological Statistics, **3**, 513–540.

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975) *Discrete Multivariate Analysis*. MIT Press, Cambridge, Massachusetts, USA.

Brooks, S.P., Catchpole, E.A. & Morgan, B.J.T. (2000) Bayesian animal survival estimation. *Statistical Science*, **15**, 357–376.

Brooks, S.P., King, R. & Morgan, B.J.T. (2004) A Bayesian approach to combining animal abundance and demographic data. *Animal Biodiversity and Conservation*, **27**, 515–529.

Burnham, K.P. & Rexstad, E. (1993) Modeling heterogeneity in survival rates of banded waterfowl. *Biometrics*, **49**, 1194–1208.

Catchpole, E.A. (1995) Matlab: an environment for analysing ring-recovery and recapture data. *Journal of Applied Statistics*, **22**, 801–816.

Chandler, R.B. & Clark, J.D. (2014) Spatially explicit integrated population models. To appear, *Methods in Ecology and Evolution*. doi: 10.1111/2041-210X.12153.

Durbin, J. & Koopman, S.J. (2001) *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford.

Freeman, S.F. & Morgan, B.J.T. (1992) A modelling strategy for recovery data from birds ringed as nestlings. *Biometrics*, **48**, 217–235.

Freeman, M.F. & Tukey, J.W. (1950) Transformations related to the angular and square root. *Annals of Mathematical Statistics*, **221**, 607–611.

Gelman, A., Meng, X. & Stern, H. (1996) Posterior predictive assessment of model fitness via realised discrepancies (with discussion). *Statistica Sinica*, **6**, 733–807.

Johnson, V.E. (2004) A Bayesian chi-square test for goodness-of-fit. *Annals of Statistics*, **32**, 2361–2384.

King, R. (2011) Statistical Ecology. *Handbook of Markov chain Monte Carlo* (eds S. Brooks, A. Gelman, G. Jones, & X.-L. Meng), pp. 419–447. CRC Press, Boca Raton, Florida, USA.

King, R., Morgan, B.J.T., Gimenez, O., & Brooks, S.P. (2009) *Bayesian Analysis for Population Ecology*. Chapman and Hall, Boca Raton, Florida, USA.

Mazzetta, C., Morgan, B.J.T. & Coulson, T. (2010) A state-space modelling approach to population size estimation. University of Kent Technical Report, UKC/SMSAS/10/025, 1–30.

McCrea, R.S., Morgan, B.J.T., Gimenez, O., Besbeas, P., Bregnballe, T. & Lebreton, J.-D. (2010) Multisite integrated population modelling. *Journal of Biological, Agricultural and Environmental Statistics*, **15**, 539–561.

Millar, R.B. & Meyer, R. (2000) Non-linear state space modelling of fisheries biomass dynamics by using Metropolis-Hastings with-Gibbs sampling. *Applied Statistics*, **49**, 327–342.

Newman, K.B., Buckland, S.T., Morgan, B.J.T., King, R., Borchers, D.L., Cole, D.J., Besbeas, P., Gimenez, O. & Thomas, L. (2014) *Modelling Population Dynamics: Model Formulation, Fitting and Assessment Using State-Space Methods*. Springer, Boca Raton, Florida, USA.

Pledger, S. & Schwarz, C.J. (2002) Modelling heterogeneity of survival in band-recovery data using mixtures. *Journal of Applied Statistics*, **29**, 315–327.

Schaub, M. & Abadi, F. (2011) Integrated population models: a novel analysis framework for deeper insights into population dynamics. *Journal of Ornithology*, **152**, 227–237.

Stoffer, D.S. & Wall, K.D. (1991) Bootstrapping state space models: Gaussian maximum likelihood estimation and the Kalman filter. *Journal of the American Statistical Association*, **86**, 1024–1033.

Tavecchia, G., Besbeas, P., Coulson, T., Morgan, B.J.T. & Clutton-Brock, T.H. (2009) Estimating population size and hidden demographic parameters with state-space modelling. *The American Naturalist*, **173**, 722–733.

Zhang, J.L. (2014) Comparative investigation of three Bayesian *p* values. *Computational Statistics and Data Analysis*, **79**, 277–291.

Received 19 November 2013; accepted 1 September 2014
Handling Editor: Richard Barker

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Simulations to check on marginal univariate normality.

Appendix S2. Additional comparisons of chi-square and Freeman-Tukey discrepancy measures.

Appendix S3. A preliminary study of alternative discrepancy measures for univariate time series, $SMAPE_1$ and $SMAPE_2$.