Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017, **8,** 408–419 doi: 10.1111/2041-210X.12761

TECHNOLOGICAL ADVANCES AT THE INTERFACE BETWEEN ECOLOGY AND STATISTICS

Graphical diagnostics for occupancy models with imperfect detection

David I. Warton*[,1], Jakub Stoklosa[1], Gurutzeta Guillera-Arroita[2], Darryl I. MacKenzie[3] and Alan H. Welsh[4]

[1]*School of Mathematics and Statistics and Evolution & Ecology Research Centre, The University of New South Wales, NSW 2052, Australia;* [2]*School of Biosciences, University of Melbourne, Parkville, Vic. 3010, Australia;* [3]*Proteus Wildlife Research Consultants, PO Box 5193, Dunedin, New Zealand; and* [4]*Mathematical Sciences Institute, The Australian National University, Canberra, ACT 2601, Australia*

Summary

1. Occupancy-detection models that account for imperfect detection have become widely used in many areas of ecology. As with any modelling exercise, it is important to assess whether the fitted model encapsulates the main sources of variation in the data, yet there have been few methods developed for occupancy-detection models that would allow practitioners to do so.

2. In this paper, a new type of residual for occupancy-detection models is developed according to the method of Dunn & Smyth (*Journal of Computational and Graphical Statistics*, **5**, 1996, 236–244). Residuals are separately constructed to diagnose the occupancy and detection components of the model.

3. Because the residuals are quite noisy, we suggest fitting a smoother through plots of residuals against predictors of fitted values, with 95% confidence bands, to diagnose lack-of-fit.

4. The method is illustrated using Swiss squirrel data, and evaluated using simulations based on that dataset.

5. Plotting residuals against predictors or against fitted values performed reasonably well as methods for diagnosing violations of occupancy-detection model assumptions, particularly plots of residuals against a missing predictor. Relatively high false positive rates were sometimes observed, but this seems to be controlled reasonably well by fitting smoothers to these plots and being guided in interpretation by 95% confidence bands around the smoothers.

Key-words: Dunn–Smyth residuals, goodness-of-fit, imperfect detection, probability integral transforms

Introduction

Occupancy-detection models are widely used in ecology to estimate the probability a site is occupied by a species while accounting for imperfect detection. These are a type of hierarchical model usually fitted using dedicated software (White & Burnham 1999; Fiske *et al.* 2011; MacKenzie & Hines 2015), to detection/nondetection data from repeated surveys (MacKenzie *et al.* 2002; Tyre *et al.* 2003) or other types of survey data that inform about the detection process (Guillera-Arroita in press).

In occupancy-detection models, as indeed in any model-fitting exercise, a critical issue is diagnosing whether or not the chosen model structure is appropriate for the data at hand. Tools designed to interrogate model assumptions are needed, yet there is relatively little literature on methods of diagnosing goodness-of-fit of occupancy-detection models, and their effectiveness, the topic of this paper.

The problem of assessing goodness-of-fit is distinct but related to the problem of choosing the best model from a set of candidates ('model selection'). In model selection, we are interested in a relative comparison of how well models fit the data, but in model checking we are interested in absolute measures of goodness-of-fit, i.e. is the model a plausible explanation for the data at hand. Thus if using AIC (Akaike 1974) as a metric, in model selection we could choose the model with smallest AIC (Burnham & Anderson 1998). But when diagnosing goodness-of-fit we are after an *absolute* measure of fit, i.e. we would want to know if the given value of AIC is sufficiently small for the current model to be plausible (which might for example require its comparison to some null distribution).

A commonly used overall measure of goodness-of-fit, rather than AIC or a related measure, is the Pearson chi-square goodness-of-fit statistic. MacKenzie & Bailey (2004) proposed a simulation-based method of testing for lack of fit of occupancy-detection models, constructing *P*-values based on the Pearson chi-square statistic. Royle *et al.* (2007) proposed a closely related Bayesian method, using the Bayesian *P*-value (following Gelman, Meng & Stern 1996). Both methods can be understood as computer-intensive extensions of methods transferred to occupancy-detection modelling following their uptake in the capture-recapture literature (Pollock, Hines & Nichols 1985; Brooks, Catchpole & Morgan 2000; Choquet

*Correspondence author. E-mail: david.warton@unsw.edu.au

et al. 2009). While widely used, the ability of these methods to detect violations of assumptions in occupancy-detection models has not been tested extensively (although see MacKenzie & Bailey 2004; Kéry & Royle 2015, section 10·8). Further, given that a violation has been detected, methods are needed that can 'drill down' into the model to understand where the problems arise. Residual plots are a potentially useful tool here (as mentioned in Kéry & Royle 2015), but a key issue is how to define residuals for this type of model, given the discreteness of the responses, and given that the quantity of primary interest (site occupied or not) is not observed directly.

In this paper, residuals are derived for occupancy-detection models according to the ideas of Dunn & Smyth (1996). These Dunn–Smyth residuals have the special property that under repeated sampling, each residual is exactly standard normal in distribution if the model is exactly correct. These residuals are then evaluated in terms of their ability to diagnose different components of an occupancy-detection model.

The paper is organised as follows: first, existing diagnostic tools for occupancy-detection models are reviewed; then, application of Dunn–Smyth residuals to this problem is described and illustrated on an example dataset; finally, the performance of the proposed tool is evaluated, along with existing goodness-of-fit approaches by simulation.

Review of diagnostic tools for occupancy-detection modelling

Assume there are S sites, each visited T times. At site i the variable of interest is a vector $\boldsymbol{Y}_i$ of detection and non-detection events. A particular realisation of it is written as $\boldsymbol{y}_i$, consisting of binary values where at time t, y_{it} is one if the species was detected and zero otherwise. Assuming exactly T surveys were conducted everywhere, there are 2^T possible sequences of ones and zeros in $\boldsymbol{y}_i$, which will be referred to as a detection history. A statistic used frequently in this paper is the total number of detections at a site, $n_i = \sum_{t=1}^{T} y_{it}$.

There are many variations and extensions of occupancy-detection models (MacKenzie *et al*. 2006; Guillera-Arroita in press). Here we work with the original model structure (as in MacKenzie *et al*. 2002; Tyre *et al*. 2003). This model assumes that the occupancy status of the sites does not change during the survey period (i.e. the species either occupies the site at all or none of the sampling occasions), and that there is independence among sites. Sites are occupied with a probability ψ_i at the ith site, where ψ_i can be modelled as a function of environmental variables ($\boldsymbol{x}_i$ at site i). We will use a logit link function (although there are other possibilities):

$$\text{logit}(\psi_i) = \mathbf{x}_i'\beta$$

In this paper, this will be referred to as the *occupancy component* of the model.

Detection events are assumed independent and with no possibility of false positives (Chambert, Miller & Nichols 2015). The probability of detection at the ith site at time t (p_{it}), given that the site is occupied, can be modelled using logistic regression (for example) against explanatory variables ($\boldsymbol{z}_{it}$ for site i and time t):

$$\text{logit}(p_{it}) = \mathbf{z}_{it}'\gamma$$

In this paper, this will be referred to as the *detection component* of the model. Under the above assumptions, the probability of observing a particular detection history $\boldsymbol{y}_i$ at site i can be written as:

$$P(\boldsymbol{Y}_i = \boldsymbol{y}_i) = \begin{cases} \psi_i \prod_{t=1}^{T} p_{it}^{y_{it}} (1 - p_{it})^{1-y_{it}} & \text{if } n_i > 0 \\ (1 - \psi_i) + \psi_i \prod_{t=1}^{T} (1 - p_{it}) & \text{if } n_i = 0 \end{cases}$$

and the likelihood is the product of $P(\boldsymbol{Y}_i = \boldsymbol{y}_i)$ across the S sites. Parameters can be estimated by maximum likelihood or using a Bayesian approach.

There are several key aspects of this model that should be considered when designing a study and fitting an occupancy-detection model:

1. Whether the logistic regression models for probability of occupancy and of detection are appropriate. This could include for example consideration of missing predictors, linearity assumptions, and the choice of link function.

2. Whether it is reasonable to assume independence of detection events across surveys and of occupancy across sites.

3. Does occupancy status remain unchanged across the T sampling times? (i.e. Do the species either occupy the site at all or none of the sampling occasions?)

4. Other sources of heterogeneity in measurement, such as whether abundance is changing across sites and whether this affects detection or not. This issue will be returned to in the Discussion section.

The key methods that have been proposed to collectively check assumptions will be reviewed below.

ESTABLISHED DIAGNOSTIC TOOLS

To get a sense for which diagnostic tools are commonly used in practice, Google Scholar was searched in April 2016 for articles containing 'occupancy', 'imperfect detection' and 'goodness-of-fit'. The first 60 papers, as automatically sorted by relevance, were screened to look specifically for journal articles applying an occupancy-detection model to binary (detected/not detected) data. A total of 28 papers remained, and of these more than half (16 of the 28) used the simulation-based test of MacKenzie & Bailey (2004), and two used a Bayesian P-value analogue of this test. Several more used variations on these techniques (e.g. using a deviance statistic rather than a Pearson chi-square statistic). Many did not assess absolute goodness-of-fit at all, rather they used model selection on a set of candidate models (*relative* goodness-of-fit), primarily using AIC or AICc (18 of the 28), or occasionally a related information metric (4 papers) or predictive performance on hold out samples (3 papers). Further details are available in Data S1, Supporting Information. Below we describe in more detail the most commonly used methods of goodness-of-fit for occupancy-detection models, the test due to MacKenzie & Bailey (2004) and its Bayesian P-value analogue (a modification of Royle *et al*. 2007).

As explained in Kéry & Royle (2015, section 10·8), it is possible to construct a Pearson chi-square statistic by aggregating

detection/non-detection data to detection counts over sites, surveys or detection histories. They undertook a small simulation which suggested that aggregating over detection history tended to be the most powerful of these options. From T surveys there are 2^T possible detection histories, and MacKenzie & Bailey (2004) originally proposed comparing observed and expected counts for each possible detection history. A count of the number of times a given detection history $\boldsymbol{y}$ is observed ($O_{\boldsymbol{y}}$) is compared to its expected number ($E_{\boldsymbol{y}}$), computed as a sum across sites of the estimated probability of observing detection history $\boldsymbol{y}$:

$$E_{\boldsymbol{y}} = \sum_{i=1}^{S} P(\boldsymbol{Y}_i = \boldsymbol{y})$$

where $P(\boldsymbol{Y}_i = \boldsymbol{y})$ is estimated from the fitted occupancy-detection model. Then the usual Pearson chi-square statistic is computed by summing across the set $\mathcal{D}$ of all 2^T possible detection histories:

$$\chi^2 = \sum_{\boldsymbol{y} \in \mathcal{D}} \frac{(O_{\boldsymbol{y}} - E_{\boldsymbol{y}})^2}{E_{\boldsymbol{y}}} \qquad \text{eqn 1}$$

This statistic is used to check that the observed counts are not too far from expected counts, i.e. that the chi-square statistic is not unusually large. But doing this requires knowledge of the range of values of χ^2 to expect when the model is correct. Because not all detection histories will be observed a large number of times, χ^2 can be poorly approximated by a chi-square distribution, and MacKenzie & Bailey (2004) suggested estimating its null distribution by simulation. Repeatedly, new sets of detection histories are generated from the fitted model (a 'parametric bootstrap', Davison & Hinkley 1997), the model is refitted to each, and the chi-square statistic recomputed, in order to get a sense for what range of values for χ^2 to expect if the model were true. This process tends to be computationally intensive. The test has been found to be capable of finding problems with the detection component of the model, but it has little success finding violations of the occupancy component (MacKenzie & Bailey 2004) – since aggregating over detection histories puts the focus on diagnosing patterns in detection. The approach can be implemented in the `R` package `AICcmodavg` in the `mb.gof.test` function which allows for fitted `unmarked` occupancy-detection model objects, or a faster version of the code can be found in the PRESENCE software which can be accessed from the `RPresence` package (MacKenzie & Hines 2015, version 1.1.2 or later), downloadable from http://www.mbr-pwrc.usgs.gov/software/presence.html (under Downloads, install both PRESENCE and RPresence).

Royle *et al.* (2007) also used a Pearson chi-square statistic, but in a posterior predictive check (known as the Bayesian *P*-value) of a model for abundance accounting for imperfect detection. The approach has since become relatively common in models for abundance. When using a MCMC approach to estimation, at each iteration model parameter estimates are simulated, and new response values for $\boldsymbol{Y}_i$ can then be simulated in order to compute a simulated value for the test statistic under the assumption that the model is correct. Doing this many times generates a null distribution, to compare to the observed chi-square statistic to assess goodness-of-fit. This is something of a Bayesian analogue of MacKenzie & Bailey (2004), but additionally accounting for sampling uncertainty estimating model parameters, instead of conditioning on the observed values. This idea has been adapted to occupancy-detection models (Kery & Schaub 2011; Tobler *et al.* 2015), using the Pearson statistic on the binary detection/non-detection response directly without aggregation, and to the multi-species context (Broms, Hooten & Fitzpatrick 2016). Using binary response without aggregation has since been found to be problematic (Kéry & Royle 2015, section 10·8) because in binary responses it is not possible to estimate overdispersion, hence it is not possible to diagnose lack-of-fit in a Pearson statistic of binary data. Thus, aggregating to counts in some way seems to be necessary prior to constructing a Pearson statistic. While MacKenzie & Bailey (2004) aggregated over detection history, an alternative (Wright, Irvine & Rodhouse 2016) is to aggregate across neighbours (in space or time) and count pairs of detection events amongst neighbours ('joins'). Such a test would be especially suited to testing for temporal dependence, counting consecutive detection events.

Dunn–Smyth residuals for occupancy-detection models

Residual plots have potential as a diagnostic tool for occupancy-detection models. A difficulty however is that it is not obvious in this context what to construct residuals of, and how. In this paper, two different types of residual are proposed, to separately diagnose the detection and occupancy components of the model. Both make use of residuals as introduced in Dunn & Smyth (1996), who resolved the problem of constructing residuals from a discrete random variable, such that residuals across observations can be expected to have (approximately) the same distribution, a property that can be exploited to check different aspects of a fitted model.

Dunn & Smyth (1996) proposed residuals by making use of the probability integral transform, i.e. the result that if a continuous random variable X has cumulative distribution function (or 'quantile function') $F(x)$, then transforming the random variable using this function will lead to a standard uniform variable (equally likely to take any value between zero and one). For discrete random variables, some random noise is introduced to 'jitter' residuals and remove the discreteness. A similar idea can be found elsewhere (Smith 1985; Brockwell 2007), but Dunn & Smyth (1996) additionally proposed mapping these onto the standard normal distribution, such that they could be plotted and interpreted like residuals from ordinary linear regression. For a discrete random variable X that takes integer values, residuals z are constructed to satisfy:

$$\Phi(z) = (1-u)F(x) + uF_-(x) \qquad \text{eqn 2}$$

where $\Phi(\cdot)$ and $F(\cdot)$ are the cumulative distribution functions of a standard normal variable and of X (respectively), $F_-(\cdot)$ is the previous value of $F(\cdot)$, and u is a value randomly generated from the standard uniform distribution. If the model for

X is correct, these residuals will have a standard normal distribution. Thus they can be used just like residuals from linear regression – they can be plotted against predictors or fitted values to check for no trend (Fig. 1a), and they can be compared to the standard normal using a (standard) normal quantile plot (Fig. 1b). If using a categorical predictor, residuals can be plotted against different levels of it, or a boxplot of residuals could be constructed for each factor level.

Dunn & Smyth (1996) illustrated the use of these residuals in the context of generalised linear models, and in an ecological context they have been extended to multivariate problems in the `mvabund` (Wang *et al.* 2012) and `boral` (Hui 2016) packages, to mixture models that classify species according to their environmental response (Dunstan *et al.* 2013), and to capture-recapture models (Stoklosa, Dann & Huggins 2014). We propose extending this technique for residual construction to occupancy-detection models, but in two different ways, to diagnose the two components of the model.

DETECTION RESIDUALS

Information on the detection component is contained in the pattern of repeat detections at a site, across surveys. Thus, it is proposed here that detection residuals be constructed using the total number of detections at a site n_i, conditional on there having been some detections. This will be compared to the distribution expected under the fitted model.

Note that residuals computed from n_i will be more likely to find some assumption violations than others – this aggregates detection/nondetection events within a site, so one might expect it to be more sensitive to problems with the detection component across sites (e.g. a missing covariate that varies across site) than across surveys within a site (e.g. a missing covariate that varies across time periods). This is usually the more appropriate focus, because cross-site problems with the detection component are more likely to lead to bias in the occupancy component.

If the detection probability parameter is constant across surveys and equal to p_i, then the total number of detections across T surveys would be distributed as a truncated binomial with parameters T and p_i (truncated at one), so its cumulative distribution function is:

$$F(n_i|n_i > 0) = \frac{1}{1-(1-p_i)^T}\sum_{k=1}^{n_i}\binom{T}{k}p_i^k(1-p_i)^{T-k}$$

which is used to construct Dunn–Smyth detection residuals using eqn (2).

If on the other hand, the detection probability is not constant across surveys, $F(n_i)$ will need to be computed as a summation across all possible histories with n_i or fewer detections. The cumulative distribution function of n_i can be written as:

$$F(n_i|n_i > 0) = \frac{1}{1-\prod_{t=1}^T(1-p_{it})}\sum_{\boldsymbol{y}|1\le n_y\le n_i}\prod_{t=1}^T p_{it}^{y_t}(1-p_{it})^{1-y_t}$$

where $n_y = \sum_{t=1}^T y_t$ is the total number of detections in detection history $\boldsymbol{y}$. This function can be computationally intensive to enumerate if T and n_i are not small, because the number of possible histories with n_i or fewer detections becomes large. Nevertheless, our experience to date has been that this method requires less computation time than previously introduced simulation-based measures of goodness-of-fit (MacKenzie & Bailey 2004; Royle *et al.* 2007; see Data S5).

A technical difficulty arises when plotting detection residuals against covariates – covariates for detection may take different values across surveys at a site, yielding ST values, whereas the above method gives only one detection residual at each site, yielding only S values. In the plot produced later of detection residuals against a covariate (Fig. 2c), the covariate values were averaged across surveys within each site. Alternatives would be to replicate the detection residual T times, or to base residuals on the binary detection/nondetection events.

OCCUPANCY RESIDUALS

To diagnose the occupancy component of the model, the relevant available information is in the binary indicator for whether or not any detections are observed at a site, $J_i = I(n_i > 0)$. This is not solely a function of the occupancy component model, it is also related to the detection component (given that observing a species requires it to not just occupy a site, but also to be detected), which is reflected in the method of residual construction used here. The cumulative distribution function for any detections at a site, J_i, is as follows:

$$F(j_i) = \begin{cases} 1-\psi_i+\psi_i\prod_{t=1}^T(1-p_{it}) & \text{if } j_i = 0 \\ 1 & \text{if } j_i = 1 \end{cases}$$

in the case where detection probability is not constant across surveys. If it is constant, the product in the expression at $j_i = 0$ simplifies to $(1-p_i)^T$. Dunn–Smyth occupancy residuals are constructed using the binary site-level detection variable J_i, by substituting the above expression into eqn (2).

OMNIBUS RESIDUALS

Optionally, a single 'omnibus' residual could be constructed to simultaneously diagnose the detection and occupancy components of the model, by taking number of detections across all T surveys at a site (which would take the value zero if there were no detections). This has cumulative distribution function:

$$F(n_i) = \begin{cases} 1-\psi_i+\psi_i\prod_{t=1}^T(1-p_{it}) & \text{if } n_i = 0 \\ 1-\psi_i+\psi_i\sum_{\boldsymbol{y}|0\le n_y\le n_i}\prod_{t=1}^T p_{it}^{y_t}(1-p_{it})^{1-y_t} & \text{if } n_i = 1,2,\ldots,T \end{cases}$$

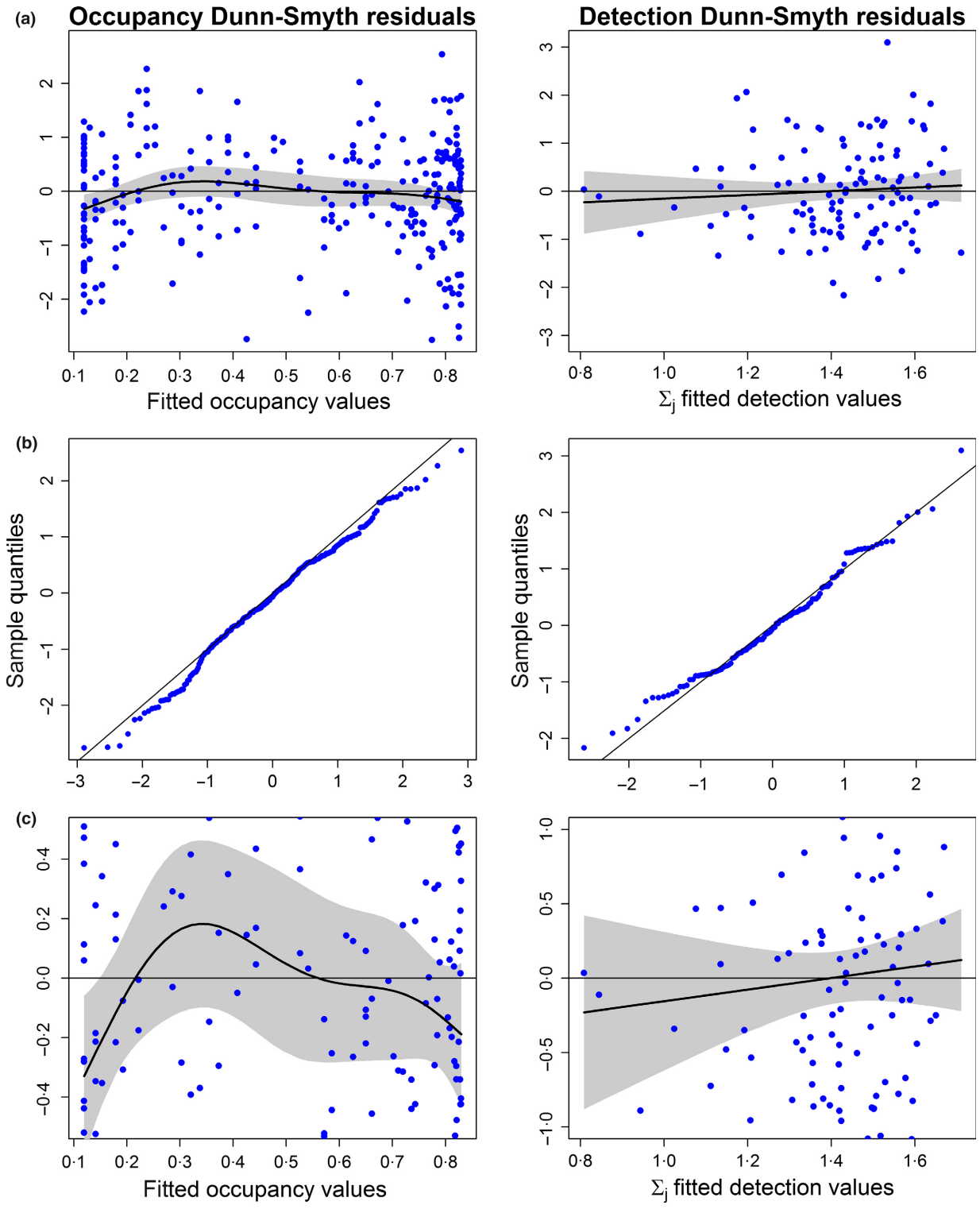

Fig. 1. Dunn–Smyth residual plots (with overlaid smooth curves and 95% confidence bands) for occupancy (left column) and detection (right column) for the 'No Elevation' model (see Table 1). (a) Residuals vs. fitted values, (b) Standard normal quantile plots, (c) the residual vs. fits plot from (a) 'zoomed in' around the line residual = 0 to see if the fitted smoother shows a trend. In this case there is a suggestion of lack of fit of the occupancy-detection model (from which the `elev` predictor had been omitted), although in this particular randomisation of residuals, the smoother still covers zero.

In this paper however the focus is on the use of detection and occupancy residuals only, as our aim is to 'drill down' into the model to attempt to identify the source of any assumption violations.

IMPLICATIONS OF 'JITTERING' FOR INTERPRETATION

Both detection and occupancy residuals involve random number generation (or 'jittering'), through u in eqn (2), in order to

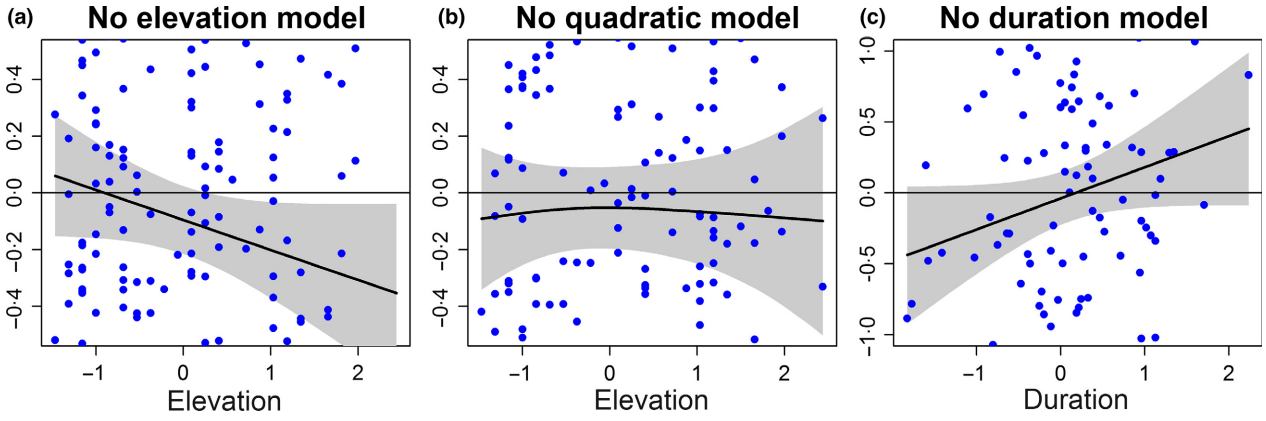

Fig. 2. Plotting residuals against covariates to detect potential violations of the fitted model. (a–b) Occupancy Dunn–Smyth residuals are plotted against the missing predictor elevation, for the 'No Elevation' and 'No Quadratic' Models respectively, (c) detection residuals against duration for the 'No Duration' model. There is a suggestion from (a) that elevation be included in the occupancy-detection model, from (b) that a quadratic for elevation may be warranted as well as the linear term, and from (c) that duration be included in the detection model.

avoid artifacts in analyses that could otherwise arise in residuals due to their discreteness. But this comes at the cost of introducing randomness. The amount of random variation that is introduced depends on how discrete the data are. For example, occupancy residuals are based on a binary response, and as such are highly discrete, and jittering can introduce a lot of variation into corresponding residuals. Detection residuals, based on a count response, can be expected to show less variability under jittering as a general rule, particularly if the number of surveys T is large. (This is one reason detection residuals were based on n_i, as these counts were expected to be more informative than binary detection/nondetection events.) The random variation introduced to residuals by jittering has a couple of consequences.

A first consequence is that it tends to be harder to see patterns in Dunn–Smyth residual plots involving jittering than is the case for usual continuous residuals, given the additional random variation that has been introduced. This can be understood as arising not because of the randomness *per se*, but because of the underlying discreteness in the data, which obscures patterns. A tool we have found especially useful is to add a smoother to residual plots, with 95% confidence bands around it, and to use the smoother as a basis for diagnosing lack of fit rather than the residuals themselves (we used the gam function from the MGCV package Wood 2011). This smooths out much of the noise in residual plots and can effectively find structure that is not obvious to the naked eye, as in Fig. 1a. Sometimes patterns in smoothers can be significant but quite small relative to the range of variation in residuals, thus we sometimes plot only part of the residual plot near zero in order to diagnose lack-of-fit solely from a smoother fitted to residuals (Fig. 1c). A useful diagnostic check is to see if 95% confidence bands around the smoother cover zero for the entire range of the data.

A second consequence of random variation in residuals is that different plots of the same data can look different, so at times it is advisable to repeat residual plots instead of just basing interpretation on a single plot. This is especially the case if there are just a few points driving a pattern, or if a smoother is close to missing zero at some points, because individual points can appear in quite different places when using different random values for jittering. Differences on repeat plots are typically subtle, e.g. compare Fig. 2 with the third row of Fig. 3 – these are plots of the same models on the same data but using different random values in jittering. There are subtle differences in residual patterns; the most conspicuous changes are in the smoothers, which now overlap zero in all cases, contrary to Fig. 2. Our interpretation in such a case is that there is marginal evidence of a violation of model assumptions.

Example

The above techniques will be illustrated on single season occupancy-detection models for the European red squirrel (*Sciurus vulgaris*) in Switzerland. Squirrel data were collected as part of the breeding bird survey by the Swiss Ornithological Institute (Schmid, Zbinden & Keller 2004), the 2007 dataset is available as Data S2. There were $S = 265$ sites surveyed, each one square kilometre in size, and surveyed $T = 3$ times during the bird breeding season (15 April–15 July). These data were previously analysed in chapter 10·9 of Kéry & Royle (2015).

Code for the below analyses, and an introductory tutorial on using Dunn–Smyth residuals in detection-occupancy models, are available in Data S4–S6.

A shortlist of four potential models were fitted, as listed in Table 1. Elevation and forest cover were used as occupancy covariates, and survey date and duration were used as detection covariates. All four models are a subset of the first 'Full' model, the 'No Elevation' model drops the elevation covariate from the occupancy component, the 'No Quadratic' model drops quadratic terms from the occupancy component (but keeps a quadratic term in the detection component), and the 'No Duration' model drops the duration covariate from the detection component of the model. Each model was fitted using `RPresence` version 1.1.2 (MacKenzie & Hines 2015). The chi-square parametric bootstrap test is reported (M&B GoF test, MacKenzie & Bailey 2004, using 1000 resamples), and AIC to compare across models (Table 1). Coefficients of

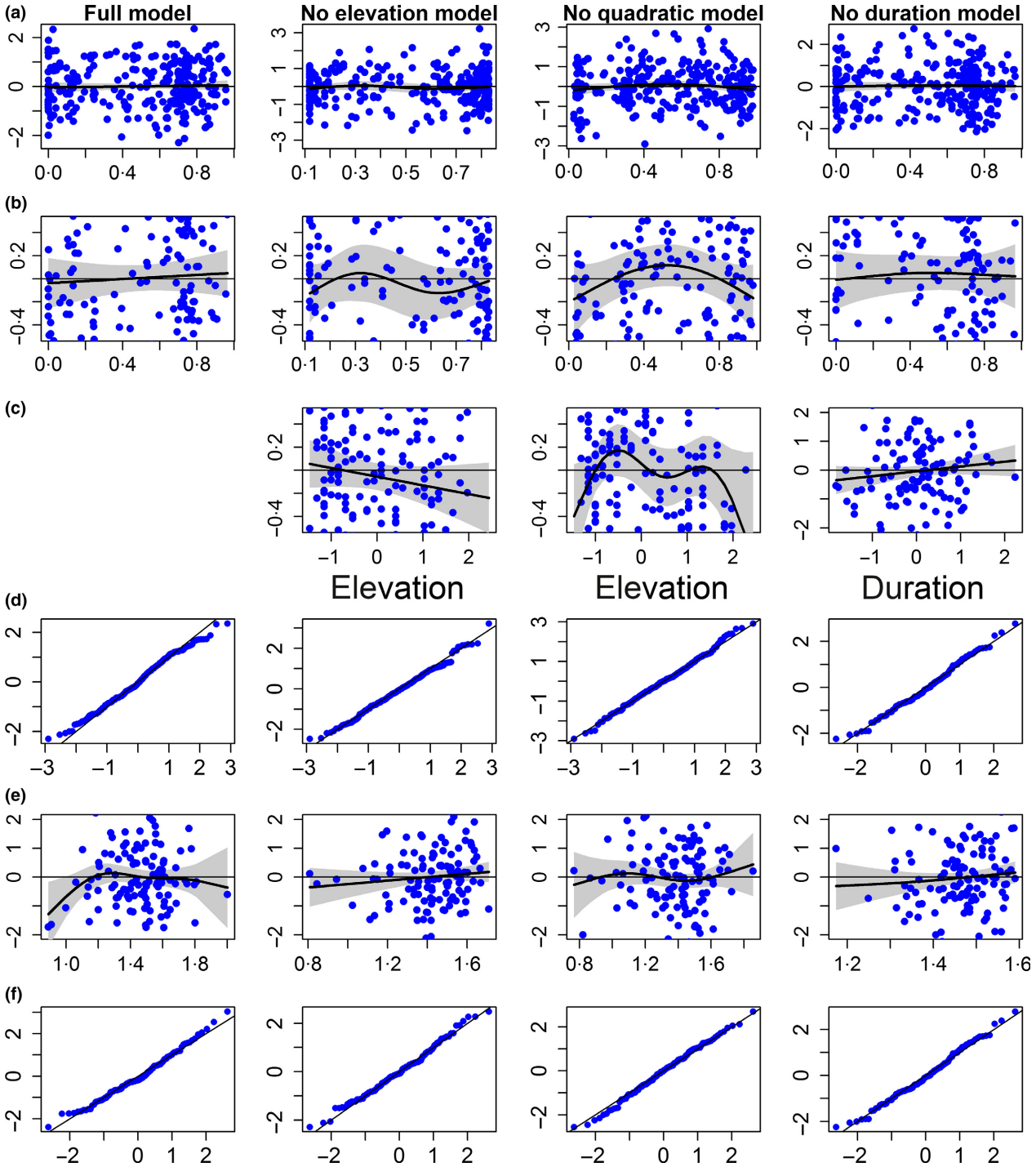

Fig. 3. Example display of the full set of diagnostic plots based on Dunn–Smyth residuals, for all four models from Table 1 (one column for each model). In simulation, this is the display of plots that was used to diagnose model goodness-of-fit. (a, b) Occupancy residual plots (as in Fig. 1a and c, left), (c) residuals against covariates (as in Fig. 2), (d) occupancy residual quantile plot (as in Fig. 1b, left), (e, f) detection residual plots (as in Fig. 1a, b, right).

the fitted models can be found in Data S3. The AIC metric suggests the 'Full' and 'No Duration' models had the best fit of the four considered, and that dropping elevation or quadratic terms from the occupancy component of the model is not advisable. The model was refitted in a Bayesian framework to estimate Bayesian *P*-values, using the same priors as given in Kéry & Royle (2015) (with 50 000 MCMC samples with a burn-in sample of 10 000 for three chains and thinning set to 10, see Data S3 for further details), based on R code available in Tobler *et al.* (2015). Neither goodness-of-fit test indicated lack of fit of any of the four models fitted to the data (Table 1).

Table 1. Model descriptions and goodness-of-fit results for four selected occupancy-detection models fitted to the European red squirrel data

Model	Model structure
Full	$p \sim date + dur + dur^2, \psi \sim elev + elev^2 + for + for^2 + elev : for + elev : for^2 + elev^2 : for$
No Elevation	$p \sim date + dur + dur^2, \psi \sim for + for^2$
No Quadratic	$p \sim date + dur + dur^2, \psi \sim elev + for + elev : for$
No Duration	$p \sim date, \psi \sim elev + elev^2 + for + for^2 + elev : for + elev : for^2 + elev^2 : for$

Model	ΔAIC	M&B GoF test statistic	M&B GoF test (*P*-value)	Bayesian *P*-value
Full	0·0	3·11	0·87	0·28
No Elevation	8·7	6·19	0·45	0·29
No Quadratic	7·9	3·49	0·90	0·31
No Duration	0·9	3·17	0·89	0·37

Elevation (*elev*) and forest cover (*for*) were used as occupancy (ψ) covariates, and survey date (*date*) and duration (*dur*) were used as survey/detection (p) covariates. Note that ':' represents an interaction between covariates.

Using the 'No Elevation' model as an illustrative example, residual vs. fits plots were constructed for each of occupancy and detection residuals, and normal quantile plots, to check for lack of fit (Fig. 1). The trend in the smoother for occupancy residuals (Fig. 1c, left) has a slight suggestion of a problem with the occupancy component of the model, but the smoother still covers zero at all values, so it is not conclusive.

The relevant Dunn–Smyth residuals can also be plotted against covariates, irrespective of whether or not they have been included in the model. Figure 2a plots occupancy residuals against elevation for the 'No Elevation' model, and is suggestive that elevation should be included in the model, although marginally so. Occupancy residuals from the 'No Quadratic' model seem to have a non-linear trend against elevation (Fig. 2b), motivating inclusion of a quadratic term in the model. Figure 2c has an increasing trend but still covers zero at all values, under other randomisations it does not. Thus, there is marginal evidence that duration should be included in the detection model.

Diagnosing goodness-of-fit for all four models requires a larger panel of residual plots as in Fig. 3. From the plots for the 'Full' model, there is no strong evidence for lack of fit, other results are largely as previously. Notice that for the model with a missing term from the occupancy component of the model (Fig. 3, second column), it is the occupancy residuals that are problematic (Fig. 3a–d, but especially b, c), whereas for the model with a missing term from the detection component (Fig. 3, fourth column) it is the detection residuals that are problematic (Fig. 3c and e, f, but especially c). Thus, the combination of occupancy and detection residuals seems to have some potential to identify where problems in a model may be. Further, plotting residuals against predictors suggests how the model can be improved (Fig. 3c).

Simulation studies

How effective are the above diagnostic tools at detecting lack of fit? This question was explored, using simulations mimicking the properties of the Swiss squirrel data. Parameter estimates from the four models of Table 1 (Data S3) were used to simulate data. The effect sizes tended to be larger for linear and quadratic coefficients of elevation as compared to duration, as suggested by the larger relative change in AIC when dropping from the full model to the 'No Elevation' and 'No Quadratic' models rather than when dropping to the 'No Duration' model (Table 1). Sample size was varied across two settings (using the existing sample size, $S = 265$, or doubling to $S = 530$) and number of surveys per site was similarly varied (using the existing value, $T = 3$, or doubling to $T = 6$). In total 160 datasets were generated (via the `simulate` function in the `unmarked` package, Fiske *et al.* 2011), 40 for each combination of S and T, choosing the true model at random from the four candidate models of Table 1.

For each dataset, the two simulation-based chi-square goodness-of-fit tests (MacKenzie & Bailey 2004; Royle *et al.* 2007) were computed for all four candidate models, using the same method described in the previous section, but in the case of Bayesian *P*-values, with the number of MCMC samples reduced to 5000 with a burn-in sample of 1000 for three chains and thinning set to 10, to reduce computation time. Bayesian *P*-values never detected assumption violations (as in Kéry & Royle 2015) and will not be considered further. Additionally, AIC was computed to see how reliably one could choose between the four candidate models. To assess the power of residual plots as a diagnostic tool, a display was constructed with detection and occupancy residual vs. fits plots for each of the four candidate models, as well as relevant plots against covariates, as in Fig. 3. These were generated by the second author and sent 'blind' to the four co-authors, 40 datasets per person. Those viewing the plots were told sample size information and which four models were being compared, but were not told the true model nor given access to the raw data nor results of goodness-of-fit tests. The MacKenzie & Bailey (2004) test was fitted to 200 separate simulated datasets to better estimate rejection rates. Code for these simulations is in Data S7.

Table 2 presents overall results (aggregated across the four model types due to small sample sizes). The graphical diagnostic tools correctly detected assumption violations at least two thirds of the time ('true positive'), but incorrectly declared and assumption violation when there was none as much as 30% of the time ('false positive', Table 2a). The false positive rate

Table 2. Simulation results (aggregated across the four model types) at each combination of S and T, with rates presented for: (a) Assumption violations correctly identified by goodness-of-fit checks (and false positives, when the model was in fact true, in parentheses); (b) Model selection results using AIC or diagnostic plots

	$S = 265$		$S = 530$	
	$T = 3$	$T = 6$	$T = 3$	$T = 6$
(a) Assumption violations identified				
Graphs	66 (28)	77 (19)	64 (29)	82 (15)
MacKenzie & Bailey (2004)	12 (12)	10 (14)	16 (12)	12 (12)
(b) Rate at which the selected model had no lack-of-fit				
Graphs	68	74	71	89
AIC	98	83	90	100

All rates presented to the nearest percentage, and estimated from 40 simulated datasets for graphical techniques, and from 200 separate datasets for the MacKenzie & Bailey (2004) test.

varied across the four analysts participating in the study (15–40%) as did the rate of correct detection of assumption violations (50–95%). Analysts more likely to declare an assumption violation tended to have higher false and true positive rates. There was some suggestion of an improvement when the number of surveys (T) was larger, with a slight increase in true positives and decrease in false positives (Table 2a).

The performance of the MacKenzie & Bailey (2004) test was disappointing, with false and true positive rates close to the chosen significance level of 10%. The MacKenzie & Bailey (2004) test has greater sensitivity to violations in the detection component of the model, further simulations exploring this are in the next section.

While the graphical diagnostic plots were developed to assess goodness-of-fit rather than for model selection, it is interesting to note that they had some power in this context, selecting a correct model out of a choice of four more than half the time, but did not perform nearly as well as AIC (Table 2b).

A number of different types of residual plots were used to diagnose assumption violations, as in Fig. 3, and it is interesting to study which types of plots were most effective at diagnosing problems with the various models considered (Table 3). The most useful plots were of residuals against selected covariates (as in Figs 2 or 3c), these correctly identified assumption violations as much as 75% of the time, although with a false positive rate as high as 35%. Residual vs. fits plots were next most effective, especially of occupancy residuals when 'zoomed in' on the smoother to see if it moved away from zero. This plot was particularly useful at diagnosing problems with the 'No Elevation' and 'No Quadratic' models, which had violations specifically of the occupancy component of the model (22 and 37%, respectively). Normal quantile plots had little effectiveness.

In Table 3, violations of the Full Model were sometimes declared, even though data were always simulated under some subset of the Full Model, thus violations of it were impossible. Analysts were aware of this, but if a residual plot had the appearance of an assumption violation (e.g. a smoother not covering zero), an assumption violation was declared irrespective of the fitted model, because this is what the analyst would have done in practice, when the true model is unknown.

POWER SIMULATION FOR VIOLATIONS OF THE DETECTION MODEL

An additional power simulation was constructed to better understand the poor performance of the MacKenzie & Bailey (2004) test and detection residual vs. fits plots – both had close to 10% false positives irrespective of whether the model was correct (Tables 2 and 3). Both procedures were designed to be sensitive to violations of the detection component in particular, and the only such scenario in simulations (fitting the 'No Duration' model when it was not correct) had a relatively small effect size – data were simulated with standardised Duration coefficients of 0·3 or less, and a relatively small difference in AIC between the 'No Duration' and 'Full' models (Table 1).

Table 3. Effectiveness of different types of residual plot for the four different models, from 160 simulated datasets (aggregating across the different values of S and T)

	Full model	No Elevation	No Quadratics	No Duration
Occupancy residual vs. fits plot	(2)	3 (0)	3 (0)	2 (5)
Occupancy res. vs. fits 'zoomed in'	(12)	22 (8)	37 (12)	20 (21)
Residual vs. covariate plots	(–)	75 (35)	29 (16)	65 (15)
Occupancy residual quantile plot	(2)	3 (0)	5 (0)	7 (3)
Detection residual vs. fits plot	(8)	11 (19)	14 (2)	14 (10)
Detection residual quantile plot	(4)	3 (3)	6 (2)	4 (3)

Plots are ordered down rows as they appear in Fig. 3. Reported values are 'true positive' rates to the nearest percentages, i.e. rates of detected assumption violations, with false positives in parentheses. Note that true positives were not possible for the Full model, and there was no residual vs. covariate plot for the Full model.

Table 4. Power simulation of the effectiveness of different methods of interrogating the detection component of the model

	$\beta_{dur} = \beta_{dur^2} =$				
	0	0·3	0·6	0·9	1·2
MacKenzie & Bailey (2004)	14	9	19	30	41
Residual vs. covariate plot	15	80	98	100	100
Detection residual vs. fits plot	8	10	15	18	30
Occupancy res. vs. fits (zoomed in)	5	8	10	8	15

Data were simulated under the 'Full' model with linear (β_{dur}) and quadratic (β_{dur^2}) coefficients for Duration taking one of five values, then the 'No Duration' model was fitted. The proportion of times an assumption violation was declared is reported to the nearest percentage, based on 250 datasets for MacKenzie & Bailey (2004), and from a separate 40 datasets for the residual plots.

So an additional simulation focussed on methods designed to interrogate the detection component of the model, as the effect size in the detection component was varied.

Data were simulated using parameters from the 'Full' Model, but with the linear and quadratic coefficients for Duration set to each of {0, 0·3, 0·6, 0·9, 1·2}. The 'No Duration' model was then fitted to the data, in which case the first setting ($\beta_{dur} = \beta_{dur^2} = 0$) was a Type I error simulation, and the second setting ($\beta_{dur} = \beta_{dur^2} = 0{\cdot}3$) had an effect of a similar size to that used in the original simulation. The subsequent values increased the effect size progressively to assess when lack-of-fit in the detection component was identified. The original dataset design and size was used ($S = 265$, $T = 3$). For each of 250 simulated datasets, the test due to MacKenzie & Bailey (2004) used 250 bootstrap samples and a significance level of 0·1. For each of a separate set of 40 simulated datasets, detection residual vs. fits and detection residual vs. duration plots were constructed. Occupancy residual vs. fits plots were also constructed for comparison. Each plot was viewed by the first author, who recorded the proportion of plots for which the confidence band on the smoother did not cover zero for the full range for which data were observed. Code for these simulations is in Data S8.

The MacKenzie & Bailey (2004) test had increasing power (Table 4), with little power for small effect sizes but increasing to about 40% when the effect size was large. The most striking result, however, was the high power of the missing covariate plot (Fig. 2c) – when plotting detection residuals against the missing Duration variable, an assumption violation was correctly detected on most occasions, with a false positive rate of about 15% (Table 4). The detection residual vs. fits plot was less successful, having increasing power, but at a slower rate than the MacKenzie & Bailey (2004) test. The occupancy residual vs. fits plot had lower power again, emphasising that this plot is better at detecting effects in the occupancy component of the model.

Discussion

Residuals have been developed, along the lines of those proposed by Dunn & Smyth (1996), that can be used for diagnostic checks of occupancy-detection models. The main advantage of this approach is that visual displays can assist in model construction as well as diagnosis, e.g. a curvilinear trend on a residual vs. fits plot suggests non-linearity, or a trend against a covariate not included in the model suggests that it should be included. By constructing separate residuals for occupancy and detection components of the model, the component(s) of the model that have problematic assumptions can be more readily identified, with problems in the occupancy component more readily identified by occupancy residuals (Table 3, 'No Elevation' and 'No Quadratic' models) and problems with the detection component more readily identified by detection residuals (Table 4), albeit some ambiguity and false positives. We think Dunn–Smyth residuals are a potentially useful tool for model builders, and plan to include them in the freely available `RPresence` package (MacKenzie & Hines 2015).

Simulations in this paper compared plots based on these residuals to a common tool for diagnostic checking (MacKenzie & Bailey 2004) and found residual plots to have noticeable power advantages when diagnosing the occupancy component of the model, or when plotting against a covariate missing from the detection component of the model. While decisions about model adequacy were made from a panel of five to six residual plots per model (columns of Fig. 3), a couple of individual plots performed quite well in their own right. Without taking into account *a priori* information about which types of assumption violations to focus on, the residual vs. fits plot was the most useful single plot with a smoother and 95% confidence bands to check if the trend deviated significantly from zero at any point. When these confidence bands were narrow we found it helpful to 'zoom in' around zero (as in Fig. 1c), necessary for occupancy residuals because the high level of discreteness in the data (these were based on a binary response) dulled the signal such that quite subtle changes in the mean of residuals could be important. Detection residual vs. fits plots were less effective, and did not seem as capable of finding violations of the detection component as the MacKenzie & Bailey (2004) test. We also found that normal quantile plots (Fig. 1b) were not particularly useful, perhaps because of difficulties assessing when a deviation from normality was 'significant'. In future work, perhaps simulation envelopes (Baddeley *et al.* 2014) around the one-to-one line would assist in interpreting quantile plots.

In the original set of simulations (Tables 2 and 3), model violations were detected as much as 80% of the time, but this came at the cost of a false positive rate sometimes as large as

35%. One reason for this seemed to be 'multiple testing' – individual graphs (Table 3) typically had lower false positive rates than when making a decision collectively from all graphs (Table 2). One solution would be to use smoothers with wider confidence bands to reflect the number of plots to be viewed (e.g. use 99% instead of 95% when viewing five residual plots with smoothers). This would reduce the false positive rate, but at the cost of reducing power to detect violations of model assumptions.

We found in simulations we were able to make more effective decisions when exploiting *a priori* knowledge about potentially missing covariates or missing terms (Fig. 2). Without the missing covariate plots it was more difficult to see model problems and they would often go underdetected (Table 3). If there is information about the types of assumption violations that might be expected, or if a specific type of change to the model is being considered (variable deletion or addition), it is advisable to construct diagnostic checks specifically tailored to interrogate the relevant assumptions (as in Fig. 2). As an aside, it seems to be easier to detect problems with the functional form assumed in a model when plotting residuals against each covariate separately, rather than plotting them once against fitted values, so there is an argument that plotting against each covariate should be done routinely.

Graphical tools were also considered as a model selection technique to choose between a set of candidate models, and found to be useful but generally less effective than AIC. Instead the main role we see for graphical tools is in suggesting different ways to build the model outside of a preconceived set of candidate models, e.g. suggesting a quadratic term be added to the model to deal with nonlinearity on a residual plot.

In this paper only a few types of assumption violation were actually considered, in simulations and in residual construction. Violations of independence assumptions were not a focus here. Spatial autocorrelation could be diagnosed from Dunn–Smyth residuals using standard techniques (e.g. a correlogram, or a spatial plot of residuals), although we expect that the discreteness in the data, converted into randomness by jittering, will dull the signal, as seen earlier. Dependence across repeated surveys in time could be diagnosed using the time between detections as a basis for a residual, along the lines of Guillera-Arroita *et al.* (2011), or using join statistics (Wright, Irvine & Rodhouse 2016). A related issue is overdispersion, which could in principle be identified from detection residuals, although simulations did not specifically test this. MacKenzie & Bailey (2004) found their parametric bootstrap test to have some power to detect overdispersion.

Another core assumption of occupancy-detection models is that imperfect detection is modelled as an event collectively across all individuals in a site survey. Often it would be more natural to assume detection events operate at the level of the individual (and as such detection of a species in a survey should be a function of its abundance). This problem is best addressed at the design stage by collecting additional data to model abundance counts directly, while accounting for imperfect detection of individuals (as in Royle *et al.* 2007), or using other extensions that account for more generic detection heterogeneity (e.g. MacKenzie *et al.* 2006; Royle 2006a).

Because of the limited information available in a binary response, it is difficult to model detection/nondetection data and to separate out different sources of effects. This is even more difficult if there is insufficient information in the data to estimate all the parameters, known as model unidentifiability (Carroll, Ruppert & Stefanski 1995). A simple way to diagnose models that are near-unidentifiable in the non-Bayesian setting is to look at the standard errors of model coefficients, and whether they are unduly large (more specifically, whether the variance-covariance matrix of coefficients is near-singular, which inflates standard errors). Problems can arise when sites are not sufficiently intensively sampled (e.g. a small number of survey visits when detection probability is small), or if predictors used for occupancy and detection components are highly correlated. In both cases a large number of sites may be needed to estimate parameters reliably. Increasing sample size does not resolve all problems, e.g. if there is a need to include parameters for false positives in the model, additional information on false positive rates is needed to estimate them (Chambert, Miller & Nichols 2015). Finally, when all that is available is detection/nondetection data, and we wish to tease apart the occupancy and detection processes, some core assumptions are necessary that can be difficult to check, such as the 'closure' assumption (that occupancy status does not change across surveys). Sometimes the study design immediately suggests the closure assumption is reasonable, e.g. if surveys are simultaneously conducted by different observers, but in other settings it is not guaranteed and inferences about occupancy and detection can then be unreliable (Rota *et al.* 2009). Robust designs can be used in collecting data to test (Rota *et al.* 2009) or relax (MacKenzie *et al.* 2006) the closure assumption, or it can be relaxed by assuming staggered arrival and departure times (Kendall *et al.* 2013). Designing studies with this assumption in mind is the best way to guard against problems.

Dunn–Smyth residuals were developed here for simple occupancy-detection models, and a number of extensions to other models for imperfect detection are possible and indeed desirable. Mixture models are sometimes used to account for heterogeneity in detection that is not explained by covariates (Royle 2006b), a relatively straightforward extension, indeed such residuals have been used in a mixture model previously (Dunstan *et al.* 2013). They could also be generalised to models for abundance (as in Royle *et al.* 2007).

Authors' contributions

D.W., D.M. and A.W. conceived the ideas for this paper, D.W. and J.S. led the programming, J.S. led the data analysis, all authors contributed to evaluation of residual diagnostic tools on simulated data, J.S. and G.G. led the literature review, D.W. led the write-up. All authors contributed critically to the drafts.

Acknowledgements

Our thanks to the Die Schweizerische Vogelwarte for permission to use the Swiss squirrel data in this paper. D.I.W. is supported by an Australian Research Council Future Fellowship (FT120100501), G.G.A. by an Australian Research Council Discovery Early Career Researcher Award (DE160100904).

Data accessibility

The Swiss squirrel dataset used in this paper is available as Data S2.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Baddeley, A., Diggle, P.J., Hardegen, A., Lawrence, T., Milne, R.K. & Nair, G. (2014). On tests of spatial pattern based on simulation envelopes. *Ecological Monographs*, **84**, 477–489.

Brockwell, A. (2007). Universal residuals: a multivariate transformation. *Statistics and Probability Letters*, **77**, 1473–1478.

Broms, K.M., Hooten, M.B. & Fitzpatrick, R.M. (2016). Model selection and assessment for multi-species occupancy models. *Ecology*, **97**, 1759–1770.

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

Burnham, K.P. & Anderson, D.R. (1998). *Model Selection and Inference – A Practical Information-Theoretic Approach*. Springer-Verlag, New York City, NY, USA.

Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall, London, UK.

Chambert, T., Miller, D. & Nichols, J. (2015). Modeling false positive detections in species occurrence data under different study designs. *Ecology*, **96**, 332–339.

Choquet, R., Lebreton, J.-D., Gimenez, O., Reboulet, A.-M. & Pradel, R. (2009). U-CARE: utilities for performing goodness of fit tests and manipulating CApture-REcapture data. *Ecography*, **32**, 1071–1074.

Davison, A.C. & Hinkley, D.V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge, UK.

Dunn, P.K. & Smyth, G.K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.

Dunstan, P.K., Foster, S.D., Hui, F.K.C. & Warton, D.I. (2013). Finite mixture of regression modelling for high-dimensional count and biomass data in Ecology. *Journal of Agricultural, Biological and Environmental Statistics*, **18**, 357–375.

Fiske, I. & Chandler, R. (2011). unmarked: an `R` package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, **43**, 1–23.

Gelman, A., Meng, X.-L. & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, **6**, 733–760.

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography*, **40**, 281–295.

Guillera-Arroita, G., Morgan, B.J.T., Ridout, M.S. & Linkie, M. (2011). Species occupancy modeling for detection data collected along a transect. *Journal of Agricultural, Biological, and Environmental Statistics*, **16**, 301–317.

Hui, F.K. (2016). boral – Bayesian ordination and regression analysis of multivariate abundance data in R. *Methods in Ecology and Evolution*, **7**, 744–750.

Kendall, W.L., Hines, J.E., Nichols, J.D. & Grant, E.H.C. (2013). Relaxing the closure assumption in occupancy models: staggered arrival and departure times. *Ecology*, **94**, 610–617.

Kéry, M. & Royle, J.A. (2015). *Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS*. Academic Press, London, UK.

Kery, M. & Schaub, M. (2011). *Bayesian Population Analysis Using WinBUGS: A Hierarchical Perspective*. Elsevier, Waltham, MA, USA.

MacKenzie, D.I. & Bailey, L.L. (2004). Assessing the fit of site-occupancy models. *Journal of Agricultural, Biological, and Environmental Statistics*, **9**, 300–318.

MacKenzie, D. & Hines, J. (2015). RPresence: R Interface for Program PRESENCE. R package version 1.1.1.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Andrew Royle, J. & Langtimm, C.A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006). *Occupancy Estimation and Modelling – Inferring Patterns and Dynamics of Species Occurrence*. Elsevier, Burlington, MA, USA.

Pollock, K.H., Hines, J.E. & Nichols, J.D. (1985). Goodness-of-fit tests for open capture-recapture models. *Biometrics*, **11**, 399–410.

Rota, C.T., Fletcher Jr, R.J., Dorazio, R.M. & Betts, M.G. (2009). Occupancy estimation and the closure assumption. *Journal of Applied Ecology*, **46**, 1173–1181.

Royle, J.A. (2006a). Site occupancy models with heterogeneous detection probabilities. *Biometrics*, **62**, 97–102.

Royle, J.A. (2006b). Site occupancy models with heterogeneous detection probabilities. *Biometrics*, **62**, 97–102.

Royle, J.A., Kry, M., Gautier, R. & Schmid, H. (2007). Hierarchical spatial models of abundance and occurrence from imperfect survey data. *Ecological Monographs*, **77**, 465–481.

Schmid, H., Zbinden, N. & Keller, V. (2004). Überwachung der Bestandsentwicklung häufiger Brutvögel in der Schweiz.

Smith, J.Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting*, **4**, 283–291.

Stoklosa, J., Dann, P. & Huggins, R.M. (2014). Semivarying coefficient models for capturerecapture data: colony size estimation for the little penguin Eudyptula minor. *Mathematical Biosciences*, **255**, 43–51.

Tobler, M.W., Zúñiga Hartley, A., Carrillo-Percastegui, S.E. & Powell, G.V.N. (2015). Spatiotemporal hierarchical modelling of species richness and occupancy using camera trap data. *Journal of Applied Ecology*, **52**, 413–421.

Tyre, A.J., Tenhumberg, B., Field, S.A., Niejalke, D., Parris, K. & Possingham, H.P. (2003). Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications*, **13**, 1790–1801.

Wang, Y., Naumann, U., Wright, S.T. & Warton, D.I. (2012). `mvabund` – an `R` package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, **3**, 471–474.

White, G. & Burnham, K.P. (1999). Program MARK: survival estimation from populations of marked animals. *Bird Study*, **46** Supplement, 120–138.

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 3–36.

Wright, W.J., Irvine, K.M. & Rodhouse, T.J. (2016). A goodness-of-fit test for occupancy models with correlated within-season revisits. *Ecology and Evolution*, **9**, 5404–5415.

Received 8 November 2016; accepted 30 January 2017
Handling Editor: Robert B. O'Hara

Supporting Information

Details of electronic Supporting Information are provided below.

Data S1. Literature Search.

Data S2. Squirrels data.

Data S3. Additional analysis details.

Data S4. Tutorial for plotting Dunn-Smyth residuals from fitting occupancy-detection models.

Data S5. Swiss Squirrel Analysis: Goodness-of-fit tests for occupancy-detection models.

Data S6. ResidFunctions.R.

Data S7. DiagnosticSims.R – Simulations.

Data S8. DetectionSims.R.