ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2017)

A new strategy for diagnostic model assessment in capture–recapture

Rachel S. McCrea and Byron J. T. Morgan

University of Kent, Canterbury, UK

and Olivier Gimenez

Centre d'Ecologie Functionelle et Evolutive, Montpellier, France

[Received December 2015. Final revision September 2016]

Summary. Common to both diagnostic tests used in capture–recapture and score tests is the idea that starting from a simple base model it is possible to interrogate data to determine whether more complex parameter structures will be supported. Current recommendations advise that diagnostic tests are performed as a precursor to a model selection step. We show that certain well-known diagnostic tests for examining the fit of capture–recapture models to data are in fact score tests. Because of this direct relationship we investigate a new strategy for model assessment which combines the diagnosis of departure from basic model assumptions with a step-up model selection, all based on score tests. We investigate the power of such an approach to detect common reasons for lack of model fit and compare the performance of this new strategy with the existing recommendations by using simulation. We present motivating examples with real data for which the extra flexibility of score tests results in an improved performance compared with diagnostic tests.

Keywords: Goodness-of-fit tests; Model selection; Power; Transience; Trap dependence; U-CARE

1. Introduction

This paper considers model selection for capture–recapture data that are obtained from open populations of wild animals. Capture–recapture studies involve the capture and unique marking of individuals, which are then released into the population and subsequent attempts are made to recapture them. The resulting data can be recorded as individual encounter histories for each animal, which take the form of vectors with elements 0 and 1, indicating non-capture and capture respectively. The encounter history data can often be conveniently summarized in terms of an upper triangular matrix, which is known as an m-array, with elements $m_{i,j}$ denoting the number of individuals released at occasion t_i and next recaptured at occasion t_j, concatenated with a column vector with elements v_i denoting the numbers of individuals released at occasion t_i which were never captured again. The ith row of the matrix has a multinomial distribution with index R_i denoting the number of individuals released at occasion t_i, $i = 1, \ldots, T$. We write $\mathbf{m} = \{m_{ij}\}$ and $\mathbf{v} = \{v_i\}$.

The Cormack–Jolly–Seber (CJS) model is the benchmark model for such data when age structure is not considered. It is defined in terms of two sets of parameters: ϕ_i is the probability

Address for correspondence: Rachel S. McCrea, National Centre for Statistical Ecology, School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, CT2 7NF, UK.
E-mail: R.S.McCrea@kent.ac.uk

 0035–9254/17/66000
Published by John Wiley & Sons Ltd on behalf of the Royal Statistical Society.

that an individual that is alive at time t_i survives until time t_{i+1}, and p_i is the probability that an individual that is alive at time t_i is captured at that time. We write $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_{T-1}\}$ and $\mathbf{p} = \{p_2, \ldots, p_T\}$. The likelihood is then product multinomially distributed, over the rows of the m-array, defined by

$$L(\boldsymbol{\phi}, \boldsymbol{p}; \mathbf{m}, \mathbf{v}) \propto \prod_{i=1}^{T-1} \left(\prod_{j=i+1}^{T} \eta_{i,j}^{m_{i,j}} \right) \times \chi_i^{v_i} \tag{1}$$

where

$$\eta_{ij} = \phi_i \left\{ \prod_{k=i+1}^{j-1} \phi_k (1 - p_k) \right\} p_j \qquad \text{for } i < j,$$

and $\eta_{ij} = 0$ for $i \geqslant j$. We define $\chi_i = 1 - \Sigma_{j=i+1}^{T} \eta_{ij} = 1 - \phi_i \{1 - (1 - p_{j+1}) \chi_{i+1}\}$ for $i < T$, and $\chi_T = 1$.

If all $m_{ij} > 0 \; \forall \; i < j$ then the CJS model is parameter redundant with deficiency of 1, since ϕ_{T-1} and p_T only ever occur in the cell probabilities as a product. The other parameters and this product have explicit maximum likelihood estimates; see for example McCrea and Morgan (2014), page 70. We note that, if some $m_{ij} = 0$, for $i < j$, the parameter redundancy of the model may change; see for example Cole *et al.* (2012).

The likelihood of equation (1) can be factorized to give a term involving the model parameters and one which provides the distribution of data conditional on a set of sufficient statistics. The second of these terms may be used to assess model adequacy; see Davison (2003), page 177. Pollock *et al.* (1985) derived a goodness-of-fit test for the Jolly–Seber capture–recapture model. The CJS model is a special case of the Jolly–Seber model and thus this goodness-of-fit test is also a goodness-of-fit test for the CJS model. Burnham (1991) showed that the Jolly–Seber goodness-of-fit test can be expressed as the product of two conditionally independent terms, which lead to the diagnostic tests that are now known as test 2 and test 3. We describe these in detail in the next section. The diagnostic tests do not require any model fitting and it is thus recommended that these are performed as a preliminary step, before model selection, which may result in a simplified set of models for consideration—see Lebreton *et al.* (1992) and Pradel *et al.* (2005).

The CJS model of equation (1) has been extended in many directions, which creates a problem for model selection. Two generalizations which relate to the diagnostic tests which we shall encounter later are a model incorporating trap dependence and a model accommodating transient individuals, which are individuals which pass through the study area and are therefore encountered only once. Structurally the transience model is equivalent to a capture–recapture model with two age classes for survival, with all individuals marked as young.

The trap-dependent model is defined in terms of three sets of parameters: $\{\phi_i\}$ and $\{p_i\}$ as before, and p_i^* is the probability that an individual alive at time t_i is captured at that time, given that it was also caught at occasion t_{i-1}. We write $\mathbf{p}^* = \{p_2^*, \ldots, p_T^*\}$. The likelihood is then a product multinomial distribution, over the rows of the m-array, defined by

$$L(\boldsymbol{\phi}, \mathbf{p}, \mathbf{p}^*; \mathbf{m}, \mathbf{v}) \propto \prod_{i=1}^{T-1} \left\{ \prod_{j=i+1}^{T} (\eta_{i,j}^{\mathrm{TD}})^{m_{i,j}} \right\} \times \chi_i^{v_i} \tag{2}$$

where

$$\eta_{i,i+1}^{\mathrm{TD}} = \phi_i p_{i+1}^*,$$

$$\eta_{ij}^{\mathrm{TD}} = (1-p_{i+1}^*)\phi_i\phi_{i+1}\left\{\prod_{k=i+2}^{j-1}\phi_k(1-p_k)\right\}p_j \qquad \text{for } i<j+1,$$

and $\eta_{ij}^{\mathrm{TD}}=0$ for $i \geqslant j$ and $\chi_i = 1-\Sigma_{j=i+1}^{T}\eta_{ij}^{\mathrm{TD}}$.

The standard m-array is not sufficient for fitting a model for transience. We generalize the m-array by defining $m_{\{0\}ij}$ to be the number of individuals that are captured for the first time at occasion t_i and next recaptured at occasion t_j and $m_{\{1\}ij}$ to be the number of previously captured individuals which are captured at occasion t_i and next recaptured at occasion t_j. $v_{\{0\}i}$ denotes the numbers of newly marked individuals that were released at occasion t_i which were never captured again, and $v_{\{1\}i}$ denotes the numbers of previously marked individuals that were released at occasion t_i which were never captured again. We write $\mathbf{m}_{\{\mathbf{0}\}}=\{m_{\{0\}ij}\}$, $\mathbf{m}_{\{\mathbf{1}\}}=\{m_{\{1\}ij}\}$, $\mathbf{v}_{\{\mathbf{0}\}}=\{v_{\{0\}i}\}$ and $\mathbf{v}_{\{\mathbf{1}\}}=\{v_{\{1\}i}\}$. The transience model is then defined in terms of three sets of parameters: $\{\phi_i\}$ and $\{p_i\}$ as before, and ϕ_i^* is the probability that a newly marked individual that is alive at time t_i survives until time t_{i+1}. We write $\boldsymbol{\phi}^*=\{\phi_1^*,\ldots,\phi_{T-1}^*\}$. The likelihood is then a product multinomial distribution, over the rows of the extended m-array, defined by

$$\begin{aligned} L(\boldsymbol{\phi}^*,\boldsymbol{\phi},\mathbf{p};\mathbf{m}_{\{\mathbf{0}\}},\mathbf{m}_{\{\mathbf{1}\}},\mathbf{v}_{\{\mathbf{0}\}},\mathbf{v}_{\{\mathbf{1}\}}) \propto & \prod_{i=1}^{T-1}\left\{\prod_{j=i+1}^{T}\eta_{\{0\}i,j}^{m_{\{0\}i,j}}\right\}\times\chi_{\{0\}i}^{v_{\{0\}i}} \\ & \times\prod_{i=1}^{T-1}\left\{\prod_{j=i+1}^{T}\eta_{\{1\}i,j}^{m_{\{1\}i,j}}\right\}\times\chi_{\{1\}i}^{v_{\{1\}i}} \end{aligned} \tag{3}$$

where

$$\eta_{\{0\}} = \phi_i^*\left\{\prod_{k=i+1}^{j-1}\phi_k(1-p_k)\right\}p_j \qquad \text{for } i<j,$$

$$\eta_{\{1\}} = \phi_k\left\{\prod_{k=i+1}^{j-1}\phi_k(1-p_k)\right\}p_j \qquad \text{for } i<j,$$

and $\eta_{\{0\}}=\eta_{\{1\}}=0$ for $i \geqslant j$, $\chi_{\{0\}i}=1-\Sigma_{j=i+1}^{T}\eta_{\{0\}ij}$ and $\chi_{\{1\}i}=1-\Sigma_{j=i+1}^{T}\eta_{\{1\}ij}$.

In current practice, tests of whether trap dependence or transience are required within the model use appropriately constructed contingency tables, which have the benefit of reducing model fitting but the weakness of low power and disconnection from the parametric modelling framework. These tests can alternatively be considered as diagnostics regarding the omission of particular components or as steps in a selection procedure of which components should be included. Within this paper we propose alternative likelihood-based methods.

We might expect diagnostic tests to be related to score tests and we demonstrate that, whereas two important diagnostic tests are, others are not. In addition the two approaches that we compare within this paper differ in mode of application and thus have the potential to produce different results. Model selection procedures using these two approaches are compared in this paper, and clear conclusions result.

The methods that are proposed in this paper can be applied to any capture–recapture data; the approach is shown to be at least as good as existing methods, and in fact it often outperforms other approaches because of the improvement in statistical power.

Motivating examples are introduced in Section 2 and within Section 3 the connection between score and diagnostic tests is established. Section 4 describes the two model selection strategies and compares them by using simulation. The analyses of the two case-studies that are described in Section 2 are presented in Section 5 and the paper ends with discussion and recommendations in Section 6.

The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Motivating examples

We consider two motivating capture–recapture data sets. The first is a large study of breeding great cormorants *Phalacrocorax carbo sinensis* from Denmark. The cormorant data have been fully analysed in Hénaux *et al*. (2007). The cormorants provide a complex case-study for which it is unknown *a priori* what behavioural traits may be exhibited by the population. The data consist of capture histories from 862 breeding birds, captured at an established single colony over a period of 11 breeding seasons. The cormorants are only initially captured at the time of marking and are then subsequently resighted in the breeding colony.

The second is a set of capture–recapture data on the humpback whale *Megaptera novaeangliae* population in the South Pacific. These data have been analysed by Madon *et al*. (2013). The capture–recapture data are compiled from genetic records and here we consider just the female genetic data for illustration, which have capture histories from 101 individuals, collected over a period of seven encounter occasions.

In both cases, identifying behavioural responses, such as transience or trap response, may provide important biological insight into the animals being studied. If such responses are ignored within a model, then biases would result in the estimates of the parameters of interest, and therefore it is essential to fit appropriate models to the data.

3. Equivalence of score tests and diagnostic tests

Diagnostic tests for capture–recapture data have become a standard preliminary tool before model fitting and consist of a number of contingency table tests based on summary statistics. They are commonly used because of readily available computer software, RELEASE, which can be run from within program MARK (White and Burnham, 1999) and U-CARE (Choquet *et al.*, 2009). Once the preliminary diagnostic tests have been conducted, the traditional approach then relies on fitting all biologically plausible models (excluding those which have been ruled out by the diagnostic tests), comprising a model set which can be prohibitively large for successful implementation. An alternative step-up model selection strategy using score tests has been successfully used for ring recovery models (Catchpole and Morgan, 1996) and multistate capture–recapture models (McCrea and Morgan, 2011). For comparing nested models, score tests are asymptotically equivalent to likelihood ratio tests under the null hypothesis, but they are simpler in not requiring models to be fitted under the alternative hypothesis to conduct tests. See for example Morgan (2008), page 101.

Both diagnostic and score tests share the common feature of checking whether particular aspects of models need to be included in a model selection procedure, starting from a simple base model and without fitting more complex models unless the data suggest otherwise. It is therefore natural to explore the relationships that might exist between the two types of test.

Smyth (2003) showed that the Pearson goodness-of-fit test for a 2×2 contingency table is mathematically equivalent to a score test and we outline a proof of this in Appendix A.1. We now use this result to demonstrate how specific important diagnostic tests for capture–recapture data can be expressed as score tests. Throughout the paper we adopt the notation that is used in the software U-CARE.

3.1. Diagnostic test 2

Test 2 involves comparing the future histories of individuals that are captured and not captured at a given capture occasion, and thus tests whether capturing individuals affects the probability of future encounters (Pradel, 1993). This test is performed through a series of paired contingency table tests, examining differences between individuals that were captured at occasion t_i and those not captured at occasion t_i but which are known to be alive then, thus detecting a behavioural response to capture. The tests for capture occasion t_i are denoted by test 2.CT(i) and test 2.CL(i), for $i = 2, \ldots, T-1$. The contingency table corresponding to test 2.CT(i) compares whether capture at occasion t_i affects time of subsequent capture and is generally given by Table 1.

We now consider the model probabilities that are associated with this test. If p^*_{i+1} denotes the probability that an individual is captured at occasion t_{i+1} given that it was also captured at occasion t_i, a score test for immediate trap dependence at occasion t_i would examine $H_0 : p^*_{i+1} = p_{i+1}$. An X^2-test of homogeneity based on the expected values of the contingency table tests whether

$$\frac{p^*_{i+1}}{p^*_{i+1} + \sum_{j=i+2}^{T} p_j \left\{ \prod_{k=i+1}^{j-1} \phi_k (1 - p_k) \right\}}$$

is equivalent to

$$\frac{p_{i+1}}{p_{i+1} + \sum_{j=i+2}^{T} p_j \left\{ \prod_{k=i+1}^{j-1} \phi_k (1 - p_k) \right\}}.$$

These expressions are equal if and only if $p^*_{i+1} = p_{i+1}$, and therefore, by Smyth (2003), test 2.CT(i) is equivalent to a score test. The peeling–pooling algorithm of Burnham (1991) demonstrates how p_{i+1} is estimated solely from the components of the m-array that is used within test 2.CT(i), which means that the score test of $H_0 : \phi_t, p_2, \ldots, p_{i+1} = p^*_{i+1}, \ldots, p_T$ *versus* $H_1 : \phi_t, p_2, \ldots, p_{i+1}, p^*_{i+1}, \ldots, p_T$, where $p_{i+1} \neq p^*_{i+1}$, is equivalent to test 2.CT(i).

Test 2.CL(i) tests for differences between the expected time of recapture between those captured and not captured at occasion t_i, for those individuals that were captured after time t_{i+1}. Thus, this component test should intuitively be equivalent to a score test of a delayed trap dependence, such that capture at occasion t_i affects capture at occasion t_{i+2}, as the test compares whether capture at occasion t_i affects the probability of capture at occasion t_{i+2} or later. However, in this case the score test of a long-term trap effect and test 2.CL(i) are not equivalent. This is due to the parameter p_{i+2} that appears in cell probabilities corresponding to cells which are not included in the contingency table for test 2.CL(i). It is, however, possible to perform a

Table 1. Contingency table for test 2.CT(i)

	Individuals captured at t_{i+1}	*Individuals captured after* t_{i+1}
Individuals not captured at t_i	$\sum_{k=1}^{i-1} m_{k,i+1}$	$\sum_{k=1}^{i-1} \sum_{h=i+2}^{T} m_{k,h}$
Individuals captured at t_i	$m_{i,i+1}$	$\sum_{h=i+2}^{T} m_{i,h}$

score test of long-term trap effect following capture and one approach of how this can be done is discussed in Appendix A.2.

3.2. Diagnostic test 3

Test 3 compares the future encounter histories of 'new' and 'old' individuals, where new individuals are those which have not been previously captured and old individuals are those which have been encountered before their current capture and thus will test for differences in survival probability of new and old individuals. The standard m-array that was presented earlier conditions on the time of last capture, and therefore the past encounters of particular individuals are not recorded within this format. It is therefore necessary to use the generalized m-array that was introduced in Section 1, which includes information on whether individuals are new or old. We note that, at occasion t_1, all released individuals will be new.

Test 3 is constructed as a series of contingency table tests based on the generalized m-array components, and comparisons are made between new and old individuals that are released at occasion t_i through tests 3.SR(i) and 3.Sm(i). The contingency table that is associated with component test 3.SR(i) is given by Table 2.

The probabilities that are associated with the contingency table for test 3.SR(i) are

$$\phi_i^* \sum_{j=i+1}^{T} p_j \left\{ \prod_{k=i+1}^{j-1} \phi_k (1 - p_k) \right\}$$

for the newly marked individuals, and

$$\phi_i \sum_{j=i+1}^{T} p_j \left\{ \prod_{k=i+1}^{j-1} \phi_k (1 - p_k) \right\}$$

for the previously marked individuals. Pradel *et al*. (1997) described this as a test for transient individuals.

Therefore, test 3.SR(i) is equivalent to a score test of $H_0 : \phi_i^* = \phi_i$. As with test 2.CL(i), there is no clear score test relationship with remaining component test 3.Sm(i), which is in line with the lack of ecological interpretation for this component test (Pradel *et al*., 2005).

Because of independence of the component diagnostic tests at occasion t_i, test statistics can be summed over i, resulting in tests 2.CT, 2.CL, 3.SR and 3.Sm. It is these summed test statistics which are often presented in practice. Component test statistics 2.CT and 2.CL can also be added, which result in test 2, and similarly test statistics 3.SR and 3.Sm can be added to form test 3. A global goodness-of-fit test results from the sum of the four tests; however, generally they are reported individually to diagnose departures from model assumptions. Further description of diagnostic tests for capture–recapture data can be found in McCrea and Morgan (2014),

Table 2. Contingency table for test 3.SR(i)

	Individuals captured after t_i	*Individuals not captured after t_i*
Individuals newly marked and captured at occasion t_i	$\sum_{j=i+1}^{T} m_{\{0\}i,j}$	$v_{\{0\}i}$
Individuals previously marked and captured at occasion t_i	$\sum_{j=i+1}^{T} m_{\{1\}i,j}$	$v_{\{1\}i}$

Table 3. Summary of relationship between diagnostic tests and the equivalent score tests†

Diagnostic test	*Score test*	
	Null hypothesis	*Alternative hypothesis*
2.CT(i)	$\phi_1, \ldots, \phi_{T-1}, p_2, \ldots, \{p_{i+1} = p^*_{i+1}\}, \ldots, p_T$	$\phi_1, \ldots, \phi_{T-1}, p_2, \ldots, p_{i+1}, p^*_{i+1}, \ldots, p_T$
2.CT	$\phi_1, \ldots, \phi_{T-1}, \{p_2 = p^*_2, \ldots, p_T = p^*_T\}$	$\phi_1, \ldots, \phi_{T-1}, p_2, p^*_2, \ldots, p_T, p^*_T$
3.SR(i)	$\phi_1, \ldots, \{\phi_i = \phi^*_i\}, \ldots, \phi_{T-1}, p_2, \ldots, p_T$	$\phi_1, \ldots, \phi_i, \phi^*_i, \ldots, \phi_{T-1}, p_2, \ldots, p_T$
3.SR	$\{\phi_1 = \phi^*_1, \ldots, \phi_{T-1} = \phi^*_{T-1}\}, p_2, \ldots, p_T$	$\phi_1, \phi^*_1, \ldots, \phi_{T-1}, \phi^*_{T-1}, p_2, \ldots, p_T$
2.CL(i) and 2.CL	No equivalent score test	
3.Sm(i) and 3.Sm	No equivalent score test	

†The parameters under the null and alternative hypothesis are provided for the score tests.

chapter 9. Table 3 summarizes the equivalences between diagnostic tests and score tests and presents the parameter structures under the null and alternative hypotheses.

4. Simulation comparison of different model selection procedures

4.1. Model selection strategies

In Section 2 we demonstrated the equivalence of components of two important diagnostic tests to specific score tests, and this relationship motivates us to examine whether the diagnoses of trap dependence and transience can be incorporated in a step-up model-selection approach. We shall compare the performance of two alternative strategies.

(a) The traditional diagnostic tests based on the CJS model are conducted and then the potential model set is determined by the results of these tests. If none of the diagnostic tests are significant, the model set will consist of the CJS model with all combinations of time dependent and constant parameters. If any of the diagnostic tests is significant, then the model set will incorporate potential trap dependence (if test 2 was significant) or transience (if test 3 was significant), or combinations of both if tests 2 and 3 were each significant. Once the model set has been determined, all models in the set are fitted and are compared by using the Akaike information criterion (AIC).

(b) The second strategy is a score test approach which tests for trap dependence and transience during the step-up algorithm that is adopted. The score test approach starts with the simplest model with constant survival and capture parameters and tests for each parameter dependence in turn, including tests for trap-dependent capture probabilities and transience in survival probabilities as well as time dependence in parameters. This is an important difference compared with strategy (a) which assumes time dependence throughout. Starting with a CJS model with constant parameters, a path is followed through the model set by selecting the model with the most significant score test and then fitting that model, which becomes the model under the null hypothesis for the next level of tests. The procedure stops at the stage when all score tests are non-significant.

The simulation study compares the powers of these two strategies and investigates the power of the score test approach to detect trap dependence and transience for a variety of parameter structures.

The simulations that we present here and the applications in the next section have generally

used a level of significance of 0.05 for each of the score tests, although different significance levels are examined in Section 4. 2. As discussed in McCrea and Morgan (2011) there is an issue of multiple testing with step-up approaches; however, within the model set that we consider here the number of models being compared is relatively small and therefore not formally correcting significance levels, e.g. through a Bonferroni correction, is unlikely to cause problems in practice. Further, McCrea and Morgan (2011) suggested the use of step-down tests in conjunction with step-up tests because of the complexity of the model space that they were working in. Again, this is unlikely to be a problem for the models of this paper.

We present illustrative simulation results for diagnostic and score tests; however, we have drawn the same conclusions for a wide range of parameter values, and the power simulation results for the diagnostic tests which we have run as part of our performance comparisons are in line with the results of Pollock *et al*. (1985).

We note that throughout the remainder of the paper we use standard capture–recapture notation; for example a model which includes trap-dependent capture probabilities (as described in Section 3.1) is denoted by p(trap), and ϕ(trans) denotes that the model incorporates transient survival probabilities, as described in Section 3.2. Time dependence in capture and survival is denoted by $p(t)$ and $\phi(t)$ respectively. Interactions of parameter-dependence are denoted by '$*$'.

4.2. Simulation investigating power

We have shown that performing component diagnostic tests 2.CT and 3.SR is equivalent to performing score tests where the model under the null hypothesis is the CJS model, with time-dependent survival and capture probabilities. However, for some data the survival and/or capture probability parameters may not vary with time, resulting in some of the parameters of the null model for these two diagnostic tests being superfluous. We therefore investigate the effect of such superfluous parameters on the power of the tests.

4.2.1. Detecting trap dependence

We simulate data with $R_i = 500$, for $i = 1, \ldots, T = 10$, assuming a constant survival probability, $\phi = 0.6$, and we assume that the capture probability p is constant for individuals that were captured at the previous occasion, and $p^* = p + \beta$ for individuals that were captured at the previous occasion. Therefore, β determines the 'trap effect': $\beta < 0$ indicates trap shyness, whereas $\beta > 0$ indicates trap happiness. We define the structure of the capture–recapture models that we are considering by using a '$\cdot$' to denote a probability which is constant over time and a 't' to denote time-dependent probabilities. We consider the performance of two tests of trap dependence:

(a) a score test of $H_0 : \phi(\cdot), \{p(\cdot) = p^*(\cdot)\}$ *versus* $H_1 : \phi(\cdot), p(\cdot), p^*(\cdot)$ and
(b) the diagnostic test 2.CT, which is equivalent to a score test of $H_0 : \phi(t), \{p(t) = p^*(t)\}$ *versus* $H_1 : \phi(t), p(t), p^*(t)$.

We observe from Fig. 1 that the score test has a much higher power to detect trap happiness than the diagnostic test under these conditions, with β ranging from 0 to 0.1 in increments of 0.01 for values of $p = 0.2, 0.4, 0.6, 0.8$.

We have also looked at the power of the score test $H_0 : \phi(\cdot), \{p(\cdot) = p^*(\cdot)\}$ *versus* $H_0 : \phi(\cdot), p(\cdot), p^*(\cdot)$, when the survival and/or capture probabilities are time dependent, with additive trap happiness β. For each iteration of each simulation run, the time-dependent survival probability was simulated as $\phi_t \sim U(0.5, 0.7)$ and time-dependent $p_t \sim U(0.2, 0.5)$. When constant, $p = 0.2$ and $\phi = 0.7$. The power results in this case are displayed in Fig. 2. We observe that there is an increased type 1 error for the score test of trap dependence (when $\beta = 0$) when there is time-

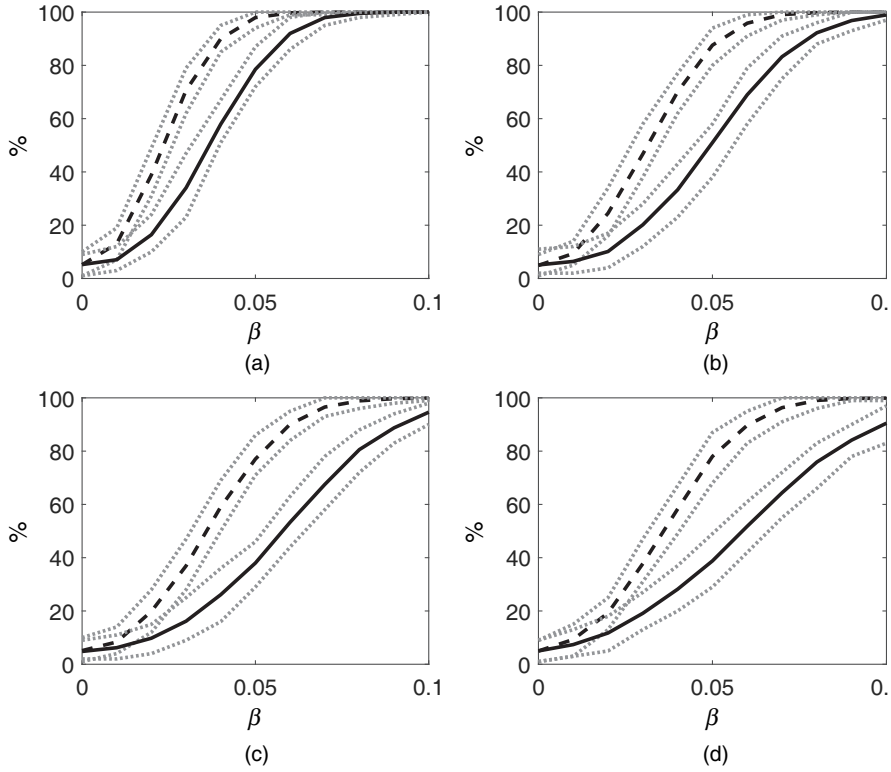

Fig. 1. Percentage of significant score test results (- - - - - -) and diagnostic test 2.CT results (———) from 100 simulation runs, repeated 100 times to provide 2.5% and 97.5% percentiles (· · · · · ·), for values of trap happiness, β, and for various values of capture probability p (the sample size is $R_i = 500$, for all i): (a) p = 0.2; (b) p = 0.4; (c) p = 0.6; (d) p = 0.8

dependent capture probability, because the model under the null hypothesis does not account for the time dependence.

However, in practice, within the step-up strategy the score test of trap dependence is performed at the same time as the score test for time dependence, and the path resulting from the most significant test statistic would be followed. We display boxplots of the p-values resulting from the score tests for time-dependent survival, trap dependence and time-dependent capture probability when $\beta = 0$ in Fig. 2 and we note that the score test for time dependence is more significant than the score test for trap dependence and therefore time dependence will be included first, and a subsequent test for trap dependence at the next step will not have an inflated type 1 error. We note that, if the step-up score test selects time dependence in both capture and survival probabilities, the model under the null hypothesis becomes $H_0 : \phi(t), p(t)$ and the score tests for the next set of tests will be exactly equivalent to the diagnostic tests for trap dependence and transience and so the two model selection strategies coincide.

4.2.2. Pooling the diagnostic test

Diagnostic tests may lose power owing to the assumption of time-dependent parameters under the null hypothesis when that may not be necessary. Therefore we have considered a pooled diagnostic test, which results from pooling the contingency table values for each component test 2.CT(i) with respect to i and performing a single contingency table test. Diagnostic tests are computed in terms of component contingency table tests partitioned by time of previous capture

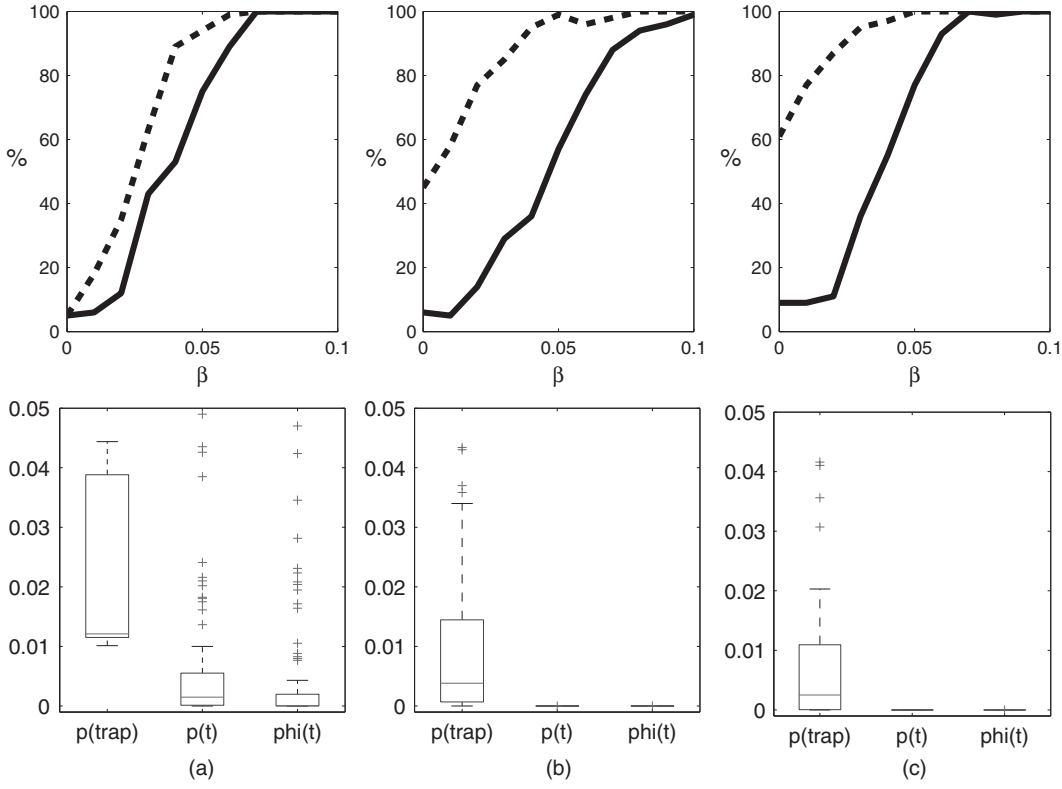

Fig. 2. Percentage of significant score test results (------) and diagnostic test results (———) under models (a) time-dependent survival, (b) time-dependent survival and capture probability and (c) constant survival and time-dependent capture probability, from 100 simulation runs with trap happiness, β (we take $R_i = 500$, for all i): the boxplots show significant p-values (at the 5% level) when $\beta = 0$, for p(trap), $p(t)$ and $\phi(t)$

(for test 2) or occasion of first capture (for test 3). When a stepwise score test approach is carried out, the initial null model assumes no time dependence, and so we constructed a contingency table test which ignored temporal effects. We devised a pooled contingency table test, which adds the cell entries of each of the 2×2 2.CT(i) contingency tables, and then computed a single test statistic from the pooled data. A similar pooled test can be constructed for transience, by pooling the 2×2 3.SR(i) contingency tables.

The power curves for the case of $p = 0.8$, for $-0.1 \leqslant \beta \leqslant 0.1$ are displayed in Fig. 3. We see that the pooled contingency table approach has an intermediate power to detect trap dependence, with an improvement compared with the standard diagnostic tests, but has less power than the score test approach.

4.2.3. Detecting transience

The power of the tests for transience is presented in Fig. 4. We simulate data, with $R_i = 500$, for $i = 1, \ldots, T = 10$, assuming a constant capture probability $p = 0.8$ and constant survival probability $\phi = 0.7$ for individuals that were previously captured and $\phi^* = \phi + \gamma$ for newly captured individuals. Since we assume that transient individuals are less likely to be caught again, we consider values of γ between -0.1 and 0. We observe that the power of the diagnostic test is lower than that of the equivalent score test, and interestingly the power of the pooled diagnostic test is very similar to the power of the score test in this case.

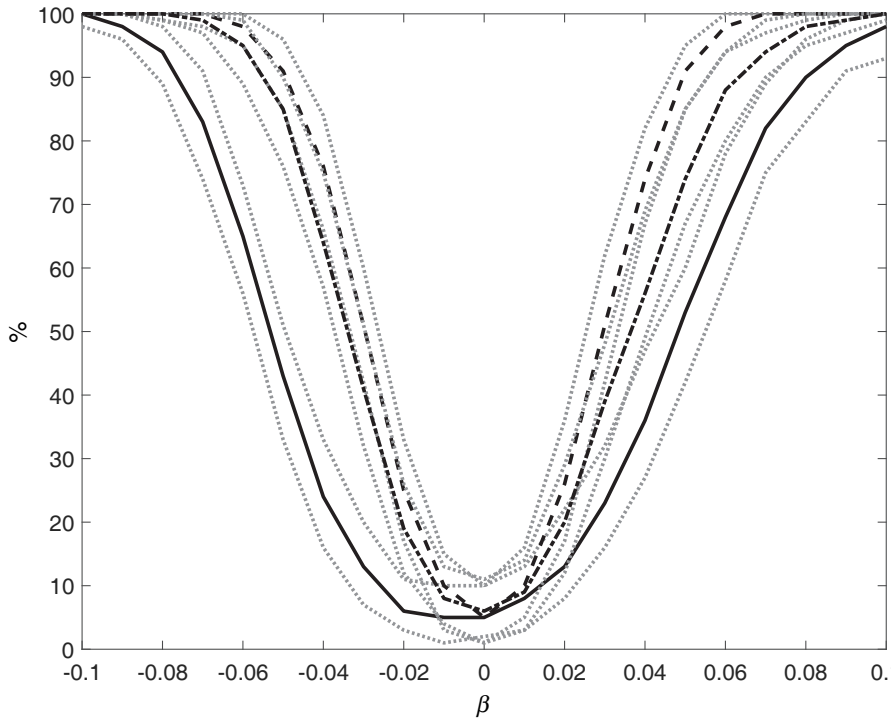

Fig. 3. Percentage of significant score test results (------), diagnostic test results (———) and pooled diagnostic test results (·-·-·-·) from 100 simulation runs with $p = 0.8$, repeated 100 times to provide 2.5% and 97.5% percentiles (·····), for values of trap effect, β (we take $R_i = 500$, for all i)

4.3. Simulation comparing strategies

A simulation study has been run to compare the overall performance of the two alternative model selection approaches for varying sample sizes and levels of significance. Data were simulated from a model with constant capture probability of 0.4; previously marked individuals had a survival probability of 0.7, and new individuals had a marginally higher survival probability of 0.8. The sample size was varied through the values of R_i and varied from 100 to 500. At smaller sample sizes, the power of the diagnostic test was not as good as the score test approach (in line with the earlier power simulations), and in over 50% of cases failed to detect the difference in survival probabilities between new and old individuals (Table 4). Only when the ecologically unrealistic sample size of $R_i = 500$ and the level of significance of 5% were used did the diagnostic test outperform the score test. We see that the 5% level of significance should be reduced as sample size increases considerably.

These simulations, and others that we have run, suggest that current recommendations promoting the use of diagnostic tests to rule out the need for trap dependence or transience within a candidate model set may result in important effects being ignored.

5. Applications

5.1. Cormorants

The results from the stepwise score test approach are displayed in Table 5. We note that the AIC values and likelihood ratio tests have been computed only for comparison. Tests that were conducted within a single level of the model selection procedure are denoted with the same letter

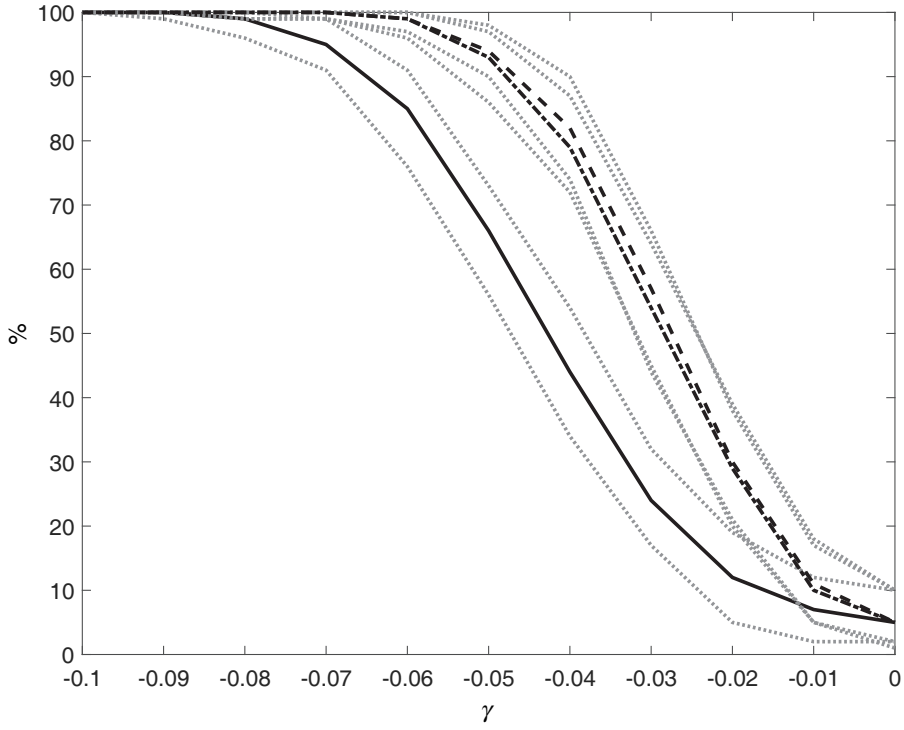

Fig. 4. Percentage of significant score test results (------), diagnostic test results (———) and pooled diagnostic test results (·-·-·-·) from 100 simulation runs with $p = 0.8$ and $\phi = 0.7$, repeated 100 times to provide 2.5% and 97.5% percentiles (· · · · · ·), for values of transience, γ (we take $R_i = 500$, for all i)

(with A representing the first stage of models, B the second stage etc.) and the model under the null hypothesis at each level is denoted with a 0. The procedure selects a model with transience, time-dependent survival probability and trap-dependent capture probability. We note that the p-values for the significant score tests are highly significant and thus the choice of a conservative level of significance is not important.

The diagnostic tests indicate that both trap dependence and transience are significant (Table 6), which is identified by the significance of tests 2.CT and 3.SR respectively. Consideration of the AICs of the models incorporating both trap dependence and transience indicates the optimal model to be $\phi(\text{trans}*t)$, $p(\text{trap})$, agreeing with the score test approach. The score test approach has been more straightforward since only four models have been fitted, compared with nine for the diagnostic approach, and the diagnosis of trap dependence and transience has been conducted within the model selection stage rather than during a preliminary testing step.

5.2. Humpback whales

Performing the diagnostic tests results in non-significant diagnostic tests. In particular test 3.SR results in $p = 0.38$; however, some evidence of transience is provided by a one-sided test of the signed square root of the Pearson X^2-statistics ($p = 0.04$); see Madon *et al.* (2013) for details. Using the standard diagnostic test conclusions, the relevant model set for consideration would require the four models $\phi(\cdot), p(\cdot)$, $\phi(t), p(\cdot)$, $\phi(\cdot), p(t)$ and $\phi(t), p(t)$ to be compared, and the model with the smallest AIC is the simplest model with constant capture and survival probabilities. The AIC values for three of these four models are presented in Table 7 for comparison.

Table 4. Proportions of simulations that result in the correct model, ϕ(trans), $p(\cdot)$, being chosen by the score test method using a 5%, 2% and 1% level of significance and diagnostic tests and AIC model comparison†

Sample size R_i	*Results for the score tests and the following levels of significance:*			*Diagnostic*
	0.01	*0.02*	*0.05*	
100	0.45	0.53	0.57	0.38
200	0.76	0.77	0.73	0.66
500	0.92	0.87	0.76	0.83

†The parameter values are $\phi = 0.6$, $\phi^* = 0.7$, $p = 0.4$ and $T = 10$.

Table 5. Cormorant model selection by using score tests†

Model code	*Model*	*k*	*s*	*p*	$-log(L)$	*AIC*	*LR*
A0	$\phi(\cdot), p(\cdot)$	2			1918.41	3840.82	
A1	$\phi(t), p(\cdot)$	11	33.65	0.0001	1901.40	3824.80	34.02
A2	$\boldsymbol{\phi}$**(trans)**, $\boldsymbol{p}(\cdot)$	3	76.02	2.81×10^{-18}	1879.61	3765.21	77.61
A3	$\phi(\cdot), p(t)$	11	21.99	0.0089	1906.57	3835.14	23.68
A4	$\phi(\cdot)$, p(trap)	3	23.75	1.10×10^{-6}	1906.91	3819.81	23.01
B0	ϕ(trans), $p(\cdot)$	3			1879.61	3765.21	
B1	$\boldsymbol{\phi}$**(trans**$*$**t)**, $\boldsymbol{p}(\cdot)$	20	66.72	7.84×10^{-8}	1845.07	3730.14	69.07
B2	ϕ(trans), $p(t)$	12	27.83	0.0010	1863.92	3751.84	31.38
B3	ϕ(trans), p(trap)	4	28.52	9.26×10^{-8}	1865.32	3738.64	28.58
C0	ϕ(trans$*t$), $p(\cdot)$	20			1845.07	3730.14	
C1	ϕ(trans$*t$), $p(t)$	28	PR		1840.21	3736.42	9.72
C2	$\boldsymbol{\phi}$**(trans**$*$**t)**, $\boldsymbol{p}$**(trap)**	21	28.56	9.07×10^{-8}	1831.27	3704.54	27.60
D0	ϕ(trans$*t$), p(trap)	21			1831.27	3704.54	
D1	ϕ(trans$*t$), p(trap$*t$)	36	11.21	0.7376	1825.61	3723.21	11.33

†The model codes are explained in the text, k denotes the number of parameters in the model, s denotes score test statistics, p is the p-value corresponding to the score test of the model *versus* the null model of that level of test, denoted by 0 in the model code. AIC and likelihood ratio test statistics LR are computed for comparison. $-\log(L)$ denotes the minimized negative log-likelihood value. PR denotes that the model is parameter redundant and hence the score test cannot be computed because of the singularity of the information matrix. The AIC comparisons for this level indicate that model C2 is preferred to model C1. Models selected at each stage of the step-up score test procedure are displayed in bold.

Using a stepwise score test approach the transience is detected at the first stage of model selection ($p = 0.02$) and the model selected has a very simple structure, of transience in survival probabilities and a constant capture probability (Table 7). This model also has the lowest AIC value of all the fitted models. Here it is clear that there is insufficient evidence that the parameters in the model are time dependent and therefore the score test approach has greater power to detect the transience than the diagnostic test approach.

6. Discussion and conclusions

We have demonstrated the equivalence of components of the diagnostic tests to specific score tests, which has motivated an alternative strategy for detecting trap dependence and transience.

Table 6. Cormorant model selection using diagnostic tests followed by AIC model selection†

	df	X^2	*p*	*k*	$-log(L)$	*AIC*
Test						
2.CT	8	31.00	0.00			
2.CL	7	9.63	0.21			
3.SR	9	110.64	0.00			
3.Sm	8	16.78	0.03			
Model						
$\phi(t*\text{trans}), p(\text{trap})$				21	1831.27	3704.54
$\phi(t+\text{trans}), p(\text{trap})$				13	1840.77	3707.54
$\phi(t+\text{trans}), p(t+\text{trap})$				22	1833.62	3711.24
$\phi(t*\text{trans}), p(t+\text{trap})$				29	1828.64	3715.28
$\phi(t+\text{trans}), p(t*\text{trap})$				28	1832.48	3720.96
$\phi(t*\text{trans}), p(t*\text{trap})$				36	1825.61	3723.21
$\phi(\text{trans}), p(t+\text{trap})$				13	1852.13	3730.26
$\phi(\text{trans}), p(\text{trap})$				4	1865.32	3738.64
$\phi(\text{trans}), p(t*\text{trap})$				20	1850.28	3740.56

†$-\log(L)$ denotes the minimized negative log-likelihood value. Models are listed in order of increasing AIC value.

Table 7. Whale model selection using score tests†

Model code	*Model*	*k*	*s*	*p*	$-log(L)$	*AIC*	*LR*
A0	$\phi(\cdot), p(\cdot)$	2			55.86	115.73	
A1	$\phi(t), p(\cdot)$	7	4.78	0.44	53.11	120.23	5.50
A2	$\boldsymbol{\phi(\text{trans}), p(\cdot)}$	3	5.80	0.02	53.25	112.50	5.23
A3	$\phi(\cdot), p(t)$	7	2.22	0.82	54.49	122.97	2.75
A4	$\phi(\cdot), p(\text{trap})$	3	1.59	0.21	55.05	116.10	1.63
B0	$\phi(\text{trans}), p(\cdot)$	3			53.25		
B1	$\phi(\text{trans}*t), p(\cdot)$	12	12.53	0.19	49.18	122.36	8.14
B2	$\phi(\text{trans}), p(t)$	8	3.69	0.59	50.94	117.89	4.61
B3	$\phi(\text{trans}), p(\text{trap})$	4	0.81	0.37	52.82	113.63	0.87

†The model codes are explained in the text, *k* denotes the number of parameters in the model, *s* denotes score tests, *p* is the *p*-value corresponding to the score test of the model *versus* the null model of that level of test, denoted by 0 in the model code. $-\log(L)$ denotes the minimized negative log-likelihood value. AIC and likelihood ratio test statistics LR are computed for comparison. Models selected at any stage of the step-up score test procedure are displayed in bold.

Drawing conclusions from diagnostic tests can be challenging for particular applications. For example, a significant test for trap dependence within a population which is not physically captured may in fact be due to spatial heterogeneity of the survey region; see for example Lahoz-Monfort *et al.* (2011). We note that overdispersion may be calculated based on the significant diagnostic tests and then a modified AIC might be used for model selection. Using our new strategy means that such an initial evaluation is not possible; however, McCrea *et al.* (2011) have presented a general method for assessing absolute goodness of fit following a step-up model selection procedure and appropriate corrections can be made at this stage to the resulting standard errors in the model.

McCrea *et al.* (2014) extended the basic diagnostic tests to diagnostic tests for joint recapture

and recovery data. Similarly there are tests for multistate capture–recapture data as presented in Pradel *et al.* (2003). None of these tests will have a direct equivalence to a score test because the contingency tables are generally larger than 2×2 for the joint recapture and recovery case and contingency table tests for mixtures being used for the multistate case. However, the strategy that is proposed in this paper still holds for these more complex data structures, as the tests for effects on recovery probability, emigration, memory, trap effects and transience can all be included in the basic model set and a step-up approach can be used to explore the large model space. The lack of power of the diagnostic test of memory for multistate capture–recapture data was detected in Cole *et al.* (2014) and the lack of power of diagnostic tests for single-site capture–recapture data has been demonstrated here by using simulation.

The stepwise score test approach has been shown to work well on both simulated and real data sets and may detect important biological traits which diagnostic tests lack the power to identify. Consequently, our recommendation is to incorporate all possible parameter dependences (time, trap dependence, transience and possibly age if known) within a candidate model set and to explore that model set during the model selection procedure. An efficient way to proceed is to use score tests; however, likelihood ratio tests or the AIC could be used as comparative measures, although they would require the fitting of more models.

Acknowledgements

We thank Thomas Bregnballe for providing the cormorant data and Claire Garrigue from Opération Cétacés (`www.operationcetaces.nc`) for providing the whale data. McCrea is funded by Natural Environment Research Council fellowship grant NE/J018473/1 and all three authors were funded by the Royal Society international joint project grant JP090515, 'New statistical methods for wildlife population demography'.

Appendix A.

A.1. Equivalence of Pearson X^2- and score tests

Consider observations from two binomial distributions $\{m_1, m_2\}$ and $\{n_1, n_2\}$ with associated probabilities $\{\pi, 1-\pi\}$ and $\{\pi^*, 1-\pi^*\}$. Suppose that we wish to test the null hypothesis defined by $H_0 : \pi = \pi^*$ against the alternative hypothesis $H_1 : \pi \neq \pi^*$. A contingency table of the observed values is given by Table 8.

Then the expected cell counts can be constructed as Table 9.

The Pearson X^2 goodness-of-fit test can then be computed:

$$\begin{aligned} X^2 &= \frac{\left(m_1 - M\dfrac{m_1+n_1}{M+N}\right)^2}{M\dfrac{m_1+n_1}{M+N}} + \frac{\left(m_2 - M\dfrac{m_2+n_2}{M+N}\right)^2}{M\dfrac{m_2+n_2}{M+N}} + \frac{\left(n_1 - N\dfrac{m_1+n_1}{M+N}\right)^2}{N\dfrac{m_1+n_1}{M+N}} + \frac{\left(n_2 - N\dfrac{m_2+n_2}{M+N}\right)^2}{N\dfrac{m_2+n_2}{M+N}} \\ &= \frac{(m_1+n_2)(m_1N - Mn_1)^2 + (m_1+n_1)(m_2N - Mn_2)^2}{MN(m_1+n_1)(m_2+n_2)} \\ &= \frac{(m_2+n_2)\{m_1(n_1+n_2) - (m_1+m_2)n_1\}^2 + (m_1+n_1)\{m_2(n_1+n_2) - (m_1+m_2)n_2\}^2}{MN(m_1+n_1)(m_2+n_2)} \\ &= \frac{(n_1m_2 - n_2m_1)^2(M+N)}{MN(m_1+n_1)(m_2+n_2)}. \end{aligned} \tag{4}$$

The log-likelihood function is given by

$$l = \text{constant} + m_1 \log(\pi) + m_2 \log(1-\pi) + n_1 \log(\pi^*) + n_2 \log(1-\pi^*).$$

The score test statistic is defined by $S = U' I^{-1} U$ where $U = (\partial l/\partial \pi \;\; \partial l/\partial \pi^*)'$ and I is the Fisher information matrix. Both U and I are evaluated at $\pi = \pi^* = \hat{\pi}$ where $\hat{\pi}$ is the maximum likelihood estimate of π under the

Table 8

m_1	m_2	$m_1+m_2=M$
n_1	n_2	$n_1+n_2=N$
m_1+n_1	m_2+n_2	$M+N$

Table 9

$\frac{(m_1+n_1)M}{M+N}$	$\frac{(m_2+n_2)M}{M+N}$
$\frac{(m_1+n_1)N}{M+N}$	$\frac{(m_2+n_2)N}{M+N}$

null hypothesis $\pi=\pi^*$. In this case, $\hat{\pi}=(m_1+n_1)/(M+N)$. Following calculation of the partial derivatives, and substitution of $\hat{\pi}$,

$$U=\begin{pmatrix} \dfrac{(M+N)(m_1n_2-m_2n_1)}{(m_1+n_1)(m_2+n_2)} & \dfrac{(M+N)(n_1m_2-n_2m_1)}{(m_1+n_1)(m_2+n_2)} \end{pmatrix}.$$

The expected information matrix is given by

$$J=\begin{pmatrix} \dfrac{1}{\pi(1-\pi)} & 0 \\ 0 & \dfrac{1}{\pi^*(1-\pi^*)} \end{pmatrix}$$

and, when substituting $\pi=\pi^*=\hat{\pi}$, we obtain

$$J^{-1}=\begin{pmatrix} \dfrac{(m_1+n_1)(m_2+n_2)}{M(M+N)^2} & 0 \\ 0 & \dfrac{(m_1+n_1)(m_2+n_2)}{N(M+N)^2} \end{pmatrix}.$$

Then the score statistic S is given by

$$\begin{aligned} S &= U'J^{-1}U \\ &= \frac{(M+N)^2(m_1n_2-m_2n_1)^2(m_1+n_1)(m_2+n_2)}{(m_1+n_1)^2(m_2+n_2)^2M(M+N)^2}+\ldots+\frac{(M+N)^2(n_1m_2-n_2m_1)^2(m_1+n_1)(m_2+n_2)}{(m_1+n_1)^2(m_2+n_2)^2N(M+N)^2} \\ &= \frac{N(m_1n_2-m_2n_1)^2+M(n_1m_2-n_2m_1)^2}{MN(m_1+n_1)(m_2+n_2)} \\ &= \frac{(N+M)(n_1m_2-n_2m_1)^2}{MN(m_1+n_1)(m_2+n_2)}. \end{aligned} \tag{5}$$

It is then clear that equations (4) and (5) are the same. Therefore, the 2×2 contingency table X^2-test statistic is exactly the same as the score test statistic. This means that we can present certain diagnostic tests of the paper as appropriately parameterized score tests.

A.2. Using score tests to detect long-term trap effects

Although test 2.CL does not have a direct equivalence to a CJS parameterized score test, it is often intuitively described as a test for long-term trap effect on capture probability. Test 2.CT and the equivalent score test examine differences in capture probability at occasion t_{i+1} between individuals which were captured at occasion t_i and those which were not captured at occasion t_i. However, biologically, the effect of capture may last for more than one sampling occasion, and such effects were considered for closed populations in Cormack (1989).

One possible way of modelling such a trap effect is through the use of a logistic–linear relationship between the capture probability and the length of time since previous capture. To specify such a model, suppose that we define the probability that an individual is captured at occasion t_j, given that it was last captured at occasion t_i, as

$$p_{ij}^{*} = \frac{1}{1+\exp[-\{\alpha+\beta(j-i)\}]}.$$

Under $H_0 : \beta = 0$, the model assumes that the capture probability does not depend on the occasion of last capture; however, under $H_1 : \beta \neq 0$, the model includes either increasing probability with time since last capture (trap shyness) or decreasing probability with time since the last capture (trap happiness). Other models for a long-term trap effect would be possible. The use of score tests for examining the significance of temporal covariates for ring recovery models was considered in Catchpole *et al*. (1999) and the formulation extends to capture–recapture models.

References

Burnham, K. P. (1991) On a unified theory for release-resampling of animal populations. *Proc. 1990 Taipei Symp. Statistics*, pp. 11–36. Taipei: Academia Sinica.

Catchpole, E. A. and Morgan, B. J. T. (1996) Model selection in ring recovery models using score tests. *Biometrics*, **52**, 664–672.

Catchpole, E. A., Morgan, B. J. T., Freeman, S. N. and Peach, W. J. (1999) Modelling the survival of British Lapwings, Vanellus vanellus using ring-recovery data and weather covariates. *Brd Stud.*, **46**, suppl., S5–S13.

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

Cole, D. J., Morgan, B. J. T., Catchpole, E. A. and Hubbard, B. A. (2012) Parameter redundancy in mark-recovery models. *Biometr. J.*, **54**, 507–523.

Cole, D. J., Morgan, B. J. T., Choquet, R., McCrea, R. S., Pradel, R. and Gimenez, O. (2014) Does your (study) species have memory?: Analysing capture-recapture data with memory models. *Ecol. Evoln*, **4**, 2124–2133.

Cormack, R. M. (1989) Log-linear models for capture-recapture. *Biometrics*, **45**, 395–413.

Davison, A. C. (2003) *Statistical Models*. Cambridge: Cambridge University Press.

Hénaux, V., Bregnballe, T. and Lebreton, J.-D. (2007) Dispersal and recruitment during population growth in a colonial bird, the great cormorant *Phalacrocorax carbo sinensis*. *J. Avn Biol.*, **38**, 44–57.

Lahoz-Monfort, J. J., Morgan, B. J. T., Harris, M. P., Wanless, S. and Freeman, S. (2011) A capture-recapture model for exploring multi-species synchrony. *Meth. Ecol. Evoln*, **2**, 116–124.

Lebreton, J.-D., Burnham, K. P., Clobert, J. and Anderson, D. R. (1992) Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecol. Monogr.*, **62**, 67–118.

Madon, B., Garrigue, C., Pradel, R. and Gimenez, O. (2013) Transience in the humpback whale population of New Caledonia and implications for abundance estimation. *Mar. Mamml Sci.*, **29**, 669–678.

McCrea, R. S. and Morgan, B. J. T. (2011) Multi-site mark-recapture model selection using score tests. *Biometrics*, **67**, 234–241.

McCrea, R. S. and Morgan, B. J. T. (2014) *Analysis of Capture-recapture Data*. Boca Raton: Chapman and Hall–CRC.

McCrea, R. S., Morgan, B. J. T. and Pradel, R. (2014) Diagnostic goodness-of-fit tests for joint recapture and recovery models. *J. Agric. Biol. Environ. Statist.*, **19**, 338–356.

Morgan, B. J. T. (2008) *Applied Stochastic Modelling*, 2nd edn. London: Chapman and Hall–CRC.

Pollock, K. H., Hines, J. E. and Nichols, J. D. (1985) Goodness-of-fit tests for open capture-recapture models. *Biometrics*, **41**, 399–410.

Pradel, R. (1993) Flexibility in survival analysis from recapture data: handling trap-dependence. In *Marked Individuals in the Study of Bird Population* (eds J. D. Lebreton and P. M. North), pp. 29–37. Basel: Birkhäuser.

Pradel, R., Gimenez, O. and Lebreton, J.-D. (2005) Principles and interest of GOF tests for multistate capture-recapture models. *Anim. Biodivers. Conservn*, **28**, 189–204.

Pradel, R., Hines, J. E., Lebreton, J.-D. and Nichols, J. D. (1997) Capture-recapture survival models taking account of transients. *Biometrics*, **53**, 60–72.

Pradel, R., Wintrebert, C. M. and Gimenez, O. (2003) A proposal for a goodness-of-fit test to the Arnason-Schwarz multistate capture-recapture model. *Biometrics*, **59**, 43–53.

Smyth, G. K. (2003) Pearson's goodness of fit statistic as a score test statistic. In *Science and Statistics: a Festschrift for Terry Speed* (ed. D. R. Goldstein), pp. 115–126. Institute of Mathematical Statistics.

White, G. C. and Burnham, K. P. (1999) Program MARK: survival estimation from populations of marked animals. *Brd Stud.*, **46**, suppl., S120–S138.