For consideration in *Methods in Ecology and Evolution*

Dealing with false positive and false negative errors about species occurrence at multiple levels

Guillera-Arroita, G. [1], Lahoz-Monfort, J.J. [1], van Rooyen, A.R. [2], Weeks, A.R. [1,2], Tingley, R. [1]

1. School of BioSciences, University of Melbourne, Parkville, 3010, Victoria, Australia

2. cesar Pty Ltd, 293 Royal Pde, Parkville, 3052, Victoria, Australia

Summary

1. Accurate knowledge of species occurrence is fundamental to a wide variety of ecological, evolutionary, and conservation applications. Assessing the presence or absence of species at sites is often complicated by imperfect detection, with different mechanisms potentially contributing to false negative and/or false positive errors at different sampling stages. Ambiguities in the data mean that estimation of relevant parameters might be confounded unless additional information is available to resolve those uncertainties.
2. Here we consider the analysis of species detection data with false positive and false negative errors at multiple levels. We develop and examine a two-stage occupancy-detection model for this purpose. We use profile likelihoods for identifiability analysis and estimation, and study the types of additional data required for reliable estimation. We test the model with simulated data, and then analyse data from environmental DNA (eDNA) surveys of four Australian frog species. In our case study, we consider that false positives may arise due to contamination at the water-sample and qPCR-sample levels, whereas false negatives may arise due to eDNA not being captured in a field sample, or due to the sensitivity of lab tests. We augment our eDNA survey data with data from aural surveys and lab calibration experiments.
3. We demonstrate that the two-stage model with false positive and false negative errors is not identifiable if only survey data prone to false positives are available. At least two sources of extra information are required for reliable estimation (e.g., records from a survey method with unambiguous detections, and a calibration experiment). Alternatively, identifiability can be achieved by setting plausible bounds on false detection rates as prior information in a Bayesian setting. The results of our case study matched our simulations with respect to data requirements, and revealed false positive rates greater than zero for all species.
4. We provide statistical modelling tools to account for uncertainties in species occurrence survey data when false negatives and false positives could occur at multiple sampling stages. Such data are often needed to support management and policy decisions. Dealing with these uncertainties is relevant for traditional survey methods, but also for promising new techniques, such as eDNA sampling.

Keywords (max 10): detectability, environmental DNA, identifiability, imperfect detection, monitoring, multiple stages, sensitivity, species occupancy, specificity

Introduction

Understanding spatial patterns and drivers of species occurrence is fundamental to a wide variety of ecological, evolutionary, and conservation applications. An important consideration in such applications is that detection of species is often imperfect in wildlife surveys. Occurrence data are likely to include false absence records, as a species may remain undetected in surveys conducted at sites where it is, in fact, present. One way to deal with this type of error is to collect survey data that are informative about a species' detectability, for instance, by conducting repeat surveys at a set of sites, or by collecting times to detection (Fig. 1 in Guillera-Arroita 2016). Given this type of data, statistical models can be built (MacKenzie *et al.* 2002; Tyre *et al.* 2003) to estimate separately, and simultaneously, the probability of species presence (hereafter, occupancy), and the probability of detection of a species at sites where it is present (hereafter, detectability). We refer to these models and their extensions as site occupancy-detection models (SODM).

In addition to false absences, wildlife surveys may also suffer from false positive errors. For instance, such errors can be important in aural surveys (McClintock *et al.* 2010a; Miller *et al.* 2012b), and in surveys based on the detection of environmental DNA (eDNA; Darling & Mahon 2011; Ficetola *et al.* 2015; Lahoz-Monfort, Guillera-Arroita & Tingley 2016). There are a number of approaches to simultaneously account for false positive and false negative errors in species occupancy estimation (Chambert, Miller & Nichols 2015). These approaches are based on the mixture model in Royle & Link (2006), which describes detections as being generated by either true detections at occupied sites, or false detections at unoccupied sites. This model yields two sets of parameter values as solutions, which are indistinguishable unless prior knowledge about them is available. This identifiability issue can be resolved by augmenting the analysis with additional data that directly informs one or more of the parameters (Miller *et al.* 2011). One such approach involves the use, at least at some of the sites, of a second survey method that produces unambiguous detection data (i.e., is not prone to false positives, though false negatives may still be present). These additional data inform the model by unequivocally *confirming* the presence of the species at those sites with unambiguous detections. Another approach is to conduct independent calibration experiments, to inform the model directly about rates of true and/or false detection. For instance, this idea has been recently applied to account for potential false positive errors in citizen science data (Ruiz-Gutiérrez, Hooten & Campbell Grant 2016).

In their most basic formulation, SODMs assume that a species is always available for detection during surveys conducted at sites that it uses. In some cases, however, the presence or absence of a species at a site may be somewhat dynamic ("random emigration"), which affects the interpretation of the estimates that a standard SODM yields. In such cases (and assuming here there are no false positives), a standard SODM will estimate different quantities as "occupancy" ($\hat{\psi}$) and "detectability" ($\hat{p}$), depending on the timing of surveys in relation to the movement of the species. For instance, if separate surveys to a site are timed such that the species had the chance to enter or exit the site between visits, then the model will estimate ($\hat{\psi}$, $\hat{p}$) = (ψ, θp), where we can think of θ as the proportion of time that the species is present at a site that it uses (or how often it is "available" for detection at the site). In this case, the occupancy parameter represents the probability that the species *uses* a site, and the detectability parameter captures the combined effect of the species being available for detection (i.e. present) at the site, and our ability to detect it at a site when it is present. At the other extreme, in which repeat surveys are conducted close in time, the model estimates ($\hat{\psi}$, $\hat{p}$) = ($\psi\theta$, p). The estimated occupancy now reflects the combined effect of how likely it is that the species uses the site and how often it is present in it, while the detectability parameter only reflects our ability to detect the species once it is available for detection. For cases between these two extremes, parameter interpretation is less clear.

Multi-stage SODMs have been developed to estimate separately these three components, ψ, θ and p (Nichols *et al.* 2008). To do so, data that allow teasing apart the different components are needed, which requires an extra dimension in terms of replication. For instance, Nichols *et al.* (2008) analysed occurrence of skinks and salamanders at two scales, by applying multiple independent detection methods within each of several survey occasions at each site. Another approach to tease apart ψ, θ and p is to follow a so-called "robust design" (by analogy to mark-recapture studies), in which replication is applied at two time-scales; a number of "primary survey occasions" separated in time are defined and, within each of those occasions, a number of repeat surveys are conducted relatively close in time ("secondary survey occasions").

The concept of availability for detection can be expressed in different ways, depending on the type of survey. For example, in the case of eDNA surveys, availability for detection may be interpreted as whether or not a water sample collected at a site where the species is present contains the species' eDNA. Even if eDNA is

indeed captured in a water sample, it is not guaranteed that it will be detected by PCR (i.e., there is a probability of detection p that is less than one). Where eDNA surveys are conducted following a sampling protocol with multiple PCR runs on multiple water samples, these processes can be separated following the methods of Nichols *et al.* (2008). Here, the water samples can be interpreted as primary survey occasions, and the nested PCR assays as secondary survey occasions, in a robust design. Schmidt *et al.* (2013) followed this approach to study the occurrence of the amphibian chytrid fungus *Batrachochytrium dendrobatidis* in ponds using eDNA. Similarly, Hunter *et al.* (2015) applied this approach to study Burmese python occurrence. In both cases, availability probability in water samples and per-PCR detection probability were less than one.

A key limitation of previous multi-stage analyses is that they have assumed that data are free from false positive errors. Although this might be a reasonable assumption in some circumstances, numerous studies have illustrated that some types of surveys can be prone to false positive errors (e.g., McClintock *et al.* 2010a; Miller *et al.* 2012b; Ruiz-Gutiérrez, Hooten & Campbell Grant 2016). In this paper, we address this limitation, by developing and testing a model for analysing species detection data with false positive and false negative errors at multiple levels. We first provide a probabilistic and hierarchical description of the model. We then study the ambiguities inherent in the survey data, and the types of additional data that can resolve those ambiguities. We illustrate the applicability of the model through a case study, in which we analyse eDNA detection data from a study of four frog species in southern Australia, in combination with detections from aural surveys and data from calibration experiments. We conclude by discussing limitations and potential future extensions of the methods.

Model construction and estimation

In this section, we describe the construction of the model and the approach we take to parameter estimation. To start, we consider only data coming from an ambiguous survey method, i.e. data which may contain false positive errors in addition to false negative errors. Without loss of generality, we consider this to be the main survey method in the study, and hence we refer to its probabilistic description as the "core model". We then explore how the analysis can be augmented with other data sources to resolve structural identifiability issues. In what follows, we assume a two-stage survey protocol (we later discuss its generalization to more than two

stages). In our description, we use eDNA surveys as an example, as the analysis of such data is the motivation for model development.

CORE MODEL

We consider a sampling situation in which sites are sampled on a number of secondary survey occasions, which are nested within a number of primary survey occasions (Fig 1a). In eDNA surveys, this could represent several PCR assays conducted on each of several water samples collected at a site. We can conceptualize the probabilistic description of how such data are generated using a tree diagram, with the parameters characterizing the different stages of the detection process shown in the corresponding branches (Fig. 1b). At the highest level, there is a probability ψ that the species (or its eDNA) occupies/uses the site (and therefore a probability $1-\psi$ that it does not). Where present, there is a probability θ_{11} that the species is available for detection during the primary survey occasion. In our eDNA example, this could represent the probability that eDNA of the species is captured in a water sample. There may also be a chance (θ_{10}) that eDNA ends up in water samples from sites where the species is absent, due to equipment contamination during the sampling or handling process (in the field and/or in the lab). Where the species is available for detection (e.g. eDNA in the water sample), there is a probability p_{11} of indeed detecting it in a secondary survey occasion (e.g. a PCR run detects the species' eDNA). We consider that (false) detection may also happen where the species is not available for detection, with a probability p_{10}. In the eDNA example, this represents cases where the PCR yields a positive result but the water sample did not contain the species' eDNA. In what follows we use the notation above, denoting the probability of presence/detection at the site, primary sampling occasion level, and secondary sampling occasion levels as ψ, θ and p, respectively. We use subscripts to indicate whether the probability of presence/detection is conditional on presence ('11') or absence ('10') at the immediate upper level. For instance, p_{11} is the probability of detecting the species (first '1') given that it was available for detection (second '1').

Assuming independence between sampling sites and between sampling occasions (primary and secondary), we can write the probabilistic description of the data as follows:

$$L = \prod_{i=1}^{S}\left[\psi_i \left\{ \prod_{j=1}^{V_i} \left(\theta_{11} \prod_{l=1}^{L_{ij}} p_{11}^{d_{ijl}} (1-p_{11})^{\bar{d}_{ijl}} + (1-\theta_{11}) \prod_{l=1}^{L_{ij}} p_{10}^{d_{ijl}} (1-p_{10})^{\bar{d}_{ijl}} \right) \right\} \right.$$

$$\left. + \quad (1-\psi_i) \left\{ \prod_{j=1}^{V_i} \left(\theta_{10} \prod_{l=1}^{L_{ij}} p_{11}^{d_{ijl}} (1-p_{11})^{\bar{d}_{ijl}} + (1-\theta_{10}) \prod_{l=1}^{L_{ij}} p_{10}^{d_{ijl}} (1-p_{10})^{\bar{d}_{ijl}} \right) \right\} \right], \quad \text{eqn1}$$

where S is the number of sites surveyed; V_i is the number of primary sampling occasions (e.g. water samples) at site i; L_{ij} is the number of secondary sampling occasions at site i on primary sampling occasion j; and d_{ijl} is the detection/non-detection of the species (a 1 or a 0) at site i on secondary sampling occasion l within primary sampling occasion j. The model parameters ($\psi, \theta_{11}, \theta_{10}, p_{11}, p_{10}$) are as described above and in Fig 1b. We use the bar to denote the complement of an event (i.e. if $d_{ijl} = 1$, then $\bar{d}_{ijl} = 0$). This expression (eqn 1) is the likelihood function for the model. Within the maximum-likelihood framework of inference, we can estimate model parameter values with their uncertainty by identifying which parameter values lead to the largest values of the function. For an example of the likelihood function for a particular site detection history see Appendix S1.

A perhaps more intuitive hierarchical representation of the model is:

```
for i in 1:S {
    z1[i] ~ Bernoulli(ψ)
    for j in 1:V[i] {
        z2[i,j] ~ Bernoulli(z1[i] * θ11 + (1 - z1[i]) * θ10)
        for l in 1:L[i,j] {
            d[i,j,l] ~ Bernoulli(z2[i,j] * p11 + (1 - z2[i,j]) * p10)
        }
    }
}
```

Here, z_1 and z_2 represent latent (imperfectly observed) binary states of the system: z_1 indicates whether the species (or its eDNA) occupies or not the site, and z_2 indicates whether the species is available for detection or not during the primary survey period (e.g. whether eDNA is present in the water sample). The binary variable z_2 can take value 1 through two mutually exclusive paths (Fig 1b), and hence the probability governing its distribution is composed of two additive terms: (1) if the species is present at the site ($z_1 = 1$),

z_2 takes value 1 with probability θ_{11}; (2) alternatively, if the species is absent from the site ($z_1 = 0$), then z_2 takes value 1 with probability θ_{10}. Similarly, the description of how detections (d) are generated involves two mutually exclusive paths depending on the state of z_2. Note that, in the above, data could be summarized by the number of detections per site and primary survey occasion (i.e. $d_{ij} = \sum_l d_{ijl}$) and the third nested loop (for l in 1:$L[i,j]$) removed, replacing the Binomial by a Bernoulli distribution. Hereafter, we summarize the data this way for simplicity.

IDENTIFIABILITY ISSUES AND PARAMETER ESTIMATION

The possibility of false positive errors in a single-stage occupancy-detection model leads to two sets of possible solutions that cannot be distinguished from the data alone (Royle & Link 2006). The likelihood function has the same support at (ψ, p_{11}, p_{10}) and at ($1-\psi$, p_{10}, p_{11}). Imagine that data are generated following the "correct" model structure (i.e. no model violations) with parameters $\psi = 0.3$, $p_{11} = 0.7$ and $p_{10} = 0.1$. Regardless of the amount of data we collect, our analysis will not be able to distinguish whether (ψ, p_{11}, p_{10}) = (0.3, 0.7, 0.1) or (ψ, p_{11}, p_{10}) = (0.7, 0.1, 0.7).

Structural identifiability issues extend in the 2-stage setting. By looking at Fig. 1b, we can see that there are ways in which we can relabel the parameters that yield exactly the same tree diagram. For instance, this is achieved by relabelling ψ as $1-\psi$, and then interchanging θ_{11} and θ_{10}. Two other alternatives are shown in Table 1. Without additional information, there is possible ambiguity about all five model parameters. The symmetries observed in the tree diagram mean that the likelihood function has several peaks of exactly the same height, and therefore that several solutions are a priori equally plausible.

Standard maximum-likelihood estimation methods often obtain estimates by identifying a single peak in the likelihood function via numerical optimization, and derive associated standard errors around that point using asymptotic approximations. This approach is not appropriate when fitting models that may suffer from identifiability issues, as results may overlook that a set of alternative competing solutions exist. A safer option is to look at *profile likelihood* functions, and derive confidence intervals based on the concept of likelihood-ratio tests (Morgan 2008, pp. 89-94). The profile log-likelihood function for a given parameter ϕ, is the value of the log-likelihood function for each value of ϕ, maximized with respect to the other

parameters. Confidence intervals can be computed by identifying the parameter values for which the profile log-likelihood function takes values within a given range from the maximum, that is, slicing the log-likelihood at a given depth (e.g. within 3.84/2 units for a 95% confidence interval of a scalar parameter, where 3.84 is the 5% upper point in a chi-square distribution with one degree of freedom). This approach allows handling situations where confidence intervals may include disjoint sections, as it explores the shape of the likelihood surface away from the maximum. Hence it provides a more appropriate way to analyse data that may suffer from identifiability issues, where the likelihood function can be multimodal. Benefits come at the expense of a greater computational burden, as the method requires a range of function optimizations to construct the profile. However, with current processing capabilities, this is no longer a major limitation.

In this paper, we use profile log-likelihoods for estimation. We plot the functions and compute associated (potentially disjoint) confidence intervals. We use custom-made code written in R (Appendix S2). We construct the profiles in steps of 0.03 (or 0.01 in the frog analyses; see below). At each step, we run the optimization procedure 25 times (30 for the frog analyses) with different starting values (normally distributed around zero, on the logit scale), to ensure that the true maximum in the function is identified. This is important given that multimodalities in the function can result in numerical optimization procedures getting stuck in local maxima.

To illustrate the structural non-identifiability of the core model, here we analyse a simulated data set with S = 250 sampling sites, V = 3 primary samples per site, and L = 3 nested secondary samples. We set $\psi = 0.7$, $\theta_{11} = 0.7$, $\theta_{10} = 0.1$, $p_{11} = 0.3$ and p_{10} = 0.05. Despite the relatively large amount of data, the model cannot identify the true parameter values (Fig. 2). The profile likelihoods for parameters θ_{11} and θ_{10} display four maxima, and symmetry with respect to 0.5. These four peaks are around 0.1, 0.3, 0.7 and 0.9, that is, the four values that are confounded for each of those parameters (θ_{10}, $1 - \theta_{11}$, θ_{11} and $1 - \theta_{10}$; Table 1). The profile likelihoods for parameters p_{11} and p_{10} display two maxima, around 0.05 and 0.3; these peaks reflect the fact that the model cannot distinguish whether each of these parameters take the value of the true p_{11} or that of the true p_{10} (Table 1). The profile likelihood function for ψ is broad and quite flat on top, but in this example there is a hint of two maxima (at ψ and $1 - \psi$).

RESOLVING IDENTIFIABILITY ISSUES: ADDITIONAL INFORMATION

To resolve the structural non-identifiability in the core model, additional information is required. In what follows, we evaluate how different data sources can inform the model to resolve identifiability issues, with a focus on the types of data relevant to our case study. We assess this intuitively, by considering which parameter values are informed by each of the proposed data types, and how this can help distinguish the confounded solutions identified in Table 1. We complement this assessment with the analyses of simulated datasets and inspection of the corresponding profile likelihoods. The conclusions reached are summarised in Table 2.

One way to help resolve identifiability issues induced by false positives is to collect some data with a second survey method that yields unambiguous detections (i.e. a method with no possibility for false positives; Miller *et al.* 2011). This method could be deployed at a subset of the sites. For instance, the unambiguous detections could be collected via trapping, where individuals are handled and the species can be identified unequivocally, or simply by direct observation. Detections for an unambiguous method confirm the presence of the species at some of the sites, and directly inform about the occupancy probability parameter ψ. This information helps rule out two of the four potential solutions in Table 1, as it indicates whether it is more likely that occupancy takes a given value (ψ, solutions 1 & 3), or its complement ($1-\psi$, solutions 2 & 4). Yet, with that information alone, we cannot fully identify the detection parameters: we still cannot distinguish the parameters θ from their complement, nor p_{11} from p_{10}.

Another approach that helps resolve identifiability issues is to run calibration experiments that inform about rates of true and/or false detection. For instance, in the context of our eDNA example, a possible calibration experiment is to run PCR blanks with no template DNA, and then record how frequently PCRs yield a positive result. The data obtained directly informs about the parameter p_{10}, and therefore again help rule out two of the four potential solutions. This time we gain knowledge about which pair of solutions has more support: solutions 1 & 2, or solutions 3 & 4 (Table 1). These calibration data, in combination with data from an unambiguous survey method, resolve the structural identifiability of the core model, i.e. with sufficient data, the true values of all five model parameters can be identified.

A calibration experiment can also be run that encompasses the handling of water samples across stages. Such an experiment could involve collecting blank water samples in the field (e.g., from a site where the species is known to be absent) or filtering and extracting blank water samples in the lab, then recording how many of the PCR assays yield a positive result. Such an experiment does not provide direct information about a single parameter, but instead provides indirect information about θ_{10}, p_{11} and p_{10}. Where the calibration data consist of a single PCR per water sample, they inform about $\theta_{10}p_{11} + (1 - \theta_{10})p_{10}$. This is the combined probability for the two paths that can ultimately lead to a false detection. It can be shown that this probability takes the same value for solutions 1 & 3, and for solutions 2 & 4. Where the calibration experiment is designed with more than one PCR per water sample, the same ambiguity remains, as the calibration data does not allow distinguishing (θ_{10}, p_{11}, p_{10}) from ($1 - \theta_{10}$, p_{10}, p_{11}). Therefore, combining this type of calibration experiment with an unambiguous survey method is not enough to resolve the structural non-identifiability in the core model.

We verified our expectations by augmenting our simulated dataset with records from an unambiguous method (two replicates per site), a calibration experiment at the water sample level (50 blank water samples, with one PCR per water sample) and a calibration experiment at the PCR level (50 PCR blanks). We ran analyses with different combinations of data (seven in total). Corresponding plots are provided in Appendix S3; the construction of the extended model used for analysis is presented next.

EXTENDED MODEL

We extend the likelihood function of the core model to incorporate the three additional data sources described above, leading to a joint analysis of all of the data. The data from the unambiguous survey method are treated as detections from an independent detection method, as in other multi-method analyses (Miller *et al.* 2011). We assume closure in the occupancy status of sites with respect to the unambiguous method, but the model could be extended easily to account for potential random emigration with data collected following a robust design. The calibration data are considered as a completely independent dataset, and hence their contribution to the likelihood is in the form of separate multiplicative terms. The likelihood of this extended model can be written as follows:

$$L = \prod_{i=1}^{S}\Bigg[\quad \psi_i \left\{\prod_{j=1}^{V_i}\left(\theta_{11}p_{11}^{d_{ij}}(1-p_{11})^{L_{ij}-d_{ij}} + (1-\theta_{11})p_{10}^{d_{ij}}(1-p_{10})^{L_{ij}-d_{ij}}\right)\right\}\left\{r_{11}^{\delta_i}(1-r_{11})^{K_i-\delta_i}\right\}$$

$$+ \quad (1-\psi_i)\left\{\prod_{j=1}^{V_i}\left(\theta_{10}p_{11}^{d_{ij}}(1-p_{11})^{L_{ij}-d_{ij}} + (1-\theta_{10})p_{10}^{d_{ij}}(1-p_{10})^{L_{ij}-d_{ij}}\right)\right\}\left\{0^{\delta_i}\right\}\Bigg]$$

$$\times \quad \prod_{j=1}^{V^{[c1]}}\left(\theta_{10}p_{11}^{d_j^{[c1]}}(1-p_{11})^{L_j^{[c1]}-d_j^{[c1]}} + (1-\theta_{10})p_{10}^{d_j^{[c1]}}(1-p_{10})^{L_j^{[c1]}-d_j^{[c1]}}\right)$$

$$\times \quad \left[p_{10}^{d^{[c2]}}(1-p_{10})^{L^{[c2]}-d^{[c2]}}\right], \quad \text{eqn 2}$$

where, in addition to the notation already defined, we use K_i and δ_i for the number of replicates and detections with the unambiguous method at site i; r_{11} for the probability of detecting the species using the unambiguous method; $V^{[c1]}$ for the number of trials in the calibration experiment at the water sample level; $L_j^{[c1]}$ and $d_j^{[c1]}$ for the number of PCRs and detections in blank water sample j; and $L^{[c2]}$ and $d^{[c2]}$ for the number of PCR blanks and detections in the calibration experiment at the PCR level. For simplicity, here the part of the likelihood corresponding to the core model is expressed in terms of number of detections per site and primary survey occasion (i.e., $d_{ij} = \sum_l d_{ijl}$); similarly, the contribution of the unambiguous method is in terms of number of detections per site (δ_i). Note that the unambiguous method does not need to be conducted necessarily at all of the sites (i.e., K_i can be zero in some sites).

The hierarchical description of the same model is:

for i in $1{:}S$ {

$z_1[i] \sim$ Bernoulli(ψ)

for j in $1{:}V[i]$ {

$z_2[i,j] \sim$ Bernoulli$(z_1[i] * \theta_{11} + (1 - z_1[i]) * \theta_{10})$

$d[i,j] \sim$ Binomial$(L[i,j],\ z_2[i,j] * p_{11} + (1 - z_2[i,j]) * p_{10})$

}

$\delta[i] \sim$ Binomial$(K[i],\ z_1[i] * r_{11})$ # unambiguous method

}

for j in $1{:}V^{[c1]}$ { # calibration experiment 1

$z_2^{[c1]}[j]$ ~Bernoulli(θ_{10})

$d^{[c1]}[j]$~Binomial$(L^{[c1]}[j],\ z_2^{[c1]}[j] * p_{11} + \left(1 - z_2^{[c1]}[j]\right) * p_{10})$

}

$d^{[c2]}$~Binomial$(L^{[c2]}, p_{10})$ # calibration experiment 2

This is the model that we use for our case study.

In addition to the above, there are other ways in which information can be included in the model to resolve identifiability issues. For instance, calibration experiments confirming detection at sites where the species is present are possible. Furthermore, where available, prior knowledge about parameter values can be formalized in the form of a prior distribution; analyses can then be conducted within the Bayesian modelling framework to make use of that prior information.

Once parameter values are estimated, one can compute the probability of species presence at each of the surveyed sites, conditional on the observed data (equivalent to eqn 1 in Lahoz-Monfort et al. 2016). This quantity is computed as Pr(species presence & observed data)/Pr(observed data). We provide code for this purpose, as calculations are somewhat cumbersome (essentially, they involve the first two lines in the likelihood function in our eqn 2).

Case study: using eDNA to survey frogs

We applied the multi-stage model to a case study of concurrent eDNA sampling and nocturnal aural surveys for four frog species (Striped Marsh Frog *Limnodynastes peronii*; Spotted Marsh Frog *L. tasmaniensis*, Southern Brown Tree Frog *Litoria ewingii*; Common Froglet *Crinia signifera*) at seven ephemeral roadside drains in suburban Melbourne, Victoria, Australia (see Tingley *et al.* 2015 for a full description of the study area). Sites were surveyed over a five-month period (August–December 2013). Two visits were generally conducted each month, at each site (n = 67 surveys in total, about 10 per site). On each visit, the same two observers conducted an aural survey of the site, by listening for calls during three 3-minute time intervals at the beginning, middle, and end of each site (six of the seven sites were 135 m in length). A species was

classified as detected during a survey if it was heard calling at least once. Environmental DNA sampling occurred earlier on the same day that aural surveys were conducted. On each visit, three 500-mL water samples were taken from approximately the same locations where the aural surveys were conducted (n = 182 water samples in total, about 30 per site). The only exception was one site, which was considerably smaller than the others; only a single water sample was taken and only a single aural survey was conducted at that site, on each visit. We ran four quantitative PCRs (qPCRs) on each water sample (see Appendix S4 for details of eDNA analysis). Field sampling protocols are described in Tingley *et al.* (2015) and Smart *et al.* (2015).

We treated aural surveys as an unambiguous detection method in our analyses. Some aural surveys can be prone to false positive detections (Miller *et al.* 2012b), but this was extremely unlikely in our case study. Only four species were present in the study area, as confirmed by simultaneous trapping and visual surveys (Smart *et al.* 2015; Tingley *et al.* 2015); each species has a unique call; and the same two observers conducted all surveys. We assumed our system to be closed to changes in occupancy over the sampling period, and treated all water samples and call surveys as survey replicates without temporal considerations.

To estimate the rate of false positives potentially introduced during lab operations, we ran both water sample and qPCR blanks. In the case of the former, 50 samples of distilled water were filtered and extracted in the same manner as field samples. Four qPCRs were run on each blank sample, for each of the four species (n = 800 qPCRs). For both *C. signifera* and *L. ewingii*, 1/50 water blanks amplified the species DNA (four and two positive qPCRs, respectively). For *L. peronii* and *L. tasmaniensis*, 4/50 and 2/50 water blanks led to detection (six and three qPCRs, respectively). These records were used as calibration data in the model, indirectly informing about θ_{10}, p_{11} and p_{10}. Our case study assumes that water samples were not contaminated with a species' eDNA in the field.

To control for potential qPCR contamination, we ran 48 qPCR blanks for each of the four species. None of those qPCRs tested positive for the target species' DNA. This rate (0/48) was used as calibration data for p_{10} in the model.

The results of our case study perfectly matched those of our simulations (Table 3 and Appendix S5). Parameters were unidentifiable when only the uncertain survey data were used (row 1 in Appendix S5 figures). Identifiability issues remained when models only included the unambiguous survey method, calibration data from water sample blanks, or calibration data for p_{10} (rows 2-4). Similarly, there was equivalent support for multiple parameter values when only the unambiguous survey method and calibration data from water sample blanks were included in the model (row 5). Identifiability was not an issue, however, when models included the unambiguous survey method in combination with both types of calibration data (row 8; Table 3). Combining the unambiguous survey method with calibration data for p_{10}, or combining both sources of calibration data, also solved identifiability problems in the core model (rows 6-7).

Final estimates of false positive detection probabilities for each of the four frog species were obtained by incorporating all available data (i.e., unambiguous survey data, and calibration data at the water sample and qPCR level). This analysis revealed that false positive detection probabilities were greater than zero at both levels, for all four frog species (Table 3).

Discussion

We have developed a novel extension of the site occupancy-detection modelling framework (Guillera-Arroita 2016) that can be used to estimate species occupancy probabilities that account for potential false positive and false negative errors originating at two different sampling stages. Single-stage occupancy models that account for false positive errors (Miller *et al.* 2011), and multi-stage occupancy models that assume no false positives (Nichols *et al.* 2008), are special cases of the model presented here (Fig. 3).

We have shown that multiple sources of uncertainty in the data lead to ambiguities in estimation where only simple data sets are available. To avoid parameters being confounded, replication at multiple levels is required, as is the collection of additional data that help tease apart alternative solutions that arise due to the symmetries in the likelihood function. We have shown how, in the two-stage scenario, at least two additional data sources are required to resolve the *structural* non-identifiability in the core model. For instance, this can be achieved by including data from an unambiguous detection method and data from a calibration experiment conducted at the lowest sampling level. In contrast, in the single-stage scenario, one source of additional data is enough

for the model to be identifiable. Obviously, the degree to which parameter values can be well estimated in practice depends on the amount of data available. *Practical* non-identifiability may remain where survey data are of limited quantity or quality. In the absence of system-specific data, researchers can also address identifiability issues by setting plausible bounds on false positive rates using prior information (e.g., from the literature) in a Bayesian setting.

Care must be exercised when attempting estimation from models that may suffer from non-identifiability issues. In this study, we showcase the application of profile likelihood confidence intervals as a safe method for estimation in such situations. Standard maximum-likelihood approaches run the risk of overlooking relevant parts of the parameter space, leading to potentially biased estimation and misreporting of associated uncertainty. There is evidence favouring the use of profile likelihood confidence intervals over those based on point estimates in the context of other ecological modelling applications (e.g., Morgan & Freeman 1989; Cormack 1992). We advocate that this approach be used more frequently, given that nowadays computation time is normally not a limiting factor. Profile likelihood confidence intervals are not only useful when the likelihood function has multiple maxima; another more general benefit is that they can capture asymmetries in the likelihood function that are misrepresented by standard symmetric confidence intervals. Where standard maximum-likelihood estimation approaches are used, analyses should be run multiple times using different initial values for the numerical optimization algorithm that searches for the maximum of the function. Assessing whether different runs lead to different estimated parameter values with similar associated values of the likelihood function provides a sound diagnosis of whether the likelihood function is indeed multimodal, and whether other estimation methods would be more appropriate. We have conducted our analyses within the maximum-likelihood framework of inference, but alternatively one could follow a Bayesian approach. Implementation of the model in programs such as BUGS or JAGS (Plummer 2003; Lunn *et al.* 2009) is in principle straightforward, but applying these general-purpose tools can potentially lead to the same inference problems discussed above (Appendix S6). Where the likelihood is strongly multimodal, the MCMC algorithm may get stuck exploring one of the peaks; results may suggest convergence, even when the algorithm has missed important parts of the posterior distribution. Running several chains with different initial values helps diagnose such situations.

Our case study of frog eDNA surveys illustrates how the model developed here can be used to account for both false negative and false positive errors in eDNA studies. Several eDNA-based occupancy studies have accounted for false negative errors (Schmidt *et al.* 2013; Hunter *et al.* 2015); in contrast, false positive errors have received far less attention (c.f. Ficetola *et al.* 2015; Lahoz-Monfort, Guillera-Arroita & Tingley 2016), despite early recognition that false positives pose a significant threat to the credibility of the technique (Darling & Mahon 2011). False positive errors in eDNA studies can originate from a variety of sources. Our case study accounted for potential eDNA contamination of water samples due to contamination during filtering and DNA extraction in the lab, as well as for potential false positives due to contamination at the qPCR level. To deal with these errors, we complemented our eDNA survey data with field data collected via an unambiguous detection method and information about lab contamination rates (from calibration experiments). An alternative approach to account for false positives in eDNA studies would be to collect calibration data at sites where species are, and are not, known to occur (Chambert, Miller & Nichols 2015). It is important, however, that the sources of additional data considered indeed inform about the same processes defined by the parameters to be estimated. For instance, a calibration experiment at the water sample-level conducted in the lab cannot inform about the rates of water sample contamination in the field. In practice, some sources of potential eDNA false positives that originate in the field will be particularly difficult to account for. Calibrating the probability of a non-target species transferring the DNA of a target species to an unoccupied site, for example, will likely be difficult (if not impossible) in practice. In contrast, calibration experiments in the lab are relatively easy to carry out, and should become standard practice in eDNA studies. Indeed, it was such experiments that led us to develop a refined protocol for water filtration and DNA extraction that is less prone to false positives (use of disposable syringes and fully-encased filter units).

The potential range of applications of the multi-stage model extend beyond our specific eDNA case study, and include other types of surveys and study objectives. The multi-stage model could, for instance, be applied to estimate disease prevalence, while accounting for test sensitivity and specificity, as well as for the risk of sample contamination and uncertainty associated with the sampling of individuals. Occupancy-detection modelling has been previously advocated as a useful framework for dealing with uncertainty in disease ecology (McClintock *et al.* 2010b), where it has been applied to study malaria prevalence in wild birds (Lachish *et al.* 2012) and chytridiomycosis prevalence in amphibians (Miller *et al.* 2012a), under the assumption of no false

positive errors. This assumption, however, might not hold, depending on the type of survey, test, and pathogen. Indeed, research incorporating false positive errors has been identified as an important theme for disease ecology (McClintock *et al.* 2010b). Our model provides a useful step in that direction.

The ideas presented here could also prove particularly useful for aural surveys of bird species. Temporary random emigration is a common feature of bird surveys, and some researchers have followed a "robust design" to account for it (Nichols *et al.* 2008). Surveys based on aural detections are, however, also prone to false positives, even when conducted by trained surveyors (Miller *et al.* 2012b). The work in this paper thus opens a promising avenue for the analysis of this type of data. For this setting, the structure would need some adjustment. For instance, it would make sense to fix $\theta_{10} = 0$. In our eDNA example, this parameter reflects the probability that a water sample becomes contaminated, that is, the chance that the "species" is present during a primary sampling occasion at a site where the species does not really occur. In contrast, in a bird survey, the species cannot be available for detection at a site where it is absent. Misidentification of other species are possible, and those events could be captured with parameter p_{10}. This simplification of the model structure implies that there are fewer ambiguities in the data. In fact, we believe that in this case there would not be *structural* identifiability issues, even in the absence of additional data; yet, we expect that *practical* identifiability issues are likely to remain, and that robust estimation will indeed still require some additional information. Investigating and applying this modelling approach for bird survey data is an interesting area of future research.

The work presented here could be extended in other ways. Our code could be used to simulate data and investigate trade-offs in survey design. It is currently unclear, for instance, how best to distribute a fixed survey budget between an ambiguous method, and an unambiguous but costlier or less sensitive survey method. Little guidance is available, and previous recommendations do not reflect the role that unambiguous detections play in resolving identifiability issues (Clement 2016). In addition, with trivial adjustments, our code for estimation can be applied to studies where other arrangements of additional data sources are used to augment ambiguous survey records; for instance, where more than two types of unambiguous detection methods are used, or where data can be classified in different detection states based on their certainty (Miller *et al.* 2011). Future applications of the model could also usefully explore the consequences of disregarding the two-stage model

structure presented here, as well as the value of learning about false positive rates in different contexts. In some cases, incorporating effects of covariates on model parameters may provide useful insight into the drivers of false positive and false negative errors, as well as about the drivers of species occurrence.

Finally, we have focused on studies where sampling is arranged at two levels, but the same principles could be applied to modelling survey data that are collected at three or more levels. Incorporation of additional sampling levels (with their corresponding potential for errors) will increase identifiability issues in the model; therefore, more sources of additional data, informative at different levels, are likely to be required to resolve estimation problems in such cases. Our work showcases how to approach such modelling challenges, in terms of (i) studying and visualising identifiability issues in the model (e.g. by detecting symmetries in the likelihood through tree diagrams and/or plotting of profile likelihoods), and consequently (ii) assessing what forms of additional data can help resolve structural uncertainties. As models become more sophisticated, more care and data are required to make sure that estimation is indeed meaningful. Importantly, these modelling developments should not overshadow the importance of careful survey design and reliable data collection protocols. It is advisable to reduce potential errors during data collection (e.g., by training surveyors and keeping a high standard in laboratory procedures), rather than simply dealing with them later via modelling.

Acknowledgements

We thank Adam Smart for his contribution to the case study. Museum Victoria kindly provided tissue samples that were used in the case study. G. Guillera-Arroita is the recipient of a Discovery Early Career Research Award from the Australian Research Council (DE160100904). R. Tingley and A.R. Weeks are supported by the Australian Research Council Linkage Scheme (LP140100731). J. Lahoz-Monfort receives support from an Australian Research Council Discovery Project (DP160101003).

Author Contributions statement

GGA and JJLM conceived the idea and conducted the statistical developments; GGA wrote the code, ran simulations and led the writing of the manuscript; JJLM and RT contributed critically to the drafts; RT

collected and analysed the survey data; ARW and AvR conducted eDNA extraction and analyses; All authors gave final approval for publication.

Supporting information

Appendix S1: Example of likelihood construction with annotated explanations

Appendix S2: Data, R code for parameter estimation (based on profile likelihoods), and guided examples

Appendix S3: Additional plots (simulated data analysis)

Appendix S4: eDNA analyses

Appendix S5: Additional plots (survey data analysis)

Appendix S6: JAGs code and issues with Bayesian analyses

Data accessibility

All data used in this manuscript are provided as Supporting information.

References

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

Clement, M.J. (2016) Designing occupancy studies when false-positive detections occur. *Methods in Ecology and Evolution***,** doi:10.1111/2041-1210X.12617.

Cormack, R.M. (1992) Interval Estimation for Mark-Recapture Studies of Closed Populations. *Biometrics,* **48,** 567-576.

Darling, J.A. & Mahon, A.R. (2011) From molecules to management: Adopting DNA-based methods for monitoring biological invasions in aquatic environments. *Environmental Research,* **111,** 978-988.

Ficetola, G.F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., Gielly, L., Lopes, C.M., Boyer, F., Pompanon, F., Rayé, G. & Taberlet, P. (2015) Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources,* **15,** 543-556.

Guillera-Arroita, G. (2016) Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography*, **TBD,** TBD.

Hunter, M.E., Oyler-McCance, S.J., Dorazio, R.M., Fike, J.A., Smith, B.J., Hunter, C.T., Reed, R.N. & Hart, K.M. (2015) Environmental DNA (eDNA) Sampling Improves Occurrence and Detection Estimates of Invasive Burmese Pythons. *Plos One,* **10**.

Lachish, S., Gopalaswamy, A.M., Knowles, S.C.L. & Sheldon, B.C. (2012) Site-occupancy modelling as a novel framework for assessing test sensitivity and estimating wildlife disease prevalence from imperfect diagnostic tests. *Methods in Ecology and Evolution,* **3,** 339–348.

Lahoz-Monfort, J.J., Guillera-Arroita, G. & Tingley, R. (2016) Statistical approaches to account for false-positive errors in environmental DNA samples. *Molecular Ecology Resources,* **16,** 673–685.

Lunn, D., Spiegelhalter, D., Thomas, A. & Best, N. (2009) The BUGS project: Evolution, critique and future directions. *Statistics in Medicine,* **28,** 3049-3067.

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

McClintock, B.T., Bailey, L.L., Pollock, K.H. & Simons, T.R. (2010a) Experimental investigation of observation error in anuran call surveys. *Journal of Wildlife Management,* **74,** 1882-1893.

McClintock, B.T., Nichols, J.D., Bailey, L.L., MacKenzie, D.I., Kendall, W.L. & Franklin, A.B. (2010b) Seeking a second opinion: uncertainty in disease ecology. *Ecology Letters,* **13,** 659-674.

Miller, D.A., Nichols, J.D., McClintock, B.T., Grant, E.H.C., Bailey, L.L. & Weir, L.A. (2011) Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology,* **92,** 1422-1428.

Miller, D.A.W., Talley, B.L., Lips, K.R. & Grant, E.H. (2012a) Estimating patterns and drivers of infection prevalence and intensity when detection is imperfect and sampling error occurs. *Methods in Ecology and Evolution,* **3,** 850-859.

Miller, D.A.W., Weir, L.A., McClintock, B.T., Grant, E.H.C., Bailey, L.L. & Simons, T.R. (2012b) Experimental investigation of false positive errors in auditory species occurrence surveys. *Ecological Applications,* **22,** 1665-1674.

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

570 Morgan, B.J.T. & Freeman, S.N. (1989) A Model with 1st-Year Variation for Ring-Recovery Data.
571 *Biometrics*, **45,** 1087-1101.
572 Nichols, J.D., Bailey, L.L., O'Connell Jr., A.F., Talancy, N.W., Campbell Grant, E.H., Gilbert, A.T.,
573 Annand, E.M., Husband, T.P. & Hines, J.E. (2008) Multi-scale occupancy estimation and modelling
574 using multiple detection methods. *Journal of Applied Ecology*, **45,** 1321–1329.
575 Plummer, M. (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.
576 *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna*.
577 Royle, J.A. & Link, W.A. (2006) Generalised site occupancy models allowing for false positive and false
578 negative errors. *Ecology*, **87,** 835-841.
579 Ruiz-Gutiérrez, V., Hooten, M.B. & Campbell Grant, E.H. (2016) Uncertainty in biological monitoring: a
580 framework for data collection and analysis to account for multiple sources of sampling bias.
581 *Methods in Ecology and Evolution*, **7,** 900-909.
582 Schmidt, B.R., Kéry, M., Ursenbacher, S., Hyman, O.J. & Collins, J.P. (2013) Site occupancy models in the
583 analysis of environmental DNA presence/absence surveys: A case study of an emerging amphibian
584 pathogen. *Methods in Ecology and Evolution*, **4,** 646-653.
585 Smart, A., Tingley, R., Weeks, A., van Rooyen, A. & McCarthy, M. (2015) Environmental DNA sampling is
586 more sensitive than a traditional survey technique for detecting an aquatic invader. *Ecological*
587 *Applications*, **25,** 1944-1952.
588 Tingley, R., Weeks, A.R., Smart, A.S., van Rooyen, A.R., Woolnough, A.P. & McCarthy, M.A. (2015)
589 European newts establish in Australia, marking the arrival of a new amphibian order. *Biological*
590 *Invasions*, **17,** 31-37.
591 Tyre, A.J., Tenhumberg, B., Field, S.A., Niejalke, D., Parris, K. & Possingham, H.P. (2003) Improving
592 precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological*
593 *Applications*, **13,** 1790-1801.

594

595

Figures

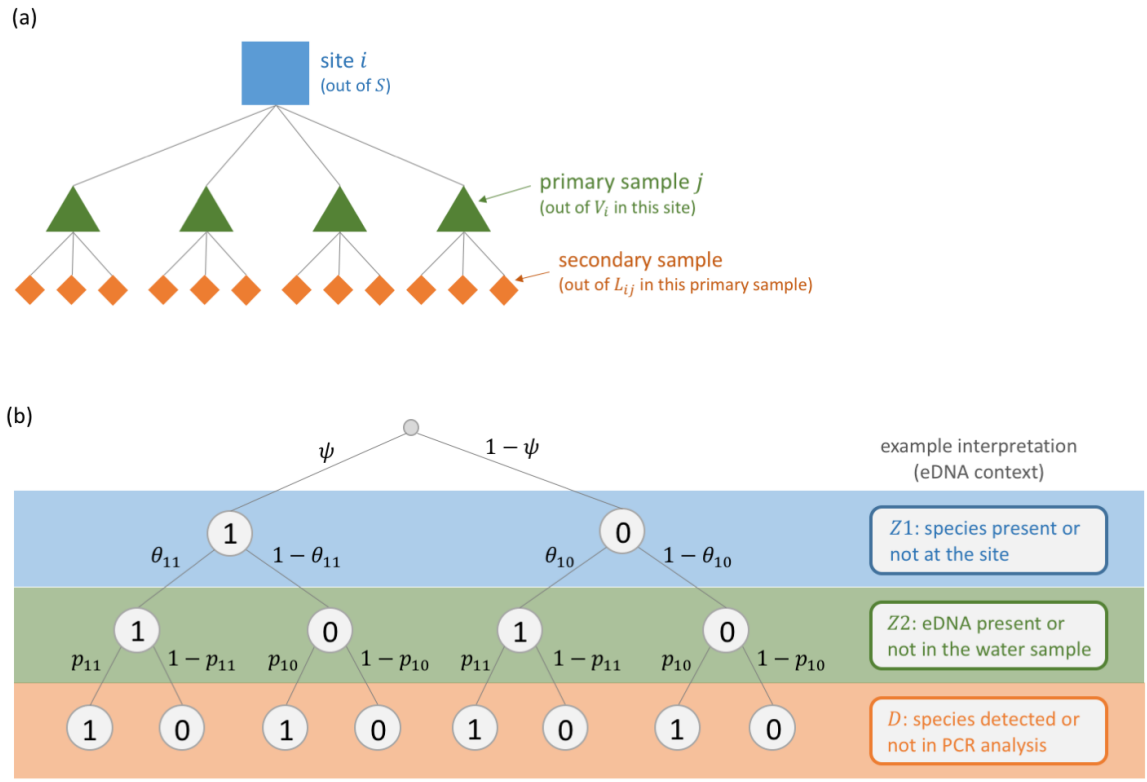

Figure 1 Sampling protocol, latent states and parameters: (a) a sampling situation in which a number of secondary survey occasions (orange diamonds) are embedded within a number of primary survey occasions (green triangles); (b) tree diagram representing data generation, underlying latent states, and associated parameters; the latter are the probabilities of the events indicated by the corresponding branches, e.g. θ_{11} is the probability that eDNA material is present in a water sample that comes from a site where the species was indeed present.

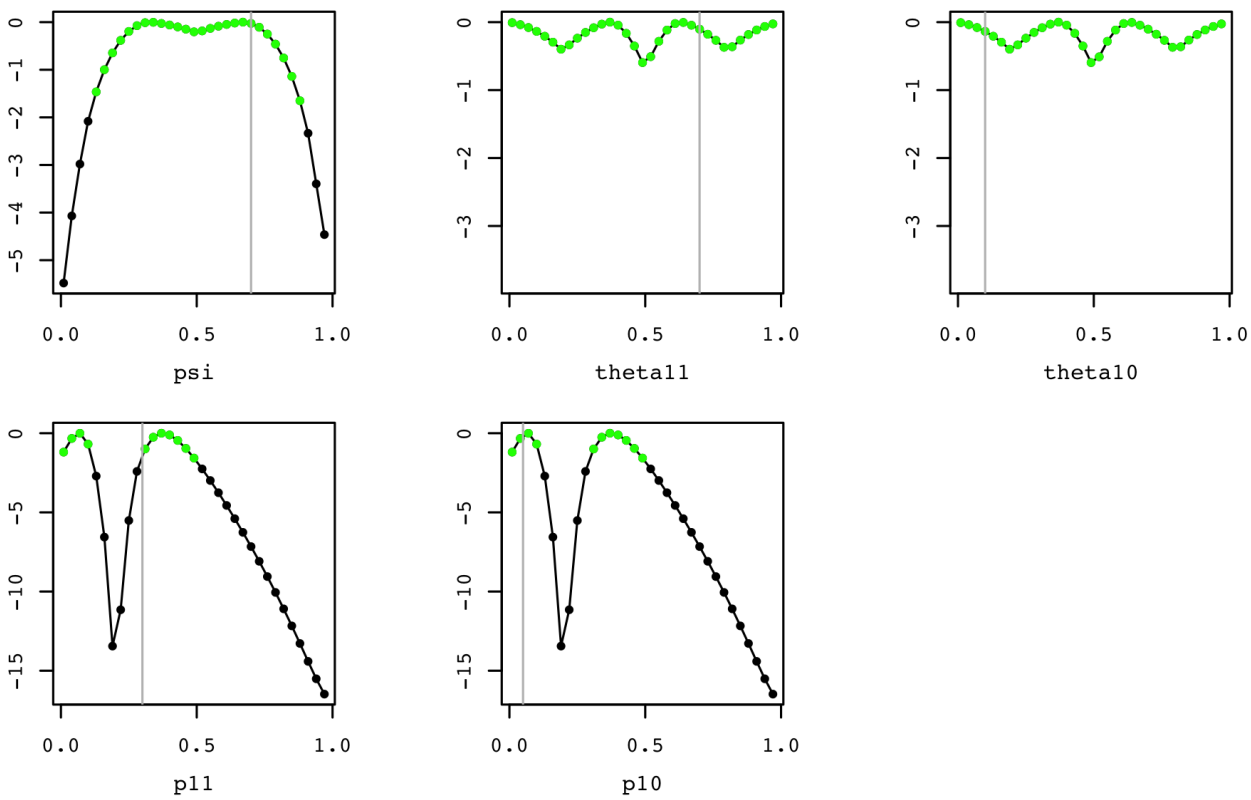

Figure 2 Structural non-identifiability in the core model. The plots show the profile log-likelihood functions for each of the five model parameters (as defined in Fig 1b), for a simulated dataset with 250 sites, 3 water samples per site, and 3 PCR runs per water sample. The true parameters values used in the simulation are indicated by the grey vertical line. The region of parameter values indicated by the green dots correspond to 95% (profile likelihood) confidence intervals.

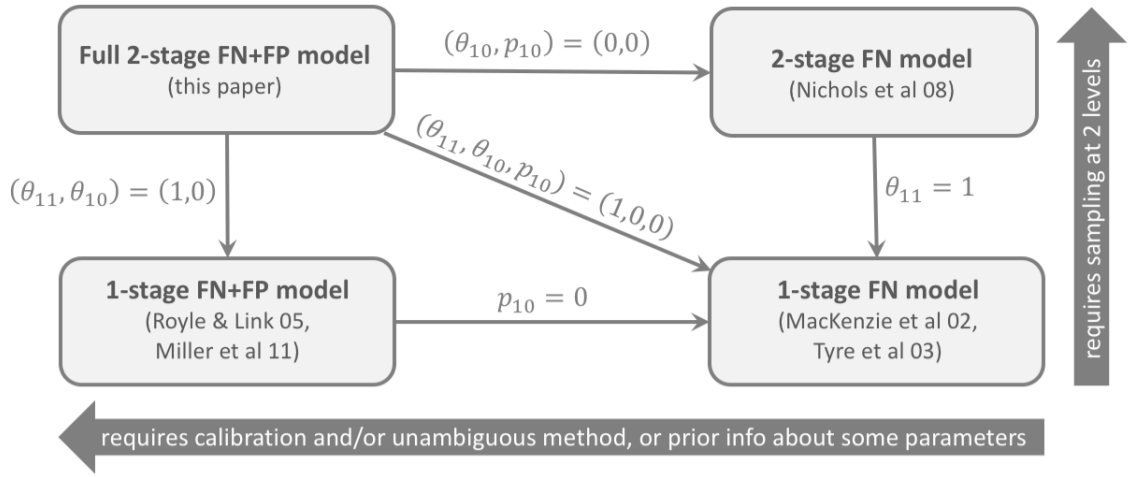

Figure 3 Other models as special cases of the general model described in this paper. Parameter names are as defined in Figure 1.

Tables

Table 1 Solutions that have the same support in the core model.

(1)	ψ	θ_{11}	θ_{10}	p_{11}	p_{10}
(2)	$1-\psi$	θ_{10}	θ_{11}	p_{11}	p_{10}
(3)	ψ	$1-\theta_{11}$	$1-\theta_{10}$	p_{10}	p_{11}
(4)	$1-\psi$	$1-\theta_{10}$	$1-\theta_{11}$	p_{10}	p_{11}

Table 2 Summary of how additional data sources help resolve the structural identifiability problems of the core model.

additional data source	informs about	ambiguity that remains^	parameters confounded*	example plot
U	ψ	(1)(3) or (2)(4)	$\theta_{11}, \theta_{10}, p_{11}, p_{10}$	Fig S3.1
C1	$\theta_{10}p_{11} + (1-\theta_{10})p_{10}$	(1)(3) or (2)(4)	$\theta_{11}, \theta_{10}, p_{11}, p_{10}$	Fig S3.2
C2	p_{10}	(1)(2) or (3)(4)	$\psi, \theta_{11}, \theta_{10}$	Fig S3.3
U+C1	$\psi,\ \theta_{10}p_{11} + (1-\theta_{10})p_{10}$	(1)(3) or (2)(4)	$\theta_{11}, \theta_{10}, p_{11}, p_{10}$	Fig S3.4
U+C2	ψ, p_{10}	none	none	Fig S3.5
C1+C2	$\theta_{10}p_{11} + (1-\theta_{10})p_{10}, p_{10}$	none	none	Fig S3.6

U: unambiguous method, C1: calibration experiment of false positive errors across the two stages (water sample level blanks in the eDNA example), C2: calibration experiment of false positive errors at the lower stage (PCR level blanks in the eDNA example); ^ numbers refer to the solutions in Table 1, e.g. (1)(3) or (2)(4) indicates that, of the four initial solutions supported by the core model, the additional data allows narrowing down the estimation to two of them, either the pair (1)(3) or the pair (2)(4) ; *as per Table 1; Note: the above assumes that for C1, one PCR is run per water sample. Nested designs are also possible (i.e. more than one PCR per water sample), and the implications about identifiability still hold.

Table 3. Model parameter estimates for four frog species at seven sites in Melbourne, Australia.

	Crinia signifera		*Limnodynastes tasmaniensis*		*Litoria ewingii*		*Limnodynastes peronii*	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
ψ	0.86	0.51, 1.00	0.71	0.36, 0.94	0.57	0.23, 0.98	1.00	0.76, 1.00
θ_{11}	0.38	0.31, 0.46	0.60	0.51, 0.68	0.44	0.27, 0.54	0.57	0.50, 0.64
θ_{10}	0.01^	0.01^, 0.06	0.10	0.06, 0.16	0.07	0.01^, 0.12	0.01^	0.01^, 0.04
p_{11}	0.86	0.81, 0.91	0.96	0.94, 0.98	0.88	0.82, 0.93	0.95	0.93, 0.97
p_{10}	0.04	0.03, 0.05	0.05	0.04, 0.06	0.03	0.02, 0.04	0.05	0.03, 0.06
r_{11}	0.44	0.32, 0.56	0.47	0.34, 0.60	0.19	0.06, 0.33	0.54	0.42, 0.65

^ assessment started at 0.01, so these values should be interpreted as ≤ 0.01.