


ORIGINAL RESEARCH

Use of hidden Markov capture–recapture models to estimate abundance in the presence of uncertainty: Application to the estimation of prevalence of hybrids in animal populations

Nina Luisa Santostasi^{1,2}  | Paolo Ciucci¹ | Romolo Caniglia³ | Elena Fabbri³ | Luigi Molinari⁴ | Willy Reggioni⁴ | Olivier Gimenez²

¹Department of Biology and Biotechnologies “Charles Darwin”, University of Rome La Sapienza, Rome, Italy

²CEFE, CNRS, University of Montpellier, University Paul Valéry Montpellier 3, EPHE, IRD, Montpellier, France

³Italian Institute for Environmental Protection and Research (ISPRA), Unit for Conservation Genetics (BIO-CGE), Ozzano dell’Emilia, Bologna, Italy

⁴Wolf Apennine Center, Appennino Tosco-Emiliano National Park, Ligonchio, Italy

Correspondence

Nina Luisa Santostasi, Department of Biology and Biotechnologies “Charles Darwin”, University of Rome La Sapienza, Rome, Italy.
Email: n.santostasi@gmail.com

Funding information

French National Research Agency, Grant/Award Number: ANR-16-CE02-0007.

Abstract

Estimating the relative abundance (prevalence) of different population segments is a key step in addressing fundamental research questions in ecology, evolution, and conservation. The raw percentage of individuals in the sample (naive prevalence) is generally used for this purpose, but it is likely to be subject to two main sources of bias. First, the detectability of individuals is ignored; second, classification errors may occur due to some inherent limits of the diagnostic methods. We developed a hidden Markov (also known as multievent) capture–recapture model to estimate prevalence in free-ranging populations accounting for imperfect detectability and uncertainty in individual's classification. We carried out a simulation study to compare naive and model-based estimates of prevalence and assess the performance of our model under different sampling scenarios. We then illustrate our method with a real-world case study of estimating the prevalence of wolf (*Canis lupus*) and dog (*Canis lupus familiaris*) hybrids in a wolf population in northern Italy. We showed that the prevalence of hybrids could be estimated while accounting for both detectability and classification uncertainty. Model-based prevalence consistently had better performance than naive prevalence in the presence of differential detectability and assignment probability and was unbiased for sampling scenarios with high detectability. We also showed that ignoring detectability and uncertainty in the wolf case study would lead to underestimating the prevalence of hybrids. Our results underline the importance of a model-based approach to obtain unbiased estimates of prevalence of different population segments. Our model can be adapted to any taxa, and it can be used to estimate absolute abundance and prevalence in a variety of cases involving imperfect detection and uncertainty in classification of individuals (e.g., sex ratio, proportion of breeders, and prevalence of infected individuals).

KEYWORDS

anthropogenic introgression, capture–recapture, hidden Markov models, hybridization, multievent models, prevalence, Viterbi algorithm

1 | INTRODUCTION

The relative abundance (prevalence) of different population segments is a fundamental piece of information to understand processes in ecology, evolution, and conservation. For example, the prevalence of infected individuals is critical to understand the mechanisms driving disease dynamics (Jennelle, Cooch, Conroy, & Senar, 2007; Moreno-Torres, Wolfe, Saville, & Garabed, 2016); the prevalence of key demographic categories (e.g., mature females) is needed to assess the viability of endangered populations (Caswell, 2000) and when hybridization represents a threat, the prevalence of admixed individuals is needed to evaluate the appropriate management option (Allendorf, Leary, Spruell, & Wenburg, 2001).

However, estimating prevalence for wildlife populations is challenging and the raw percentage of individuals of a class in the sample (naive prevalence) is often used as a proxy (Jennelle et al., 2007). However, this approach overlooks two main sources of bias. First, imperfect and/or heterogeneous detection leads to biased abundance estimates when it is ignored (Cubaynes et al., 2010; Jennelle et al., 2007). Second, uncertainty in the classification of individuals (e.g., diseased/healthy, breeder/nonbreeder, male/female) is common in wildlife population studies where individuals are assigned to a specific status based on imperfect clues. Examples include determining sex or breeding status based on the behavior of individuals (Genovart, Pradel, & Oro, 2012) or establishing health status from the observation of outer symptoms only (Conn & Cooch, 2009). Another less explored but intriguing situation is assigning individuals to genetic classes (subpopulations) based on a limited number of genetic markers (Vähä & Primmer, 2006).

Reliable estimates of wildlife abundance can be obtained by correcting field counts by the proportion of undetected individuals (i.e., the ratio between the number of observed individuals and the probability of detection; Nichols, 1992). The probability of detection is usually estimated by using capture–recapture models (CR) from a sample of individual encounter histories (Otis, Burnham, White, & Anderson, 1978). In particular, multistate CR models estimate the probability of detection for different classes of individuals by assigning individuals to static or dynamic states. However, multistate CR models assume the correct assignment of all individuals to their state (Lebreton, Burnham, Clobert, & Anderson, 1992). Multievent models relax this assumption by acknowledging the uncertainty of the observation process in the model structure (Pradel, 2005). In these models, a hidden biological process (e.g., survival or dispersal) is modeled as a Markov chain of states (Pradel, 2005). The observation process (the data) arises from the underlying states through the probability of detection. To include uncertainty in state assignment, the observation process is further split into two steps: detection and state assignment conditional on detection (Gimenez, Lebreton, Gaillard, Choquet, & Pradel, 2012; Pradel, 2005). This formulation includes a probability of assignment (besides the probability of detection and the probabilities associated with the biological process), allowing for the inclusion of individuals classified with uncertainty (Pradel, 2005).

Multievent models have been used to estimate a variety of population parameters in the presence of uncertainty in state assignment. Examples of that include the rates of entry and exit from disease states (Conn & Cooch, 2009), the probability of skipping reproduction (Sanz-Aguilar et al., 2011), and the probability of survival of different age classes (Gervasi et al., 2017). However, multievent models have never been used to estimate the abundance of individuals in different states because the numerator of the abundance estimator (the number of observed individuals) is contaminated by uncertain observations.

Here, we develop a capture–recapture approach to estimate the prevalence of admixed individuals (hereafter “hybrids”) in a population while simultaneously accounting for both imperfect detection and classification uncertainty. Specifically, we show how to use the multievent CR framework to estimate abundance of individuals in different states (i.e., “Parental,” “Hybrid,” “Dead”) in the presence of uncertainty in state assignment. We first use multievent models to estimate survival and detection parameters; second, we use the Viterbi algorithm to assign the uncertain observed individuals to the most likely state (Rouan, Gaillard, Guédon, & Pradel, 2009; Zucchini, MacDonald, & Langrock, 2009), and lastly, we estimate prevalence via a Horvitz–Thompson-like estimator combined with a bootstrapping procedure to produce standard error and confidence intervals (Davison & Hinkley, 1997).

We assess the importance of incorporating detectability and uncertainty in state assignment by comparing the performance of model-based and naive prevalence under different scenarios. The accuracy of CR parameters’ estimators depends on the recapture rate and on the number of capture occasions (Otis et al., 1978). Increasing the detectability and/or the number of occasions requires different sampling strategies and generates different costs in terms of financial and human resources (Lieury et al., 2017). Therefore, we also explore how different sampling strategies may affect the model performance.

Despite the increasing attention that researchers are devoting to hybridization cases (Schwenk, Brede, & Streit, 2008; Todesco et al., 2016), there have been only few attempts to estimate prevalence of hybrids in wild populations (Vaz Pinto, Beja, Ferrand, & Godinho, 2015). We illustrate our method with a case study by estimating the prevalence of wolf (*Canis lupus*) × dog (*Canis lupus familiaris*) hybrids in a wolf population in northern Italy. This is a case of anthropogenic hybridization (i.e., the interbreeding of individuals from genetically distinct populations due to human action; Allendorf et al., 2001) and is considered a major threat to wolf genomic integrity (Boitani, 2000). Therefore, accurately estimating the prevalence of hybrids in wolf populations is a priority to elaborate conservation strategies (Hindrickson et al., 2017). We show that in this case using naive prevalence as a proxy underestimates the prevalence of hybrids.

2 | METHODS

2.1 | Hidden Markov model

We assumed that animals are individually recognized at discrete encounter occasions, therefore obtaining an encounter history for each individual. Individuals can be in one of three possible nonobservable

states: alive and parental (P), alive and hybrid (H), or dead (D). We underline that hereafter the term “hybrid” refers to all categories of admixture and not only to first-generation hybrids. Upon its first encounter, an individual has a probability π_p to be a parental and the complementary probability $\pi_h = 1 - \pi_p$ to be a hybrid. The initial state probabilities describe the probability that an individual is in one or another state when first encountered. Then, the states change over time according to a first-order Markov process, with the state process being governed by apparent survival probabilities φ_p and φ_h . More specifically, the state process, which summarizes the underlying biological process, is represented by a transition probability matrix with states at time t in rows (“Parental,” “Hybrid,” “Dead”) and $t + 1$ in columns (“Parental,” “Hybrid,” “Dead”):

$$\begin{matrix} & P & H & D \\ \begin{matrix} P \\ H \\ D \end{matrix} & \begin{pmatrix} \varphi_p & 0 & 1 - \varphi_p \\ 0 & \varphi_h & 1 - \varphi_h \\ 0 & 0 & 1 \end{pmatrix} \end{matrix}$$

where parameter φ_p (resp. φ_h) is the probability that an individual alive and in state “Parental” (resp. “Hybrid”) at time t is still alive in the study area and in state “Parental” (resp. “Hybrid”) at time $t + 1$ and corresponds to the apparent survival probability of parental (resp. “Hybrid”) individuals.

The second time series (or event process) is generated from the states at each occasion and describes the observation process. Individuals are detected at time t with probability of detection p_p for parental and p_h for hybrid individuals. Upon detection, an attempt is made of assigning the individuals to a hybridization state based on genetic and/or morphologic diagnostic features and there is a probability δ that an individual is assigned to the state “Parental” or “Hybrid”. If the diagnostic clues are not sufficient to ascertain the hybridization state, the individual is classified as uncertain with the complementary probability $1 - \delta$. The observation process is summarized by a matrix with states in rows (“Parental,” “Hybrid,” “Dead”) and events in columns (0 = “Not detected,” 1 = “Detected as parental,” 2 = “Detected as hybrid,” 3 = “Detected as uncertain”):

$$\begin{matrix} & 0 & 1 & 2 & 3 \\ \begin{matrix} P \\ H \\ D \end{matrix} & \begin{pmatrix} 1 - p_p & p_p \delta_p & 0 & p_p (1 - \delta_p) \\ 1 - p_h & 0 & p_h \delta_h & p_h (1 - \delta_h) \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{matrix}$$

In the first row, the $(1 - p_p)$ term is the probability that an individual in state “Parental” is not detected, and $p_p \delta_p$ is the probability that an individual in state “Parental” is detected and assigned to the category “Parental” while $p_p (1 - \delta_p)$ is the probability that an individual in state “Parental” is detected (p_p) and not assigned to any category ($1 - \delta_p$). Note that an individual in state “Parental” cannot be detected as a hybrid; hence, the corresponding probability is 0. The second row is similar to the first one, except that it refers to the hybrid individuals. In the third row, all individuals are nondetected because they are

dead. Equivalently, the event process can be decomposed as the product of a detection matrix by an assignment matrix which expresses the probability that an individual is assigned to a state given that it has been detected:

$$\begin{matrix} & 0 & 1 & 2 & 3 \\ \begin{matrix} P \\ H \\ D \end{matrix} & \begin{pmatrix} 1 - p_p & p_p & 0 & 0 \\ 1 - p_h & 0 & p_h & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \delta_p & 0 & 0 & 0 \\ 0 & \delta_p & 0 & 1 - \delta_p \\ 0 & 0 & \delta_h & 1 - \delta_h \end{pmatrix} \end{matrix}$$

when an animal is first encountered, the capture process is not modeled because an animal must be encountered at least once to enter the dataset, but the state ascertainment remains valid.

To illustrate the calculation of an encounter history, let us consider the case of a 3-year CR experiment. For instance, the encounter history “303” denotes an individual encountered at the first and third occasions but not at the second occasion. The state of this individual is never assigned. Assuming parameters are constant, we have

$$\begin{aligned} \Pr(303) &= \pi_h (1 - \delta_h) \varphi_h (1 - p_h) \varphi_h p_h (1 - \delta_h) \\ &\quad + \pi_p (1 - \delta_p) \varphi_p (1 - p_p) \varphi_p p_p (1 - \delta_p) \end{aligned}$$

On the right side of the equation, the first element of the sum is the probability of the observed history if the underlying state is “Hybrid”, while the second element is the probability of the observed history given that the underlying state is “Parental”. The likelihood of the entire dataset is obtained as the product of the probabilities of all individual encounter histories assuming independence. In this paper, estimates of initial states (π_p and π_h), apparent survival (φ_p and φ_h), detection (p_p and p_h), and assignment probabilities (δ_p and δ_h) are obtained by maximizing the likelihood function. Different models can be built by allowing the parameters to vary according to, for example, state, time, or group of individuals classified based on discrete variables (e.g., age class or sex).

2.2 | Abundance and prevalence estimation

Naive prevalence (P_{naive}) at a given occasion is calculated as the proportion of observed hybrids in the sample:

$$P_{\text{naive}} = \frac{n_h}{n_p + n_h}$$

where n_h is the number of observed hybrids and n_p is the number of observed parentals.

Here, we propose a model-based prevalence estimate (P_{model}) as the ratio between hybrids’ abundance in the population and total population abundance. Assuming that marked and unmarked individuals have the same probability of detection, total abundance at a given occasion (\hat{N}_{tot}) is then estimated with the Horvitz-Thompson estimator (McDonald & Amstrup, 2001) as:

$$\hat{N}_{\text{tot}} = \frac{n_p}{\hat{p}_p} + \frac{n_h}{\hat{p}_h} = \hat{N}_p + \hat{N}_h$$

where n_h is the number of observed hybrids, and n_p is the number of observed parental individuals, \hat{p}_h is the recapture probability of hybrids, and \hat{p}_p is the recapture probability of parental individuals. Model-based prevalence is then estimated as:

$$P_{\text{model}} = \frac{\hat{N}_h}{\hat{N}_p + \hat{N}_h}$$

The main difficulty is therefore in determining the number of observed parental individuals and hybrids, because the uncertain individuals have to be assigned to one of the two states ("Hybrid" or "Parental"). To do so, we use the Viterbi algorithm (Zucchini et al., 2009) which, given any observation sequence (here, the encounter histories) and the parameters estimated by the hidden Markov model, finds the most probable underlying sequence of states that has generated the observed data (Rouan et al., 2009). Once the uncertain observations are assigned to the most likely state, the number of observed hybrid and parental individuals can be reconstructed and the formula can be used to estimate prevalence. We obtain confidence intervals for the hybrids' prevalence by using a nonparametric bootstrap procedure (Davison & Hinkley, 1997).

2.3 | Evaluation of model performance and sampling strategy

To test the model performance, we generated encounter histories with known prevalence mimicking our case study on wolves and dogs in the Northern Apennines (see next section) and compared model-based prevalence and naive prevalence. Using the R (R Core Team, 2017) package HMM (Himmelman, 2010), we simulated a cohort of 100 individuals that we split into 2 states "Hybrid" and "Parental". For all the scenarios, we set the initial proportion of wolves $\pi_p = 0.7$ and initial proportion of hybrids $\pi_h = 1 - \pi_p = 0.3$ as in our case study sample (see next section). Using values estimated by Caniglia et al. (2012) for the same wolf population, we considered state-dependent survival (constant over time) with parental survival ($\varphi_p = 0.8$) higher than hybrid survival ($\varphi_h = 0.7$). We then considered three hypothetical scenarios for detectability and assignment probability: (see Supporting information Table S1 for a complete list of parameters for the three scenarios) (1) state-dependent detectability ($p_p > p_h$) and homogeneous assignment probability, (2) homogeneous detectability and state-dependent assignment probability ($\delta_p > \delta_h$) and (3) homogeneous detectability and assignment probability. Within these three main scenarios, we evaluated the effect of lower and higher sampling intensities by comparing sampling strategies with low and high detectability and with 5 and 10 capture occasions. We simulated 100 datasets for each combination of parameters within the three main scenarios. We fitted constant and state-dependent models to the simulated data, and we calculated the relative bias and the root mean squared error (RMSE) for the model-based and the naive prevalence estimators. We

also calculated the confidence interval coverage for the model-based estimator of prevalence.

2.4 | Case study

We collected fresh wolf scats in the Appennino Tosco-emiliano National Park from August 2016 to May 2017. We extracted, amplified and sequenced DNA from the scats following the procedures described in Caniglia, Fabbri, Galaverni, Milanese, and Randi (2014). We identified wolves, dogs and putative hybrids based on the analysis of molecular markers listed in Randi et al. (2014) and using Bayesian genetic clustering procedures implemented in STRUCTURE 2.3.4 (Falush, Stephens, & Pritchard, 2003; Pritchard, Stephens, & Donnelly, 2000). We distinguished wolves, hybrids, and uncertain individuals based on their membership proportions to the wolf cluster (q_{wolf}) and 90% Bayesian credible intervals (BCI). We set the thresholds for the three categories based on the genetic clustering analyses performed on simulated genotypes by Pacheco et al. (2017), see Fig. S1 in the Supplementary Materials of Pacheco et al. (2017). We considered as pure wolves those individuals whose q_{wolf} was included in the range of q_{wolf} values of simulated pure wolves genotypes and did not overlap with that of simulated backcrosses (hybrid \times parental). We classified as uncertain those individuals whose q_{wolf} was included in the range in which the q_{wolf} values of simulated pure wolves and backcrosses to wolves (first and second generation) overlapped. Finally, we considered as hybrids those individuals whose q_{wolf} overlapped with the range of q_{wolf} values of simulated backcrosses (first and second generation) and/or hybrids (first and second generation). We additionally considered as hybrids those individuals which presented a Y haplotype of canine origin or the deletion at the K-locus (Caniglia et al., 2013; Randi et al., 2014) regardless of their q_{wolf} values.

The CR data were pooled in 2-month capture occasions, with a total of five capture occasions. Thirty-nine individuals were sampled (19 wolves, 12 hybrids, 8 uncertain based on their q_{wolf} value). We applied the multievent CR model described above to test for differences in detectability and assignment probability between hybrid and parental individuals and to estimate prevalence of wolf \times dog hybrids in the population. Since the hybridization assessment is performed only once for each genotype, we constrained the assignment probability to be estimated only upon first capture (see the models details in the Supporting information). We fitted models with state-dependent and constant parameters and a combination of the two. The models were fitted in the E-SURGE software (Choquet, Rouan, & Pradel, 2009) and the Viterbi algorithm was implemented using the R (R Core Team, 2017) package HMM (Himmelman, 2010). We used the Akaike Information Criterion (AIC) for model selection, considering models within $\Delta\text{AICc} \leq 2$ as the most supported and used model averaging to account for uncertainty in model selection (Burnham & Anderson, 2002).

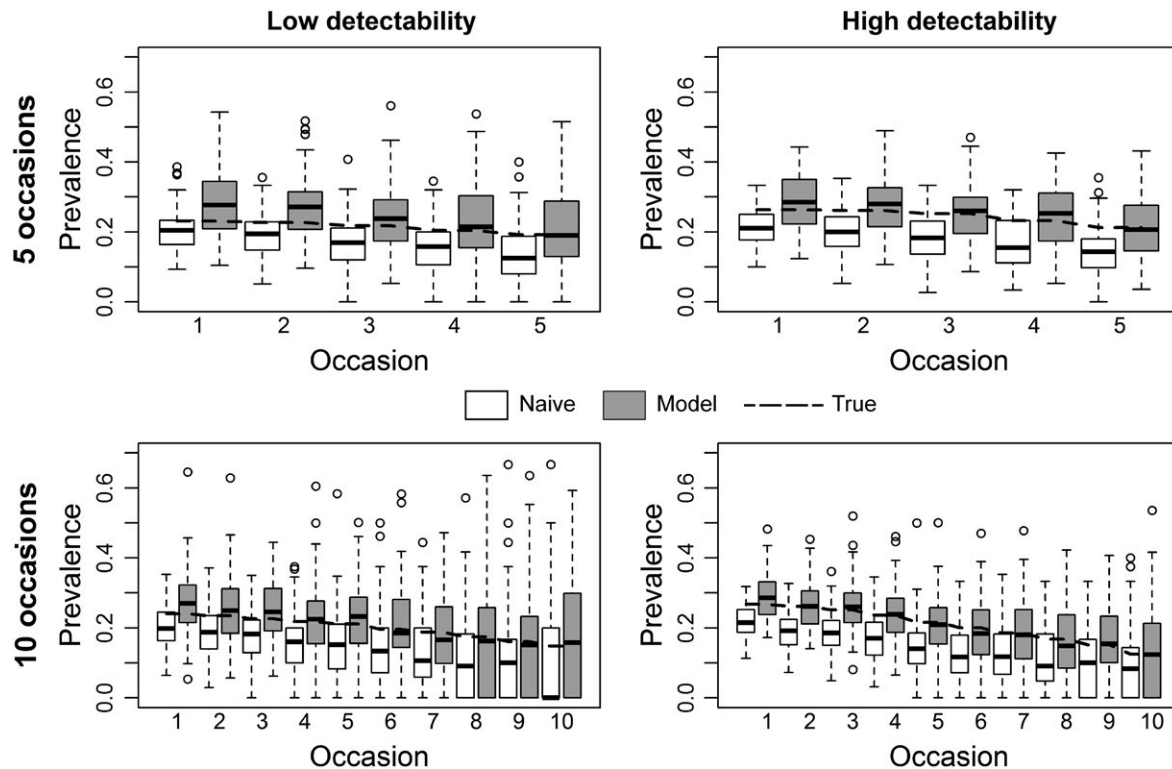


FIGURE 1 Scenario 1 of simulation (state-dependent detectability and homogeneous assignment probability). Sampling strategies with 5 (upper panels) versus 10 (lower panels) capture occasions and low (left-column panels) versus high (right-column panel) detectability. True prevalence is represented as a dashed line while the 100 values of naive and model-based prevalence are displayed in the white and gray boxplots, respectively

3 | RESULTS

3.1 | Evaluation of model performance

Naive prevalence had higher RMSE and percent relative bias than model-based prevalence in scenario 1 (state-dependent detectability and homogeneous assignment probability; Figure 1, Tables 1,2). The same occurred in scenario 2 (homogeneous detectability and state-dependent assignment probability; Supporting information Figure S1 and Tables S2, S3). Naive and model-based prevalence had similar RMSE and relative bias only in scenario 3 (homogeneous detectability and assignment probability; Supporting information Figure S1 and Tables S4, S5).

The bias associated with model-based prevalence tended to 0 in scenarios with high detectability (Tables 1,2; Supporting information Tables S2, S5). Interestingly, the bias associated with naive prevalence had the opposite behavior, as it increased at higher detectability (Tables 1,2; Supporting information Tables S2–S5). The negative bias in naive prevalence is due to the simulation settings for the true detection and assignment probabilities, and in particular to the fact that $p_p > p_h$ for scenario 1 and $\delta_p > \delta_h$ for scenario 2. Switching the true values for detectability in scenario 1 and for assignment probability in scenario 2 would cause naive prevalence to be positively-biased. The bias associated to the estimates of apparent survival, detectability, and probability of assignment was negligible for all scenarios (Figures 2–4; Supporting information Figures S3–S9), while

TABLE 1 Scenario 1 of simulation (state-dependent detectability and homogeneous assignment probability)

	Occ. 1	Occ. 2	Occ. 3	Occ. 4	Occ. 5
Root mean squared error					
Low detectability					
Naive	0.07	0.14	0.23	0.23	0.30
Model-based	0.30	0.15	0.06	0.07	0.02
High detectability					
Naive	0.26	0.36	0.48	0.41	0.48
Model-based	0.04	0.02	0.00	0.01	0.00
Percent relative bias					
Low detectability					
Naive	−0.03	−0.04	−0.05	−0.05	−0.05
Model-based	0.05	0.04	0.03	0.03	0.01
High detectability					
Naive	−0.05	−0.06	−0.07	−0.06	−0.07
Model-based	0.02	0.01	0.00	0.01	0.00

Note. Root mean squared error and relative bias of naive and model-based prevalence for sampling strategies with 5 capture occasions (Occ).

the estimates of the initial state probability were slightly biased for scenario 1 (Figure 5). Estimates of parameters for the category with fewer individuals (in this case, the hybrids) were less precise, as

TABLE 2 Scenario 1 of simulation (state-dependent detectability and homogeneous assignment probability)

	Occ. 1	Occ. 2	Occ. 3	Occ. 4	Occ. 5	Occ. 6	Occ. 7	Occ. 8	Occ. 9	Occ. 10
Root mean squared error										
Low detectability										
Naive	0.14	0.22	0.21	0.33	0.33	0.30	0.40	0.36	0.26	0.78
Model-based	0.1	0.04	0.05	0.01	0.01	0.00	0.00	0.00	0.02	0.02
High detectability										
Naive	0.23	0.51	0.37	0.45	0.49	0.50	0.39	0.36	0.16	0.16
Model-based	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Percent relative bias										
Low detectability										
Naive	-0.04	-0.05	-0.05	-0.06	-0.06	-0.05	-0.06	-0.06	-0.05	-0.03
Model-based	-0.03	-0.02	-0.02	-0.01	-0.01	-0.00	-0.00	-0.01	-0.01	-0.01
High detectability										
Naive	-0.05	-0.07	-0.06	-0.07	-0.07	-0.07	-0.06	-0.06	-0.04	-0.04
Model-based	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.00

Note. Root mean squared error and relative bias of naive and model-based prevalence for sampling strategies with 10 capture occasions (Occ).

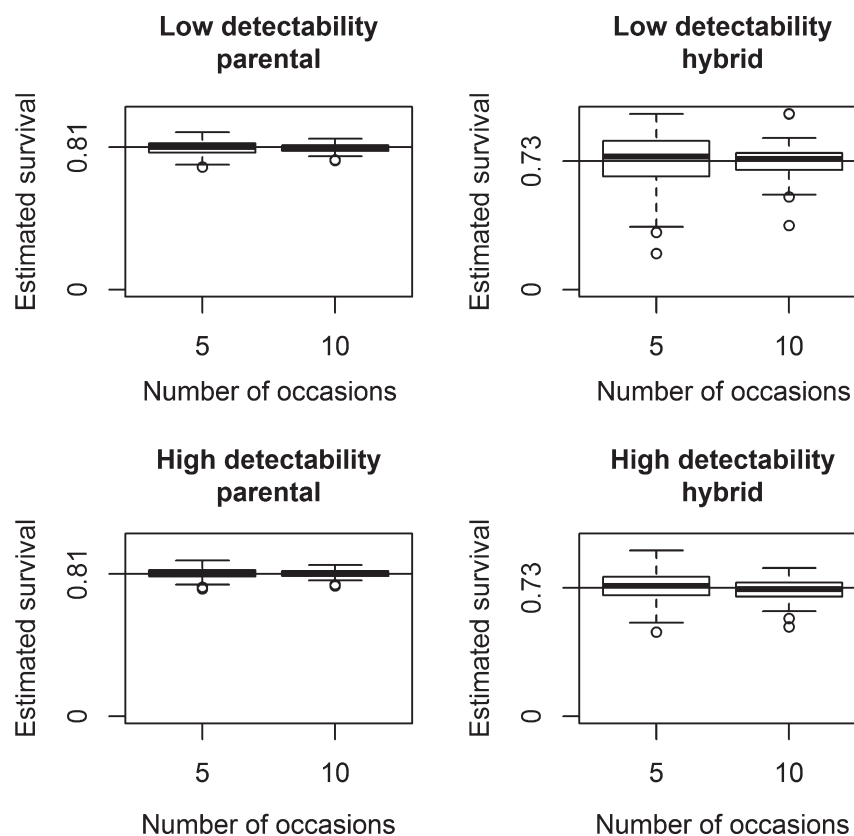


FIGURE 2 Scenario 1 of simulation (state-dependent detectability and homogeneous assignment probability). Boxplots of 100 simulated survival estimates for parentals (left two panels) and hybrids (right two panels) for each sampling strategy. Sampling strategies with low detectability are in the top row, and sampling strategies with high detectability are in the bottom row. Estimates obtained with sampling strategies with 5 and 10 capture occasions are compared in each panel

showed by the boxplot larger ranges (Figures 2–4; Supporting information Figures S2–S10).

Confidence interval coverage was always ≥ 0.89 for all scenarios with 5 sampling occasions (Table 3; Supporting information Tables S6, S7). For scenarios with 10 occasions (Table 4; Supporting information Tables S8, S9) coverage decreased after the 7th–8th occasion. This is because the low number of individuals present at the

end of the study (due to a low apparent survival) affects the accuracy of the estimates (Figure 1; Supporting information Figures S1, S2).

3.2 | Evaluation of sampling strategies

The performance of model-based prevalence improved more at higher detectability than with an increasing number of capture

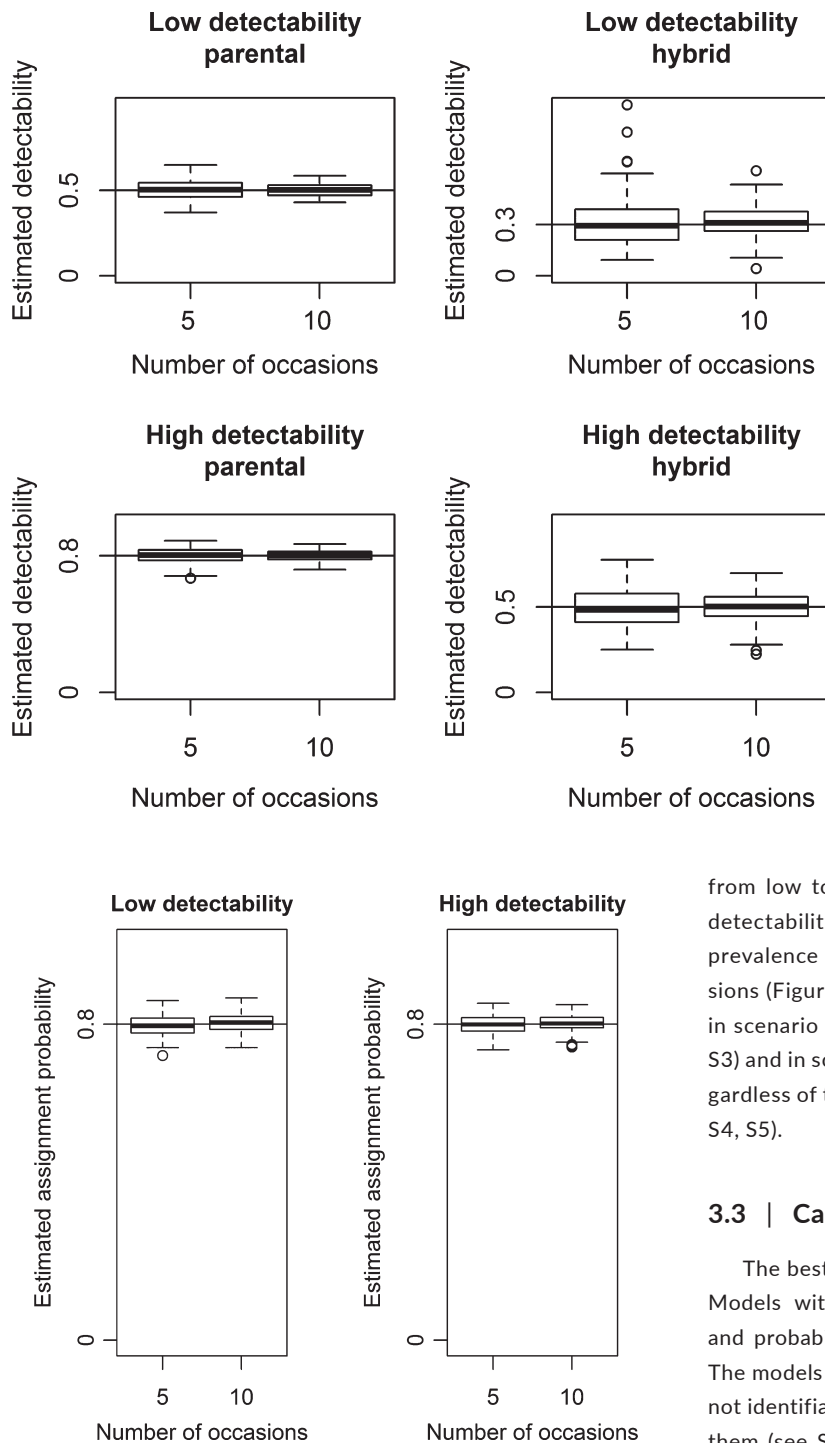


FIGURE 4 Scenario 1 of simulation (state-dependent detectability and homogeneous assignment probability). Boxplots of 100 simulated assignment probability estimates. Sampling strategies with low detectability are on the left panel, and sampling strategies with high detectability are on the right panel. Estimates obtained with sampling strategies with 5 and 10 capture occasions are compared in each panel

occasions. Specifically, in scenario 1 the RMSE and the relative bias approached 0 for both sampling strategies with 5 (Table 1) and 10 capture occasions (Table 2) when detectability changed

FIGURE 3 Scenario 1 of simulation (state-dependent detectability and homogeneous assignment probability). Boxplots of 100 simulated detectability estimates for parentals (left two panels) and hybrids (right two panels) for each sampling strategy. Sampling strategies with low detectability are in the top row, and sampling strategies with high detectability are in the bottom row. Estimates obtained with sampling strategies with 5 and 10 capture occasions are compared in each panel

from low to high. In contrast, in sampling strategies with low detectability, the RMSE and the relative bias of model-based prevalence decreased less rapidly from 5 to 10 capture occasions (Figure 1, Tables 1,2). We observed the same pattern both in scenario 2 (Supporting information Figures S1, S2, Tables S2, S3) and in scenario 3, although in the latter the bias was small regardless of the sampling strategy (Supporting information Tables S4, S5).

3.3 | Case study

The best-supported model had constant parameters (Table 5). Models with state-dependent apparent survival, detectability, and probability of assignment were also supported ($\Delta AIC_c < 2$). The models with state-dependent probability of assignment were not identifiable due to the reduced sample size, and we discarded them (see Supporting information Table S10 for a complete list of fitted models). According to the model-averaged estimates of

TABLE 3 Scenario 1 of simulation (state-dependent detectability and homogeneous assignment probability)

Confidence interval coverage 5 occasions						
	Occ. 1	Occ. 2	Occ. 3	Occ. 4	Occ. 5	Average
Low p	0.94	0.97	0.96	0.94	0.89	0.94
High p	1.00	0.98	0.95	0.95	0.99	0.97

Note. Confidence interval coverage for sampling strategies with 5 capture occasions (Occ.).

TABLE 4 Scenario 1 of simulation (state-dependent detectability and homogeneous assignment probability)

Confidence interval coverage 10 occasions											
	Occ. 1	Occ. 2	Occ. 3	Occ. 4	Occ. 5	Occ. 6	Occ. 7	Occ. 8	Occ. 9	Occ. 10	Average
Low p	0.98	0.95	0.97	0.96	0.94	0.92	0.84	0.65	0.62	0.64	0.84
High p	1.00	0.99	0.97	0.97	0.95	0.91	0.92	0.85	0.91	0.75	0.92

Note. Confidence interval coverage for sampling strategies with 10 capture occasions (Occ.).

parameters (Table 6), model-based prevalence was consistently higher (range: 0.23–0.53) than naive prevalence (range: 0.20–0.46), with the latter always included in the confidence interval of the former (Table 7).

4 | DISCUSSION

We presented a hidden Markov model to estimate prevalence in wildlife population taking into account the imperfect detectability and uncertainty in individuals' classification. We compared model-based and naive prevalence showing that the latter can be severely biased when detectability is state-dependent, in agreement with Jennelle et al. (2007). In addition, we identified another source of bias in naive prevalence which was related to the state-dependent probability of assignment. We demonstrated that, if naive prevalence has to be used as a proxy for population prevalence, the burden of proof should be placed on demonstrating homogeneity in detection and assignment probabilities (Jennelle et al., 2007; MacKenzie & Kendall, 2002).

4.1 | Model assumptions

Our approach provides a framework to statistically test differences in detectability and probability of assignment and take them into account to produce unbiased estimates of prevalence. However, a series of assumptions must be met (Lebreton et al., 1992; Otis et al., 1978): (a) parameter and processes estimated for the marked individuals can be applied to the unmarked ones. In particular, because multievent CR models are conditional on first capture, we assume that capture probability is the same for unmarked and marked individuals in order to obtain abundance estimates, (b) marks do not affect the behavior of the individuals, (c) marks are not lost, and they are correctly recognized, (d) individuals alive in the population at time t have homogeneous detectability and apparent survival probability, (e) individuals are independent from each other, and (f) no births, deaths, emigration, or immigration occur during the capture occasions. The assumption of homogeneous detectability and survival can be relaxed by including different sources of heterogeneity in the model structure (Cubaynes et al., 2010; Pradel, 1993; Pradel, Hines, Lebreton, & Nichols, 1997). In particular when

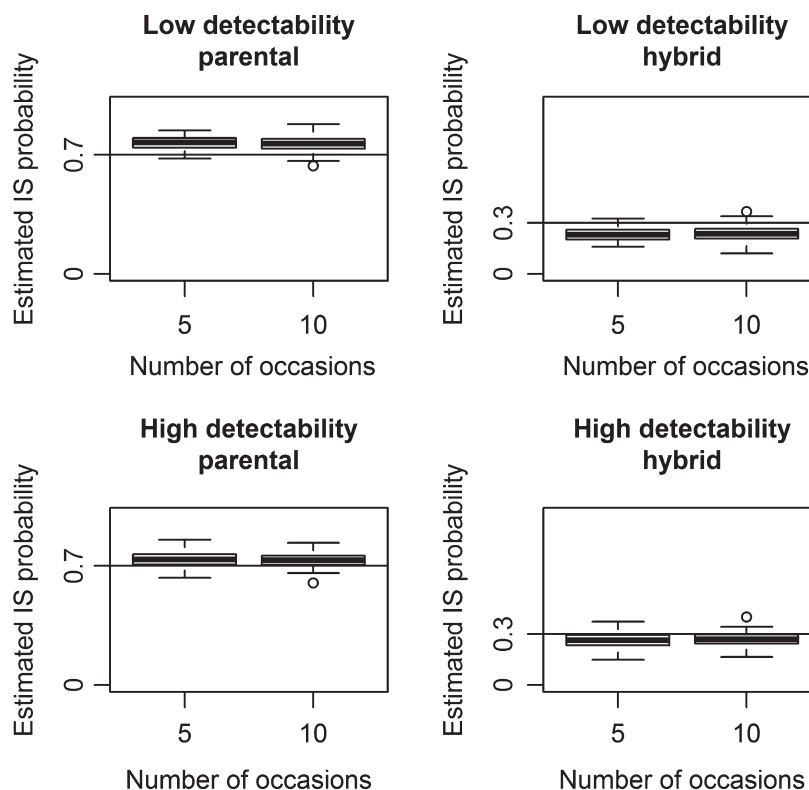


FIGURE 5 Scenario 1 of simulation (state-dependent detectability and homogeneous assignment probability). Boxplots of 100 simulated initial state probabilities estimates for parentals (left two panels) and hybrids (right two panels) for each sampling strategy. Sampling strategies with low detectability are in the top row, and sampling strategies with high detectability are in the bottom row. Estimates obtained with sampling strategies with 5 and 10 capture occasions are compared in each panel

Model	n_{par}	Deviance	QAICc	ΔQAICc
$\pi(i)\varphi(i)p(i)\delta(a1 + a2_fix)$	4	183.48	192.16	0
$\pi(i)\varphi(f)p(i)\delta(a1 + a2_fix)$	5	181.41	192.44	0.28
$\pi(i)\varphi(i)p(f)\delta(a1 + a2_fix)$	5	181.76	192.79	0.63

Note. The notation (.) indicates constant parameters while (state) indicates state-dependent parameters. π = initial state probability, φ = survival probability, p = detection probability, δ = assignment probability, n_{par} = number of parameters. The term (a1 + a2_fix) indicates that we constrained the model to have fixed assignment probabilities after first capture.

TABLE 6 Model-averaged parameter estimates for the case study on wolf \times dog hybridization

Parameter	Estimate	95% C.I.
Initial proportion of individuals in state "Wolf" π_w	0.60	0.43–0.76
Detection probability of wolves p_w	0.42	0.17–0.67
Detection probability of hybrids p_h	0.50	0.26–0.77
Survival probability of wolves φ_w	0.72	0.39–0.91
Survival probability of hybrids φ_h	0.84	0.46–0.99
Assignment probability δ	0.79	0.64–0.89

estimating abundance attention must be paid to unaccounted for heterogeneity in capture probabilities, which is known to bias estimates of abundance (Pollock, Winterstein, Bunck, & Curtis, 1989).

4.2 | Limitations of the model

A potential drawback of our approach lies in convergence issues that might occur when there is a high proportion of uncertain individuals in the sample (Pradel et al., 2007). This problem can be overcome by confirming the state of just a handful of individuals with some error-free method. Specifically, Pradel et al. (2007) showed that in a case study in which the sex of about 80% of the individuals was uncertain, adding few genetic confirmations (24 individuals over 4,025) greatly improved the shape of the likelihood and hence the convergence of the optimization algorithm. This is particularly relevant because the model does not handle assignment errors, so having a high proportion of uncertain individuals is preferable to taking the risk of making assignment errors to reduce such proportion.

4.3 | Sampling guidelines

Through the simulations we showed that the precision of model-based prevalence increased proportionally more by enhancing

detection probability than by increasing the number of sampling occasions. The precision of survival estimates increased by the same amount with increasing detectability and number of occasions, confirming that a cost-effective sampling strategy should maximize detectability within capture occasions instead of sampling more frequently. For a specific study, our simulation framework can be used to determine the best trade-off to obtain a cost-effective level of accuracy and precision of parameter estimates (Lieury et al., 2017).

4.4 | Case study

For the hybridization case study, potential differences in the detectability of parental and admixed individuals generate bias in naive prevalence and can originate from various reasons. For example, differences in vocalization behavior (Derégnaucourt, Guyomarc'h, & Spanò, 2005), migratory tendency (Derégnaucourt, Guyomarc'h, & Belhamra, 2004), and social status (Battocchio, Iacolina, Canu, & Mori, 2017) were documented between admixed parental individuals of different species and may cause differential detectability. In wolf packs, in particular, differences in detectability through scat sampling are related to social status and marking behavior (Cubaynes et al., 2010; Marucco et al., 2009).

Previous hybridization studies on wolves (Godinho et al., 2011; Pacheco et al., 2017) acknowledged that, due to uncertainty of classification, a proportion of backcrosses was assigned to the parental cluster (i.e., wolves), leading to an underestimation of their prevalence in the population. With our approach, a conservative q_{wolf} threshold can be used for the parental cluster, greatly reducing the chance of type II errors (erroneously classifying hybrids as parentals), as showed by the higher model-based prevalence values in the case study (Table 7). These results should be considered as a warning that relying on naive prevalence underestimates the hybridization-related risks for the conservation of the parental populations.

Prevalence	Occ. 1	Occ. 2	Occ. 3	Occ. 4	Occ. 5
Naive	0.27	0.33	0.20	0.46	0.27
Model-based	0.32	0.42	0.23	0.53	0.43
95% CI	0.07–0.55	0.09–0.63	0.06–0.48	0.12–0.72	0.09–0.67

TABLE 7 Naive and model-based prevalence of hybrids for the case study on wolf \times dog hybridization

4.5 | Model extensions and broader applications of the model

The current model formulation contains two states ("Hybrid" and "Parental"). It might be desirable, however, to further split the "Hybrid" state into two distinct states "Hybrid" and "Backcross," or to include not just one but both parental species. This can be done by increasing the number of states and corresponding parameters depending on the available data and the power of the genetic tests. However, attention should be paid to avoid over-parameterization (Gimenez, Choquet, & Lebreton, 2003). Multievent models can quickly become parameter-rich and thus result in nonidentifiability in studies with small sample sizes.

The Viterbi algorithm has been previously applied in the multievent CR framework to reconstruct the reproductive life of individual roe deers (*Capreolus capreolus*; Rouan et al., 2009). We used the Viterbi algorithm to reconstruct the abundance of individuals in different states. This approach can be used in any case study that requires the estimation of abundance and prevalence of individuals in the presence of imperfect detection and uncertainty in state assignment, for example, the estimation of sex ratios in monomorphic species (Genovart et al., 2012; Pradel et al., 2007). Moreover, the model can be extended to include dynamic states by adding a transition probability matrix in the formulation. Such extension would expand its applications to other fields such as epidemiology (e.g., the estimation of the number of infected individuals in a population; Marescot et al., 2018) or reproductive biology studies (e.g., the number of breeders in a population; Desprez, McMahon, Hindell, Harcourt, & Gimenez, 2013).

ACKNOWLEDGMENTS

We are grateful to M. Canestrini and F. Moretti for sample collection and to M. Galaverni for assistance with genetic analyses. We thank Chris Grieves, Paul B. Conn, and one anonymous reviewer for their insightful suggestions during the editorial process. We also thank Rémi Choquet and Roger Pradel for their help in the model development. OG and NLS were supported by the French National Research Agency with a Grant ANR-16-CE02-0007. NLS was supported by a PhD grant by Sapienza University of Rome.

CONFLICT OF INTEREST

None declared.

AUTHOR CONTRIBUTIONS

NLS, OG, and PC conceived the ideas and designed methodology and led the writing of the manuscript; PC and WR organized and supervised the field data collection and logistics; LM collected the data; NLS and OG constructed the models, performed the simulations, and analyzed the data; RC and EF performed the genetic

laboratory and cluster analyses. All authors contributed critically to the drafts and gave final approval for publication.

DATA ACCESSIBILITY

Codes of the models, simulations, and data are available from the Dryad Digital Repository (<https://doi.org/10.5061/dryad.8g8r675>).

ORCID

Nina Luisa Santostasi  <https://orcid.org/0000-0001-9148-7069>

REFERENCES

- Allendorf, F. W., Leary, R. F., Spruell, P., & Wenburg, J. K. (2001). The problems with hybrids: setting conservation guidelines. *Trends in Ecology & Evolution*, 16, 613–622. [https://doi.org/10.1016/S0169-5347\(01\)02290-X](https://doi.org/10.1016/S0169-5347(01)02290-X)
- Battocchio, D., Iacolina, L., Canu, A., & Mori, E. (2017). How much does it cost to look like a pig in a wild boar group? *Behavioural Processes*, 138, 123–126. <https://doi.org/10.1016/j.beproc.2017.03.001>
- Boitani, L. (2000). *Action plan for the conservation of Wolves in Europe (Canis lupus)*. Strasbourg, France: Council of Europe Publishing.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York, NY: Springer.
- Caniglia, R., Fabbri, E., Cubaynes, S., Gimenez, O., Lebreton, J.-D., & Randi, E. (2012). An improved procedure to estimate wolf abundance using non-invasive genetic sampling and capture – recapture mixture models. *Conservation Genetics*, 13, 53–64. <https://doi.org/10.1007/s10592-011-0266-1>
- Caniglia, R., Fabbri, E., Galaverni, M., Milanese, P., & Randi, E. (2014). Noninvasive sampling and genetic variability, pack structure, and dynamics in an expanding wolf population. *Journal of Mammalogy*, 95, 41–59. <https://doi.org/10.1644/13-MAMM-A-039>
- Caniglia, R., Fabbri, E., Greco, C., Galaverni, M., Manghi, L., Boitani, L., & Randi, E. (2013). Black coats in an admixed wolf × dog pack is melanism an indicator of hybridization in wolves? *European Journal of Wildlife Research*, 59, 543–555. <https://doi.org/10.1007/s10344-013-0703-1>
- Caswell, H. (2000). *Matrix population models: Construction, analysis and interpretation*. Sunderland, MA: Sinauer.
- Choquet, R., Rouan, L., & Pradel, R. (2009). Program E-Surge: A software application for fitting multievent models. In D. L. Thompson, E. G. Cooch, & M. J. Conroy (Eds.), *Modeling demographic processes in marked populations*, volume 3 of Environmental and Ecological Statistics (pp. 845–865). New York, NY: Springer series.
- Conn, P. B., & Cooch, E. G. (2009). Multistate capture – recapture analysis under imperfect state observation: an application to disease models. *Journal of Applied Ecology*, 46, 486–492. <https://doi.org/10.1111/j.1365-2664.2008.01597.x>
- Cubaynes, S., Pradel, R., Choquet, R., Duchamp, C., Gaillard, J.-M., Lebreton, J.-D., ... Gimenez, O. (2010). Importance of accounting for detection heterogeneity when estimating abundance: the case of French wolves. *Conservation Biology*, 24, 621–626. <https://doi.org/10.1111/j.1523-1739.2009.01431.x>
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.
- Derégnaucourt, S., Guyomarc'h, J.-C., & Belhamra, M. (2004). Comparison of migratory tendency in European Quail *Coturnix coturnix*, domestic Japanese Quail *Coturnix japonica* and their hybrids. *Ibis*, 147, 25–36. <https://doi.org/10.1111/j.1474-919x.2004.00313.x>

- Derégnaucourt, S., Guyomarc'h, J.-C., & Spanò, S. (2005). Behavioural evidence of hybridization (Japanese \times European) in domestic quail released as game birds. *Applied Animal Behaviour Science*, 94, 303–318. <https://doi.org/10.1016/j.applanim.2005.03.002>
- Desprez, M., McMahon, C. R., Hindell, M. A., Harcourt, R., & Gimenez, O. (2013). Known unknowns in an imperfect world : incorporating uncertainty in recruitment estimates using multievent capture – recapture models. *Ecology and Evolution*, 3, 4658–4668. <https://doi.org/10.1002/ece3.846>
- Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164, 1567–1587. <https://doi.org/10.1111/j.1471-8286.2007.01758.x>
- Genovart, M., Pradel, R., & Oro, D. (2012). Exploiting uncertain ecological fieldwork data with multievent capture – recapture modelling: an example with bird sex assignment. *Journal of Animal Ecology*, 81, 970–977. <https://doi.org/10.1111/j.1365-2656.2012.01991.x>
- Gervasi, V., Boitani, L., Paetkau, D., Posillico, M., Randi, E., & Ciucci, P. (2017). Estimating survival in the Apennine brown bear accounting for uncertainty in age classification. *Population Ecology*, 59, 119–130. <https://doi.org/10.1007/s10144-017-0587-0>
- Gimenez, O., Choquet, R., & Lebreton, J.-D. (2003). Parameter redundancy in multistate capture-recapture models. *Biometric Journal*, 45, 704–722.
- Gimenez, O., Lebreton, J.-D., Gaillard, J.-M., Choquet, R., & Pradel, R. (2012). Estimating demographic parameters using hidden process dynamic models. *Theoretical Population Biology*, 82, 307–316. <https://doi.org/10.1016/j.tpb.2012.02.001>
- Godinho, R., Llaneza, L., Blanco, J. C., Lopes, S., Álvares, F., García, E. J., & Ferrand, N. (2011). Genetic evidence for multiple events of hybridization between wolves and domestic dogs in the Iberian Peninsula. *Molecular Ecology*, 20, 5154–5166. <https://doi.org/10.1111/j.1365-294X.2011.05345.x>
- Himmelman, L. (2010). HMM: HMM-hidden Markov Models. Scientific Software Development, 2010. R Package Version 1.0.
- Hindrickson, M., Remm, J., Pilot, M., Godinho, R., Stronen, A. V., Baltrunaitė, L., ... Saarma, U. (2017). Wolf population genetics in Europe: A systematic review, meta-analysis and suggestions for conservation and management. *Biological Reviews*, 92, 1601–1629.
- Jennelle, C. S., Cooch, E., Conroy, M. J., & Senar, J. C. (2007). State specific detection probability and disease prevalence. *Ecological Applications*, 17, 154–167. [https://doi.org/10.1890/1051-0761\(2007\)017\[0154:SDPADP\]2.0.CO;2](https://doi.org/10.1890/1051-0761(2007)017[0154:SDPADP]2.0.CO;2)
- Lebreton, J., Burnham, K., Clobert, J., & Anderson, D. (1992). Modelling survival and testing biological hypothesis using marked animals: a unified approach with case studies. *Ecological Monographs*, 62, 67–118. <https://doi.org/10.2307/2937171>
- Lieury, N., Devillard, S., Besnard, A., Gimenez, O., Hameau, O., Ponchon, C., & Millon, A. (2017). Designing cost-effective capture – recapture surveys for improving the monitoring of survival in bird populations. *Biological Conservation*, 214, 233–241. <https://doi.org/10.1016/j.biocon.2017.08.011>
- MacKenzie, D. I., & Kendall, W. L. (2002). How should detection probability be incorporated into estimates of relative abundance? *Ecology*, 83, 2387–2393. <https://doi.org/10.2307/3071800>
- Marescot, L., Benhaïem, S., Gimenez, O., Hofer, H., Lebreton, J.-D., Olarte-Castillo, X. A., & East, M. L. (2018). Social status mediates the fitness costs of infection with canine distemper virus in Serengeti spotted hyenas. *Functional Ecology*, 32(1237–1250), 1–14. <https://doi.org/10.1111/1365-2435.13059/epdf>
- Marucco, F., Pletscher, D. H., Boitani, L., Schwartz, M. K., Pilgrim, K. L., & Lebreton, J. (2009). Wolf survival and population trend using non-invasive capture – recapture techniques in the Western Alps. *Journal of Applied Ecology*, 46, 1003–1010. <https://doi.org/10.1111/j.1365-2664.2009.01696.x>
- McDonald, T. L., & Amstrup, S. C. (2001). Estimation of population size using open capture-recapture models. *Journal of Agricultural, Biological and Environmental statistics*, 6, 206–220. <https://doi.org/10.1198/108571101750524553>
- Moreno-Torres, K., Wolfe, B., Saville, W., & Garabed, R. (2016). Estimating *Neospora caninum* prevalence in wildlife populations using Bayesian inference. *Ecology and Evolution*, 6, 2216–2225. <https://doi.org/10.1002/ece3.2050>
- Nichols, J. D. (1992). Capture-recapture models: using marked animals to study population dynamics. *Bio Science*, 42, 94–102. <https://doi.org/10.2307/1311650>
- Otis, D. L., Burnham, K. P., White, G. C., & Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, 62, 3–135. <https://doi.org/10.2307/2287873>
- Pacheco, C., López-bao, J. V., García, E. J., Lema, F. J., Llaneza, L., Palacios, V., & Godinho, R. (2017). Spatial assessment of wolf-dog hybridization in a single breeding period. *Scientific Reports*, 7, 1–10. <https://doi.org/10.1038/srep42475>
- Pollock, K. H., Winterstein, S. R., Bunck, C. M., & Curtis, P. D. (1989). Survival analysis in telemetry studies: The staggered entry design. *Journal of Wildlife Management*, 53, 7–15.
- Pradel, R. (1993). Flexibility in survival analysis from recapture data: Handling trap-dependence. In J.-D. Lebreton, & P. M. North (Eds.), *Marked individuals in the study of bird populations* (pp. 29–37). Basel, Switzerland: Birkhäuser Verlag.
- Pradel, R. (2005). Multievent: an extension of multistate capture – recapture models to uncertain states. *Biometrics*, 61, 442–447. <https://doi.org/10.1111/j.1541-0420.2005.00318.x>
- Pradel, R., Hines, J. E., Lebreton, J.-D., & Nichols, J. D. (1997). Capture – recapture survival models taking account of transients. *Biometrics*, 53, 60–72. <https://doi.org/10.2307/2533097>
- Pradel, R., Maurin-Bernier, L., Gimenez, O., Genovart, M., Choquet, R., & Oro, D. (2007). Estimation of sex-specific survival with uncertainty in sex assessment. *The Canadian Journal of Statistics*, 35, 1–14. <https://doi.org/10.1002/cjs.5550360105>
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945–959. <https://doi.org/10.1111/j.1471-8286.2007.01758.x>
- R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Randi, E., Hulva, P., Fabbri, E., Galaverni, M., Galov, A., Kusak, J., ... Caniglia, R. (2014). Multilocus detection of wolf \times dog hybridization in Italy, and guidelines for marker selection. *PLoS ONE*, 9, e86409. <https://doi.org/10.1371/journal.pone.0086409>
- Rouan, L., Gaillard, J.-M., Guédon, Y., & Pradel, R. (2009). Estimation of lifetime reproductive success when reproductive status cannot always be assessed. In D. L. Thompson, E. G. Cooch, & M. J. Conroy (Eds.), *Modeling demographic processes in marked populations*, volume 3 of Environmental and Ecological Statistics (pp. 845–865). New York, NY: Springer series.
- Sanz-Aguilar, A., Tavecchia, G., Genovart, M., Igual, J. M., Oro, D., Rouan, L., & Pradel, R. (2011). Studying the reproductive skipping behavior in long-lived birds by adding nest inspection to individual-based data. *Ecological Applications*, 21, 555–564. <https://doi.org/10.1890/09-2339.1>
- Schwenk, K., Brede, N., & Streit, B. (2008). Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. *Philosophical Transactions of the Royal Society B*, 363, 2805–2811. <https://doi.org/10.1098/rstb.2008.0055>
- Todesco, M., Pascual, M. A., Owens, G. L., Ostevik, K. L., Moyers, B. T., Hübner, S., & Riesberg, L. H. (2016). Hybridization and extinction. *Evolutionary Applications*, 9, 892–908. <https://doi.org/10.1111/eva.12367>
- Vähä, J. P., & Primmer, C. R. (2006). Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization

- scenarios and with different numbers of loci. *Molecular Ecology*, 15, 63–72. <https://doi.org/10.1111/j.1365-294X.2005.02773.x>
- Vaz Pinto, P., Beja, P., Ferrand, N., & Godinho, R. (2015). Hybridization following population collapse in a critically endangered antelope. *Scientific Reports*, 6, 1–9. <https://doi.org/10.1038/srep18788>
- Zucchini, W., MacDonald, I. L., & Langrock, R. (2009). *Hidden Markov models for time series: an introduction using R*. Boca Raton, FL: Chapman & Hall/CRC Press.

How to cite this article: Santostasi NL, Ciucci P, Caniglia R, et al. Use of hidden Markov capture–recapture models to estimate abundance in the presence of uncertainty: Application to the estimation of prevalence of hybrids in animal populations. *Ecol Evol*. 2019;9:744–755. <https://doi.org/10.1002/ece3.4819>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.