Diversity and Distributions, (Diversity Distrib.) (2014) **20**, 223–235

BIODIVERSITY RESEARCH

Accounting for false positives improves estimates of occupancy from key informant interviews

Rajeev Pillay[1,2]*, David A. W. Miller[3,4], James E. Hines[3], Atul A. Joshi[1,5] and M. D. Madhusudan[1,6]

[1]*Nature Conservation Foundation, 3076/5, 4th Cross, Gokulam Park, Mysore 570002, India,* [2]*Department of Wildlife Ecology and Conservation, University of Florida, 110 Newins-Ziegler Hall, PO Box 110430, Gainesville, FL 32611-0430, USA,* [3]*United States Geological Survey, Patuxent Wildlife Research Center, 12100 Beech Forest Road, Laurel, MD 20708-4039, USA,* [4]*Department of Ecosystem Science and Management, Pennsylvania State University, 411 Forest Resources Building, University Park, PA 16802, USA,* [5]*National Centre for Biological Sciences, Tata Institute of Fundamental Research, GKVK Campus, Bellary Road, Bangalore 560065, India,* [6]*Department of Environmental and Forest Biology, State University of New York, College of Environmental Science and Forestry, 1 Forestry Drive, Syracuse, NY 13210-2724, USA*

*Correspondence: Rajeev Pillay, Nature Conservation Foundation, 3076/5, 4th Cross, Gokulam Park, Mysore 570002, India.
E-mail: rajeev@conservation.in

ABSTRACT

Aim Much research in conservation biogeography is fundamentally dependent on obtaining reliable data on species distributions across space and time. Such data are now increasingly being generated using various types of public surveys. These data are often integrated with occupancy models to evaluate distributional patterns, range dynamics and conservation status of multiple species at broad spatio-temporal scales. Occupancy models have traditionally corrected for imperfect detection due to false negatives while implicitly assuming that false positives do not occur. However, public survey data are also prone to false-positive errors, which when unaccounted for can cause bias in occupancy estimates. We test whether false positives in a dataset collected from public surveys lead to overestimation of species site occupancy and whether estimators that simultaneously account for false-positive and false-negative errors improve occupancy estimates.

Location Western Ghats, India.

Methods We fit occupancy models that simultaneously account for false positives and negatives to data collected from a large-scale key informant interview survey for 30 species of large vertebrates. We tested their performance against standard occupancy models that account only for false negatives.

Results Standard occupancy models that correct only for false negatives tended to overestimate species occupancy due to false-positive errors. Occupancy models that simultaneously accounted for false positives and negatives had greater support [lower Akaike's information criterion (AIC)] and, consistent with predictions, generated systematically lower occupancy estimates than standard models. Furthermore, accounting for false positives improved the accuracy of occupancy estimates despite the added complexity to the statistical estimator.

Main conclusions Integrating large-scale public surveys with occupancy modelling approaches is a powerful tool for informing conservation and management. However, in many if not most cases, it will be important to explicitly account for false positives to ensure the reliability of occupancy estimates obtained from public survey datasets such as key informant interviews, volunteer surveys, citizen science programmes, historical archives and acoustic surveys.

Keywords

Citizen science, detectability, misclassification, misidentification, multiple detection method model, non-detection, overestimation, public survey, single-season occupancy model, species distribution modelling.

Diversity and Distributions
A Journal of Conservation Biogeography

DOI: 10.1111/ddi.12151
http://wileyonlinelibrary.com/journal/ddi

INTRODUCTION

Understanding the distribution of species and the drivers underlying distributional patterns is of prime interest in conservation biogeography (Brown, 1984; Whittaker *et al.*, 2005; Lomolino *et al.*, 2010; Richardson & Whittaker, 2010). The accurate assessment of patterns of species distributions requires that imperfect detection be accounted for (MacKenzie *et al.*, 2002, 2006; Tyre *et al.*, 2003). Imperfect detection due to false negatives yields underestimates of the extent of species distributions (MacKenzie *et al.*, 2002). Occupancy modelling, a versatile tool in applied ecological research for estimating the proportion of area occupied by a species and assessing spatio-temporal trends in species distributions (MacKenzie *et al.*, 2006; Royle & Dorazio, 2008; Karanth *et al.*, 2011; Pillay *et al.*, 2011), has traditionally corrected for false negatives. False negatives or non-detection errors occur when surveyors fail to detect a species or any evidence of its presence at a site even when the site is occupied by that species (MacKenzie *et al.*, 2002, 2006). Much less attention, however, has been paid to imperfect detection due to false positives, which occur when organisms are detected but misidentified or when detections are wrongly recorded at sites where species are truly absent (Royle & Link, 2006; McClintock *et al.*, 2010a,b; Miller *et al.*, 2011). Imperfect detection due to false positives can yield overestimates of the extent of species distributions (Royle & Link, 2006; Miller *et al.*, 2011; Molinari-Jobin *et al.*, 2012).

Accounting for imperfect detection in an occupancy modelling framework involves either conducting multiple surveys to collect replicate observations (minimum two) of detections and non-detections for a species in a site (MacKenzie *et al.*, 2006) or modelling detection probability as a continuous process (Garrard *et al.*, 2008; Guillera-Arroita *et al.*, 2011). Collecting the necessary data to fit these models is often a major limiting factor impeding the widespread application of occupancy models for investigating species distributions, especially over large spatio-temporal scales. To get around the costs of obtaining replicate detection/non-detection data at large spatio-temporal scales, scientists have come up with innovative and economical public survey methods. Detection histories generated from key informant interviews (Karanth *et al.*, 2009; Pillay *et al.*, 2011; Zeller *et al.*, 2011), volunteer surveys (Kéry *et al.*, 2010a; Sewell *et al.*, 2010), citizen science programs (Kéry *et al.*, 2010b; Yu *et al.*, 2010), historical archives (Karanth *et al.*, 2010) and acoustic call surveys (Simons *et al.*, 2007; McClintock *et al.*, 2010a,b) are increasingly being integrated with occupancy models. The novel integration of these public survey methods to occupancy models has vastly expanded the scope of occupancy estimation, offering efficient and cost-effective means of modelling the distributions of multiple species at broad spatio-temporal scales. However, in addition to false-negative error, these methods may be prone to imperfect detection due to misidentification and false-positive error (McKelvey *et al.*, 2008; Fitzpatrick *et al.*, 2009; Molinari-Jobin *et al.*, 2012). If unaccounted for, even small rates of false-positive error can lead to substantial bias in estimators of occupancy (Royle & Link, 2006; Miller *et al.*, 2011; Molinari-Jobin *et al.*, 2012) as well as in estimators of the vital rates of colonization and extinction when dynamic occupancy models are used (McClintock *et al.*, 2010a; Miller *et al.*, 2013).

The aim of this study is to draw attention to the fundamental issue of species misidentification and overestimation of species distributions when occupancy modelling is integrated with public survey datasets. We present the first application of the recently developed single-season false-positive occupancy models (Miller *et al.*, 2011) to simultaneously correct for false-positive and false-negative errors in detection histories from key informant interviews. These interviews recorded sightings and indirect detections (Pillay *et al.*, 2011) of 30 species of large vertebrates (Table 1) in the Western Ghats biodiversity hotspot in India (Fig. 1) and were obtained from three different key informant groups: Forest Department field personnel (FD), forest-dwelling local communities (LC) and experts (comprising wildlife scientists and professional wildlife photographers/filmmakers) (EX). These data were analysed using the estimator proposed by Miller *et al.* (2011), which we refer to as the full occupancy estimator because it accounts for both false-positive and false-negative errors when estimating occupancy probabilities. This estimator can be applied to improve occupancy estimates when detections can be classified into two or more subsets: at least one subset of a type or method for which false positives can be assumed not to occur (certain detections) and one or more subset(s) of a type or method for which false positives could occur (uncertain detections). Thus, the full estimator always uses both certain and uncertain data types. We made the assumption that the subset of detection histories collected from experts was free from misidentification errors (Fitzpatrick *et al.*, 2009; Yu *et al.*, 2010; Molinari-Jobin *et al.*, 2012). We classified this subset as certain due to the familiarity of the experts we interviewed with the species of interest, our sampling units in the Western Ghats as well as our survey methods. We assumed that false positives could occur in the remaining observations from Forest Department field personnel and forest-dwelling local communities (collectively referred to as non-experts) and classified these as uncertain. We compare these results to estimates using the standard occupancy estimator originally outlined by MacKenzie *et al.* (2002).

Our objectives in this study were to: (1) determine whether the data were consistent with the occurrence of false positives in uncertain observations by non-experts, (2) determine whether including uncertain observations together with certain observations could improve the accuracy of occupancy estimates when false positives were accounted for in statistical models and (3) summarize our results for estimated true- and false-positive detection probabilities using key informant surveys and provide insights into others implementing similar analyses with different types of public survey datasets. To meet our first objective, we tested the

Table 1 Species surveyed in the Western Ghats, total number of detections and non-detections (replication) for each species collected from each key informant group and naïve estimates of occupancy (i.e. proportion of sites with at least one detection)

Species			Total detections/Total non-detections*			Naïve occupancy estimates		
Common name	Scientific name	Code	Forest Department	Local community	Expert	Uncertain†	Certain‡	Both§
Class Mammalia – Order Carnivora								
Tiger	*Panthera tigris* (Linnaeus, 1758)	TGR	253/353	509/1059	183/591	0.60	0.32	0.64
Leopard	*Panthera pardus* (Linnaeus, 1758)	LPD	398/208	840/727	265/509	0.83	0.46	0.86
Dhole	*Cuon alpinus* (Pallas, 1811)	DHL	399/207	922/645	155/619	0.85	0.29	0.85
Golden jackal	*Canis aureus* (Linnaeus, 1758)	JKL	234/372	699/868	79/695	0.76	0.20	0.77
Striped hyena	*Hyaena hyaena* (Linnaeus, 1758)	HYN	5/288	94/1296	3/730	0.14	0.01	0.14
Sloth bear	*Melursus ursinus* (Shaw, 1791)	SLB	311/295	730/837	295/583	0.71	0.38	0.72
Class Mammalia – Order Proboscidea								
Asian elephant	*Elephas maximus* (Linnaeus, 1758)	ELP	383/223	815/752	260/514	0.69	0.45	0.71
Class Mammalia – Order Artiodactyla								
Gaur	*Bos gaurus* (C.H. Smith, 1827)	GAR	443/163	1083/484	307/467	0.87	0.48	0.89
Sambar	*Rusa unicolor* (Kerr, 1792)	SAM	493/113	1163/404	313/461	0.93	0.50	0.94
Chital	*Axis axis* (Erxleben, 1777)	CHT	289/317	829/738	203/570	0.69	0.33	0.70
Indian muntjac	*Muntiacus vaginalis* (Boddaert, 1785)	MJK	456/150	1259/308	150/557	0.96	0.39	0.96
Indian chevrotain	*Moschiola indica* (Gray, 1852)	MDR	399/207	1103/464	92/682	0.92	0.21	0.91
Four-horned antelope	*Tetracerus quadricornis* (de Blainville, 1816)	FHA	51/242	218/1172	24/709	0.28	0.08	0.30
Nilgiri tahr	*Nilgiritragus hylocrius* (Ogilby, 1838)	NTR	131/475	90/1477	16/758	0.18	0.03	0.17
Wild pig	*Sus scrofa* (Linnaeus, 1758)	WPG	583/23	1442/125	315/459	1.00	0.56	1.00
Class Mammalia – Order Primates								
Nilgiri langur	*Trachypithecus johnii* (J. Fischer, 1829)	NLG	237/369	303/1264	45/729	0.37	0.11	0.37
Southern plains/Tufted/Black-footed grey langur	*Semnopithecus dussumieri* (I. Geoffroy, 1843)/*S priam* (Blyth, 1844)/*S hypoleucos* (Blyth, 1841)	CLG	299/307	1103/464	316/458	0.74	0.47	0.75
Lion-tailed macaque	*Macaca silenus* (Linnaeus, 1758)	LTM	126/480	166/1400	64/710	0.21	0.11	0.24
Bonnet macaque	*Macaca radiata* (E. Geoffroy, 1812)	BNT	546/60	1382/185	296/478	0.99	0.56	0.99
Class Mammalia – Order Rodentia								
Indian crested porcupine	*Hystrix indica* (Kerr, 1792)	POR	172/121	1089/301	199/534	0.96	0.41	0.95
Indian giant squirrel	*Ratufa indica* (Erxleben, 1777)	IGS	480/125	1225/342	266/508	0.95	0.49	0.95
Class Aves – Order Bucerotiformes								
Great hornbill	*Buceros bicornis* (Linnaeus, 1758)	GHB	229/353	561/986	53/721	0.55	0.14	0.57
Malabar pied hornbill	*Anthracoceros coronatus* (Boddaert, 1783)	MPH	87/300	331/1092	54/716	0.35	0.12	0.36
Malabar grey hornbill	*Ocyceros griseus* (Latham, 1790)	MGH	218/254	773/725	85/687	0.74	0.21	0.73
Indian grey hornbill	*Ocyceros birostris* (Scopoli, 1786)	IGH	56/322	294/1108	49/721	0.31	0.14	0.34
Class Aves – Order Galliformes								
Grey junglefowl	*Gallus sonneratii* (Temminck, 1813)	GJF	544/62	1440/127	219/555	0.99	0.45	0.99
Indian peafowl	*Pavo cristatus* (Linnaeus, 1758)	PFL	392/214	1202/365	203/571	0.87	0.39	0.87

Table 1 Continued.

Species			Total detections/Total non-detections*			Naïve occupancy estimates		
Common name	Scientific name	Code	Forest Department	Local community	Expert	Uncertain†	Certain‡	Both§
Class Reptilia – Order Squamata								
King cobra	*Ophiophagus hannah* (Cantor, 1836)	KCB	163/443	440/1127	28/746	0.47	0.07	0.48
Asian rock python	*Python molurus* (Linnaeus, 1758)	PYT	133/160	868/522	55/678	0.83	0.10	0.82
Class Reptilia – Order Crocodilia								
Marsh crocodile/Mugger	*Crocodylus palustris* (Lesson, 1831)	MUG	107/499	254/1312	25/747	0.28	0.07	0.28

*The first number indicates the total number of detections, while the second number indicates the total number of non-detections for a given species by a key informant group.
†Uncertain: Naïve occupancy estimated from data collected from non-experts (Forest Department field personnel and forest-dwelling local communities) only.
‡Certain: Naïve occupancy estimated from data collected from experts only.
§Both: Naïve occupancy estimated from combined uncertain and certain data.

Figure 1 The Western Ghats of India depicting study area (outlined in black, Inset) and the 395 ranges sampled to generate detection histories from key informant interviews.

following two specific predictions. First, we predicted that if uncertain data included false-positive errors, occupancy estimates using standard occupancy models that account only for false negatives should be greater when only uncertain data are used than when only certain data are used (Prediction 1A). This prediction results because unaccounted for false-positive errors lead to positive bias when estimating occupancy (Royle & Link, 2006; Miller *et al.*, 2011). Second, we predicted that if uncertain data included false-positive errors, occupancy estimates obtained from the full estimator that explicitly accounts for these false positives should be lower than estimates obtained from the standard estimator with combined certain and uncertain data. Further, models using the full estimator should have greater support [lower Akaike's information criterion (AIC) values] than equivalent models using the standard estimator (Prediction 1B).

Accounting for false positives adds complexity to the statistical estimator, which may negate any added benefit of collecting and including uncertain data. To meet our second objective, we tested whether this is indeed the case by comparing the precision and bias of occupancy estimates generated from the full estimator to the precision and bias of estimates generated from the standard estimator with (1) only certain data and (2) combined certain and uncertain data. We predicted that in comparison to the approach using

the full estimator, standard estimates using only certain data would be less precise due to reduced sample size while standard estimates using combined data would be biased due to false positives in the uncertain subset (Prediction 2).

With respect to the third objective, we present estimates of true- and false-positive detection probabilities as a function of key informant group. By distinguishing true- and false-positive detection types, we were able to explicitly account for variation in detection probability among key informant groups or multiple observer types, which may have an influence on occupancy estimates (Kéry *et al.*, 2010b; McClintock *et al.*, 2010a). We discuss the wide applications of this statistical modelling technique to remove bias due to non-detection and misidentification errors (McKelvey *et al.*, 2008; Dickinson *et al.*, 2010; Hochachka *et al.*, 2012) and improve species site occupancy estimates when integrated with various types of public survey datasets.

METHODS

Study area and design

The Western Ghats biodiversity hotspot (8°N–21°N) (Myers *et al.*, 2000) is a 1600-km-long mountain chain, which runs all along the west coast of the Indian peninsula (Fig. 1). These mountains (300–2700 m a.s.l.) stretch across an area exceeding 100,000 km^2, although there is considerable variation in how different authors define its precise ecological boundaries, and hence, its exact area (Das *et al.*, 2006; CEPF, 2007; Gadgil *et al.*, 2011). For the purpose of this study, we chose the ecological boundary defined by CEPF (2007) (Western Ghats Portal www.thewesternghats.in/map) within which our 89439.43 km^2 study area extended from Kanyakumari in Tamil Nadu as the southern limit to Mahabaleshwar in Maharashtra as the northern limit (Fig. 1 Inset, outlined in black). We used forest ranges (hereinafter called sites), which are standard forest administrative units both within and outside protected areas as our sampling units. Our study area contained 395 sites [mean area (±SE) = 226.43 km^2 (± 9.15)]. We carried out key informant interviews within each site between April 2008 and February 2012.

We conducted 1760 structured interviews in the local language with 2318 knowledgeable key informants comprising Forest Department field personnel (610), forest-dwelling local communities (1680) and experts (28) to generate detection histories for 30 species of large vertebrates (body mass > 2 kg; Table 1). Non-experts provided data (1–10 detections/non-detections) for 394 sites while experts provided data (1–6 detections/non-detections) for 303 sites. Twenty-one experts provided data for more than one site. Our final combined dataset from non-experts and experts comprised 3–14 replicates (detection/non-detection data) in each site [mean replicates/site (±SE) = 7.46 (± 0.14)]. 398 of 1760 interviews were conducted as group interviews. We ensured that key informants comprising each group were all from the same category, that is, either Forest Department field personnel or forest-dwelling local communities. Experts were not interviewed in groups. We were careful not to average detections reported by multiple informants during group interviews. Rather, if any one informant in a group reported detecting a species, it was recorded. Thus, different informants in a group could report detections of different species. Consequently, group interviews functioned in a manner similar to individual interviews at a species by species level, that is, detection probability for each species in a group interview was a function of an individual informant in the group rather than the group as a whole. Table 1 shows the replication obtained, that is, the total number of detections and non-detections for each species by each key informant group.

Field interviews were conducted in an identical manner to the methods detailed in Pillay *et al.* (2011). However, despite exercising appropriate care to ensure quality and reliability of data (see Pillay *et al.*, 2011), we acknowledge that there will remain a nonzero probability of misidentification error entering into detection histories generated from key informant interviews. This nonzero false-positive probability may be caused by various factors beyond the control of surveyors. These include: uncertainty about the actual extent of the boundaries of a site, deliberate falsification of detections of species even if not encountered, and inadvertent errors on the part of the surveyor in perfectly discerning true detections from inadvertent/deliberate false detections in all interviews.

Analysis

We used recently developed false-positive occupancy models (Miller *et al.*, 2011), which incorporate additional information about the degree of certainty in a detection and allow data from multiple survey/detection methods to be modelled together. Miller *et al.* (2011) describe two general single-season modelling approaches to estimate occupancy when false-positive detections occur. The first approach, the multiple detection state model, is suitable for cases wherein a single detection method is used which may result in two types of detections: uncertain (may include false positives) and certain (does not include false positives). The second approach, the multiple detection method model, tackles cases wherein multiple survey/detection methods are used and each method differs in the degree of certainty that a given detection is true. For instance, in a case where two methods are used, the first may include false-positive detections whereas the second is assumed not to contain false positives. In our example, we use the multiple detection method model and consider interviews of non-experts as the first (uncertain) method where detections may include false positives. Interviews of experts comprise the second (certain) method where detections are assumed not to contain false positives.

Following the notation of Miller *et al.* (2011), we use p_{11} to denote the true-positive detection probability for non-experts (uncertain data) and r_{11} to denote the true-positive detection probability for experts (certain data). We denote false-positive detection probability using p_{10} for non-experts

and assume this probability to be zero for experts. Following standard notation, we use ψ to denote the occupancy probability for a site. In the full model described by Miller *et al.* (2011), all of these parameters are estimated. The standard occupancy estimator of MacKenzie *et al.* (2002) can be seen as a special case of the full model where p_{10} is fixed to be zero for all key informant groups. Thus, both full and standard models can be estimated using the same basic likelihood equation, making it possible to compare the two estimators using AIC (Burnham & Anderson, 2010).

We tested Prediction 1A by fitting standard occupancy models (MacKenzie *et al.*, 2002) separately to uncertain (non-expert) and certain (expert) data. To test Prediction 1B, we fitted and compared five models for each species: three standard models in which false positives were assumed not to occur and two full models in which false positives were explicitly modelled. For the scenario involving standard models, the first model was parameterized to have equal true-positive detection probability for all three key informant groups [$\psi(.)$, $p_{11}(.) = r_{11}(.)$, p_{10}(fixed = 0)]. The second model was parameterized to have equal true-positive detection probabilities for both non-expert key informant groups but different true-positive detection probability for experts [$\psi(.)$, $p_{11}(.)$, $r_{11}(.)$, p_{10}(fixed = 0)]. The third model in this scenario was parameterized to have different true-positive detection probabilities for both non-expert key informant groups as well as for experts [$\psi(.)$, p_{11}(FD $\neq$ LC), $r_{11}(.)$, p_{10}(fixed = 0)]. For each of these three standard estimators, false positives were assumed not to occur and thus p_{10} was fixed to zero. When false positives were explicitly modelled using the full estimator, the first model was parameterized to have equal true- and false-positive detection probabilities for both non-expert key informant groups [$\psi(.)$, $p_{11}(.)$, $r_{11}(.)$, $p_{10}(.)$], while the second was parameterized to have different true and false-positive detection probabilities for both non-expert key informant groups [$\psi(.)$, p_{11}(FD $\neq$ LC), $r_{11}(.)$, p_{10}(FD $\neq$ LC)]. In both of the above full models, true-positive detection probability (r_{11}) using the second method (certain detections from experts) was allowed to differ from the non-expert key informant groups. Finally, we tested Prediction 2 by comparing the precision and bias of occupancy estimates generated by full models to the precision and bias of occupancy estimates generated by standard models fitted to (1) only certain data and (2) combined certain and uncertain data.

We fitted all models in the maximum-likelihood framework of inference using the software PRESENCE 4.4 (Hines, 2006) called using R (v. 2.15.1; R Development Core Team, 2012). We used AIC (Burnham & Anderson, 2010) to choose among alternative parameterizations of models allowing and not allowing for false positives.

RESULTS

Occupancy estimates for non-experts and experts were consistent with Prediction 1A that false positives occur in the non-expert data. For 26 of 30 species, mean occupancy estimates generated from standard estimators fitted to uncertain data were greater than mean occupancy estimates generated from the same estimators fitted to certain data (Fig. 2). The probability of this occurring by chance was < 0.0001 ($\chi^2 = 16.13$, 1 d.f.). For 18 of the 30 species, 95% confidence intervals did not overlap for the approach. On average, occupancy estimates were 0.16 greater for non-experts than for experts (SD = 0.19), providing strong support for Prediction 1A. Certain data from experts were sparse for some species (Table 1). In some (e.g. FHA, MPH, IGH and MUG) although not all (e.g. HYN and NTR) cases, this led to very low precision of occupancy estimates (Fig. 2).

Comparisons of statistical models using the combined dataset were also consistent with false positives occurring in

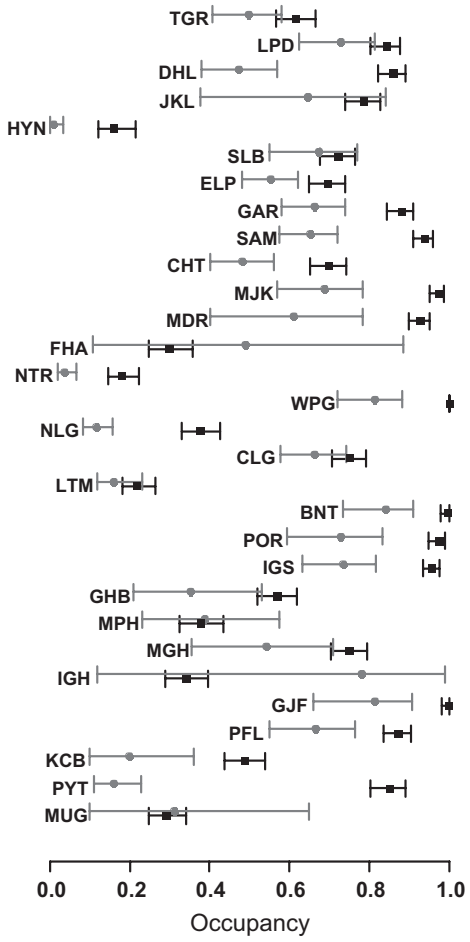

Figure 2 To test the prediction that unaccounted for false-positive detections among non-expert observations would bias estimates high (Prediction 1A), we used standard occupancy models to compare occupancy estimates based on only expert observations (grey) to estimates based on only non-expert observations (black). Consistent with the prediction, 26 of 30 species had greater occupancy estimates from the non-expert data than the expert data. Error bars represent 95% confidence intervals.

the non-expert data. When unaccounted for, these detection errors caused systematic overestimation of occupancy with estimates from the standard model being 0.09 greater, on average, than those from the full model (SD = 0.09). For 29 of 30 species, a full model, where false positives were accounted for, had the lowest AIC score (Table 2). Figure 3 shows variation in occupancy estimates between the most parsimonious full and standard models. For all species, occupancy estimates generated from full models are systematically lower than those generated from standard models. The consistent support for models that accounted for false positives and systematically higher occupancy estimates when false positives were not accounted for indicate strong support for Prediction 1B.

Figure 4 compares the precision and potential bias of occupancy estimates generated by full models to the precision and

Table 2 ΔAIC values for competing models used to estimate occupancy of 30 species of large vertebrates in the Western Ghats. The standard occupancy model only accounted for false-negative errors (MacKenzie *et al.*, 2002), while the full occupancy model accounted for both false-positive and false-negative errors (Miller *et al.*, 2011). In all but one case, the best model with the lowest Akaike's information criterion (AIC) value used the full occupancy parameterization (Prediction 1B). Both uncertain and certain datasets were used to run all standard and full models for testing Prediction 1B

	ΔAIC				
	Standard occupancy model*			Full occupancy model†	
Species	Model 1	Model 2	Model 3	Model 1	Model 2
TGR	130.19	70.59	61.23	1.06	0.00
LPD	197.15	63.39	45.48	18.59	0.00
DHL	589.15	124.64	124.26	0.00	2.55
JKL	492.58	151.64	145.79	52.00	0.00
HYN	83.72	36.00	18.00	20.43	0.00
SLB	249.67	52.72	54.69	9.26	0.00
ELP	288.09	108.16	108.95	0.00	1.51
GAR	381.72	109.14	110.24	0.00	2.43
SAM	411.03	59.20	55.51	0.00	1.84
CHT	421.19	138.02	138.51	12.49	0.00
MJK	739.76	72.33	66.63	11.81	0.00
MDR	937.53	54.93	48.06	11.53	0.00
FHA	161.10	48.05	49.99	16.93	0.00
NTR	62.28	44.60	46.02	0.00	1.98
WPG	908.33	28.36	39.79	18.24	0.00
NLG	126.76	43.74	45.74	3.55	0.00
CLG	360.21	45.42	39.70	3.31	0.00
LTM	35.67	18.61	17.96	7.89	0.00
BNT	781.28	51.09	51.79	0.00	1.44
POR	612.44	146.14	51.69	100.30	0.00
IGS	595.32	67.98	69.60	0.00	0.38
GHB	350.87	21.24	16.44	2.91	0.00
MPH	210.03	59.38	35.64	40.11	0.00
MGH	611.17	112.87	77.19	47.48	0.00
IGH	169.20	30.07	2.40	26.93	0.00
GJF	1152.53	22.08	21.93	38.39	0.00
PFL	769.86	80.57	65.23	16.06	0.00
KCB	385.25	74.52	39.20	46.01	0.00
PYT‡	767.77	30.41	0.00	118.20	10.21
MUG	271.27	106.80	108.46	14.62	0.00

*Model 1 [$\psi(.)$, $p_{11}(.) = r_{11}(.)$, $p_{10}(\text{fixed} = 0)$] – true-positive detection probability equal across all three key informant groups; Model 2 [$\psi(.)$, $p_{11}(.)$, $r_{11}(.)$, $p_{10}(\text{fixed} = 0)$] – true-positive detection probability equal across both non-expert key informant groups but differed for experts; Model 3 [$\psi(.)$, $p_{11}(\text{FD} \neq \text{LC})$, $r_{11}(.)$, $p_{10}(\text{fixed} = 0)$] – true-positive detection probability differed between all three key informant groups.

†Model 1 [$\psi(.)$, $p_{11}(.)$, $r_{11}(.)$, $p_{10}(.)$] – true-positive detection probability equal across both non-expert key informant groups but differed for experts. False-positive detection probability equal across both non-expert key informant groups; Model 2 [$\psi(.)$, $p_{11}(\text{FD} \neq \text{LC})$, $r_{11}(.)$, $p_{10}(\text{FD} \neq \text{LC})$] – true-positive detection probability differed between all three key informant groups. False-positive detection probability differed between both non-expert key informant groups.

‡This was the only species for which the standard occupancy model that did not account for false positives had a lower AIC value than the full occupancy model that did account for them.

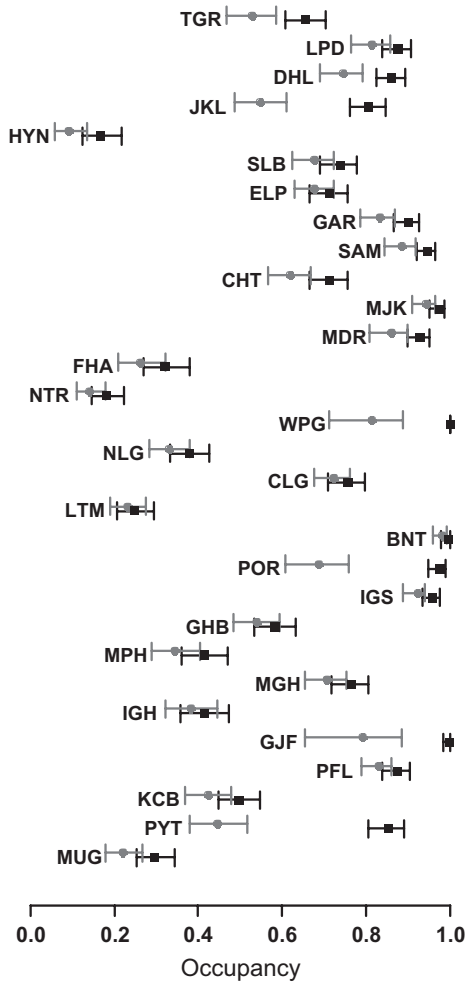

Figure 3 To test the prediction that occupancy estimates from full occupancy models should be lower than those from standard occupancy models after correcting for false positives (Prediction 1B), we compared occupancy estimates from full (grey) and standard occupancy models (black). Consistent with the prediction, all species had lower estimates of occupancy from full occupancy models. Error bars represent 95% confidence intervals. Both uncertain and certain datasets were used to run all standard and full models for testing Prediction 1B.

potential bias of occupancy estimates generated by standard models fitted to (1) only certain data and (2) combined certain and uncertain data. When the standard estimator was used with only certain data, we generally observed similar occupancy estimates to the full model. However, standard errors were smaller on average when the full model was used, in some cases dramatically so. In the case where we used the standard estimator with the combined data, precision was similar to the full model. This result is likely to be misleading, however, as there was evidence of systematic overestimation of occupancy by the standard estimator. Overall, these results affirmed Prediction 2 that including uncertain data will improve the accuracy of occupancy estimates.

The estimates of true- and false-positive detection probabilities for the most parsimonious full and standard occupancy models are summarized in Table 3.

DISCUSSION

Species site occupancy data are integral to several key areas of research in conservation biogeography such as species inventories and mapping surveys, species distribution modelling, understanding distributional dynamics (colonization, extinction, persistence) and conservation planning (Richardson & Whittaker, 2010). While species site occupancy datasets collected from public surveys have enormous potential to address crucial issues pertaining to conservation biogeography over large spatial and temporal scales (Devictor *et al.*, 2010), parameter estimates and inferences from these undertakings may be biased by uncertainty due to two types of observational errors: non-detection and misidentification. Occupancy models were initially developed to address the issue of non-detection or false negatives, while an implicit assumption was that misidentification or false-positive errors did not occur (Royle & Link, 2006). The integration of occupancy modelling with detection histories generated from key informant interviews is becoming an increasingly popular and economical method to estimate species distributions (Karanth *et al.*, 2009; Zeller *et al.*, 2011) and range dynamics (Karanth *et al.*, 2010; Pillay *et al.*, 2011) over broad spatio-temporal extents. This popularity of public surveys to generate species site occupancy data is likely to increase in the future as they represent cost-effective tools in conservation biogeography (Devictor *et al.*, 2010). We are aware of several studies that are currently in progress and are using similar methods to collect occupancy data. While all possible precautions must be taken by surveyors in the field to ensure quality and reliability of data from key informant interviews and other types of public surveys, it is important to acknowledge and statistically account for false-positive errors that such data may include despite all precautions.

Our estimates of naïve occupancy are much higher for uncertain data than for certain data (Table 1), possibly indicating the presence of false positives in the former dataset. However, naïve estimates have to be interpreted with caution because they are likely to reflect underlying biases in detection probabilities of different key informant groups. To obtain realistic occupancy estimates, statistical estimators need to distinguish heterogeneity in true- and false-positive detection probabilities among sites by incorporating additional information about the detection process such as multiple survey methods and the degree of certainty in a detection method (Miller *et al.*, 2011). We devise a simple protocol for assessing the reliability of detection histories from key informant interviews by classifying detection histories into certain and uncertain categories conditional on the key informant group they were obtained from. We show that standard occupancy models fitted to uncertain data lead to a higher estimate of the proportion of area occupied by a species

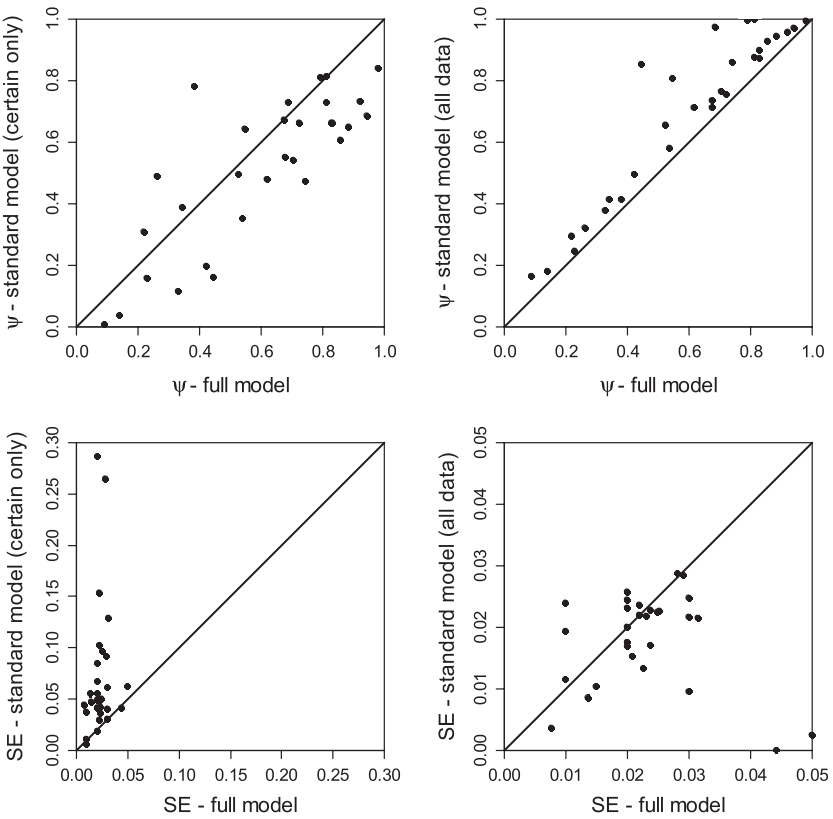

Figure 4 We compared occupancy estimates (ψ) and standard errors (SE) for the full model to cases where only certain data were included and where combined data were included with the standard model where false positives are assumed not to occur. When uncertain data were excluded from the analysis, standard occupancy estimates were similar to those from the full estimator. However, standard errors were much larger indicating a loss of precision due to smaller sample sizes. Alternatively, when false positives were unaccounted for using combined data with the standard estimator, precision was similar or better to the full model. However, there was evidence of a systematic bias leading to overestimation of occupancy.

compared with estimates from the same models fitted to certain data. This indicates a higher probability of false positives in uncertain data, which in turn lead to overestimation of the proportion of sites occupied by a species and introduce substantial bias even after false negatives are accounted for by standard models.

We then demonstrate that integrating detection histories from key informant interviews with full occupancy models (Miller *et al.*, 2011) has greater support (lower AIC) and substantially improves occupancy estimates as opposed to fitting standard models (MacKenzie *et al.*, 2002) to unclassified data. Statistical models for assessing occupancy with data containing false positives have previously been developed (Royle & Link, 2006) but have only been applied in a limited manner. The statistical estimator we use here (Miller *et al.*, 2011) has greater utility in dealing with uncertain detections by allowing multiple survey methods and distinguishing heterogeneity in true- and false-positive detection probabilities among key informant groups to generate more reliable occupancy estimates.

Traditional occupancy modelling approaches would either throw out uncertain data to meet the assumption of zero false-positive errors and generate estimates with only certain data or use both data types and an estimator that does not account for false positives. The former approach reduces the precision of occupancy estimates as, in many cases, certain data are likely to be more difficult and/or costly to obtain and therefore sparse. The latter approach violates a key assumption of the standard estimator and can cause bias in occupancy estimates. We demonstrate that including uncertain data and using the full model adds value in terms of improving the precision of occupancy estimates, as opposed to discarding such data to estimate occupancy with only sparse certain data and the standard model. While our estimates from the standard model with combined data are as precise as those from the full model, the standard model systematically overestimates occupancy compared with the full model (Fig. 4), which is consistent with estimator bias. The fact that occupancy estimates for some species with sparse certain data (Table 1) had very low precision when only certain data were modelled with the standard estimator (Fig. 2) is consistent with our point that including uncertain data from non-experts is useful. As long as false positives are explicitly accounted for in a statistical estimator that can distinguish between true- and false-positive detections among sites (Royle & Link, 2006; Miller *et al.*, 2011), including uncertain data are likely to yield much more precise occupancy estimates than simply relying on certain data (Miller *et al.*, 2011).

In most cases, the estimated false-positive error rate is relatively small and reasonable (Table 3). However, for four species (highlighted in Table 3), this value appears to be similar to or greater than the true-positive detection probability. These species are very common in the Western Ghats and have been reported widely. Consequently, estimated occupancy for these species approaches 1 when only the uncertain data are used and thus there are few or no unoccupied sites to

Table 3 Parameter estimates for true and false-positive detection probabilities. We present estimates from the standard occupancy model (accounts for false negatives only) and the full occupancy model (accounts for both false positives and negatives). Estimates of detection probability (95% confidence intervals) are provided for the most parsimonious [i.e. model with the lowest Akaike's information criterion (AIC)] standard and full model for each species. Both uncertain and certain datasets were used to run all standard and full models for testing Prediction 1B

	Standard occupancy model			Full occupancy model				
	True-positive detection probabilities			True-positive detection probabilities			False-positive detection probabilities	
Species	FD	LC	EX	FD	LC	EX	FD	LC
TGR	0.60 (0.55–0.65)	0.50 (0.47–0.53)	0.33 (0.29–0.37)	0.61 (0.57–0.64)	0.61 (0.57–0.64)	0.39 (0.35–0.44)	0.06 (0.04–0.09)	0.06 (0.04–0.09)
LPD	0.72 (0.68–0.76)	0.61 (0.59–0.64)	0.39 (0.35–0.42)	0.75 (0.71–0.79)	0.65 (0.62–0.68)	0.41 (0.37–0.45)	0.17 (0.09–0.28)	0.05 (0.02–0.13)
DHL	0.74 (0.70–0.77)	0.70 (0.68–0.73)	0.23 (0.20–0.27)	0.79 (0.77–0.82)	0.79 (0.77–0.82)	0.27 (0.24–0.31)	0.13 (0.10–0.18)	0.13 (0.10–0.18)
JKL	0.47 (0.43–0.52)	0.55 (0.52–0.58)	0.12 (0.10–0.15)	0.48 (0.41–0.55)	0.71 (0.67–0.74)	0.17 (0.13–0.20)	0.28 (0.21–0.37)	0.10 (0.07–0.13)
HYN	0.09 (0.04–0.21)	0.40 (0.33–0.47)	0.03 (0.01–0.09)	0.06 (0.01–0.30)	0.60 (0.49–0.70)	0.05 (0.02–0.14)	0.01 (0.00–0.04)	0.01 (0.00–0.02)
SLB	0.65 (0.63–0.67)	0.65 (0.63–0.67)	0.32 (0.28–0.35)	0.68 (0.63–0.72)	0.70 (0.67–0.72)	0.33 (0.29–0.37)	0.10 (0.06–0.16)	0.02 (0.01–0.04)
ELP	0.81 (0.79–0.83)	0.81 (0.79–0.83)	0.49 (0.45–0.53)	0.84 (0.82–0.86)	0.84 (0.82–0.86)	0.51 (0.47–0.56)	0.02 (0.01–0.04)	0.02 (0.01–0.04)
GAR	0.77 (0.76–0.79)	0.77 (0.76–0.79)	0.43 (0.39–0.47)	0.82 (0.80–0.84)	0.82 (0.80–0.84)	0.46 (0.42–0.50)	0.12 (0.08–0.18)	0.12 (0.08–0.18)
SAM	0.84 (0.81–0.87)	0.80 (0.77–0.82)	0.43 (0.39–0.46)	0.84 (0.83–0.86)	0.84 (0.83–0.86)	0.45 (0.42–0.49)	0.19 (0.13–0.26)	0.19 (0.13–0.26)
CHT	0.70 (0.68–0.73)	0.70 (0.68–0.73)	0.31 (0.28–0.35)	0.73 (0.68–0.77)	0.79 (0.76–0.81)	0.34 (0.30–0.38)	0.11 (0.07–0.16)	0.03 (0.02–0.06)
MJK	0.77 (0.73–0.80)	0.82 (0.80–0.84)	0.29 (0.26–0.32)	0.78 (0.74–0.81)	0.85 (0.83–0.87)	0.30 (0.27–0.33)	0.22 (0.07–0.50)	0.11 (0.05–0.25)
MDR	0.70 (0.66–0.74)	0.77 (0.74–0.79)	0.13 (0.11–0.15)	0.73 (0.68–0.76)	0.81 (0.79–0.84)	0.14 (0.11–0.16)	0.18 (0.09–0.32)	0.13 (0.07–0.21)
FHA	0.46 (0.41–0.50)	0.46 (0.41–0.50)	0.10 (0.06–0.14)	0.45 (0.35–0.56)	0.56 (0.50–0.62)	0.12 (0.08–0.17)	0.06 (0.03–0.10)	0.01 (0.00–0.02)
NTR	0.68 (0.62–0.73)	0.68 (0.62–0.73)	0.34 (0.22–0.48)	0.81 (0.75–0.86)	0.81 (0.75–0.86)	0.41 (0.27–0.57)	0.01 (0.00–0.01)	0.01 (0.00–0.01)
WPG*	0.96 (0.94–0.97)	0.92 (0.91–0.93)	0.41 (0.37–0.44)	0.97 (0.95–0.98)	0.91 (0.89–0.93)	0.49 (0.44–0.54)	0.91 (0.80–0.96)	0.97 (0.92–0.99)
NLG	0.77 (0.74–0.80)	0.77 (0.74–0.80)	0.35 (0.27–0.44)	0.84 (0.78–0.88)	0.86 (0.81–0.90)	0.40 (0.31–0.50)	0.03 (0.02–0.06)	0.01 (0.00–0.02)
CLG	0.76 (0.71–0.80)	0.83 (0.80–0.84)	0.44 (0.40–0.48)	0.79 (0.74–0.83)	0.85 (0.82–0.86)	0.45 (0.41–0.49)	0.02 (0.01–0.06)	0.04 (0.02–0.07)
LTM	0.64 (0.57–0.71)	0.57 (0.51–0.62)	0.40 (0.33–0.48)	0.68 (0.59–0.75)	0.60 (0.54–0.66)	0.42 (0.34–0.50)	0.01 (0.01–0.03)	0.00 (0.00–0.34)
BNT	0.89 (0.88–0.90)	0.89 (0.88–0.90)	0.38 (0.35–0.42)	0.90 (0.89–0.91)	0.90 (0.89–0.91)	0.39 (0.36–0.43)	0.18 (0.07–0.38)	0.18 (0.07–0.38)
POR*	0.60 (0.54–0.65)	0.80 (0.78–0.82)	0.28 (0.24–0.31)	0.72 (0.65–0.78)	0.70 (0.67–0.73)	0.36 (0.32–0.41)	0.18 (0.08–0.33)	0.97 (0.93–0.99)
IGS	0.82 (0.80–0.84)	0.82 (0.80–0.84)	0.36 (0.32–0.39)	0.85 (0.83–0.86)	0.85 (0.83–0.86)	0.37 (0.33–0.41)	0.11 (0.06–0.20)	0.11 (0.06–0.20)
GHB	0.59 (0.54–0.63)	0.66 (0.63–0.70)	0.13 (0.10–0.17)	0.62 (0.56–0.67)	0.70 (0.66–0.74)	0.14 (0.11–0.18)	0.02 (0.01–0.07)	0.02 (0.01–0.04)
MPH	0.38 (0.31–0.44)	0.58 (0.53–0.62)	0.15 (0.12–0.19)	0.39 (0.32–0.46)	0.66 (0.61–0.71)	0.17 (0.13–0.22)	0.08 (0.04–0.14)	0.01 (0.01–0.03)
MGH	0.56 (0.51–0.61)	0.73 (0.70–0.76)	0.15 (0.12–0.18)	0.59 (0.53–0.64)	0.79 (0.76–0.82)	0.16 (0.13–0.19)	0.07 (0.03–0.15)	0.03 (0.02–0.05)
IGH	0.28 (0.22–0.35)	0.51 (0.47–0.55)	0.12 (0.09–0.16)	0.30 (0.24–0.37)	0.54 (0.49–0.59)	0.13 (0.10–0.17)	0.00 (0.00–0.60)	0.01 (0.00–0.02)
GJF*	0.90 (0.88–0.92)	0.92 (0.91–0.93)	0.28 (0.25–0.32)	0.96 (0.93–0.98)	0.90 (0.88–0.92)	0.35 (0.30–0.41)	0.48 (0.18–0.79)	0.97 (0.93–0.99)
PFL	0.76 (0.72–0.79)	0.84 (0.82–0.86)	0.28 (0.25–0.31)	0.78 (0.74–0.82)	0.87 (0.85–0.89)	0.28 (0.25–0.32)	0.04 (0.01–0.14)	0.08 (0.04–0.14)
KCB	0.45 (0.40–0.50)	0.65 (0.61–0.69)	0.08 (0.05–0.11)	0.49 (0.43–0.55)	0.72 (0.68–0.76)	0.09 (0.06–0.13)	0.06 (0.03–0.10)	0.01 (0.01–0.03)
PYT*	0.54 (0.47–0.60)	0.73 (0.70–0.75)	0.08 (0.07–0.11)	0.52 (0.44–0.59)	0.32 (0.28–0.36)	0.18 (0.14–0.23)	0.36 (0.27–0.46)	0.87 (0.84–0.90)
MUG	0.52 (0.48–0.56)	0.52 (0.48–0.56)	0.09 (0.06–0.13)	0.57 (0.48–0.65)	0.68 (0.63–0.73)	0.12 (0.08–0.17)	0.04 (0.02–0.06)	0.01 (0.00–0.02)

*In a small number of cases, estimated false-positive detection probabilities were similar or greater than true-positive probabilities. We recommend rejecting the full occupancy model in these cases. See text for more details.

FD, Forest Department field personnel; LC, Forest-dwelling local communities; EX, Experts.

estimate a false-positive error rate. Royle & Link (2006) note that for their estimator, when occupancy is 1, the p_{10} parameter instead explains heterogeneity in true-positive detections just as in a finite mixture model (Royle, 2006). We believe the estimates for these four species reflect a similar process. We recommend that when false-positive detection probabilities are near true-positive detection probabilities, full models should be discarded and inferences should be made using the standard model where the false-positive detection probability is fixed to be 0. In typical cases where most species have occupancy probabilities much lower than 1, this should not be a major issue.

A limitation with our approach is that we do not know truth. We assume that the estimates that explicitly account for false positives have lower bias and that systematic differences observed in the comparison of estimates from the full and standard model with the combined data result from bias due to false positives. However, the results of the comparisons used to test our predictions are consistent with this assumption. Furthermore, work using simulated data shows that not accounting for false positives when they occur does lead to systematic bias (McClintock *et al.*, 2010a; Miller *et al.*, 2011). Our results are also premised on false positives not occurring in detection histories collected from experts (Fitzpatrick *et al.*, 2009; Yu *et al.*, 2010; Molinari-Jobin *et al.*, 2012). We had strong reasons to believe that false-positive errors should be negligible in this key informant group, especially compared with our other groups. However, even experts can be prone to misidentification error (Miller *et al.*, 2012). While benchmarking on experts represents an improvement over completely ignoring false-positive errors, caution may still be advised.

A key obstacle to the accuracy of citizen science-based distribution modelling approaches in informing conservation planning and management (Danielsen *et al.*, 2005; Bonney *et al.*, 2009; Sullivan *et al.*, 2009; Kéry *et al.*, 2010a; Sewell *et al.*, 2010; Pillay *et al.*, 2011) is the inherent likelihood for non-detection, misidentification and misclassification in such datasets (McKelvey *et al.*, 2008; Dickinson *et al.*, 2010; Hochachka *et al.*, 2012). Despite this, analyses of these data rely in many cases on various simplifying assumptions and rarely take into account observation uncertainty. The little prior effort in dealing with false-positive errors has largely depended on approaches to reduce misidentification at the stage of field data collection (Molinari-Jobin *et al.*, 2012). Although it is imperative to make efforts to reduce false positives with proper field sampling protocols, this is unlikely to eliminate errors, thus making it important to estimate and account for false positives as part of statistical analyses (Miller *et al.*, 2012). Advances in statistical methods will continue to play a key role in improving inferences from data collected through various public survey methods where false positives are likely to be a cause for concern. As we demonstrate here (see also Hanks *et al.*, 2011), benchmarking to data collected by taxon experts is one approach to deal with potential misidentification. In many cases, it may not be possible to collect expert observations across large scales. However, it is possible to fit the type of models we fit here when only a subset of sites is sampled using a method that can be assumed to be free from misidentification errors. Uncertain data, in many cases, may be less difficult and/or costly to collect (Molinari-Jobin *et al.*, 2012), and we show that their inclusion along with certain data can improve occupancy estimates when modelled using the full estimator.

Occupancy and detection probability parameters typically vary across a landscape. These parameters can be modelled using landscape and environmental covariates to improve estimates and further reduce bias arising from detection heterogeneity (Miller *et al.*, 2011). Including covariates that predict false-positive detection probabilities may be especially useful for modelling this source of observation error (Miller *et al.*, 2013). We recommend that future work with such datasets explore the use of predictors to further improve estimates. Another potentially fruitful area for exploration is using prior information to estimate false-positive rates using Bayesian approaches. False-positive error rates may be derived from the literature (e.g. Miller *et al.*, 2012) or may be estimated using pilot studies. When studying range dynamics of species undergoing declines in distribution, false positives may lead to biased estimates of the extent of decline in occupancy and in the vital rates of colonization and extinction (McClintock *et al.*, 2010a). This in turn may lead to flawed management decisions. It thus becomes vital to obtain unbiased estimates of occupancy and vital rates, which can reliably inform ecological research and management efforts for species of conservation concern. If a subset of detections is of a type or method wherein false positives can be assumed not to occur, the statistical models we test here are widely applicable to conservation biogeography studies that harness a variety of public survey occupancy datasets prone to both non-detection and misidentification errors.

ACKNOWLEDGEMENTS

We thank the Ministry of Environment and Forests (MoEF), Government of India and the State Forest Departments of Kerala, Tamil Nadu, Karnataka, Goa and Maharashtra for providing research permits. Our work was supported through field research grants provided by the MoEF, National Fish and Wildlife Foundation (Save the Tiger Fund), WWF International and the Rufford Small Grants Foundation. Sasindra Babu, Akbar Ali, Abdul Salam, Dilan Mandanna, H.P. Ashwin, Suhas Wayangankar, Sushil Dixit, Raman Kulkarni, Amit Lale and Dhananjay Joshi provided invaluable assistance in the field. We are grateful to our 2318 key informants for their participation. In particular, we would like to thank the following scientists, professional wildlife photographers/filmmakers and Forest Department officers for providing their detections of species: Mohan Alembath, U.T. Alva, M.K. Appaiah, Nilesh Bapat, John Britto, G.N. Bulgannawar, K.M. Chinnappa, V. Deepak, V. Ganesan, D.V. Girish, P. Gowrishankar, Sanjay Gubbi,

Balachandra Hegde, Senani Hegde, Shrikant Ingalhallikar, Niren Jain, Devcharan Jathanna, A.J.T. Johnsingh, Vishwas Katdare, S. Karthikeyan, Rajendra Kerkar, Nirmal Kulkarni, Ajith Kumar, M. Ananda Kumar, H.N. Kumara, B.S. Krupakar, A.C. Lakshmana, Aaron Lobo, Vijay Mohan Raj, Rajendran Nair, Satish Chandran Nair, N.A. Naseer, S.L. Panaskar, Ankur Patwardhan, Paresh Porob, P. Pugazhendi, R. Raghunath, Parag Ranganekar, Srinivas Reddy, Chandrashekhar Salunke, Nitin Sawant, Mewa Singh, R. Sugathan, Stanley Thekaekara, Roy Thomas, P.C. Tyagi, A. Udhayan, V.K. Uniyal, P.N. Unnikrishnan, Mohan Varghese, A. Veeramani and A.N. Yellappa Reddy. R. Raghunath prepared the map. We thank Brendan Wintle and three anonymous referees for their constructive comments, which greatly improved the quality of the manuscript.

REFERENCES

Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K.V. & Shirk, J. (2009) Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience*, **59**, 977–984.

Brown, J.H. (1984) On the relationship between abundance and distribution of species. *The American Naturalist*, **124**, 255–279.

Burnham, K.P. & Anderson, D.R. (2010) *Model selection and multimodel inference: a practical information–theoretic approach*. Springer-Verlag, New York, NY.

Critical Ecosystem Partnership Fund (CEPF) (2007) *Ecosystem profile: Western Ghats and Sri Lanka Biodiversity hotspot. Western Ghats Region*. Conservation International, Arlington, VA.

Danielsen, F., Burgess, N.D. & Balmford, A. (2005) Monitoring matters: examining the potential of locally-based approaches. *Biodiversity and Conservation*, **14**, 2507–2542.

Das, A., Krishnaswamy, J., Bawa, K.S., Kiran, M.C., Srinivas, V., Kumar, N.S. & Karanth, K.U. (2006) Prioritisation of conservation areas in the Western Ghats, India. *Biological Conservation*, **133**, 16–31.

Devictor, V., Whittaker, R.J. & Beltrame, C. (2010) Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions*, **16**, 354–362.

Dickinson, J.L., Zuckerberg, B. & Bonter, D.N. (2010) Citizen science as an ecological research tool: challenges and benefits. *Annual Review of Ecology, Evolution and Systematics*, **41**, 149–172.

Fitzpatrick, M.C., Preisser, E.L., Ellison, A.M. & Elkinton, J.S. (2009) Observer bias and the detection of low-density populations. *Ecological Applications*, **19**, 1673–1679.

Gadgil, M., Krishnan, B.J., Ganeshaiah, K.N., Vijayan, V.S., Borges, R., Sukumar, R., Noronha, L., Nayak, V.S., Subramaniam, D.K., Varma, R.V., Gautam, S.P., Navalgund, R.R. & Subrahmanyam, G.V. (2011) *Report of the Western Ghats Ecology Expert Panel*. Ministry of Environment and Forests, Government of India, New Delhi.

Garrard, G.E., Bekessy, S.A., McCarthy, M.A. & Wintle, B.A. (2008) When have we looked hard enough? A novel method for setting minimum survey effort protocols for flora surveys. *Austral Ecology*, **33**, 986–998.

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

Hanks, E.M., Hooten, M.B. & Baker, F.A. (2011) Reconciling multiple data sources to improve accuracy of large-scale prediction of forest disease incidence. *Ecological Applications*, **21**, 1173–1188.

Hines, J.E. (2006) *PRESENCE 4. Software to estimate patch occupancy and related parameters*. Available at: http://www.mbr-pwrc.usgs.gov/software/presence.html (accessed 2 August 2012).

Hochachka, W.M., Fink, D., Hutchinson, R.A., Sheldon, D., Wong, W.-K. & Kelling, S. (2012) Data-intensive science applied to broad-scale citizen science. *Trends in Ecology and Evolution*, **27**, 130–137.

Karanth, K.K., Nichols, J.D., Hines, J.E., Karanth, K.U. & Christensen, N.L. (2009) Patterns and determinants of mammal species occurrence in India. *Journal of Applied Ecology*, **46**, 1189–1200.

Karanth, K.K., Nichols, J.D., Karanth, K.U., Hines, J.E. & Christensen, N.L. Jr (2010) The shrinking ark: patterns of large mammal extinctions in India. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 1971–1979.

Karanth, K.U., Gopalaswamy, A.M., Kumar, N.S., Vaidyanathan, S., Nichols, J.D. & MacKenzie, D.I. (2011) Monitoring carnivore populations at the landscape scale: occupancy modelling of tigers from sign surveys. *Journal of Applied Ecology*, **48**, 1048–1056.

Kéry, M., Gardner, B. & Monnerat, C. (2010a) Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, **37**, 1851–1862.

Kéry, M., Royle, J.A., Schmid, H., Schaub, M., Volet, B., Häfliger, G. & Zbinden, N. (2010b) Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Conservation Biology*, **24**, 1388–1397.

Lomolino, M.V., Riddle, B.R., Whittaker, R.J. & Brown, J.H. (2010) *Biogeography*. Sinauer Associates, Sunderland, MA.

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

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006) *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Academic Press, San Diego, CA.

McClintock, B.T., Bailey, L.L., Pollock, K.H. & Simons, T.R. (2010a) Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. *Ecology*, **91**, 2446–2454.

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

McKelvey, K.S., Aubry, K.B. & Schwartz, M.K. (2008) Using anecdotal occurrence data for rare or elusive species: the illusion of reality and a call for evidentiary standards. *BioScience*, **58**, 549–555.

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

Miller, D.A.W., Nichols, J.D., Gude, J.A., Rich, L.N., Podruzny, K.M., Hines, J.E. & Mitchell, M.S. (2013) Determining occurrence dynamics when false positives occur: estimating the range dynamics of wolves from public survey data. *PLoS ONE*, **8**, e65808.

Molinari-Jobin, A., Kéry, M., Marboutin, E., Molinari, P., Koren, I., Fuxjäger, C., Breitenmoser-Würsten, C., Wölfl, S., Fasel, M., Kos, I., Wölfl, M. & Breitenmoser, U. (2012) Monitoring in the presence of species misidentification: the case of the Eurasian lynx in the Alps. *Animal Conservation*, **15**, 266–273.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., Da Fonseca, G.A.B. & Kent, J. (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

Pillay, R., Johnsingh, A.J.T., Raghunath, R. & Madhusudan, M.D. (2011) Patterns of spatiotemporal change in large mammal distribution and abundance in the southern Western Ghats, India. *Biological Conservation*, **144**, 1567–1576.

R Development Core Team (2012) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Richardson, D.M. & Whittaker, R.J. (2010) Conservation biogeography - foundations, concepts and challenges. *Diversity and Distributions*, **16**, 313–320.

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

Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical modeling and inference in ecology. the analysis of data from populations, metapopulations and communities*. Academic Press, San Diego, CA, USA.

Royle, J.A. & Link, W.A. (2006) Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, **87**, 835–841.

Sewell, D., Beebee, T.J.C. & Griffiths, R.A. (2010) Optimising biodiversity assessments by volunteers: the application of occupancy modelling to large-scale amphibian surveys. *Biological Conservation*, **143**, 2102–2110.

Simons, T.R., Alldredge, M.W., Pollock, K.H. & Wettroth, J.M. (2007) Experimental analysis of the auditory detection process on avian point counts. *The Auk*, **124**, 986–999.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D. & Kelling, S. (2009) eBird: a citizen-based bird observation network in the biological sciences. *Biological Conservation*, **142**, 2282–2292.

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

Whittaker, R.J., Araujo, M.B., Jepson, P., Ladle, R.J., Watson, J.E.M. & Willis, K.J. (2005) Conservation biogeography: assessment and prospect. *Diversity and Distributions*, **11**, 3–23.

Yu, J., Wong, W.-K. & Hutchinson, R.A. (2010) Modeling experts and novices in citizen science data for species distribution modeling. *Proceedings of the 2010 IEEE International Conference on Data Mining* (ed. by G. Webb, B. Liu, C. Zhang, D. Gunopulos and X. Wu), pp. 1157–1162. IEEE Computer Society, Los Alamitos, CA.

Zeller, K.A., Nijhawan, S., Salom-Pérez, R., Potosme, S.H. & Hines, J.E. (2011) Integrating occupancy modeling and interview data for corridor identification: a case study for jaguars in Nicaragua. *Biological Conservation*, **144**, 892–901.

BIOSKETCH

As a group, we are interested in the application of conservation ecology and in improving methods to understand processes acting across large geographical and temporal scales. This study emerged from our conversations on the challenges associated with making accurate inferences from public survey data. Our goal was to show that solutions to deal with detection errors in species occupancy data are tractable and can improve estimates of occupancy. We believe public surveys hold much promise for biodiversity assessments, but ample room exists to improve the protocols under which these data are collected and the methods by which they are analysed.

Author contributions: R.P. and M.D.M. conceived the idea for this study; M.D.M., R.P. and A.A.J. obtained the funding for a larger project of which this study is part of; R.P. and A.A.J. carried out field data collection in the form of key informant interviews; R.P., D.A.W.M., J.E.H. and M.D.M. were responsible for data analyses; R.P. wrote the manuscript; D.A.W.M., M.D.M. and A.A.J. contributed to the writing.

Editor: Brendan Wintle