DEALING WITH UNCERTAINTY IN AMPHIBIAN AND REPTILE POPULATION MONITORING FOR CONSERVATION

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Sam Stephen Cruickshank

aus

Großbritannien

Promotionskommission

Prof. Dr. Arpat Ozgul (Vorsitz)

Dr. Benedikt Schmidt (Leitung der Dissertation)

Prof. Dr. Lukas Keller

Prof. Dr. Rolf Holderegger

Zürich, 2017

THIS DISSERTATION IS DEDICATED TO THE MEMORY OF SAM GLOVER

Table of Contents

SUMMARY

Successful conservation management is underpinned by a solid understanding of species distributions and population trends, which is used both to identify populations under threat, and to monitor the effectiveness of management actions. However, as humans are imperfect, observation errors are introduced when we monitor variables such as occupancy patterns or abundances. These errors, if left unaccounted for, can bias inference in ways which can be detrimental to species conservation. In this thesis I used simulations and case studies from amphibian and reptile populations in Switzerland to discuss several forms of observation error, illustrated how their presence may lead to bias, and presented methods by which such biases can be avoided or mitigated.

In *Chapter Two*, I introduced the issue of imperfect detection in species monitoring and explained how occupancy modelling can be used to estimate detection probabilities and thereby accurately assess occupancy rates. In recent years, some authors have questioned the benefit of collecting the extra information necessary to apply this method, and have argued that conservation practise is not improved by accounting for detection. By applying IUCN red-list guidelines to nationwide monitoring data for amphibians within Switzerland, I demonstrated explicitly how failing to account for detection probabilities could lead to inappropriate management decisions being made. 5 of 12 species would have been inappropriately assigned to a higher extinction threat category if detection probabilities were ignored. Using this case study, I highlight that presence-only datasets can only ever be used to calculate maximum possible declines, and are inadequate for estimating the true magnitude of population change. In order to do so, monitoring practise must change to systems in which non-detections as well as species presences are routinely recorded.

In *Chapter Three*, I expanded upon the challenges of designing monitoring programs such that species absence can be reliably inferred. I compared two different frameworks used to determine how much effort must be invested in surveying a site before it can be considered unoccupied, and argued that to correctly interpret a string of non-detections, one must have an expectation of species prevalence. Using nationwide monitoring data for reptiles in Switzerland, I showed that this is problematic because

most species have no natural scale at which prevalence should be assessed; consequently survey recommendations depend strongly on the assumptions made. Our results emphasised that for rare species, it will barely ever be possible to invest sufficient survey effort to ensure that undetected populations are not overlooked, and that by incorporating knowledge of species prevalence, monitoring for invasive species will always cease before the species is truly eradicated.

In *Chapter Four*, I challenged the claim that volunteer-collected data are of low quality by developing a dynamic occupancy model which accounts for false-positive records in addition to imperfect detection. I quantified false-positive error rates for a long-term amphibian monitoring program and demonstrated that false-positives were uncommon for rare species, yet for the most common species monitored, up to 10% of records represented false-positives. I presented guidelines for designing future volunteer programmes such that false-positive records can be readily quantified, but concluded that for our monitoring dataset, ignoring false-positive records would not lead to a quantitative change in occupancy trends for any of the species monitored.

In *Chapter Five*, I called for caution when using count data to conduct large-scale abundance estimation for amphibian populations. I argued that amphibian populations may pose particular difficulties in abundance estimation because i) detection probabilities are often low, and ii) because it is typical that only a proportion of the total population will be available for detection during surveys. Using simulated datasets, I demonstrated that if either availability for detection or detection probabilities themselves are low, abundance estimates from open N-mixture models cannot be relied upon. Furthermore, I developed a new model formulation for scenarios in which availability is expected to vary predictably over a period of time. Using simulating data representing egg-mass counts, I demonstrated that by mechanistically building the availability process into N-mixture models, it is possible to accurately derive population size estimates even when availability is low.

In *Chapter Six*, I focussed on errors within photographic datasets used in mark-recapture studies. Failing to correctly identify individuals from photographs, or wrongly matching individuals, can induce severe bias in capture-recapture datasets. In this chapter, I investigated the effects of WILD-ID, a software designed to aid photographic recognition, upon the false-positive and false-negative error rates of 63 volunteers using a test dataset of photographs of yellow-bellied toads. I found that photographic identification software greatly increased the speed of matching as well as leading to a strong decline in both the frequency of false-negative matches, and in the variation in error rates between volunteers. I used this finding to argue that such software should be routinely used in long-term photographic monitoring programmes in order to minimise inevitable variation in matching abilities caused by staff turnover.

ZUSAMMENFASSUNG

Ein fundiertes Verständnis der Verbreitung von Arten und von Populationsentwicklungen ist für ein erfolgreiches Naturschutzmanagement unerlässlich. Durch die Messung dieser beiden Kenngrössen können gefährdete Populationen zu erkannt und die Wirksamkeit der angewendeten Managementstrategien untersucht werden. Da der Mensch die Natur nicht fehlerfrei beobachtet, entstehen jedoch Erhebungsfehler, wenn Variablen wie Besetzungsmuster von Lebensräumen oder Häufigkeit gemessen werden. Werden solche Fehler ausser Acht gelassen, können Schlussfolgerungen soweit beeinflusst werden, dass Auswirkungen auf den Artenschutz entstehen. In meinen Forschungsarbeit verwendete ich sowohl Simulationen als auch Fallstudien von Schweizer Amphibien- und Reptilienpopulationen. Mit diesen Daten untersuchte ich diverse Formen von Beobachtungsfehlern, die zu Ergebnisverzerrungen führen können und präsentierte Methoden, wie Verzerrungen vermieden oder deren Ausmass gemindert werden können.

In *Kapitel 2* behandelte ich die Thematik der unvollständigen Bestandserfassung in Monitoring-Projekten und erläuterte, wie Modelle zur Habitatbesetzung genutzt werden können, um die Nachweiswahrscheinlichkeit abschätzen und infolgedessen die Verbreitung genau bestimmen zu können. In den letzten Jahre haben verschiedene Autoren den Nutzen der Erhebung zusätzlicher Variablen, die die notwendigen Informationen für die Fehlerkorrektur bereitstellen, in Frage gestellt und argumentiert, dass der angewandte Artenschutz durch das Berücksichtigen durch die Korrektur der Beobachtungsfehler nicht verbessert wird. Durch das Anwenden der IUCN Rote Listen Kriterien auf nationale Monitoring-Daten Schweizer Amphibien stellte ich dar, wie das Vernachlässigen von Nachweiswahrscheinlichkeiten zu Management-Entscheidungen führen kann, welche den Artenschutz schwächen. Fünf der 12 Schweizer Arten wären durch das ausser Acht lassen von Nachweiswahrscheinlichkeiten in eine höhere Gefährdungskategorie eingeteilt worden als dies mit Berücksichtigung der Beobachtungsfehler der Fall war. In dieser Fallstudie hob ich hervor, wie Datensätze, die ausschliesslich Nachweise von Artvorkommen beinhalten, nur für die Messung des höchstmöglichen Populationsrückgangs geeignet sind. Für das Berechnen der effektiven Populationsveränderung erwiesen sich diese Daten jedoch als unzureichend. Um diese Veränderung

exakt berechnen zu können, muss im Artmonitoring zusätzlich zu den Anwesenheitsdaten einer Art auch deren Absenz erhoben werden.

In *Kapitel 3* ging ich ausführlicher auf die Herausforderung ein, wie die Absenz von Arten nachgewiesen werden kann. . Ich verglich zwei unterschiedliche methodische Ansätze, die dazu verwendet werden können, den Beobachtungssaufwand für den Absenznachweis zu bestimmen (d.h. wieviele Nicht-Beobachtungen sind an einem Ort notwendig, dass man mit statistischer Sicherheit sagen kann, dass eine Art an diesem Ort nicht vorkommt). . Zusätzlich betonte ich, dass man wissen muss, wie häufig oder selten eine Art in einer Gegend ist, um den Beobachtungsaufwand für den Absenznachweis bestimmen zu können. . Wir konnten mit nationalen Monitoring-Datensätzen von Reptilien aufzeigen, wie problematisch dies ist, weil es für die meisten Arten keine natürliche Skala der Verbreitung gibt. Folglich sind Beobachtungs- und Monitoringsempfehlungen massgeblich von Annahmen abhängig. Unsere Resultate betonen, dass es in Bezug auf seltene Arten fast unmöglich ist ausreichend Beobachtungsaufwand zu betreiben, um mit Sicherheit sagen zu können, dass eine Art an einem Ort nicht vorkommt und dass kleine Populationen nicht übersehen worden sind. Da ein Absenznachweis für Arten mit kleiner räumlicher Verbreitung kein Absenznachweis möglich ist, hat dies Konsequenzen für das Management von Arten. Will man beispielsweise zeigen, dass eine invasive Art durch Bekämpfungsmassnahmen ausgerottet wurden, kann man das nicht zeigen weil die Modelle empfehlen, gar nicht erst nach seltenen Arten zwecks Absenznachweis zu suchen.

In *Kapitel 4* untersuche ich die Behauptung, dass von Freiwilligen gesammelte Daten von minderer Qualität sein sollen. Ich verwende dazu ein dynamischen Besetzungsmodell, das sowohl „falsch positive“ sowie unvollständige Nachweise miteinbezieht. Als Fallbeispiel diente ein Langzeit-Amphibien-Monitoring. Ich habe die Fehlerquote von „falsch positiven“ Nachweisen geschätzt und aufgezeigt, dass „falsch positive“ Nachweise sporadisch bei seltenen Arten auftauchen. Im Gegensatz dazu liegt die Fehlerquote von „falsch positiven“ Meldungen bei häufigen Arten bei 10% des Gesamtdatensatzes. Infolgedessen habe ich Leitlinien für die Planung und Ausführung von zukünftigen Monitoringprogrammen mit Freiwilligen vorgestelltt, wodurch die „falsch positiven“ Nachweise einfach gemessen werden können. Jedoch habe ich für das Fallbeispiel den Schluss gezogen, dass das

Ignorieren von „falsch positiven“ Nachweisen bei keiner untersuchten Art zu keinem mengenmässig signifikanten Unterschied in den Besiedlungstrends geführt hat.

In *Kapitel 5* habe ich beim Verwenden von Zähldaten zur Schätzun der Populationsgrösse von Amphibienpopulationen zur Vorsicht aufgerufen. Ich argumentiere, dass bei Amphibienpopulationen spezielle Schwierigkeiten bei der Schätzung der Populationsgrösse zu berücksichtigen sind: i) Nachweisswahrscheinlichkeiten sind häufig sehr gering; ii) es ist typisch, dass während eines Monitorings nur ein Bruchteil der Gesamtpopulation erfasst werden kann (d.h. reduzierte „Verfügbarkeit“).

Durch Datensimulation konnte ich zeigen, dass den Schätzwerten von sogenannten „N-Mixture“-Modellen nicht vertraut werden kann, wenn entweder die Verfügbarkeit von Individuen oder die Nachweiswahrscheinlichkeit selbst sehr gering sind. Ferner habe ich ein neues Modell entwickelt, mit dem die sich zeitlich ändernde Verfügbarkeit geschätzt und bei der Populationsgrössenschätzung berücksichtigt werden kann. Durch das Simulieren von Daten, wie sie etwa bei der Zählung von Laichballen von Amphibien generiert werden, konnte ich zeigen, dass es durch mechanistisches Einbauen die Verfügbarkeit in „N-Mixture“-Modelle möglich ist, Populationsgrössen genau schätzen zu können, auch wenn die Verfügbarkeit per se gering ist.

In *Kapitel 6* habe ich mich auf Fehler in Datensätzen, die für Fang-Wiederfang-Studien mit Wiedererkennung von Individuen anhand von Fotos verwendet werden, fokussiert. Werden Individuen auf den Fotos nicht korrekt erkannt oder für ein andere Individuum gehalten, kann dies zu starken Verzerrungen in den Fang-Wiederfang-Schätzungen führen. In diesem Kapitel untersuchte ich, ob die Verwendung von Software (WILD-ID), die zur Unterstützung von Fotoidentifikationen erstellt wurde, die Fehlerrate senkt. 63 Freiwillige haben dazu einen Test-Datensatz bestehend aus Fotos von Gelbbauchunken ausgewertet. Ich bestimmte anschliessend die „falsch positiven“ und „falsch negativen“ Fehlerraten . Ich fand heraus, dass das Verwenden einer Fotoidentifikations-Software die Geschwindigkeit, mit der Wiedererkennung gemacht wird, signifikant erhöht, die Häufigkeit der „falsch Positiven“ gesenkt sowie die Abweichungen in Fehlerraten zwischen Freiwilligen gesenkt werden kann. Dieses Ergebnis unterstützte meine Aussage, dass eine solche Software regelmässig in Langzeit-Fang-

Wiederfangprogrammen, welche Fotos zur Erkennung von Individen verwenden,genutzt werden soll. Dadurch wird die unvermeidbare, durch Mitarbeiterwechsel hervorgerufene Fehlerrate bei der Wiedererkennung von Individuen minimiert. .

CHAPTER ONE

General Introduction

GENERAL INTRODUCTION

The goal of ecology is to explain patterns and interactions between species and individuals, and as such depends fundamentally upon an ability to describe the distribution and abundance of species (Goldsmith 1991). Similarly, effective conservation requires the ability to monitor species and identify declining populations (Lindenmayer *et al.* 2013); indeed the importance of effective monitoring is recognised by international agreements dictating that countries implement monitoring programs (e.g. Convention on Biological Diversity Aichi Targets 17 and 19; Pereira *et al.* 2013; Marques *et al.* 2014). Global databases and indices such as the IUCN Red List (IUCN 2017), the Red List Index (Butchart *et al.* 2006), and the Living Planet Index (Loh *et al.* 2005) all influence policy at global and national scales, but all are constructed by collating data from countless monitoring programs at a global scale. At a more local level, land managers may choose to carry out monitoring in order to assess the success of land management (Block *et al.* 2001), to monitor the success of species reintroductions (Seddon 1999; Seddon *et al.* 2007), to screen for the presence of invasive species (Blossey 1999; Bogich *et al.* 2008), or to carry out adaptive management (McCarthy & Possingham 2007; Runge 2011; Canessa *et al.* 2016). The ability to monitor trends in distribution and abundance accurately is therefore fundamental to our abilities to successfully implement and evaluate conservation action. However, accurately monitoring populations is a challenging task.

When any system is monitored, the resulting data reflects not only the true state of the system, but also errors and uncertainties that arise from the observation process itself (Yoccoz *et al.* 2001; Pollock *et al.* 2002). Ecologists have long recognised that different observers often produce very inconsistent results when measuring the same parameter (Hope-Simpson 1940; Smith 1944; Hayne 1949; Preston 1979), and with such uncertainty it can be challenging to tease out the effects of errors to evaluate the true state. When monitoring, errors can arise as a result of mistakes made by imperfect human observers (e.g. by an observer wrongly identifying a species), or as a result of study designs that fail to adequately account for the behaviour of the system under observation (e.g. surveying a population during a period where some individuals have temporarily emigrated from the site). The recognition that ignoring such errors would likely induce bias led directly to the development of relatively crude adjustments to account for measurement errors (Petersen 1896; Lincoln 1930). Intense methodological developments over the past 50 years, paralleled by massive increases in computing power, have led to a proliferation of modelling frameworks that explicitly deal with observational errors when estimating state variables such as occupancy (MacKenzie *et al.* 2002), abundance (Buckland *et al.* 2001; Royle & Nichols 2003), or demographic rates from capture-mark-recapture (CMR) data (Cormack 1964; Jolly 1965; Seber 1965).

Monitoring programmes focus on evaluating a large number of state variables depending on the goals of the program, the spatial scope of the monitoring, and the resources available (Yoccoz *et al.* 2001). At its most simple, monitoring can ask the question “can the species be seen at a site?”. As one can never guarantee that individuals of a species will be observed, even if present (the ubiquitous situation termed imperfect detection; Kéry & Schmidt 2008), such data does not directly correspond to establishing the presence or absence of a species at a site. However, using occupancy models (MacKenzie *et al.* 2002; Tyre *et al.* 2003) repeated observation/non-observation data can be used to establish the occupancy status of sites while accounting for errors such as imperfect detection. As this form of information is relatively inexpensive to collect, and can therefore be applied at large scales over many sites (MacKenzie *et al.* 2006), occupancy data are increasingly becoming the backbone of many species monitoring programmes (Marsh & Trenham 2008).

Although occupancy data can be useful in monitoring distributional changes (Tingley & Beissinger 2009) or identifying potentially suitable habitat though species distribution models (reviewed in Guisan & Thuiller 2005), in many situations it will be desirable to obtain more detailed information. By monitoring abundances for example, species declines can be picked up before a population becomes locally extinct (MacKenzie & Nichols 2004; Joseph *et al.* 2006; Dorazio 2007). Many different methods have been developed in order to estimate population sizes from count data (Buckland *et al.* 2001; Royle & Nichols 2003; Pollock *et al.* 2004; Royle 2004). When estimating abundance, there are many issues in addition to that of imperfect detection which may need to be addressed to avoid bias. For instance, failing to take into account species phenology or movement patterns can lead to incorrect estimation of population sizes (Nichols *et al.* 2009; Wilson *et al.* 2014).

Towards the more intensive end of the monitoring spectrum, programs may be interested in monitoring the fate of individuals within a population. In addition to allowing abundance estimation, individual based monitoring such as CMR yields information on demographic parameters such as survival and reproductive rates (Nichols 1992), and can inform knowledge of movement patterns (Powell *et al.* 2000; Ovaskainen 2004). However, such methods are very resource-intensive, and as such are typically applied over a smaller spatial scale than may be the case for abundance or occupancy monitoring (Yoccoz *et al.* 2001; Pollock *et al.* 2002). In CMR monitoring, the issue of imperfect detection may be compounded by difficulties in correctly identifying individuals within a population (Morrison *et al.* 2011), which can have drastic effects on the estimation of demographic rates or population size (Stevick & Palsbøll 2001; McDonald *et al.* 2003).

The importance of any of these potentially confounding effects upon the results drawn from monitoring will depend not only on the study system in question, but also upon the objectives of the monitoring program (Yoccoz *et al.* 2001; Pollock *et al.* 2002). It is typically necessary to collect extra data in order to account for uncertainties caused by observation error, and in some situations the amount of extra data required may be high (Banks-Leite *et al.* 2014). Although monitoring is fundamental to conservation efforts, resources for monitoring are generally limited. There is therefore a need to balance the extra resources required to adjust for measurement error against the potential advantages to be gained from doing so (Field *et al.* 2005; Moore & McCarthy 2016). In this thesis I describe several forms of error and uncertainty in monitoring data, and use case studies and simulations to demonstrate the consequences of ignoring uncertainty upon the results of monitoring. I demonstrate explicitly how failing to account for errors can affect conservation management, and present methods or design recommendations to minimise bias. I show that some forms of error are ubiquitous and ignoring them has serious consequences, whereas for others, bias may only arise in certain circumstances, thus whether or not monitoring should account for them will depend on the study system and program goals. I focus on monitoring data relating to amphibians and reptiles in this thesis, however the conclusions drawn and recommendations suggested are equally applicable to monitoring for other taxa.

Thesis Outline

The chapters of this thesis are the result of collaborative efforts, and each is organised as a self-contained manuscript which has been or will be submitted to a peer-reviewed journal. Chapters two to four address issues relating to the analysis of occupancy data, chapter five deals with the analysis of count data, and chapter six is related to error rates in individual CMR data. Finally, in chapter seven I summarise the results of the preceding chapters and use these to draw some more general conclusions regarding species monitoring designs and discuss future directions for species conservation management.

In *Chapter Two*, I focus on the importance of accounting for imperfect detection when monitoring changes in species occupancy rates through time. Although imperfect detection leads to false-negative errors and consequently sites being wrongly classified as unoccupied, it is straightforward to estimate and adjust for detection rates using occupancy models when sites are visited multiple times. However, some authors have previously questioned the value of collecting the information necessary to estimate detection rates (McGill 2012; Welsh *et al.* 2013; Banks-Leite *et al.* 2014). I used nationwide monitoring data for amphibians within Switzerland to examine the conservation implications of failing to consider species detectability by showing that the same data can lead to species being classified under different IUCN extinction threat categories if imperfect detection is ignored. I also discuss the limitations of the revisitation method by which declines are typically calculated, and emphasise that the collection of species non-detection data in addition to species presences is essential in order to overcome these limitations.

In *Chapter Three*, I examine the difficulties in assessing how much effort needs to be invested before concluding that a site is unoccupied by a species. Using nationwide reptile monitoring data, I explore the differences between two frameworks used for determining the number of non-detections necessary in order to confidently classify a site as unoccupied (McArdle 1990; Wintle *et al.* 2005). Knowledge of species prevalence as well as detectability is necessary to correctly interpret a sequence of non-detections, however unless there is natural scale at which prevalence can be assessed, it is challenging to meaningfully take this into account this when planning monitoring. I use results derived under the two frameworks to make general comments about the power of monitoring programs to infer when species are absent, with particular focus on rare species and the limitations of monitoring for invasive species.

In Chapter Four, I develop an occupancy model that accounts for both false-positive errors in addition to imperfect detection while modelling changes in occupancy over time. I apply this model to a citizen science dataset representing 15 years of amphibian monitoring data collected at nearly 650 sites, and use the model to estimate false-positive error rates for each species. Furthermore, I compare the results to those from standard dynamic occupancy models to quantify the extent to which failing to account for false-positive records or imperfect detection would change trends in occupancy. As false-positives and false-negatives (i.e. imperfect together) are the most important forms of error in distributional data, our model therefore represents an elegant method both to account for such errors in volunteer datasets, and to assess the quality of citizen-science data.

In *Chapter Five*, I argue that certain characteristics of populations may pose challenges to assessing population sizes in wide-scale monitoring programs. N-mixture models (Royle 2004) allow for abundance estimation across many sites from repeated counts, however these models assume population closure and may produce unreliable results when detection probabilities are low. Temporary emigration and other factors may reduce the availability of a population during surveys- if not all individuals are present during a survey, the rate at which individuals are detected will be substantially reduced. I explore the consequences of i) reduced but constant availability, and ii) variable availability for detection, upon the reliability of both mechanistic (Dail & Madsen 2011) and phenomenological (Hostetler & Chandler 2015) open N-mixture models, and assess the biological conditions under which population sizes can be reliably quantified.

In *Chapter Six*, I compare two methods used to construct individual capture histories from photographic mark-recapture data. Using a test dataset of photographs of the yellow-bellied toad (*Bombina variegata*), I quantify the rates at which volunteers falsely matched photographs of different individuals, or failed to match images of the same individual. Volunteers completed this exercise both by hand and using photographic matching software (Bolger *et al.* 2012). By comparing differences in error rates between the two methods, I provide some conclusions as to the likely efficiency of each method, and provide some recommendations to minimise errors in long term photographic mark-recapture studies.

Finally, in Chapter Seven, I summarise the main findings from the preceding chapters and draw some final conclusions relating to the design of species monitoring.

Literature Cited

Banks-Leite, C., Pardini, R., Boscolo, D., Cassano, C.R., Püttker, T., Barros, C.S. & Barlow, J. (2014). Assessing the utility of statistical adjustments for imperfect detection in tropical conservation science. *Journal of Applied Ecology*, **51**, 849–859.

Block, W.M., Franklin, A.B., Ward, J.P., Ganey, J.L. & White, G.C. (2001). Design and Implementation of Monitoring Studies to Evaluate the Success of Ecological Restoration on Wildlife. *Restoration Ecology,* **9**, 293–303.

Blossey, B. (1999). Before, during and after: the need for long-term monitoring in invasive plant species management. *Biological Invasions*, **1**, 301–311.

Bogich, T.L., Liebhold, A.M. & Shea, K. (2008). To sample or eradicate? A cost minimization model for monitoring and managing an invasive species. *Journal of Applied Ecology,* **45**, 1134–1142.

Bolger, D.T., Morrison, T.A., Vance, B., Lee, D. & Farid, H. (2012). A computer-assisted system for photographic mark-recapture analysis. *Methods in Ecology and Evolution*, **3**, 813–822.

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. & Thomas, L. (2001). *Introduction to distance sampling*. Oxford University Press, Oxford.

Butchart, S.H.M., Akcakaya, H.R., Kennedy, E. & Hilton-Taylor, C. (2006). Biodiversity Indicators Based on Trends in Conservation Status: Strengths of the IUCN Red List Index. *Conservation Biology,* **20**, 579–581.

Canessa, S., Guillera-Arroita, G., Lahoz-Monfort, J.J., Southwell, D.M., Armstrong, D.P., Chadès, I., Lacy, R.C. & Converse, S.J. (2016). Adaptive management for improving species conservation across the captive-wild spectrum. *Biological Conservation*, **199**, 123–131.

Cormack, R.M. (1964). Estimates of Survival from the Sighting of Marked Animals. *Biometrika*, **51**, 429–438.

Dail, D. & Madsen, L. (2011). Models for Estimating Abundance from Repeated Counts of an Open Metapopulation. *Biometrics,* **67**, 577–587.

Dorazio, R.M. (2007). On the choice of statistical models for estimating occurrence and extinction from animal surveys. *Ecology*, **88**, 2773–2782.

Field, S.A., Tyre, A.J. & Possingham, H.P. (2005). Optimizing Allocation of Monitoring Effort Under Economic and Observational Constraints. *Journal of Wildlife Management*, **69**, 473–482.

Goldsmith, B. (1991). *Monitoring for Conservation and Ecology* (B. Goldsmith, Ed.). Springer

Netherlands, Dordrecht.

Guisan, A. & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Hayne, D.W. (1949). An Examination of the Strip Census Method for Estimating Animal Populations. *The Journal of Wildlife Management*, **13**, 145–157.

Hope-Simpson, J.F. (1940). On the Errors in the Ordinary use of Subjective Frequency Estimations in Grassland. *Journal of Ecology*, **28**, 193–209.

Hostetler, J.A. & Chandler, R.B. (2015). Improved state-space models for inference about spatial and temporal variation in abundance from count data. *Ecology*, **96**, 1713–1723.

IUCN. (2017). The IUCN Red List of Threatened Species. Version 2017-2. URL http://www.iucnredlist.org [accessed 14 October 2017]

Jolly, G.M. (1965). Explicit Estimates from Capture-Recapture Data with Both Death and Immigration-Stochastic Model. *Biometrika*, **52**, 225–247.

Joseph, L.N., Field, S., Wilcox, C. & Possingham, H.P. (2006). Presence-absence versus abundance data for monitoring threatened species. *Conservation Biology*, **20**, 1679–1687.

Kéry, M. & Schmidt, B.R. (2008). Imperfect detection and its consequences for monitoring for conservation. *Community Ecology*, **9**, 207–216.

Lincoln, F.C.F.C. (1930). *Calculating waterfowl abundance on the basis of banding returns*. US Department of Agriculture. Circular No.118, Washington DC.

Lindenmayer, D.B., Piggott, M.P. & Wintle, B.A. (2013). Counting the books while the library burns: Why conservation monitoring programs need a plan for action. *Frontiers in Ecology and the Environment*, **11**, 549–555.

Loh, J., Green, R.E., Ricketts, T., Lamoreux, J., Jenkins, M., Kapos, V. & Randers, J. (2005). The Living Planet Index: using species population time series to track trends in biodiversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 289–295.

MacKenzie, D.I. & Nichols, J.D. (2004). Occupancy as a surrogate for abundance estimation. *Animal Biodiversity and Conservation*, **1**, 461–467.

MacKenzie, D.I., Nichols, J.D. & Lachman, G. (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, New York.

Marques, A., Pereira, H.M., Krug, C., Leadley, P.W., Visconti, P., Januchowski-Hartley, S.R., Krug, R.M., Alkemade, R., Bellard, C., Cheung, W.W.L., Christensen, V., Cooper, H.D., Hirsch, T., Hoft, R., van Kolck, J., Newbold, T., Noonan-Mooney, K., Regan, E.C., Rondinini, C., Sumaila, U.R., Teh, L.S.L. & Walpole, M. (2014). A framework to identify enabling and urgent actions for the 2020 Aichi Targets. *Basic and Applied Ecology*, **15**, 633–638.

Marsh, D.M. & Trenham, P.C. (2008). Current trends in plant and animal population monitoring. *Conservation Biology*, **22**, 647–655.

McArdle, B.H. (1990). When are rare species not there? *Oikos*, **57**, 276–277.

McCarthy, M.A. & Possingham, H.P. (2007). Active adaptive management for conservation. *Conservation Biology*, **21**, 956–963.

McDonald, T.L., Amstrup, S.C. & Manly, B.F.J. (2003). Tag loss can bias Jolly-Seber capture-recapture estimates. *Wildlife Society Bulletin*, **31**, 814–822.

McGill, B. (2012). Statistical Machismo. *Dynamic Ecology Blog*. URL https://dynamicecology.wordpress.com/2012/09/11/statistical-machismo/

Moore, A.L. & McCarthy, M.A. (2016). Optimizing ecological survey effort over space and time. *Methods in Ecology and Evolution*, **7**, 891–899.

Morrison, T.A., Yoshizaki, J., Nichols, J.D. & Bolger, D.T. (2011). Estimating survival in photographic capture-recapture studies: overcoming misidentification error. *Methods in Ecology and Evolution*, **2**, 454–463.

Nichols, J.D. (1992). Capture- Recapture Models: Using marked animals to study population dynamics. *Bioscience*, **42**, 94–102.

Nichols, J.D., Thomas, L. & Conn, P.B. (2009). Inferences About Landbird Abundance from Count Data: Recent Advances and Future Directions. *Modeling Demographic Processes in Marked Populations* (eds D.L. Thomson, E.G. Cooch & M.J. Conroy), pp. 201–235. Springer US, Boston, MA.

Ovaskainen, O. (2004). Habitat-specific movement parameters estimated using mark-recapture data and a diffusion model. *Ecology*, **85**, 242–257.

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory, R.D., Heip, C., Hoft, R., Hurtt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J.P.W., Stuart, S.N., Turak, E., Walpole, M. & Wegmann, M. (2013). Essential Biodiversity Variables. *Science*, **339**, 277–278.

Petersen, C.G.. (1896). *The yearly immigration of young plaice into the Limfjord from the German Sea*. Report of the Danish Biological Station 6.

Pollock, K.H., Marsh, H., Bailey, L.L., Farnsworth, G.L., Simons, T.R. & Alldredge, M.W. (2004). Separating components of detection probability in abundance estimation: an overview with diverse examples. *Sampling Rare or Elusive Species: Concepts, Designs, and Techniques for Estimating Population Parameters* (ed W.L. Thompson), pp. 43–58. Island Press.

Pollock, K.H., Nichols, J.D., Simons, T.R., Farnsworth, G.L., Bailey, L.L. & Sauer, J.R. (2002). Large scale wildlife monitoring studies: statistical methods for design and analysis. *Environmetrics*, **13**, 105–119.

Powell, L.A., Conroy, M.J., Hines, J.E., Nichols, J.D. & Krementz, D.G. (2000). Simultaneous Use of Mark-Recapture and Radiotelemetry to Estimate Survival, Movement, and Capture Rates. *The Journal of Wildlife Management*, **64**, 302–313.

Preston, F.W. (1979). The Invisible Birds. *Ecology*, **60**, 451–454.

Royle, J.A. (2004). N-Mixture Models for Estimating Population Size from Spatially Replicated

Counts. *Biometrics*, **60**, 108–115.

Royle, J.A. & Nichols, J.D. (2003). Estimating abundance from repeated presence-absence data or point counts. *Ecology,* **84**, 777–790.

Runge, M.C. (2011). An Introduction to Adaptive Management for Threatened and Endangered Species. *Journal of Fish and Wildlife Management,* **2**, 220–233.

Seber, G.A.. (1965). A Note on the Multiple-Recapture Census. *Biometrika*, **52**, 249–259.

Seddon, P.J. (1999). Persistence without intervention: assessing success in wildlife reintroductions. *Trends in Ecology and Evolution*, **14**, 503–503.

Seddon, P.J., Armstrong, D.P. & Maloney, R.F. (2007). Developing the science of reintroduction biology. *Conservation Biology*, **21**, 303–312.

Smith, A.D. (1944). A study on the reliability of range vegetation estimats. *Ecology*, **25**, 441–448.

Stevick, P. & Palsbøll, P. (2001). Errors in identification using natural markings: rates, sources, and effects on capture recapture estimates of abundance. *Canadian Journal of Fisheries and Aquatic Science*, **58**, 1861–1870.

Tingley, M.W. & Beissinger, S.R. (2009). Detecting range shifts from historical species occurrences: new perspectives on old data. *Trends in Ecology and Evolution*, **24**, 625–633.

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

Welsh, A.H., Lindenmayer, D.B. & Donnelly, C.F. (2013). Fitting and Interpreting Occupancy Models. *PLoS ONE*, **8**, e52015.

Wilson, T.L., Schmidt, J.H., Thompson, W.L. & Phillips, L.M. (2014). Using double-observer aerial surveys to monitor nesting bald eagles in Alaska: Are all nests available for detection? *Journal of Wildlife Management*, **78**, 1096–1103.

Wintle, B.A., Kavanagh, R., McCarthy, M.A. & Burgman, M.A. (2005). Estimating and dealing with detectability in occupancy surveys for forest owls and arboreal marsupials. *Journal of Wildlife Management*, **69**, 905–917.

Yoccoz, N.G., Nichols, J.D. & Boulinier, T. (2001). Monitoring of biological diversity in space and time. *Trends in Ecology and Evolution*, **16**, 446–453.

CHAPTER TWO

Quantifying population declines based on presence-only records for red-list assessments

Sam S. Cruickshank [1,*], Arpat Ozgul [1], Silvia Zumbach [2], Benedikt R. Schmidt [1,2]

Published as:

Cruickshank, S.S., Ozgul, A., Zumbach, S., and Schmidt, B.R. (2016) *Quantifying population declines based on presence-only records for red-list assessments*. **Conservation Biology** 30(5): 1112-1121

---

[1] *Department of Evolutionary Biology and Environmental Studies, University of Zürich, Zürich, Switzerland*
[2] *Info Fauna karch, Neuchâtel, Switzerland*
*sam.cruickshank@ieu.uzh.ch

ABSTRACT

Accurate trend estimates are necessary for understanding which species are declining and which are most in need of conservation action. Imperfect species detection may result in unreliable trend estimates because this may lead to the overestimation of declines. As many management decisions are based on population trend estimates, such biases could have severe consequences for conservation policy. We used an occupancy modelling framework to estimate detectability and calculate nationwide population trends for 14 Swiss amphibian species both accounting for and ignoring imperfect detection. Through application of IUCN Red List criteria to the different trend estimates, we assessed whether ignoring imperfect detection could impact on conservation policy. Imperfect detection was seen in all species and detection varied substantially among species, which led to the overestimation of population declines when detectability was ignored. Consequently, accounting for imperfect detection lowered the red-list risk category for 5 of the 14 species assessed. We demonstrate that failing to consider species detectability can have serious consequences for species management, and that occupancy modelling provides a flexible framework to account for observation bias and improve assessments of conservation status. A problem inherent to most historical records is that they contain presence-only data from which only relative declines can be estimated. A move towards the routine recording of non-observation and absence data is essential if we are to move beyond this towards accurate population trend estimation.

Keywords :

amphibian, extinction risk, occupancy, detection probability, conservation, management, presence-only data, trend, monitoring

INTRODUCTION

The reliable assessment of trends in the distribution and abundance of species is a major concern for conservationists. Without an accurate measure of population trends it is challenging to identify those species most in need of conservation action or to measure the success of conservation action and policy (Butchart et al. 2010; Hoffmann et al. 2010). Given the large number of imperilled species and limited conservation funds, some sort of triage and priority setting is necessary (Bottrill et al. 2008, 2009). The use of red lists based on International Union for Conservation of Nature (IUCN) criteria to fill such a role has increased over past decades because the methods underpinning the red-listing process are transparent and quantitative and thus represent an objective method of assigning extinction threat (Lamoureux et al. 2003, Rodrigues et al. 2006, Collen et al. 2013). Red lists describe the extinction risk of species at the global scale or at smaller geographic scale (e.g., a continent or country). One of the red-list criteria used to determine extinction risk is magnitude of population decline over a defined period; species having undergone relatively greater declines receive higher red-list status (Gärdenfors et al. 2001, Mace et al. 2008). Yet, even though the quantification of declines is fundamental for red listing in particular and biodiversity conservation in general, it still remains a challenge (e.g., Yoccoz et al. 2001, Skelly et al. 2003, Tingley & Beissinger 2006).

To illustrate problems that conservation scientists and wildlife managers are likely to encounter when they attempt to quantify the decline of species in an area, we focused on changes in patch occupancy (i.e., the loss of local populations), a commonly used proxy for population declines. Patch occupancy is a state variable commonly used in monitoring programs and metapopulation ecology (Moilanen 2002, Joseph et al. 2006, Kéry & Schmidt 2008) because it is easy to collect these data relative to other metrics (e.g., abundance data [MacKenzie and Nichols 2004]). Trends in patch occupancy are often quantified using so-called revisitation studies (Drost & Fellers 1996, Fischer & Stöcklin 1997, Kéry et al. 2006). Researchers first select sites where the species was recorded in the past from biodiversity databases. Resurveys at these sites are then undertaken to detect whether the species is currently present or absent. Declines can then be estimated as

$$D = 1 - \frac{R}{H} \tag{2.1}$$

where R is the number of occupied sites in the resurvey and H is the number of sites selected where the species was known to exist in the past (Kéry et al. 2006). Unfortunately, Eq. 2.1 cannot be used directly to estimate decline because both R and H are usually measured with error. The most common error is imperfect detection (i.e., false negatives or error of omission) because detection probabilities are usually <1 (Preston 1979, Pollock et al. 2002, Kéry & Schmidt 2008). Even though detection error is the rule rather than the exception (Kellner & Swihart 2014), there is still debate as to whether consideration of this bias is necessary or even desirable in population monitoring and conservation decision making (Banks-Leite et al 2014, Guillera-Arroita et al 2014).

We used data collected for an update of the Swiss Amphibian Red List (Schmidt & Zumbach 2005) to demonstrate problems associated with the estimation of population trends. We examined variation in detection probabilities among 14 Swiss amphibian species in order to assess the bias induced by ignoring imperfect detection when estimating population trends. We also examined how these biases can be avoided using occupancy models (MacKenzie et al. 2002, Tingley & Beissinger 2006) to adjust for imperfect detection and provide more reliable estimates of decline. Furthermore, we highlight how issues relating to imperfect detection can readily influence conservation decision-making through the assigning of inappropriate extinction threat (i.e., Red List category) to species. Finally, we considered the inherent problem that historical records typically contain presence- only data and why such data sets can only be used to calculate a relative, rather than absolute magnitude of decline. We argue that this fundamental problem can be avoided in the future with a shift from the collection of presence-only biodiversity data to databases that also routinely record species non-detections.

METHODS

Data collection

Guidelines for the IUCN Red List permit the categorization of extinction risk based on the magnitude of population declines over 10 years or 3 generations (Mace et al. 2008, IUCN Standards and Petitions Subcommittee 2014). We quantified population decline (i.e. IUCN criterion A) as a change in the number of occupied sites throughout Switzerland, measured at the scale of individual amphibian breeding sites (typically a wetland or pond). We therefore assigned red-list categories to species on the basis of changes in the number of occupied ponds, which can be used as an index of abundance (Buckland et al 2005, IUCN Standards and Petitions Subcommittee 2014).

The distribution of all pond-breeding amphibian species was thoroughly surveyed in Switzerland in the 1970s and 1980s, when experienced herpetologists mapped all known amphibian breeding sites and recorded which species occurred at those sites (Grossenbacher 1988). The resulting database contains data on all amphibian species native to Switzerland and is maintained and updated by KARCH (the Swiss Amphibian and Reptile Conservation Programme, part of the Centre Suisse de Cartographie de la Faune CSCF). The database contains over 12,000 amphibian breeding sites and over 160,000 records (Schmidt & Zumbach 2016).

Sites from this database were selected for resurvey as follows. For each species, the number of the sites was randomly selected for resurvey on the basis of the species' rarity and range. Twenty sites were selected for 9 species with >400 known breeding sites, 25 sites for 3 species (*Triturus cristatus*, *Lissotriton vulgaris*, *Rana dalmatina*) with <400 known breeding sites, and 12 sites for 2 species that occur only in the southern region of Switzerland, Canton Ticino (*Triturus carnifex*, *Hyla intermedia*). This resulted in 289 resurveyed sites (Fig 2.1), for which the historical records dated back on average to 1993 (sd 9).

Although individual sites were selected for inclusion based on records of an individual species, sites were commonly occupied by multiple species. Thus, for every species, we had more presence records in the sample than the number of sites selected specifically for that species. For each species, the data set analysed included all information from all sites that were located within the biogeographic regions (Gonseth et al. 2001) in which the species was known to occur and that were within the known elevational range of the species in Switzerland (Grossenbacher 1988). Thus, all sites where a species could potentially have been present were included; individual sites were assigned a binary covariate to describe the presence or absence of historical observations of species presence (Table 2.1).

For the resurvey, each site was visited four times from March to June in either 2003 or 2004 in order to account for differences in breeding phenology between species. One-hour visits were conducted in which the surveyor used multiple methods (visual encounter surveys, call surveys, dip netting) to collect detection and non-detection data for all pond-breeding amphibians. Data for all life stages (eggs, tadpoles, juveniles, adults) were pooled for analysis. Some of the sites scheduled for resurvey had been destroyed, predominantly as a consequence of agricultural intensification and urban development. In such cases the site was recorded as destroyed and a replacement site was surveyed (selected again using the above criteria). The number of destroyed sites is shown in Table 2.1.

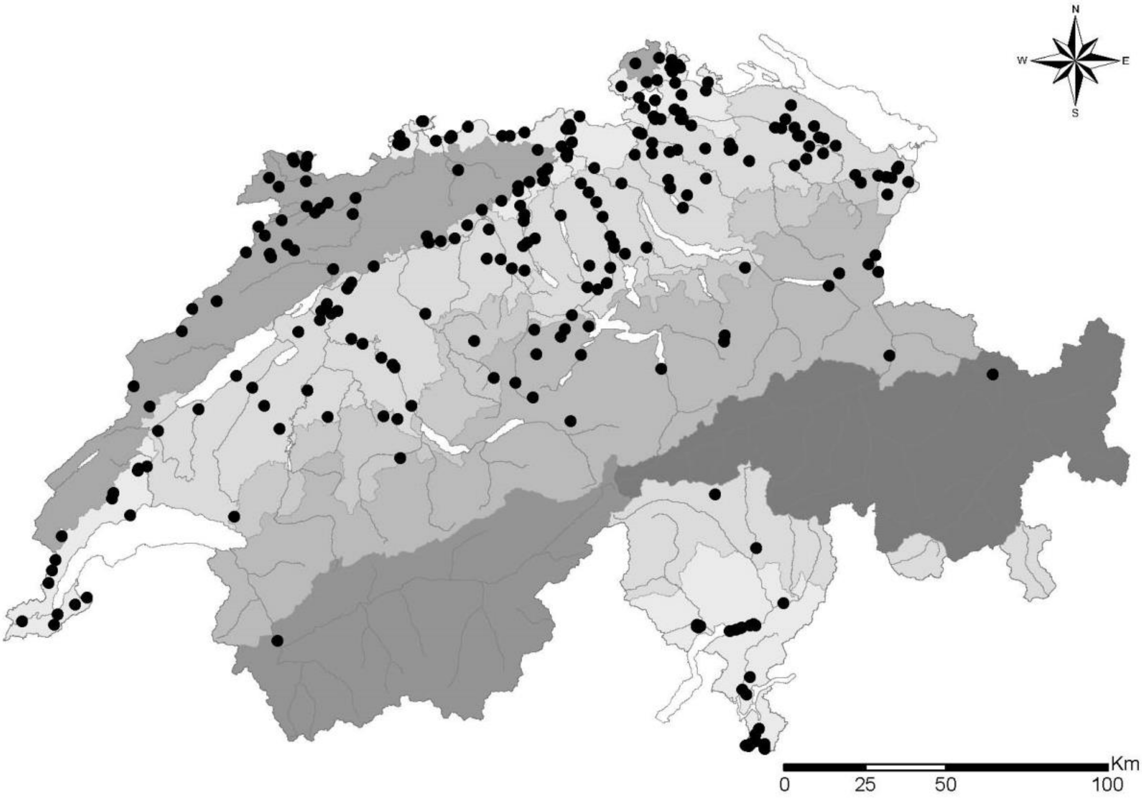


Figure 2.1. Location of the 289 sites in Switzerland included in the resurvey for the Swiss Red List update. Map shading shows the major biogeographic regions within the country. © Background of map: Swisstopo

Table 2.1. Rates of persistence and apparent colonization estimated from the occupancy model, two measures of decline, and the degree of bias induced by ignoring detectability for 14 species of amphibian found in Switzerland.

Species	Sites surveyed [a]	Sites destroyed [b]	Persistence probability (95% CI) [c]	Apparent colonization probability (95% CI) [d]	Unadjusted decline [e]	Adjusted decline [f] (95% CI)	Bias [g] (%)
Alytes obstetricans (A.obst)	212	7	0.526 (0.412-0.641)	0.040 (0.017-0.080)	0.541	0.517 (0.412-0.622)	4.6
Bombina variegata (B.vari)	211	14	0.507 (0.391-0.627)	0.093 (0.045-0.152)	0.625	0.561 (0.457-0.661)	11.4
Bufo bufo (B.bufo)	277	15	0.807 (0.728-0.876)	0.420 (0.323-0.521)	0.324	0.261 (0.198-0.333)	24.1
Epidalea calamita (E.cala)	197	13	0.503 (0.344-0.671)	0.040 (0.012-0.082)	0.661	0.602 (0.469-0.728)	9.8
Hyla arborea (H.arbo)	190	1	0.516 (0.378-0.655)	0.081 (0.038-0.137)	0.548	0.492 (0.356-0.628)	11.4
Hyla intermedia (H.inte)	42	1	0.975 (0.785-1.000)	0.732 (0.192-1.000)	0.265	0.055 (0.029-0.238)	381.8
Ichthyosaura alpestris (I.alpe)	277	15	0.904 (0.847-0.952)	0.365 (0.272-0.461)	0.209	0.173 (0.129-0.225)	20.8
Lissotriton helveticus (L.helv)	212	8	0.742 (0.635-0.842)	0.233 (0.159-0.314)	0.367	0.319 (0.226-0.417)	15.0
Lissotriton vulgaris (L.vulg)	253	1	0.547 (0.387-0.713)	0.030 (0.010-0.060)	0.536	0.463 (0.301-0.621)	15.8
Pelophylax esculentus complex (P.escu)	253	1	0.834 (0.763-0.895)	0.254 (0.166-0.355)	0.224	0.171 (0.110-0.241)	31.0
Rana dalmatina (R.dalm)	228	1	0.742 (0.603-0.877)	0.026 (0.007-0.055)	0.344	0.270 (0.137-0.407)	27.4
Rana temporaria (R.temp)	277	15	0.918 (0.870-0.957)	0.685 (0.559-0.799)	0.176	0.144 (0.108-0.189)	22.2
Triturus carnifex (T.carn)	51	3	0.695 (0.402-0.993)	0.061 (0.001-0.234)	0.556	0.330 0.047-0.614)	68.5
Triturus cristatus (T.cris)	211	3	0.449 (0.305-0.599)	0.007 (0.000-0.025)	0.617	0.573 0.431-0.710)	7.7

[a] Number of sites used to calculate population declines.

[b] Number of sites destroyed before the resurvey and replaced in the analysis with equivalent sites.

[c] Probability of population survival at a site between historic observations and the red-list resurvey.

[d] Probability that a site with no historic observations of the species was estimated to be occupied during the resurvey.
[e] Population decline ignoring the influence of imperfect detection.
[f] Estimated population decline following incorporation of imperfect detection in the contemporary resurvey.
[g] Bias induced by ignoring detection: the degree to which population declines were overestimated when imperfect detection ignored.

Data analyses

Two estimates of decline were calculated: unadjusted declines, which ignored imperfect detection, and adjusted declines, which were calculated by estimating and adjusting for species detection probabilities. Unadjusted declines were calculated using a version of Eq. 2.1 modified to incorporate the effects of the destruction of some sites (and thus loss of those populations):

$$D_{\text{unadjusted}} = 1 - \frac{N_{\text{obs}}}{H(1+L)}\,, \tag{2.2}$$

where L is the proportion of historically occupied sites that were destroyed and N_{obs} is sites where the species was observed both during the surveys for the red-list update and in the historical records.

To estimate detection and site-occupancy probabilities, we fitted site-occupancy models (MacKenzie et al. 2002, Royle & Kéry 2007, Kéry & Schaub 2012) to each species data set. The data y_{ij} consisted of binary detection and non-detection indicators at each of i sites and j visits. The model was written as

$$z_i \sim \text{Bernoulli}(\Psi_i), \tag{2.3}$$

$$y_{ij} \mid z_i \sim \text{Bernoulli}\,(z_i p)\ \ , \tag{2.4}$$

$$\text{logit}(\Psi_i) = \alpha + \beta X_i\ \ . \tag{2.5}$$

Equation 2.3 defines the latent (i.e. true) state of occurrence of the species in site i as a Bernoulli trial with success parameter Ψ_i (occupancy probability). Equation 2.4 defines the observation data as a Bernoulli trial, where the success parameter is a product of the latent occurrence state and detection probability p. Although we fixed p as a species-specific constant for simplicity and clarity of interpretation, this parameterisation could also be expanded to incorporate site- and visit-specific covariates. We used a logit link function to model the relationship between current occupancy and a binary covariate X_i , which describes the presence or absence of a historical record of species presence at the site (Eq. 2.5). The α and β are regression coefficients. Given the data and the model, α is the proportion (on the logit scale) of sites where the species was not recorded in the past but was estimated to occur now, which we refer to as an apparent colonization (see Discussion). $\alpha + \beta$ is the proportion (on the logit scale) of sites where the species was recorded in the past and is estimated to still occur at the site. It is an estimate of the probability of persistence at a site (hereafter, psi.persist). Declines adjusted for imperfect detection and site destruction were calculated using the measure of site persistence, psi.persist. This measure therefore accounted for both site destruction and imperfect detection as follows:

$$D_{adjusted} = 1 - \frac{\text{psi.persist}}{1+L} \quad . \qquad (2.6)$$

To meet the closure assumption of site-occupancy models, species must be available for sampling during each site visit (MacKenzie et al. 2002). For each species we truncated the data set to consider only the site visits carried out between the first and last observation (inclusive) for that species across all sites. This data truncation reduced the number of site visits in a species-specific manner but ensured that only survey data that occurred during the active period of each species was used. This data truncation reduced our data set to a mean across species of 3.23 site visits, except for *Hyla arborea*, *Hyla intermedia* , and *Triturus carnifex* for which the mean (across the three species) was 1.76 site visits.

We used the equation of McArdle (1990) to estimate per-visit detection probability $\hat{p}$ and then calculated the cumulative probability of detecting a species $p*$ after 1-4 visits:

$$p* = 1 - (1 - \hat{p})^n , \qquad (2.7)$$

where n is the number of site visits. This equation assumes that the species was present at the site (see Wintle et al. [2012] for the more general case that does not condition on species presence). Calculating adjusted decline and $p*$ explicitly as part of model fitting allows direct calculation of the uncertainty in parameter estimation.

Modelling was carried out in WinBUGS through the R package R2WinBUGS (Kéry & Schaub 2012). We used vague uniform priors for all model parameters, detection probability p~ dunif(0,1), and the coefficients α and β ~dunif(-10,10). Due to convergence problems, wider priors (dunif[-20,20]) were necessary for *Hyla intermedia*. Three Markov chains of 10,000 iterations were run for each model. The first 1000 iterations were discarded as burn-in and the remaining thinned by 1 in 10. We assessed convergence with the Brooks-Gelman-Rubin statistic (Gelman et al 2004).

RESULTS

Per-visit detection probabilities were well below 1 and varied among species, ranging from 0.50 (95% credible interval [CI] 0.39-0.61) (*Lissotriton vulgaris*) to 0.85 (95% CI0.73-0.93) (*Hyla arborea*) (Fig. 2.2a). Cumulative detection probabilities after multiple visits to the same site (analogous to the proportion of occupied sites detected) approached 1 by the fourth site visit (Table 2.1, Fig. 2.2b). For most species, three visits were required for 95% confidence of detecting a species if it was present. For the two tree frog species (*H. arborea* and *H. intermedia*), only two visits were necessary, whereas for *L. vulgaris* even four visits were not sufficient to reach this level of confidence (95% credible interval for *L. vulgaris* after 4 surveys: 0.86-0.98 [Fig. 2.2b]). With each consecutive visit, 95% CI shrank; thus, for most species after four visits it was possible to be highly confident that the species was detected if present.

All species exhibited some losses from sites where they had been present in the past; as such declines were recorded for all species irrespective of the method used to calculate the trend (Fig. 2.3). Declines did not appear to show any spatial patterns, except for *Alytes obstetricans*. In this species, declines were more common in eastern Switzerland, which is at the species' range edge. Unadjusted declines, whereby sites were considered occupied if the species was observed in one of the four surveys, ranged from 17.6% (*Rana temporaria*) to 66.1% (*Epidalea calamita*) (Table 2.1). When mapped to red-list criteria, these measurements resulted in 10 of 14 species being listed as threatened, 3 as vulnerable, and 7 as endangered (Fig. 2.3).

However, declines adjusted for detection probability were of a consistently lower magnitude than those not accounting for imperfect detection (Fig. 2.3, Table 2.1) (range 14.4% [*Rana temporaria*] to 59.9% [*Epidalea calamita*]). The difference between declines adjusted and not adjusted for imperfect detection was mostly small, however. Nevertheless, adjusting for detection probability resulted in changes to the ranking of species declines. For five species, the two measures of decline were sufficiently large that under IUCN Red List guidelines the different estimates would result in classification of the species under different threat categories (Fig. 2.3). These species were typically those with the lowest probabilities of detection; however, *H. arborea*, the species with the highest detection probability, was also one of those species for which the extinction risk was overestimated when detection probability was unaccounted for. Credible intervals for adjusted declines may encompass decline values higher than those observed. This is because the estimate of decline is a probability that relates not only to the sample (i.e., the sites that were surveyed) but to all populations from which the sample was drawn. Finite sample inference, as described by Royle and Kéry (2007), would lead to credible intervals that do not exceed observed decline values

For all species, a number of observations were recorded in the resurveys at sites where the species was previously not known. From these data, the occupancy model estimated an occurrence probability (α in equation 2.5) that was an apparent colonization probability (i.e., the probability that a site that was previously recorded as unoccupied in the past transitioned to an occupied state). These apparent colonization probabilities were generally higher for common species than for rare and more strongly declining species (Table 2.1). Strong negative relationships were seen between apparent colonisation probability and both measures of decline (r -0.78 and -0.81 for unadjusted and adjusted declines respectively, both $p<0.01$).

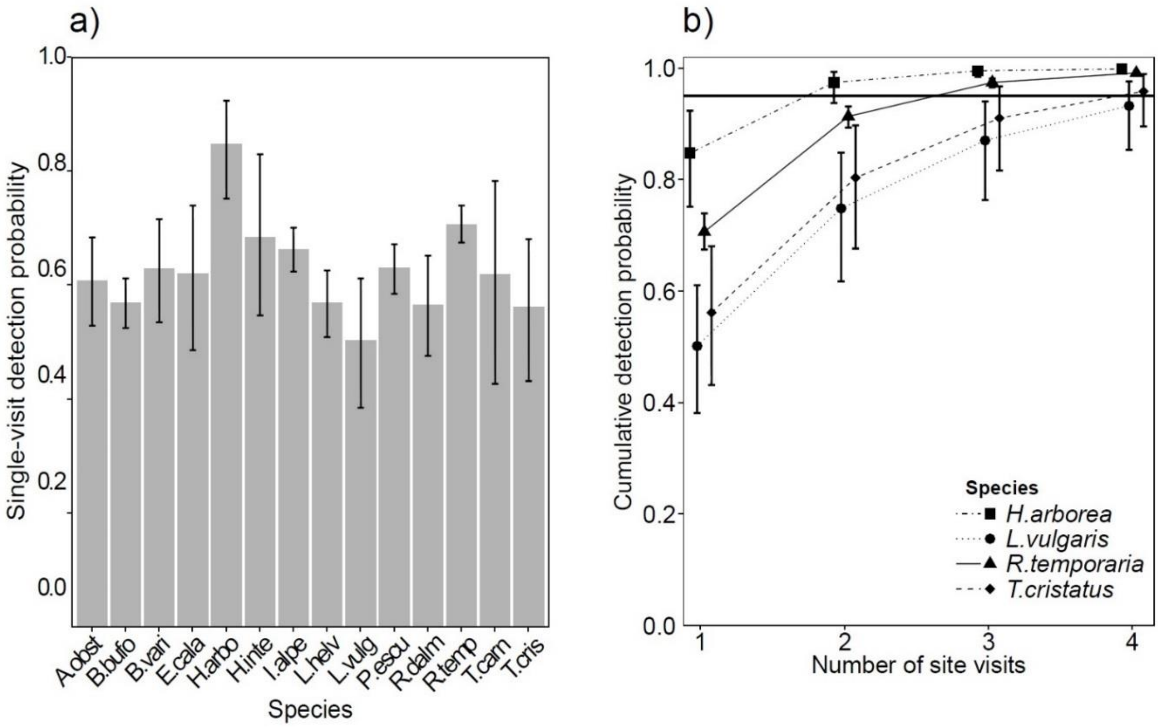


Figure 2.2. (a) . Per-visit detection probabilities and 95% credible intervals for 14 amphibian species at sites throughout Switzerland. *P.escu* represents the *Pelophylax esculentus* species complex. Species abbreviations as shown in Table. 2.1 (b) Cumulative detection curves and 95% CI for 4 amphibian species over 4 survey occasions (solid horizontal line, point at which one could have 95% confidence of detecting the species at a given site if it was present). *H. arborea* and *L. vulgaris* had the highest and lowest detection probabilities respectively.

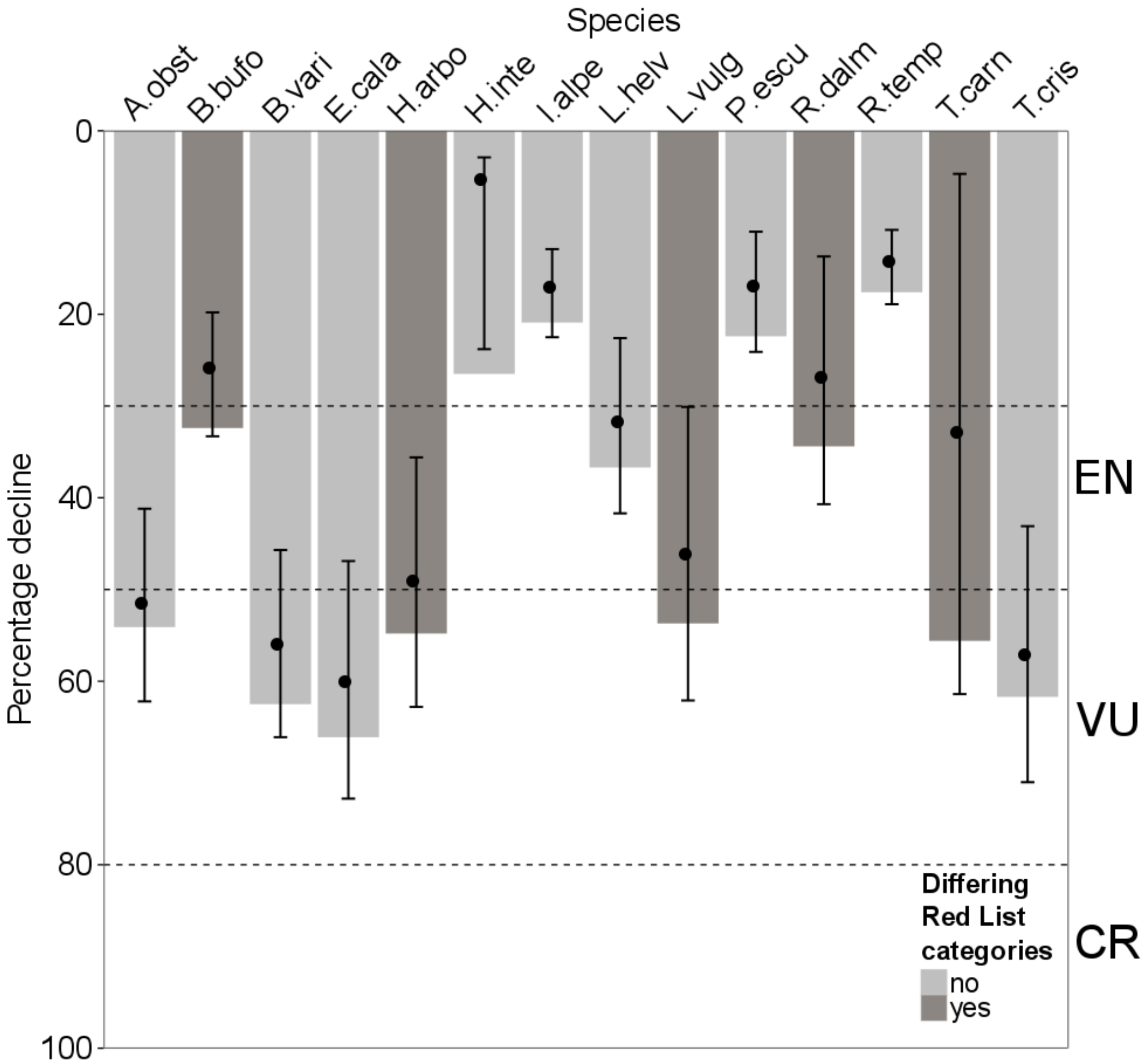


Figure 2.3. Declines in site occupancy as estimated between the time of collection of historical records and the contemporary resurvey in 2003 and 2004 (bars, unadjusted estimates of decline based on whether the species was observed at the site at any of the 4 survey occasions; points, declines adjusted for the detection probability of each species as estimated from the occupancy model; solid lines , 95% credible intervals); horizontal dashed lines, levels of decline required to warrant an International Union for Conservation of Nature classification of critically endangered [CR] [decline >80%], endangered [EN] [decline of 50-80%], or vulnerable [VU] [decline of 30-50%]).Shading represents whether declines accounting for imperfect detection result in classification of a species into a different threat category to declines calculated assuming perfect detection. Species abbreviations as shown in Table. 2.1 *P.escu* represents the *Pelophylax esculentus* species complex.

DISCUSSION

The methods that we used to calculate species trends and the assumptions, both implicit and explicit, we make when estimating changes over time have an unsurprising but often dramatic effect on the outputs of species monitoring. We focused specifically on one form of bias common to most wildlife monitoring, that of imperfect detection, and demonstrated that failing to consider such bias can easily lead to assignment of the wrong red-list category. Such inappropriate assignment may lead to management practices that fail to have the intended effect.

Our method can be extended to other taxa as well as to multiple time periods. Extending the approach to multiple periods with dynamic occupancy models (MacKenzie et al. 2003) would be a way test for a reduction or change in the rate of biodiversity loss (as for example required by the Convention on Biological Diversity). Furthermore, one could use covariates describing land use or management to understand the variation in the magnitude of trends and the factors causing population declines (e.g., Miller et al. 2012). In Switzerland one of the main reasons for the large population declines (Fig. 2.3) was the lack of temporary ponds (Schmidt and Zumbach 2005). As a consequence, pond creation has been implemented and has had promising results (Schmidt et al. 2015).

We found that detection probabilities were variable among species that were surveyed simultaneously and that detection probabilities for most amphibians tended to be substantially <1. Given the large number of studies that have estimated detection probabilities of amphibians and other animals and plants (e.g., Kéry & Schmidt 2008, Sewell et al. 2012, Chen et al 2013), these results were not unexpected. Highly vocal tree frogs had the highest detection probabilities and cryptic newts the lowest; otherwise, species traits had no apparent effect on patterns of variation in detection among species (Garrard et al. 2013). Furthermore, very similar surveys (i.e., no obvious differences in the methods sections of the papers) can result in substantially different detection probabilities. *Epidalea calamita* had a detection probability of 0.62 (95% CI 0.48 - 0.74), whereas in other studies of this species with similar survey methods, Schmidt (2005) and Pellet and Schmidt (2005) calculated the detection probability as 0.27 and 0.44 respectively (no standard errors were provided in these studies). It therefore appears that detection probabilities are not simply species or method specific. This suggests that one cannot assume a fixed value of detection probability, but that one has to estimate detection probability in every survey.

Although cumulative detection probability for some species can be very high (> 0.95) after multiple visits to a site (Fig. 2.2b), there can still be a difference between estimates of decline adjusted and not adjusted for imperfect detection. This difference can lead to an overestimation of decline and therefore species may be assigned to the wrong red-list category (e.g., the species *Hyla arborea*, *Bufo bufo*, *Rana dalmatina*, *Lissotriton vulgaris* and *Triturus cristatus*) (Fig. 2.3). Because red lists are used for conservation priority setting, conservation effort (e.g., the list of species of national priority in Switzerland [BAFU 2011]) may be allocated sub-optimally and focus on the wrong species. Thus, ignoring imperfect detection can directly impact conservation management. We therefore recommend adjusting for imperfect detection whenever possible. There may be financial or other constraints that prevent the collection of the most reliable data, but we believe one should always strive to obtain the best data (The Rolling Stones 1969). Some concerns over adjusting for imperfect detection do not seem justified because it is not always necessary to visit all sites multiple times and space-for-time substitutions may lower survey costs (MacKenzie et al. 2002, Kendall & White 2009). There are also methods that allow the joint analysis of different types of survey data (Dorazio 2014), which could also act to reduce the burden of collecting information to estimate detection.

Our estimations of population declines are based on presence records of species. That is, we know that the species was present in the recent past, and we asked whether the species was still there at the time of a resurvey or whether a local extirpation occurred. Such revisitation studies are commonly used for estimating trends, but this method is flawed because it can only measure declines. Local extirpations may be part of a turnover process in a metapopulation (Hanski 1991, Hecnar & M'Closkey 1996); thus, the losses we describe may have been compensated for by colonizations elsewhere. We included sites in our survey where the species were not recorded in the past so that we could measure colonizations (i.e., the proportion of those sites where the species was present in the resurvey) (Table 2.1). Colonization probabilities were negatively correlated with the magnitude of declines, suggesting that for the declining species, colonization of new sites was not sufficient to compensate for declines and thus declines are real rather than a methodological artefact. (There is evidence for this interpretation from later resurveys [Lupi 2015]). Importantly, colonizations are not necessarily true colonizations of previously unoccupied sites. Although the historic status of sites where observations were recorded is certain, there may be uncertainty associated with those sites where historic presence was not recorded. It may be that the species was already present in the past but was never observed and recorded in a database; thus, the apparent colonization is instead the discovery of a previously unnoticed population at a site. The true status of sites where a species was not recorded in the past is therefore uncertain.

If a distribution database holds only presence records, then one can in most cases only estimate a relative decline. Conceptually, this can be shown by rewriting Eq. 2.1 and acknowledging that we do not know R and H. We only know the sites where the species was detected and reported: $C_R = R * p_R$ and $C_H = H * p_H$, where C_i is the observed number of occupied sites and p_i is detection probability for the historical data and the resurvey. This leads to a new equation:

$$\text{decline} = 1 - \frac{C_R}{C_H} = 1 - \frac{R * p_R}{H * p_H} \quad . \qquad (2.7)$$

Typically, only C_R and C_H are available for the estimation of decline. Using the C_i for decline estimation is problematic because one has to assume $p_R = p_H$. If $p_R \neq p_H$, then the estimate of decline is biased (Pollock et al. 2002). In a resurvey, one can estimate detection probabilities such that one obtains $\hat{R}$, an estimate of R. The equation changes to

$$\text{decline} = 1 - \frac{\hat{R}}{H * p_H} \quad . \qquad (2.8)$$

This is an improvement, but because p_H is unknown and cannot be known, one can only estimate a relative decline. The expression $H * p_H$ implies there are sites where the species did occur in the past but was not recorded. Thus, some putative absences are in fact presences (this is related to the problem of apparent colonizations discussed above). Certainly, an estimate of decline can be based on C_H but with such presence-only data one can only measure a relative decline. In most cases, a decline will be overestimated. Furthermore, without ways of estimating historical detection, the sites selected for resurvey in revisitation studies may be affected by selection bias because larger populations are more likely to be detected (Tanadini & Schmidt 2011) and less likely to be extirpated (Alpizar-Jara et al. 2004). Trends derived from databases biased in this way could therefore suggest that declines are of a lower magnitude than the true trend based on all sites that were occupied in the past because the local extirpation of small populations may not be noticed. The problem outlined here applies equally to estimates of historical abundance (estimates are often not adjusted for imperfect detection [Nichols 1992]).

Uncertainty about historical absences and abundances may be the reason many indices of biodiversity decline start at an arbitrary value of 1 (e.g., the Living Planet Index [Loh et al. 2005] or estimates of global amphibian population declines [Houlahan et al. 2000]). There is a similar problem with presence-only data in species distribution modelling. With presence-only data, one can estimate only relative occurrence not absolute probability of occurrence (or species prevalence [Hastie & Fithian 2013]). Put simply, if $y = a + b*x$ describes the relationship between species occurrence, y, and a habitat characteristic, x, then one can estimate only the coefficient b while a cannot be estimated. For some applications, relative habitat suitability may be sufficient, but for others this is unlikely to be the case (Guillera-Arroita et al. 2015). What all this illustrates is that one often cannot learn from presence-only data what one would like to learn about patterns of biodiversity and changes therein.

We showed how estimation of detection probabilities can improve the estimation of declines and lead to better extinction risk assessments for conservation. If estimates of extinction risk are based on records that show where the species was known in the past, then estimates of extinction risk and decline remain relative. Similar problems exist for other indices of trends and species-distribution models. Sometimes it is possible to combine data sets from multiple species or surveys to arrive at unbiased estimates (Kéry et al. 2010, Sadoti et al. 2013). The better solution would be to complement observations of where species were observed with data on where species were not observed. One approach may be that observers should complete species lists rather than observations of a single species (Isaac & Pocock 2015). Such non-observations are not evidence for the absence of species but when analysed using appropriate statistical models (MacKenzie et al. 2002), such data would nevertheless allow the unbiased estimation of absolute population decline, extinction risk, and species distributions. Ultimately, the lack of accurate historical baselines hinder current efforts to accurately monitor trends; a change toward the routine recording of non-observations and absence data is required to stop these problems hindering future monitoring of biodiversity trends.

ACKNOWLEDGMENTS

We thank the many herpetologists who collected the data for the red-list update, as well as the two anonymous reviewers who provided valuable feedback on this manuscript. The update of the Swiss Amphibian Red List was done under a contract between the Swiss Federal Office of the Environment and KARCH. The reanalysis of the data was also supported through the Swiss Federal Office of the Environment (Forschungsvertrag 13.0007.PJ/M214-0617).

Literature Cited

Alpizar-Jara, R., Nichols, J.D., Hines, J.E., Sauer, J.R., Pollock, K.H. & Rosenberry, C.S. (2004). The relationship between species detection probability and local extinction probability. *Oecologia*, **141**, 652–660.

BAFU. (2011). *Liste der national prioritären Arten: Arten mit nationaler Priorität für die Erhaltung und Förderung* (Stand 2010). Bundesamt für Umwelt, Bern, Switzerland. 132 pp.

Banks-Leite, C., Pardini, R., Boscolo, D., Cassano, C.R., Püttker, T., Barros, C.S. & Barlow, J. (2014). Assessing the utility of statistical adjustments for imperfect detection in tropical conservation science. *Journal of Applied Ecology*, **51**, 849–859.

Bottrill, M.C., Joseph, L.N., Carwardine, J., Bode, M., Cook, C., Game, E.T., Grantham, H., Kark, S., Linke, S., McDonald-Madden, E., Pressey, R.L., Walker, S., Wilson, K.A. & Possingham, H.P. (2009). Finite conservation funds mean triage is unavoidable. *Trends in Ecology & Evolution*, **24**, 183–184.

Bottrill, M.C., Joseph, L.N., Carwardine, J., Bode, M., Cook, C., Game, E.T., Grantham, H., Kark, S., Linke, S., McDonald-Madden, E., Pressey, R.L., Walker, S., Wilson, K. A. & Possingham, H.P. (2008). Is conservation triage just smart decision making? *Trends in Ecology and Evolution*, **23**, 649–654.

Buckland, S.T., Magurran, A. E., Green, R.E. & Fewster, R.M. (2005). Monitoring change in biodiversity through composite indices. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **360**, 243–54.

Butchart, S.H.M., Walpole, M., Collen, B., van Strien, A.J., Scharlemann, J.P.W., Almond, R.E.A., Baillie, J.E.M., Bomhard, B., Brown, C., Brune, J., Carpenter, K.E., Carr, G.M., Chanson, J., Chenery, A.M., Jorge, C., Davidson, N.C., Dentener, F., Foster, M., Galli, A., Galloway, J.N., Genovesi, P., Gregory, R.D., Hockings, M., Kapos, V., Lamarque, J.-F., Leverington, F., Loh, J., McGeoch, M.A., McRae, L., Minasyan, A., Morcillo, M.H., Oldfield, T.E.E., Pauly, D., Quader, S., Revenga, C., Sauer, J.R., Skolnik, B., Spear, D., Stanwell-Smith, D., Stuart, S.N., Symes, A.J., Tierney, M., Tyrrell, T.D., Vié, J.-C. & Watson, R. (2010). Global biodiversity: indicators of recent declines. *Science,* **328**, 1164–1168.

Chen, G., Kéry, M., Plattner, M., Ma, K. & Gardner, B. (2013). Imperfect detection is the rule rather than the exception in plant distribution studies. *Journal of Ecology*, **101**, 183–191.

Collen, B., Griffiths, J., Friedmann, Y., Rodriguez, J.P., Rojas-Suàrez, F. & Baillie, J.E.M. (2013). Tracking Change in National-Level Conservation Status: National Red Lists. *Biodiversity Monitoring and Conservation*, pp. 17–44. Wiley-Blackwell, Oxford, UK.

Dorazio, R.M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, **23**, 1472–1484.

Drost, C. & Fellers, G. (1996). Collapse of a regional frog fauna in the Yosemite area of the California Sierra Nevada, USA. *Conservation Biology*, **10**, 414–425.

Fischer, M. & Stocklin, J. (1997). Local Extinctions of Plants in Remnants of Extensively Used Calcareous Grasslands 1950 -1985. *Conservation Biology*, **11**, 727–737.

Gärdenfors, U. & Hilton-Taylor, C. (2001). The application of IUCN Red List criteria at regional levels. *Conservation Biology*, **15**, 1206–1212.

Garrard, G.E., McCarthy, M.A., Williams, N.S.G., Bekessy, S.A. & Wintle, B.A. (2013). A general model of detectability using species traits (S. Ramula, Ed.). *Methods in Ecology and Evolution*, **4**, 45–52.

Gelman, A., Carlin, J.P., Stern, H.S. & Rubin, D.S. (2004). *Bayesian Data Analysis*. CRC/ Chapman & Hall, Boca Raton, FL.

Gonseth, Y., Wohlgemuth, T., Sansonnens, B. & Buttler, A. (2001). *Die biogeographischen Regionen der Schweiz. Erläuterungen und Einteilungsstandard*. Bern.

Grossenbacher, K. (1988). *Verbreitungsatlas der Amphibien der Schweiz*. Centre suisse de cartographie de la faune; Schweizerischer Bund für Naturschutz, Neuchâtel.

Guillera-Arroita, G. & Lahoz-Monfort, J. (2014). Ignoring Imperfect Detection in Biological Surveys Is Dangerous: A Response to 'Fitting and Interpreting Occupancy Models'. *PloS one*, **9**, e99571.

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R. & Wintle, B.A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*,**24** (3), 276-292.

Hanski, I. (1991). Single-species metapopulation dynamics: concepts, models and observations.

Biological Journal of the Linnean Society, **42**, 17–38.

Hastie, T. & Fithian, W. (2013). Inference from presence-only data; the ongoing controversy. *Ecography*, **36**, 864–867.

Hecnar, S. & M'Closkey, R. (1996). Regional dynamics and the status of amphibians. *Ecology*, **77** (7), 2091–2097.

Hoffmann, M., Hilton-Taylor, C., Angulo, A., Böhm, M., Brooks, T.M., Butchart, S.H.M., Carpenter, K.E., Chanson, J., Collen, B., Cox, N. a, Darwall, W.R.T., Dulvy, N.K., Harrison, L.R., Katariya, V., Pollock, C.M., Quader, S., Richman, N.I., Rodrigues, A.S.L., Tognelli, M.F., Vié, J.-C., Aguiar, J.M., Allen, D.J., Allen, G.R., Amori, G., Ananjeva, N.B., Andreone, F., Andrew, P., Aquino Ortiz, A.L., Baillie, J.E.M., Baldi, R., Bell, B.D., Biju, S.D., Bird, J.P., Black-Decima, P., Blanc, J.J., Bolaños, F., Bolivar-G, W., Burfield, I.J., Burton, J. a, Capper, D.R., Castro, F., Catullo, G., Cavanagh, R.D., Channing, A., Chao, N.L., Chenery, A.M., Chiozza, F., Clausnitzer, V., Collar, N.J., Collett, L.C., Collette, B.B., Cortez Fernandez, C.F., Craig, M.T., Crosby, M.J., Cumberlidge, N., Cuttelod, A., Derocher, A.E., Diesmos, A.C., Donaldson, J.S., Duckworth, J.W., Dutson, G., Dutta, S.K., Emslie, R.H., Farjon, A., Fowler, S., Freyhof, J., Garshelis, D.L., Gerlach, J., Gower, D.J., Grant, T.D., Hammerson, G. a, Harris, R.B., Heaney, L.R., Hedges, S.B., Hero, J.-M., Hughes, B., Hussain, S.A., Icochea M, J., Inger, R.F., Ishii, N., Iskandar, D.T., Jenkins, R.K.B., Kaneko, Y., Kottelat, M., Kovacs, K.M., Kuzmin, S.L., La Marca, E., Lamoreux, J.F., Lau, M.W.N., Lavilla, E.O., Leus, K., Lewison, R.L., Lichtenstein, G., Livingstone, S.R., Lukoschek, V., Mallon, D.P., McGowan, P.J.K., McIvor, A., Moehlman, P.D., Molur, S., Muñoz Alonso, A., Musick, J. a, Nowell, K., Nussbaum, R. a, Olech, W., Orlov, N.L., Papenfuss, T.J., Parra-Olea, G., Perrin, W.F., Polidoro, B. a, Pourkazemi, M., Racey, P. a, Ragle, J.S., Ram, M., Rathbun, G., Reynolds, R.P., Rhodin, A.G.J., Richards, S.J., Rodríguez, L.O., Ron, S.R., Rondinini, C., Rylands, A.B., Sadovy de Mitcheson, Y., Sanciangco, J.C., Sanders, K.L., Santos-Barrera, G., Schipper, J., Self-Sullivan, C., Shi, Y., Shoemaker, A., Short, F.T., Sillero-Zubiri, C., Silvano, D.L., Smith, K.G., Smith, A.T., Snoeks, J., Stattersfield, A.J., Symes, A.J., Taber, A.B., Talukdar, B.K., Temple, H.J., Timmins, R., Tobias, J. a, Tsytsulina, K., Tweddle, D., Ubeda, C., Valenti, S. V, van Dijk, P.P., Veiga, L.M., Veloso, A., Wege, D.C., Wilkinson, M., Williamson, E. a, Xie, F., Young, B.E., Akçakaya, H.R., Bennun, L., Blackburn, T.M., Boitani, L., Dublin, H.T., da Fonseca, G. a B., Gascon, C., Lacher, T.E., Mace, G.M., Mainka, S. a, McNeely, J. a, Mittermeier, R. a, Reid, G.M., Rodriguez, J.P., Rosenberg, A. a, Samways, M.J., Smart, J., Stein, B. a & Stuart, S.N. (2010). The impact of conservation on the status of the world's vertebrates. *Science*, **330**, 1503–1509.

Houlahan, J.E., Findlay, C.S., Schmidt, B.R., Meyer, A.H. & Kuzmin, S.L. (2000). Quantitative evidence for global amphibian population declines. *Nature*, **404** (6779), 752–5.

Isaac, N.J.B. & Pocock, M.J.O. (2015). Bias and information in biological records. *Biological Journal of the Linnean Society*, **115**, 522–531.

IUCN Standards and Petitions Subcommittee. (2014). *Guidelines for Using the IUCN Red List Categories and Criteria*. Prepared by the Standards and Petitions Subcommittee. Available from: http://www.iucnredlist.org/documents/RedListGuidelines.pdf.

Joseph, L.N., Field, S., Wilcox, C. & Possingham, H.P. (2006). Presence-absence versus abundance data for monitoring threatened species. *Conservation Biology*, **20**, 1679–1687.

Kellner, K.F. & Swihart, R.K. (2014). Accounting for Imperfect Detection in Ecology: A Quantitative Review. *PLoS ONE*, **9**, e111436.

Kendall, W.L. & White, G.C. (2009). A cautionary note on substituting spatial subunits for repeated temporal sampling in studies of site occupancy. *Journal of Applied Ecology*, **46**, 1182–1188.

Kéry, M., Royle, J.A., Schmid, H., Schaub, M., Volet, B., Häfliger, G. & Zbinden, N. (2010). Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Conservation Biology*, **24**, 1388–97.

Kéry, M. & Schaub, M. (2012). *Bayesian population analysis using WinBUGS*. Academic Press.

Kéry, M. & Schmidt, B.R. (2008). Imperfect detection and its consequences for monitoring for conservation. *Community Ecology*, **9**, 207–216.

Kéry, M., Spillmann, J.H.J., Truong, C. & Holderegger, R. (2006). How biased are estimates of extinction probability in revisitation studies? *Journal of Ecology*, **94**, 980–986.

Lamoreux, J., Akcakaya, H.R., Bennun, L., Collar, N.J., Boitani, L., Brackett, D., Brautigam, A., Brooks, T.M., da Fonseca, G.A.B., Mittermeier, R.A., Rylands, A.B., Gardenfors, U., Hilton-Taylor, C., Mace, G.M., Stein, B.A. & Stuart, S.N. (2003). Value of the IUCN red list. *Trends in Ecology and Evolution*, **18**, 212–214.

Loh, J., Green, R.E., Ricketts, T., Lamoreux, J., Jenkins, M., Kapos, V. & Randers, J. (2005). The Living Planet Index: using species population time series to track trends in biodiversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 289–295.

Lupi, J. (2015). Quantification and explanation of the decline in the number of populations of common toad (*Bufo bufo)*, in southern Switzerland. MSc thesis. University of Neuchâtel.

Mace, G.M., Collar, N.J., Gaston, K.J., Hilton-Taylor, C., Akcakaya, H.R., Leader-Williams, N., Milner-Gulland, E.J. & Stuart, S.N. (2008). Quantification of extinction risk: IUCN's system for classifying threatened species. *Conservation Biology*, **22**, 1424–1442.

MacKenzie, D.I. & Nichols, J.D. (2004). Occupancy as a surrogate for abundance estimation. *Animal*

Biodiversity and Conservation, **1**, 461–467.

MacKenzie, D.I., Nichols, J.D. & Hines, J. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, **84**, 2200–2207.

MacKenzie, D.I., Nichols, J.D. & Lachman, G. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

McArdle, B.H. (1990). When are rare species not there? *Oikos*, **57**, 276–277.

Miller, D.A.W., Brehme, C.S., Hines, J.E., Nichols, J.D. & Fisher, R.N. (2012). Joint estimation of habitat dynamics and species interactions: disturbance reduces co-occurrence of non-native predators with an endangered toad. *Journal of Animal Ecology*, **81**, 1288–97.

Moilanen A. (2002). Implications of empirical data quality to metapopulation model parameter estimation and application. *Oikos*, **96**, 516–530.

Nichols, J.D. (1992). Capture- Recapture Models: Using marked animals to study population dynamics. *Bioscience*, **42**, 94–102.

Pellet, J. & Schmidt, B.R. (2005). Monitoring distributions using call surveys: estimating site occupancy, detection probabilities and inferring absence. *Biological Conservation*, **123**, 27–35.

Pollock, K.H., Nichols, J.D., Simons, T.R., Farnsworth, G.L., Bailey, L.L. & Sauer, J.R. (2002). Large scale wildlife monitoring studies: statistical methods for design and analysis. *Environmetrics*, **13**, 105–119.

Preston, F.W. (1979). The Invisible Birds. *Ecology*, **60**, 451–454.

Rodrigues, A.S.L., Pilgrim, J., Lamoreux, J.F., Hoffmann, M. & Brooks, T. (2006). The value of the IUCN Red List for conservation. *Trends in Ecology & Evolution*, **21**, 71–76.

Royle, J.A. & Kéry, M. (2007). A Bayesian state-space formulation of dynamic occupancy models. *Ecology*, **88**, 1813–23.

Sadoti, G., Zuckerberg, B., Jarzyna, M.A. & Porter, W.F. (2013). Applying occupancy estimation and modelling to the analysis of atlas data. *Diversity and Distributions*, **19**, 804–814.

Schmidt, B.R. (2005). Monitoring the distribution of pond-breeding amphibians when species are detected imperfectly. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **15**, 681–692.

Schmidt, B.R. & Zumbach, S. (2016). Amphibian conservation in Switzerland. *Amphibian Biology, Volume 11: Status of conservation and declines of Amphibians: Eastern Hemisphere, Part 5: Northern Europe.* (eds H. Heatwole & J.W. Wilkinson).

Schmidt, B.R. & Zumbach, S. (2005). *Rote Liste der gefährdeten Arten der Schweiz: Amphibien*. Bern.

Schmidt, B.R., Zumbach, S., Tobler, U. & Lippuner, M. (2015). Amphibien brauchen temporäre Gewässer Einleitung Lebensweise von Amphibien in temporären Gewässern. Zeitschrift für Feldherpetologie **22**: 137-150.

Sewell, D., Guillera-Arroita, G., Griffiths, R.A. & Beebee, T.J.C. (2012). When is a species declining? Optimizing survey effort to detect population changes in reptiles. *PloS one*, **7**, e43387.

Skelly, D.K., Yurewicz, K.L., Werner, E.E. & Relyea, R. a. (2003). Estimating Decline and Distributional Change in Amphibians. *Conservation Biology*, **17**, 744–751.

Tanadini, L.G. & Schmidt, B.R. (2011). Population Size Influences Amphibian Detection Probability: Implications for Biodiversity Monitoring Programs. *PLoS ONE*, **6**, e28244.

The Rolling Stones. 1969. You can’t always get what you want. Decca Records.

Tingley, M.W. & Beissinger, S.R. (2009). Detecting range shifts from historical species occurrences: new perspectives on old data. *Trends in Ecology and Evolution*, **24**, 625–633.

Wintle, B.A., Walshe, T. V., Parris, K.M. & McCarthy, M.A. (2012). Designing occupancy surveys and interpreting non-detection when observations are imperfect. *Diversity and Distributions*, **18**, 417–424.

Yoccoz, N.G., Nichols, J.D. & Boulinier, T. (2001). Monitoring of biological diversity in space and time. *Trends in Ecology and Evolution*, **16**, 446–453.

CHAPTER THREE

Inferring the absence of rare and common species from detection data

Sam S. Cruickshank [1*], Arpat Ozgul [1], Andreas Meyer [2],

Jean-Claude Monney [2,3],Benedikt R. Schmidt [1,2]

Manuscript in preparation

---

[1] *Department of Evolutionary Biology and Environmental Studies, University of Zürich, Zürich, Switzerland*
[2] *Info Fauna karch, Neuchâtel, Switzerland*
[3] *Deceased 29/12/2012*
*sam.cruickshank@ieu.uzh.ch

ABSTRACT

Identifying when species are absent is often as important as establishing where they do exist, however surveys are often not developed with sufficient power to infer species absence with confidence. Previous studies used estimates of detection probability to estimate how many non-detections are required to establish a site is unoccupied. However, correctly interpreting the meaning of a string of non-detections at a site also requires knowledge on species prevalence. We used data from the update of the Swiss national reptile Red List to estimate detection probabilities and quantified countrywide species prevalence. Because species prevalence has no natural scale, we quantified prevalence at multiple spatial scales. Using two different approaches, we used this information to explore how survey recommendations differ in relation to species rarity and ease-of-detection. We found that for all but the most common species, incorporating information on expected prevalence reduced survey effort required in order to infer a site is unoccupied. However, for species that are either rare or hard to detect, unrealistic amounts of survey effort are necessary in order to avoid missing occupied sites. Prevalence had a strong effect on the survey effort required to infer absence. Since prevalence has no natural scale, predicting the survey effort required to infer absence depends on a definition of spatial scale. Because spatial scale ultimately depends on the objectives of a survey, no general recommendations are possible.

Keywords :

absence, detection probability, distribution, monitoring, rarity, survey

INTRODUCTION

Conservation or management of a species must begin with knowledge of its distribution. Ecologists are mostly interested in a better knowledge of where species are present, but there are many instances where it is equally important to know where a species does not occur or where it no longer occurs (McArdle 1990; Kéry 2002; Moilanen A 2002; Skelly *et al.* 2003; Kéry *et al.* 2006; Chadès *et al.* 2008; Tingley & Beissinger 2009; Collen *et al.* 2010; Wintle *et al.* 2012; Guillera-Arroita *et al.* 2015b). Unfortunately, the lack of recorded species absences makes it impossible to distinguish between true absence, local extinction, and lack of survey effort and this can hinder conservation efforts (Isaac & Pocock 2015; Boakes *et al.* 2016). For example, the presence of a rare or endangered species can constrain development of a site. If a species is present, then mitigation such as translocations may be required by law. Knowing that a rare or endangered species does not occur at a site is therefore important for conservationists, government agencies and developers (Germano *et al.* 2015) . Invasive species are another example where it is important to know that a species is absent from a site. For example, one may be interested to know that an invasive species is not yet present at a site or that a species has been successfully eradicated (Grant *et al.* 2017; Sakamoto *et al.* 2017). Last but not least, basic ecological science would benefit from absence data as ecological niche models are most powerful if there is data on both absences and presences (Guillera-Arroita *et al.* 2015a).

There are two reasons why species are not recorded at a site during a survey: the species may be truly absent or it may be present but overlooked (MacKenzie *et al.* 2002). Imperfect detection of species is ubiquitous in surveys and leads to false absences (also known as errors of omission; Preston 1979; Pollock *et al.* 2002; Kéry & Schmidt 2008; Chen *et al.* 2013). McArdle (1990, see also Reed (1996)) first addressed the influence of imperfect detection in surveys and monitoring programs. McArdle (1990) highlighted that, given that survey efforts for species are never completely efficient, it is never possible to attain complete confidence that a species is in fact absent from a site. Only probabilistic statements can be made as to the degree of certainty - with an increase in the number of site visits made, the likelihood of a population of a species remaining undetected at an occupied site decreases. McArdle formulated a series of equations describing the probability of a species with rarity p remaining undetected following visits to N sites:

$$P(undetected) = (1-p)^N \qquad (3.1)$$

Furthermore, he defined the number of sites that should be surveyed in order to be confident of detecting at least once a species of a given rarity with a permissible level of error α, as

$$N^* = \frac{\log(\alpha)}{\log(1-p)} \qquad (3.2)$$

Since the advent of occupancy modelling allowed the estimation of detection probabilities (MacKenzie *et al.* 2002) many authors have repurposed these equations- redefining p as detection probabilities, and reinterpreting N* not as the number of sites that should be visited, but as the number of visits required at a site in order to be confident of detecting the species at a site. Making these assumptions, authors used these equations to calculate cumulative detection curves describing the probability of misclassifying an occupied site in relation to the number of visits made to the site, as well as the number of site visits needed in order to detect a species presence with a maximum error rate of α (Kéry 2002; Pellet & Schmidt 2005; Jackson *et al.* 2006; Driscoll 2010; Canessa *et al.* 2012; Sewell *et al.* 2012; Sliwinski *et al.* 2016).

However, this application was questioned by Wintle et al (2005, 2012), who highlighted that species detectability is not analogous to species prevalence, and thus that when *p* represents species detection, Equation 3.1 instead gives the probability of getting N successive non-detections, given that the site is occupied (hereafter p(undetected|occupied)). Similarly, equation 3.2 gives for a pre-specified allowable error α, the number of non-detections at a site necessary for cumulative detection probability at sites which are in fact occupied to fall below α (hereafter we refer to this as N(M)). Furthermore, Wintle et al. (2012) highlighted that by failing to notice this distinction, many authors was making erroneous statements in conflating the probability of a species remaining undetected after N visits with the probability that the species occupies that site. As equations 3.1 and 3.2 are in fact conditional on species presence (Wintle et al. 2012), this amounts to committing the ‘inverse fallacy’ (Koehler 1996; Villejoubert & Mandel 2002) of concluding p(undetected|occupied)=p(occupied|undetected).

Wintle et al (2005, 2012) argued that it is impossible to estimate the true probability that a site is unoccupied (and, by extension, the number of site visits for the probability of site occupancy to drop below a given level), without incorporating information on expected species prevalence in addition to detection probabilities. Not observing a species after surveying can occur as the result of the species repeatedly remaining undetected (i.e. $(1\text{-}p)^N$), or because the site itself is unoccupied. Thus, interpreting the meaning of successive non-detections requires information on the expected rate of occupancy (i.e. prevalence) of the species across the sample of sites. With an estimate of species prevalence ψ, the probability of a site being unoccupied after N successive non-detections can then be described as:

$$P(unoccupied|undetected) = 1 - \frac{\psi\,(1-p)^N}{\psi(1-p)^N + (1-\psi)} \qquad (3.3)$$

And the number of sequential non-detections (henceforth N(W)) needed to reduce the posterior probability of presence at the site below a level α is:

$$N(W) = \frac{\log\left(\frac{\alpha}{1-\alpha}\right) - \log\left(\frac{\psi}{1-\psi}\right)}{\log(1-p)} \qquad (3.4)$$

Equation 3.4 was later modified by Guillera-Arroita et al (2015) who noted that this formulation predicted negative numbers of site visits if the expected prevalence of the species was lower than the permissible error rate α. In such circumstances, they concluded that equation 3.4 should evaluate to zero, and no site visits should be made; sites should simply be assumed to be unoccupied.

The results of Wintle et al. (2005, 2012) have strong implications for survey design and for inferring the absence of a species. Put simply, if one does not detect a rare species, then the most likely reason is that the species is absent and further site visits are unnecessary because they provide no further useful information. In contrast, if one does not detect a widespread species, then the most likely reason is that the species was overlooked and additional site visits are recommended. While the logic underlying the approach of Wintle et al. (2005, 2012) is sound, the number of sites visits that are necessary to declare absence of a species depends crucially on species prevalence. Prevalence depends on spatial scale (Joseph & Possingham 2008) but there is a natural scale only for a few species. For example, wetlands may be the natural scale for pond-breeding amphibians and patches of dry grassland the scale for a specialist plant (Hecnar & M'Closkey 1997; Fischer & Stocklin 1997). For most species, however, prevalence has no natural definition and scale (Fithian & Hastie 2013; Kéry & Royle 2015). Individuals are distributed in space and measures of prevalence depend on the scale used by an observer. Prevalence will greatly differ if presence and absence of a species is considered at scales of 10 m^2, 1 km^2 or 100 km^2; ecologists understand well that the scale of observation affects the pattern (Levin 1992).

Recommendations on how often one has to survey a site until one can confidently declare that a species is absent have become popular (Kéry 2002; Pellet & Schmidt 2005; Jackson *et al.* 2006; Driscoll 2010; Sewell *et al.* 2012; Sliwinski *et al.* 2016). No matter whether the McArdle or Wintle et al. approach is used, the recommendations will depend strongly on the assumptions that one is willing to entertain. Using a case study of data collected as part of a national Red List monitoring programme for reptiles in Switzerland, we demonstrate the differences between the two protocols and compare the recommendations derived from the McArdle and the Wintle et al. approaches. We fitted occupancy models to obtain estimates of detection probability (MacKenzie *et al.* 2002; Guillera-Arroita 2017), and calculated prevalence estimates of each species from an existing database at different grain sizes, in order to examine how the recommendations of equations 3.3 and 3.4 are impacted by spatial scale. In particular, our first goal was to compare recommendations for rare and common species and species that are easy and hard to detect. Our second goal was to highlight the impact of spatial scale at which prevalence is measured on the number of sites visits which are necessary to infer absence.

METHODS

Data collection

Two forms of information are required to explore the recommendations for the two protocols; prevalence estimates and detection probabilities for each species. We obtained estimates of species prevalence using the reptile distribution database of Info Fauna karch (Hofer *et al.* 2001; Schmidt & Zumbach 2018), which contains all reported records for native and invasive reptile species throughout Switzerland. For every reptile species native to Switzerland, we first classified which of the 6 biogeographic regions of Switzerland (Gonseth *et al.* 2001) represent suitable habitat. Some species such as *C.austriaca* are found in all regions, whereas others, such as *L.bilineata* are much more restricted in their distribution (Hofer *et al.* 2001), and thus treating the whole of Switzerland as potentially suitable habitat for the species would be inappropriate. We obtained the current distributions for each species in the relevant regions by first filtering out all pre-1980 records from the karch database (this is the time when comprehensive surveying for reptiles within Switzerland began; Hofer et al. 2001). We then aggregated the observations at a 1x1 km grid scale to produce presence/absence data for each cell. We assumed absence within a cell if no reports were present as sampling in most cells can be assumed to be quasi-exhaustive over the 35 years of records considered. Our estimates of species prevalence were then calculated as the proportion of all grid cells within the relevant biogeographic regions that were classified as occupied. To assess the importance of grid-size in determining species prevalence, we repeated this process with larger grain sizes of 5x5 km and 10x10 km.

Detection probabilities were estimated from data collected in 2003-2004 for an update of the Swiss reptile Red List (Monney & Meyer 2005). Using the karch reptile distribution database described above, 294 1km^2 quadrats were selected for surveying by stratified random sampling in order to provide representative coverage of all biogeographic regions in Switzerland. Each quadrat was visited between 1-3 times (mean 2.28 ± 0.44) by expert herpetologists who recorded all reptile species they observed. Site visits were separated by a minimum of 2 weeks, and a minimum of 2 visits were carried out at all quadrats with previous reports of snake presence. This resulting dataset therefore consisted of repeated detection/non-detection data for each species at all quadrats that fell within the species potential distributional range (as described above).

The grass snake (*Natrix natrix*) and slow worm (*Anguis fragilis*) have each been recently recognised as comprising two separate species within Switzerland (*N. natrix* and *N. helvetica*: Kindler *et al.* 2017, and *A. fragilis* and *A. veronensis*; Gvoždík *et al.* 2013), however, as these species distinctions were not made in the historical karch records or the Red-List survey data, in this study we treat each as single species. The congeneric species have non-overlapping ranges.

Data analysis

We estimated detection probabilities by fitting single-season occupancy models (MacKenzie *et al.* 2002) to the detection/non-detection data y_{ij} collected for the Red List update (there are *j* visits at *i* sites).The model can be written as:

$z_i \sim \text{Bernoulli}(\Psi)$

$y_{ij} | \, z_i \sim \text{Bernoulli}\ (z_i p)$

The first equation defines the true occurrence state of a site as a Bernoulli trial with success parameter Ψ, which is the occupancy probability. The second equation defines the y_{ij} data as a second Bernoulli trial where the success parameter is the product of the true occurrence state and detection probability. Both detection and occupancy parameters were modelled without covariates using vague uniform priors between 0 and 1. Models were run in JAGS (Plummer 2003) through R (R Core Team 2017) using the package R2jags (Su & Yajima 2015). We ran 3 Markov chains of 50,000 iterations, discarding the first 1,000 iterations as burn-in and thinning the remaining by 1 in 10. Model convergence was assessed using the Gelman-Rubin statistic (Gelman *et al.* 2004).

Based on the estimates of detection probability and prevalence, we parameterized equations 3.1 and 3.3 and to calculate the number of site visits necessary to reduce the probability of a species remaining undetected at an unoccupied site (McArdle 1990) and of occupancy (Wintle *et al.* 2012) below an α of 0.05. Finally, we explored the scale at which prevalence is estimated by applying equation 3.4 to examine occupancy curves for the three grain sizes at which prevalence was estimated.

RESULTS

Species prevalence measured at the 5x5km scale varied substantially from 0.066 (*H. viriflavus*) to 0.612 (*A. fragilis*; Table 3.1). Prevalence measures differed strongly depending on the grain size at which data were aggregated; aggregation over larger scales (e.g. 10x10 km) led to higher species prevalence (Table 3.1, Fig S3.1). However, the degree to which prevalence increased was species-specific: an increase in grain size from 1x1 km to 10x10 km led to increases in prevalence of between 420% and 1212% (mean

737%). When assessed at 1x1 km resolution, all species could be considered rare (i.e. $\Psi < 0.2$), however at the coarsest resolution only 3 of 12 species would fulfil that definition.

The observed proportion of 1x1km quadrats where the species were detected was generally much lower than the estimated proportion of occupied quadrats (Fig. 3.1), i.e. the species was not detected in all quadrats where it occurs. In some cases (e.g. *Lacerta bilineata*) the difference was small but in other cases the difference could be substantial (e.g. *Anguis fragilis* and *Coronella austriaca,* where the estimated proportion is roughly twice the observed proportion; Fig. 3.1).

Detection was far from perfect in all species. **S**nakes generally had lower detection probabilities than lizards, however. Per-visit detection probabilities were highly variable, ranging from 0.180 (95% credible interval [CRI]:0.081-0.369) for *Z. longissimus*, to 0.699 (95% CRI: 0.639-0.755) for *P. muralis* (Fig. 3.2, Table 3.1**).** Consequently, for all species, multiple non-detections would be necessary in order for the probability of not observing a species at an occupied site to fall below 0.05 (Fig. 3.3, based on equation 3.2).

When we incorporate species prevalence (i.e. move from equation 3.2 to equation 3.4), the estimated parameter becomes the number of non-detections required before occupancy probability falls below 0.05. With prevalence calculated at 5x5 km grain size, the recommended survey effort increased for two species (*Z. vivipara, A. fragilis*), remained the same for three species (*P. muralis, L. agilis, N. natrix*), and decreased for the remaining seven species (Table 3.1). The species for which extra survey effort was recommended when no longer conditioning on species presence were generally those with higher prevalence estimates. For 8 species, the differences between survey effort recommended by equation 3.2 and 3.4 were small (<2 surveys more or fewer; e.g. *L. agilis, L. bilineata* and *A. fragilis* in Fig. 3.3). However, some differences were large; if accounting for prevalence, recommended survey effort for *Z. longissimus* is reduced by over 14 surveys in order to obtain sufficient confidence that occupancy probability is less than 0.05 (Fig. 3.3).

As prevalence values were dependent upon the spatial scale at which they were calculated, so too were recommendations for survey effort when taking prevalence into account. Prevalence increased monotonically with increased grain size (i.e. decreased resolution of prevalence; Fig. 3.1); so too did recommended survey effort under equation 3.4. Although all species required more survey effort with coarser spatial scales, the magnitude of differences varied widely between species; the change from 1x1 km to 10x10 km prevalence estimates resulted in survey effort increases of between 1.0 and 16.7 site visits (Fig. 3.4, Table 3.1). Importantly, the Wintle et al. (2012) model recommended zero site visits for some species with low prevalence (e.g., *C. austriaca*; Fig. 3.4, Table 3.1) whereas between 10 and 20 site visits were necessary at larger grid sizes.

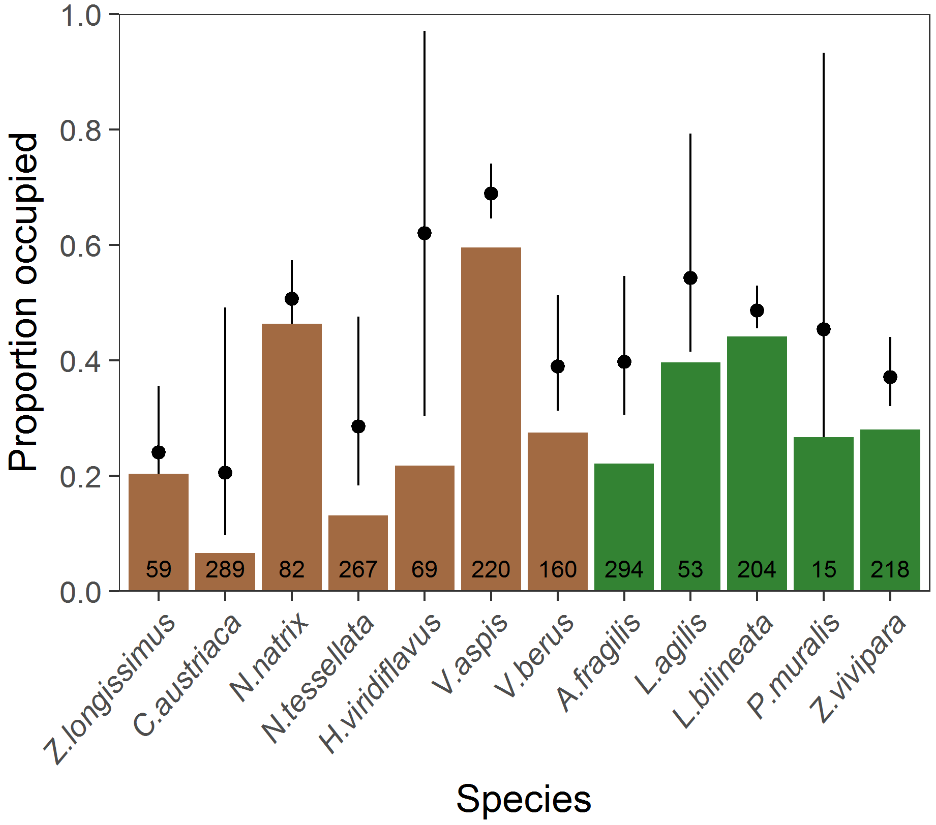


Figure 3.1. Observed (coloured bars) and estimated (dots and error bars; posterior mean and 95% credible intervals) proportion of occupied sites (1x1 km quadrats).The numbers give the total number of surveyed quadrats per species. Snakes are coloured brown and lizards green.

Tables and Figures

Table 3.1. Per-visit detection probabilities, prevalence estimates at three grain sizes, and estimates of the number of site visits required to ensure: ^{a}) p(detected | occupied) [N(McArdle)], and ^{b}) p(occupied) [N(Wintle), using prevalence calculated at 1x1, 5x5, and 10x10 km resolution] >0.05 for 12 reptile species within Switzerland.

Species	Detection probability (95% CRI)	Prevalence: 1x1 km	Prevalence 5x5 km	Prevalence 10x10 km	N(McArdle)^{a} <0.05	N(Wintle)^{b} 1x1 km	N(Wintle)^{b} 5x5 km	N(Wintle)^{b} 10x10 km
Anguis fragilis	0.336 (0.230-0.442)	0.142	0.612	0.740	7.57 (5.12-11.31)	2.83 (1.91-4.23)	8.54 (5.77-12.74)	10.03 (6.78-14.97)
Coronella austriaca	0.209 (0.069-0.389)	0.046	0.362	0.595	15.54 (6.01-38.15)	0 (0-0)	11.72 (4.53-28.77)	16.65 (6.44-40.88)
Hierophis viridiflavus	0.453 (0.280-0.622)	0.018	0.066	0.097	5.22 (3.04-9.04)	0 (0-0)	0.47 (0.27-0.81)	1.23 (0.72-2.13)
Lacerta agilis	0.681 (0.592-0.760)	0.149	0.503	0.625	2.64 (2.08-3.34)	1.05 (0.83-1.32)	2.59 (2.05-3.28)	3.01 (2.39-3.83)
Lacerta bilineata	0.673 (0.560-0.775)	0.026	0.094	0.137	2.71 (2.02-3.66)	0 (0-0)	0.58 (0.43-0.79)	0.97 (0.73-1.32)
Natrix natrix	0.278 (0.152-0.424)	0.095	0.453	0.635	9.98 (5.45-18.64)	2.13 (1.16-3.98)	9.07 (4.95-16.94)	11.53 (6.29-21.53)
Natrix tessellata	0.391 (0.133-0.719)	0.017	0.110	0.208	7.76 (2.41-21.18)	0 (0-0)	2.14 (0.66-5.83)	4.16 (1.29-11.36)
Podarcis muralis	0.699 (0.639-0.755)	0.100	0.402	0.552	2.50 (2.13-2.94)	0.61 (0.52-0.72)	2.13 (1.81-2.50)	2.63 (2.24-3.09)
Vipera aspis	0.431 (0.299-0.560)	0.049	0.257	0.360	5.48 (2.61-8.37)	0 (0-0)	3.25 (2.14-4.96)	4.12 (2.71-6.29)
Vipera berus	0.628 (0.358-0.840)	0.026	0.146	0.238	3.28 (1.62-6.85)	0 (0-0)	1.10 (0.55-2.30)	1.80 (0.89-3.76)
Zamenis longissimus	0.180 (0.081-0.369)	0.017	0.089	0.140	17.58 (6.82-34.76)	0 (0-0)	3.35 (1.30-6.62)	6.60 (2.56-13.06)
Zootoca vivipara	0.548 (0.441-0.649)	0.119	0.579	0.722	3.82 (2.85-5.14)	1.18 (0.88-1.59)	4.12 (3.07-5.54)	4.95 (3.69-6.66)

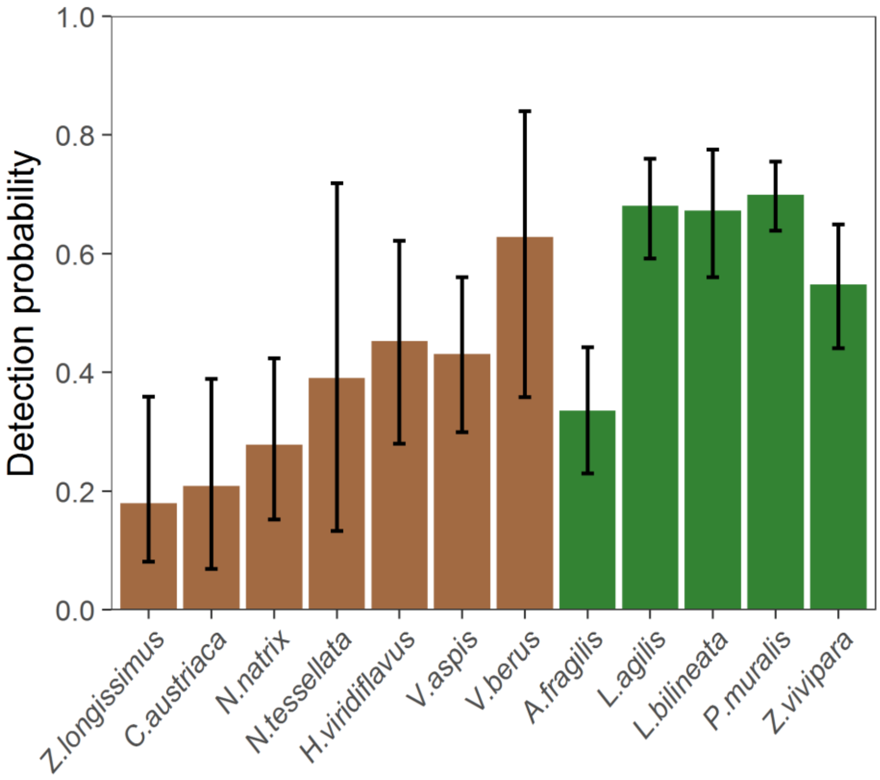


Figure 3.2. Per-visit detection probabilities with 95% credible intervals for 12 reptile species at sites (1x1 km quadrats) throughout Switzerland. Snakes are coloured brown and lizards green.

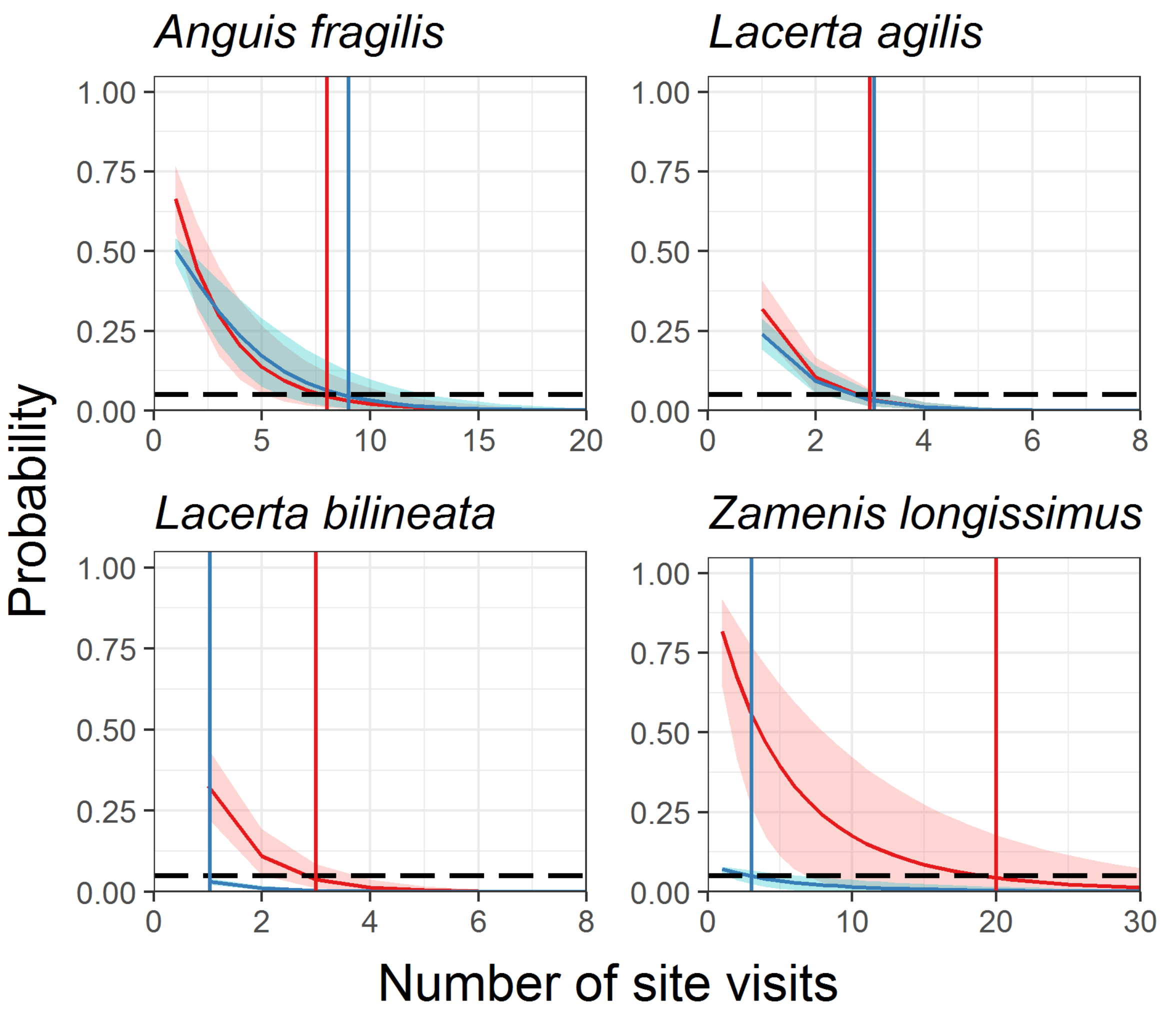


Figure 3.3. Probability of a species remaining undetected given that the sites is occupied (equation 1; p(undetected|occupied), red lines) and the probability that a site is occupied given a sequence of non-detections, based on prevalence calculated at 5x5 km resolution (equation 3.3, blue lines). Ribbons represent 95% credible intervals. Vertical lines represent the number of visits required at a site for these probabilities to drop below this threshold (i.e. equations 3.2 and 3.4). *L. agilis* is widespread and relatively easy to detect. *Z. longissimus* has a restricted range and is relatively difficult to detect. *A. fragilis* is widespread but relatively hard to detect. *L. bilineata* has a restricted distribution but is relatively easy to detect.

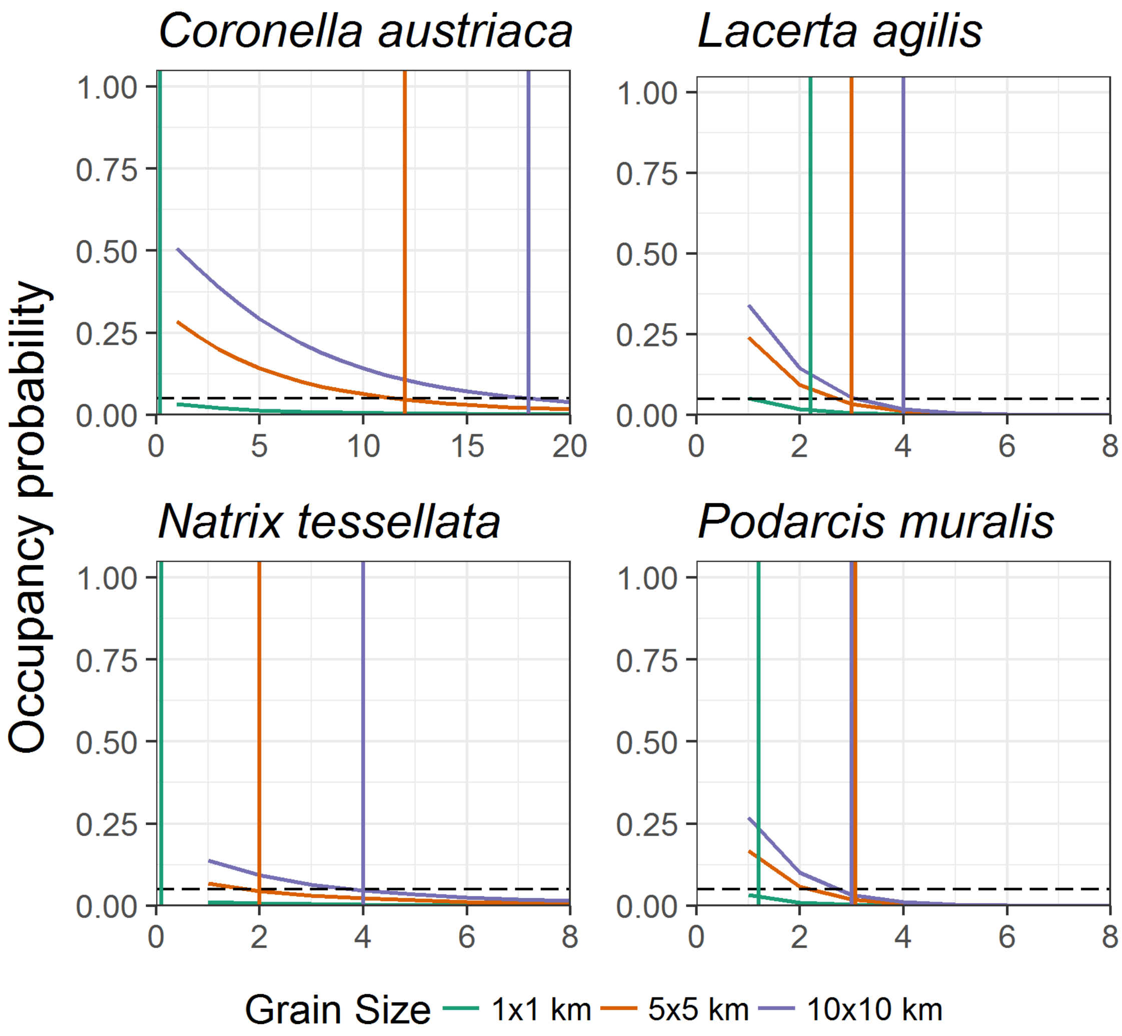


Figure 3.4. Occupancy probabilities given successive non-detections for 4 reptile species native to Switzerland. These probabilities do not condition on species presence, instead they take into account species prevalence. Colours represent probabilities based on prevalence estimates aggregated at three grain sizes: 1x1km^2 (green lines), 5x5km^2 (orange lines), 10x10km^2 (purple lines). Vertical lines represent the number of successive non-detections (i.e. N* from equations 3.2 and 3.4, rounded up to the nearest integer) required to attain a probability of occupancy of less than 0.05 (black dashed horizontal line).

DISCUSSION

Inferring the absence of a species is an important but challenging task for ecologists and conservationists. Because one can never be certain that a species is not present at a site, probabilistic approaches are necessary. The approach of McArdle (1990) tells us how many site visits are necessary until a species should have been detected at least once (given that the species is present; equations 3.1 and 3.2). As succinctly explained by Wintle et al. (2012), it is a logical fallacy to assume that the species is absent if it should have been detected if present. Wintle et al. (2012) showed that one has to account for species prevalence to infer absence.

Our results show that reptile detection probabilities are low and highly variable among species (Fig. 3.2). This is consistent with previous research on reptile detection probabilities and highlights the need for analytical methods that can account for imperfect detection (Anderson *et al.* 2001; Kéry 2002; Sewell *et al.* 2012; Schmidt *et al.* 2017). Consequently, species were not detected at many sites where the species occur (Fig. 3.1). For large-scale monitoring, where the focus of inference is identifying the proportion of occupied locations, we believe that repeated site visits and site occupancy models are the best approach to determine presence and absence of species (MacKenzie et al. 2002, Guillera-Arroita 2017). However, with the detection probabilities reported in figure 3.2, a large number of site visits may be necessary in order to obtain occupancy estimates with low standard error (Fig. 3.1; MacKenzie & Royle 2005; Bailey *et al.* 2007; Sewell *et al.* 2012).

Low detection probabilities imply that absence is not easily inferred. As shown previously by Kéry (2002), ten to twenty visits may be necessary for some species to reliably infer absence (Figures 3.3, 3.4). For both the McArdle (1990) and the Wintle et al. (2012) approaches, the number of site visits necessary to infer absence varied among species because the species have different detectabilities and prevalence. In most cases, the Wintle et al. (2012) approach required fewer sites visits but this was not always true. For the widespread slow-worm *Anguis fragilis*, which had low detection probability, the Wintle et al. (2012) approach required more site visits than the McArdle (1990) approach.

Based on our experience with planning and costing surveys, we believe that in conservation practice the differences between the two approaches may not matter much, for three reasons. For some species, the number of site visits is equal under the two approaches (Fig. 3.3, *Lacerta agilis*). For other species, the difference is small and the number of visits within a range that is realistic under most budgets (Fig. 3.3, *Lacerta bilineata*). The third group of species are the ones where the number of site visits is so large that the number of visits required is beyond what can be afforded in most surveys (Fig. 3.3, *Anguis fragilis*).

Changing the scale of prevalence has a similar effect on the number of site visits, provided that prevalence estimates remain higher than alpha. For most species, the number of site visits is small and does not differ much among estimates of prevalence (Fig. 3.4). For others, such as *Coronella austriaca*, fine-scale prevalence estimates suggest not attempting to survey, and other resolutions suggest numbers of visits which are most likely too high for any survey budget. Ultimately, the scientific or management questions and the budget will determine the number of visits that are possible (Field *et al.* 2005; Guillera-Arroita *et al.* 2015b; Moore & McCarthy 2016). When determining the number of visits, one should keep in mind that one should incorporate uncertainty in the estimate of detection probability. As figure 3.3 shows, uncertainty in the number of visits required can be considerable.

Whenever a meaningful estimate of prevalence is available, we recommend the approach to inferring absence described by Wintle et al. (2012). The method quantifies the probability of occurrence of a species (where 1 - probability of occurrence is the probability that a species is absent). This is the state variable that is required for management and conservation decisions. The drawback of the method is that an estimate of prevalence is required. This estimate is crucial as it determines the required number of site visits. As there is no “natural” spatial scale which could be used to determine prevalence (Fithian & Hastie 2013; Kéry & Royle 2015), estimating prevalence will often be difficult. Wintle et al. (2012) suggested that estimates of prevalence might be derived from a distribution model. We do not think that this is a viable approach since distribution models are not scale-free and require at least an implicit definition of spatial scale (Fithian & Hastie 2013).

Large-scale surveys and monitoring programs will reveal the presence of species at many sites where they occur but, given imperfect detection, not at all sites (Fig. 3.1). Such surveys suggest that there are many sites where a species remains undetected. Conservation managers may want to identify these unknown sites, especially for rare species. If there are historic records (as in Fischer & Stöcklin 1997; Kéry *et al.* 2006; Cruickshank *et al.* 2016), then one may assume that the species still occurs at a site and use the McArdle (1990) approach for sites with historic records. For sites with no historic records, one would have to try to infer absence using the approach of Wintle et al. (2012). For very rare species with prevalence smaller than α, the recommendation would be not to conduct surveys at all (see Guillera-Arroita et al. 2015 who provide advice on how to best decide on survey design given a management question).

A worrying problem with the approach proposed by Wintle et al. (2012) is that it can suggest that one should not survey a site at all if prevalence is low (Guillera-Arroita et al. 2015). This can also be seen in figure 3.4 where zero site visits are recommended for the snakes *Coronella austriaca* and *Natrix tessellata* at the smallest spatial scale (1x1 km) because prevalence is very low at this scale (Fig. S3.1, Table 3.1). This is particularly problematic as prevalence is likely to be lowest at the spatial scale of a patch, which is typically also the scale at which absence has to be assessed for conservation purposes (e.g., environmental impact assessments). However, such an extreme result reflects the fact that for rare species, it will typically be impossible to invest sufficient survey effort at sites to be confident that the species is not remaining unobserved. Although managers are unlikely to accept that monitoring rare species is worthless, it is important to recognise that the power of wide-scale surveys for such species may be extremely low- and indeed the majority of occupied site may be missed (as in *C.austriaca* and *H.viridiflavus;* Fig. 3.1).

If the approach of Wintle et al (2012) is followed when monitoring the success of management to remove invasive species, surveys would stop once the species prevalence drops below α. Ultimately, the Wintle et al. (2012) approach suggests that one should not attempt to prove that an invasive species was successfully eradicated. To avoid this scenario, managers could instead choose to reduce the levels of permissible error (α) to a very low level, which would result in substantial increases in the required survey effort required to establish the (almost) complete eradication of the species.

Where no reliable estimate of prevalence is available, the approach of McArdle (1990) is a valuable alternative. In the McArdle model, α is the maximum probability that a species won't be detected at an occupied site. In the context of surveys that inform conservation decisions, this can be seen as the proportion of sites that may be lost to development because the presence of the species could not be confirmed. Losing sites where rare species occur is clearly not desirable but ultimately it may be unavoidable.

In the end, which approach is used and how absence is determined cannot be determined a priori. The choice of method has to depend on the goals of a survey and the costs associated with making wrong decisions (Yoccoz *et al.* 2001; Field *et al.* 2005; Guillera-Arroita *et al.* 2015b; Moore & McCarthy 2016). There are no simple rules of thumb.

ACKNOWLEDGMENTS

We thank all the field crews for collecting data in the field, Thierry Bohnenstengel for providing the species prevalence data, and the population ecology research group for helpful discussions. The project was funded through a PhD scholarship from the Swiss Federal Office of the Environment (Forschungsvertrag 13.007PJ/M214- 0617).

Literature Cited

Anderson, D.R., Burnham, K.P., Lubow, B.C., Thomas, L., Corn, P.S., Medica, P.A. & Marlow, R.W. (2001). Field Trials of Line Transect Methods Applied to Estimation of Desert Tortoise Abundance. *The Journal of Wildlife Management*, **65**, 583.

Bailey, L.L., Hines, J.E., Nichols, J.D. & MacKenzie, D.I. (2007). Sampling design trade-offs in occupancy studies with imperfect detection: examples and software. *Ecological Applications*, **17**, 281–90.

Boakes, E.H., Fuller, R.A., McGowan, P.J.K. & Mace, G.M. (2016). Uncertainty in identifying local extinctions: the distribution of missing data and its effects on biodiversity measures. *Biology Letters*, **12**, 20150824.

Canessa, S., Heard, G.W., Parris, K.M. & McCarthy, M.A. (2012). Integrating variability in detection probabilities when designing wildlife surveys: A case study of amphibians from south-eastern Australia. *Biodiversity and Conservation*, **21**, 729–744.

Chadès, I., McCarthy, M.A., McDonald-Madden, E., Wintle, B.A., Linkie, M. & Possingham, H.P. (2008). When to stop managing or surveying cryptic threatened species. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 13936–40.

Chen, G., Kéry, M., Plattner, M., Ma, K. & Gardner, B. (2013). Imperfect detection is the rule rather than the exception in plant distribution studies. *Journal of Ecology*, **101**, 183–191.

Collen, B., Purvis, A. & Mace, G.M. (2010). When is a species really extinct? Testing extinction inference from a sighting record to inform conservation assessment. *Diversity and Distributions*, **16**, 755–764.

Cruickshank, S.S., Ozgul, A., Zumbach, S. & Schmidt, B.R. (2016). Quantifying population declines based on presence-only records for red-list assessments. *Conservation Biology,* **30**, 1112–1121.

Driscoll, D.A. (2010). Few beetle species can be detected with 95% confidence using pitfall traps. *Austral Ecology*, **35**, 13–23.

Field, S.A., Tyre, A.J. & Possingham, H.P. (2005). Optimizing Allocation of Monitoring Effort Under Economic and Observational Constraints. *Journal of Wildlife Management*, **69**, 473–482.

Fischer, M. & Stocklin, J. (1997). Local Extinctions of Plants in Remnants of Extensively Used Calcareous Grasslands 1950 -1985. *Conservation Biology*, **11**, 727–737.

Fithian, W. & Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-absence only data. *Annals of Applied Statistics*, **7**, 1917–1939.

Gelman, A., Carlin, J.P., Stern, H.S. & Rubin, D.S. (2004). *Bayesian Data Analysis*. CRC/ Chapman & Hall, Boca Raton, FL.

Germano, J.M., Field, K.J., Griffiths, R.A., Clulow, S., Foster, J., Harding, G. & Swaisgood, R.R. (2015). Mitigation-driven translocations: are we moving wildlife in the right direction? *Frontiers in Ecology and the Environment,* **13**, 100–105.

Gonseth, Y., Wohlgemuth, T., Sansonnens, B. & Buttler, A. (2001). *Die biogeographischen Regionen der Schweiz. Erläuterungen und Einteilungsstandard*. Bern.

Grant, E.H.C., Muths, E., Katz, R.A., Canessa, S., Adams, M.J., Ballard, J.R., Berger, L., Briggs, C.J., Coleman, J.T., Gray, M.J., Harris, M.C., Harris, R.N., Hossack, B., Huyvaert, K.P., Kolby, J., Lips, K.R., Lovich, R.E., McCallum, H.I., Mendelson, J.R., Nanjappa, P., Olson, D.H., Powers, J.G., Richgels, K.L., Russell, R.E., Schmidt, B.R., Spitzen-van der Sluijs, A., Watry, M.K., Woodhams, D.C. & White, C.L. (2017). Using decision analysis to support proactive management of emerging infectious wildlife diseases. *Frontiers in Ecology and the Environment*, **15**, 214–221.

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

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R. & Wintle, B.A. (2015a). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography,* 1–17.

Guillera-Arroita, G., Lahoz-Monfort, J.J., McCarthy, M.A. & Wintle, B.A. (2015b). Threatened species impact assessments: survey effort requirements based on criteria for cumulative impacts. *Diversity and Distributions*, **21**, 620–630.

Gvoždík, V., Benkovský, N., Crottini, A., Bellati, A., Moravec, J., Romano, A., Sacchi, R. & Jandzik,

D. (2013). An ancient lineage of slow worms, genus Anguis (Squamata: Anguidae), survived in the Italian Peninsula. *Molecular Phylogenetics and Evolution*, **69**, 1077–1092.

Hecnar, S. & M'Closkey, R. (1997). Spatial scale and determination of species status of the green frog. *Conservation Biology*, **11**, 670–682.

Hofer, U., Monney, J.-C. & Dusej, G. (2001). *Die Reptilien der Schweiz: Verbreitung, Lebensräume, Schutz*. Birkhäuser, Basel, Switzerland.

Isaac, N.J.B. & Pocock, M.J.O. (2015). Bias and information in biological records. *Biological Journal of the Linnean Society*, **115**, 522–531.

Jackson, J.T., Weckerly, F.W., Swannack, T.M. & Forstner, M.R.J. (2006). Inferring absence of Houston toads given imperfect detection probabilities. *Journal of Wildlife Management*, **70**, 1461–1463.

Joseph, L.N. & Possingham, H. (2008). Grid-based monitoring methods for detecting population declines: Sensitivity to spatial scale and consequences of scale correction. *Biological Conservation*, **141**, 1868–1875.

Kéry, M. (2002). Inferring the absence of a species: a case study of snakes. *The Journal of wildlife management*, **66**, 330–338.

Kéry, M. & Royle, J.A. (2015). *Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS: Volume 1: Prelude and Static Models*. Academic Press.

Kéry, M. & Schmidt, B.R. (2008). Imperfect detection and its consequences for monitoring for conservation. *Community Ecology*, **9**, 207–216.

Kéry, M., Spillmann, J.H.J., Truong, C. & Holderegger, R. (2006). How biased are estimates of extinction probability in revisitation studies? *Journal of Ecology*, **94**, 980–986.

Kindler, C., Chèvre, M., Ursenbacher, S., Böhme, W., Hille, A., Jablonski, D., Vamberger, M. & Fritz, U. (2017). Hybridization patterns in two contact zones of grass snakes reveal a new Central European snake species. *Scientific Reports*, **7**, 7378.

Koehler, J.J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, **19**, 1–53.

Levin, S.A. (1992). The Problem of Pattern and Scale in Ecology. *Ecology*, **73**, 1943–1967.

MacKenzie, D.I., Nichols, J.D. & Lachman, G. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

MacKenzie, D.I. & Royle, J.A. (2005). Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology*, **42**, 1105–1114.

McArdle, B.H. (1990). When are rare species not there? *Oikos*, **57**, 276–277.

Moilanen A. (2002). Implications of empirical data quality to metapopulation model parameter estimation and application. *Oikos*, **96**, 516–530.

Monney, J.-C. & Meyer, A. (2005). *Rote Liste der gefährdeten Arten der Schweiz: Reptilien*.

Moore, A.L. & McCarthy, M.A. (2016). Optimizing ecological survey effort over space and time (N. Lecomte, Ed.). *Methods in Ecology and Evolution*, **7**, 891–899.

Pellet, J. & Schmidt, B.R. (2005). Monitoring distributions using call surveys: estimating site occupancy, detection probabilities and inferring absence. *Biological Conservation*, **123**, 27–35.

Plummer, M. (2003). JAGS: A program for analysis of bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, **124**.

Pollock, K.H., Nichols, J.D., Simons, T.R., Farnsworth, G.L., Bailey, L.L. & Sauer, J.R. (2002). Large scale wildlife monitoring studies: statistical methods for design and analysis. *Environmetrics*, **13**, 105–119.

Preston, F.W. (1979). The Invisible Birds. *Ecology*, **60**, 451–454.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Reed, J.M. (1996). Using Statistical Probability to Increase Confidence of Inferring Species Extinction. *Conservation Biology*, **10**, 1283–1285.

Sakamoto, Y., Kumagai, N.H. & Goka, K. (2017). Declaration of local chemical eradication of the Argentine ant: Bayesian estimation with a multinomial-mixture model. *Scientific Reports*, **7**, 3389.

Schmidt, B.R., Meier, A., Sutherland, C. & Royle, J.A. (2017). Spatial capture – recapture analysis of artificial cover board survey data reveals small scale spatial variation in slow-worm Anguis fragilis density. *Royal Society Open Science*, **4**, 170374.

Schmidt, B.R. & Zumbach, S. (2018). Amphibian conservation in Switzerland. *Amphibian Biology, Volume 11: Status of conservation and declines of Amphibians: Eastern Hemisphere, Part 5: Northern Europe.* (eds H. Heatwole & J.W. Wilkinson).

Sewell, D., Guillera-Arroita, G., Griffiths, R.A. & Beebee, T.J.C. (2012). When is a species declining? Optimizing survey effort to detect population changes in reptiles. *PloS one*, **7**, e43387.

Skelly, D.K., Yurewicz, K.L., Werner, E.E. & Relyea, R. a. (2003). Estimating Decline and Distributional Change in Amphibians. *Conservation Biology*, **17**, 744–751.

Sliwinski, M., Powell, L., Koper, N., Giovanni, M. & Schacht, W. (2016). Research design considerations to ensure detection of all species in an avian community. *Methods in Ecology and Evolution*, **7**, 456–462.

Su, Y.-S. & Yajima, M. (2015). *R2jags: Using R to Run 'JAGS'*. R package version 0.5-7. Available from: https://CRAN.R-project.org/package=R2jags.

Tingley, M.W. & Beissinger, S.R. (2009). Detecting range shifts from historical species occurrences: new perspectives on old data. *Trends in Ecology & Evolution*, **24**, 625–633.

Villejoubert, G. & Mandel, D.R. (2002). The inverse fallacy: an account of deviations from Bayes's theorem and the additivity principle. *Memory & cognition*, **30**, 171–178.

Wintle, B.A. & Kavanagh, R. (2005). Estimating and dealing with detectability in occupancy surveys for forest owls and arboreal marsupials. *Journal of Wildlife Management*, **69**, 905–917.

Wintle, B.A., Walshe, T. V., Parris, K.M. & McCarthy, M.A. (2012). Designing occupancy surveys and interpreting non-detection when observations are imperfect. *Diversity and Distributions*, **18**, 417–424.

Yoccoz, N.G., Nichols, J.D. & Boulinier, T. (2001). Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution*, **16**, 446–453.

SUPPLEMENTARY INFORMATION

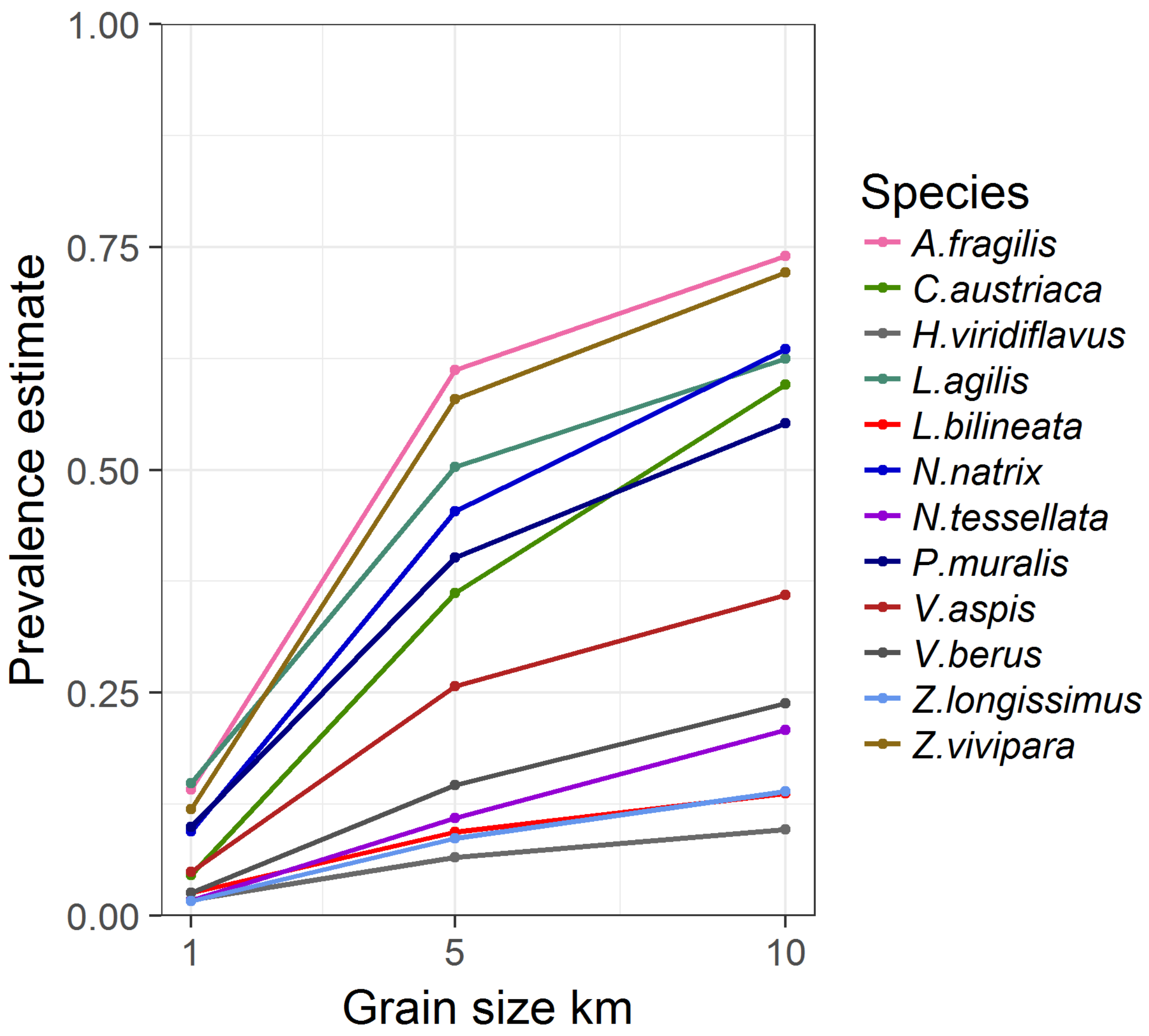


Fig S3.1. Prevalence estimates for each species, as derived from the Info Fauna Karch database (see methods), as assessed at 3 grain sizes: 1x1 km, 5x5 km, and 10x10 km.

CHAPTER FOUR

The impact of false-positive and false-negative observations on occupancy and trends derived from volunteer monitoring data

Sam S. Cruickshank [1*], Christoph Bühler [2], Benedikt R. Schmidt [1,3]

Manuscript in review

---

[1] *Department of Evolutionary Biology and Environmental Studies, University of Zürich, Zürich, Switzerland*
[2] *Hintermann & Weber AG, Reinach, Switzerland*
[3] *Info Fauna karch, Neuchâtel, Switzerland*
*sam.cruickshank@ieu.uzh.ch

ABSTRACT

1. Volunteer collected data are an important source of information used in species management decisions, yet concerns are often raised over the quality of such data. Two major forms of error exist in occupancy data; failing to observe a species when present (imperfect detection, false-negative errors, or errors of omission), and falsely reporting a species as present (false-positive errors or errors of commission). Estimating the frequency of these errors without the need for auxiliary data would allow dataset quality to be quantified and prevent the inference of biased or even erroneous trends.

2. We developed a dynamic occupancy model to estimate and adjust for both false-negative and false-positive error rates and produce unbiased estimates of occupancy. We validated this model using simulations and applied our model to 15 years of data from a volunteer monitoring scheme covering 12 species across 648 sites in Switzerland. By comparing occupancy from our model to that of standard occupancy models and the raw data, we quantified how ignoring observation errors influenced occupancy and trend estimation.

3. Simulations demonstrated that our model performed well under a range of different scenarios and required only a small number of unambiguous observations in order to resolve parameter identifiability issues. Precision of occupancy estimates was reduced at lower detection probabilities, but substantial bias in occupancy only arose when high false-positive error rates were combined with low detection probabilities.

4. Our volunteer monitoring data exhibited both imperfect detection and false-positive errors, however, not strongly enough to bias occupancy estimates. Most species had low false-positive error rates, although some common species had rates exceeding 5%. Accounting for false-positive rates as well as imperfect detection led to quantitative differences in occupancy, however annual changes (trends) remained consistent even when these effects were ignored.

5. Synthesis and applications: We present a model that represents an intuitive way of quantifying the quality of volunteer monitoring datasets, and which can produce unbiased estimates of occupancy despite the presence of multiple types of observation error. We finally present recommendations to aid the design of future monitoring programs such that they can readily estimate and account for false-positive error rates.

Keywords :

citizen science, false-positive, imperfect detection, occupancy modelling, trend, volunteer monitoring, observation bias

INTRODUCTION

A recognition that harnessing cheap and widespread volunteer networks can allow data collection at otherwise unfeasible scales has led to a dramatic increase in citizen-science projects over recent years (Silvertown 2009). Despite their increase in popularity, there are concerns over the utility of such volunteer-collected data (van Strien et al. 2001; Schmeller et al. 2009; Crall et al. 2011; Lewandowski & Specht 2015), with a major issue being the perception that such data are of lower quality than that collected by professional scientists (Foster-Smith & Evans 2003; Bird et al. 2014; Ratnieks et al. 2016). Volunteer-collected data is often the main source of information from which conservation and management decisions are made (Stem et al. 2005). There is therefore a need for methods that can evaluate the reliability of datasets collected by volunteers and account for errors, thereby increasing confidence that management decisions are made using accurate and robust information.

Attempts to assess the quality of citizen-science data typically do so by making comparisons between data collected simultaneously by volunteers and scientists (Fitzpatrick et al. 2009; Forrester et al. 2015; Dennis et al. 2017). Such approaches assume that data from expert scientists represent biological truth and any discrepancies represent errors on the behalf of volunteers. Such approaches seem flawed when considering that 'experts' may make errors as frequently as volunteers (Burgman *et al.* 2011; Austen *et al.* 2016), and indeed the experience of engaged amateur naturalists have long been recognised and valued as highly accurate by scientists (Tansley 1904; Pocock et al. 2015). Evaluating quality in relation to a benchmark with unknown accuracy seems an imperfect approach; alternative methods allowing the evaluation of data quality without relying on such comparisons are to be preferred.

Within the context of species occupancy monitoring, there are several ways in which observer data can contain biases. Imperfect detection, or the failure to detect a species when present, is acknowledged as ubiquitous in monitoring data (Kéry & Schmid 2004; Kéry & Schmidt 2008; Kellner & Swihart 2014), and statistical frameworks to estimate and account for this are well developed (Pollock *et al.* 2002; Royle & Dorazio 2008; Guillera-Arroita 2017). The relationship between the observed proportion of occupied sites, C, and the true proportion of occupied sites, N, can be considered as (Nichols 1992):

$$C = N * pT \tag{4.1}$$

with pT representing the detection probability, which can be estimated using information from a variety of methods such as repeat surveys or distance sampling (Guillera-Arroita 2017). A second form of bias, which has to date received much less attention, is that of false-positive errors. These may occur as a result of species misidentification or other mistakes that result in an observer reporting a species as present when in reality the site is not occupied. In the presence of false-positives, our conceptual equation must be modified to:

$$C = (N * pT) + \ (1 - N) * pF \tag{4.2}$$

where pF is the false-positive rate. Because the model describes the two types of observation error that may occur in species presence/absence surveys, this simple equation conceptualizes a better approach to assessing the data quality of citizen science monitoring data. Importantly, as error rates can be estimated directly from the data (Chambert et al. 2015), there is no need for comparative benchmark data.

The simultaneous estimation of false-negative and false-positive rates poses computational difficulties, as any given set of detection histories can be equally well explained by multiple sets of parameter values (Royle & Link 2006). This creates parameter identifiability issues that may be avoided through the incorporation of extra information. Royle and Link (2006) constrained false-positive rates to be lower than the true detection rate, whereas the models of Miller et al (2011) and Chambert et al (2015) incorporate information from a second detection method in which false-positive errors were considered impossible. Although well-designed survey protocols should in most cases ensure that true detection rates are higher than false-positive error rates, approaches that avoid the need to make such strong *a priori* assumptions are preferable. Furthermore, particularly for existing datasets, the availability of secondary datasets by unambiguous methods restricts the situations in which the model of Miller et al (2011) can be directly applied.

Applications of occupancy models have focussed predominantly on assessing observation error over a single season (though see Sutherland et al. 2013; Miller et al. 2013). However, a major advantage of volunteer monitoring is the possibility for data collection at spatial and temporal scales that may otherwise be logistically unfeasible. Thus, one of the best potential applications for citizen science programs is in monitoring long-term species and population trends (Dickinson et al. 2010). The development of multi-year (dynamic) occupancy models that simultaneously take into account false-positives and false-negatives is necessary to analyse such data.

In this study, we developed a dynamic occupancy model that estimates both imperfect detection and false-positive error rates to produce unbiased estimates of occupancy from which population trends can be derived. We show how some of the strong constraints of earlier models can be relaxed. We used simulations to identify both the underlying biological and detectability conditions under which our model produced reliable inference. We then applied our model to 15 years of volunteer-collected data for 12 amphibian species spanning 648 sites in Switzerland. We quantified false-positive and false-negative error rates, and assessed whether failing to adjust for false-positive errors quantitatively affected subsequent estimation of occupancy and long-term occupancy trends. Our model represents a method by which the quality of occupancy datasets can be reliably assessed without the need for a comparison with external benchmark data. Importantly, we outline the circumstances under which unmodelled errors become problematic, and demonstrate that other than in extreme situations, our model produces reliable measures of occupancy and population trends despite the presence of these errors.

MATERIALS AND METHODS

General Model

We combined elements of the single-season false-positive occupancy model of Miller et al (2011; Chambert et al. 2015) with a standard dynamic occupancy model (Royle & Kéry 2007) to produce a dynamic false-positive/false-negative occupancy model. By classifying a subset of observations as unambiguous, we can estimate the frequency of false-positive and false-negative records, and produce unbiased occupancy estimates.

We started with the initial Bayesian formulation of a standard dynamic occupancy model (MacKenzie et al. 2003; Royle & Kéry 2007):

$$z_{i,1} \sim Bernoulli\ (initial.occupancy) \quad (4.3)$$

$$z_{i,t+1} \sim Bernoulli\ \left(z_{i,t} * persistence\right) + \ \left(1 - z_{i,t}\right) * colonisation \quad (4.4)$$

$$y_{i,j,t} | z_{i,t} \sim Bernoulli\ \left(z_{i,t} * pT\right) \quad (4.5)$$

This set of equations can be considered in two parts: equations 4.3 and 4.4 describe the true state process, the occupancy states $z_{i,t}$ of *i* sites in *t* years. Equation 4.5 describes the observation process, relating these true states to the observed data, $y_{i,j,t}$, at each of *j* surveys within each year; all occupancy models require multiple visits to a site within years.

The initial occupancy state of a site is determined by the probability *initial.occupancy* (equation 4.3), which may be specified using site-specific covariates (Royle & Dorazio 2008). Occupancy states in subsequent primary periods are then described as a Markovian process, driven by patch persistence (i.e., persistence of the population in a patch) and colonisation rates (equation 4.4). Colonisation and persistence rates take constant values or vary in response to covariates, as required by design or biological considerations (Kéry & Schaub 2011). Equation 4.5 links this description of the state process to the observed data; a further Bernoulli trial specifies that a non-observation occurs if a site is not occupied, and a detection occurs with probability pT (the true detection probability) if the site is truly occupied. Thus, the model only accounts for false-negatives (i.e., imperfect detection).

Following the approach of Chambert et al (2015), several changes to the observation sub-model are necessary to allow for false-positive errors. In equation 4.5, a detection can only arise if $z_{i,t} = 1$, that is, the site is occupied. To also allow for false-positive observations, this equation must be modified to:

$$y_{i,j,t} | z_{i,t} \sim Bernoulli\ \left(z_{i,t} * pT + \ \left(1 - z_{i,t}\right) * pF\right) \quad (4.6)$$

The addition of this second term allows a detection to occur with probability pF (the false-positive error rate) if the site is unoccupied (i.e. $z_{i,t} = 0$), and with probability pT if the site is occupied.

Finally, a second dataset $w_{i,j,t}$, containing only unambiguous detections is necessary to allow the model to resolve parameter identifiability issues (Miller et al. 2011). Here we use a subset of our whole dataset $y_{i,j,t}$ to create this second dataset (details in *Case study: Long-term amphibian monitoring*). Elements of w take value 1 if the observation is unambiguous, and 0 otherwise. A final equation links this observation data to the to the underlying occupancy states. As only true positive detections can arise in this dataset, the observation process for this data resembles equation 4.5 :

$$w_{i,j,t}|z_{i,t} \sim Bernoulli\left(z_{i,t} * pU\right) \tag{4.7}$$

Here pU represents the probability of an unambiguous detection being made at an occupied site. The above modifications are sufficient to incorporate false-positive error estimation into a standard dynamic occupancy model.

We extended this core model further to allow for variation in the number of sites monitored between years (which applied for our amphibian case-study). To prevent the model estimating sites as occupied for years when they were not part of the monitoring dataset, we created an *i* by *t* matrix *extant*, which was incorporated into descriptions of the state process (equations 4.8 and 4.10). Elements take value 1 for all years when the site was part of the monitoring dataset, and 0 otherwise. Associated with this is a vector, $initial_i$ which specifies the first primary sampling period in which the site exists. These additions ensure that occupancy is deterministically zero when a site was not part of the monitoring program.

The final specification of the core model is:

$$z_{i,initial} \sim Bernoulli\left(initial.occupancy_i * \ extant_{i,initial}\right) \tag{4.8}$$

$$\Psi_{i,t+1} = \ \left(z_{i,t} * persistence\right) + \left(1 - z_{i,t}\right) * colonisation \tag{4.9}$$

$$z_{i,t+1} \sim Bernoulli\left(\Psi_{i,t+1} * extant_{i,t+1}\right) \tag{4.10}$$

$$y_{i,j,t}|z_{i,t} \sim Bernoulli\left(z_{i,t} * pT + \ \left(1 - z_{i,t}\right) * pF\right) \tag{4.6}$$

$$w_{i,j,t}|z_{i,t} \sim Bernoulli\left(z_{i,t} * pU\right) \tag{4.7}$$

Simulations

We ran simulations to assess the performance of the model under a range of scenarios. Our primary goal was to identify regions of parameter space in which the model performs poorly - for example by inducing large bias or low precision (large credible intervals; CI) around occupancy estimates. Additionally, we wanted to see whether assuming constant persistence and colonisation rates substantially influenced parameter estimation when the underlying data contained variation in these parameters. Finally, we wanted to assess whether uncertainty in parameter estimates was increased when survey data were not available for all sites in all years- a feature seen in the rotating panel design of our case study.

We simulated data under four scenarios, each covering a wide range of values for all key parameters (see SI 4.1 for further details on the 729 parameter combinations):

1. Surveys carried out in all years, no variation in colonisation and persistence rates.
2. Surveys carried out in all years, time-varying colonisation and persistence rates.
3. Scenario 1, with data for 66.6% of the years converted to missing data.
4. Scenario 2, with data for 66.6% of the years converted to missing data.

Models estimating false-positive error rates may suffer from a lack of parameter identifiability. Royle and Link (2006) recommend constraining parameter values such that false-positive detection rates are lower than the true positive detection rate. We relaxed this hard constraint by instead using informative priors. We assigned true detectability pT the standard vague prior uniform(0,1). To reflect our prior belief that false-positive rates are likely low, we assigned false-positive rates a beta(1,2) prior. We took a similar approach to priors for colonisation and persistence rates, assigning beta(1,2) and beta(2,1) priors, respectively, to reflect our belief that for our study species, patch colonisation rates are more likely to be lower than patch persistence rates (Cruickshank et al. 2016).

We examined our simulation results and assessed the degree of non-convergence (defined as Rhat values greater than 1.05), bias, and uncertainty (CI width) in key parameters (occupancy rates, pT, pF, colonisation and persistence), as well as the presence of any trends in occupancy bias through time.

Case study: Long-term amphibian monitoring

We applied our model to a long-term volunteer monitoring dataset that is used to inform conservation management in Canton Aargau, Switzerland (Meier & Schelbert 1999). We used 15 years of detection/non-detection data at 648 amphibian breeding sites distributed among 10 core areas within the study region. Core areas contained 21-96 sites, which represent all potential breeding sites within each area for 8 target amphibian species (Meier & Schelbert 1999). During the course of the study, new ponds were built and added to the inventory of sites monitored; some others were destroyed or otherwise removed from the program (SI. 4.2).

The survey protocol followed a rotating panel design (McDonald 2003), whereby 1-5 core areas were selected each year, and all sites within these core areas were surveyed. Thus, each core area was comprehensively surveyed 4-5 times (mean 4.3) throughout the 15 years of the study (SI. 4.3). When surveyed, each site was visited on 3 occasions during pre-defined time windows coinciding with the amphibian breeding season. To ensure comparable survey effort between sites, the durations of surveys were allocated to be constant relative to the water surface area of the site. Volunteers were asked to record all amphibian species seen; in addition, they assigned a population size class for each species, relating to the combined number of adults, juveniles, and calls heard (see SI 4.4 for species-specific population size classes and species codes). Further information on the survey protocol can be found in Roth *et al.* (2016). In total, data were collected for 11 species, plus one hybridogenetic species complex (*Pelophylax esculentus*: hereafter *P.escu* complex).

As complete checklists were reported, we were able to infer non-observation of species not explicitly reported during a visit (Kéry et al. 2010). For each species, we created 2 datasets; one containing observation/non-observation data, and another containing population size class data.

Case study: model

Surveys were carried out throughout our study region in both the 1980s and 1990s to identify all breeding sites (Flory 1999). Not all species were found in each core area, thus we used information from these surveys to create a binary covariate (*historic*) to allow for different initial occupancy rates among core areas. This covariate took value 1 if any observations of the species were made within a core area in the historic surveys, and this was used to estimate initial occupancy states $z_{i,initial}$ for site i in initial survey year *initial* as:

$$logit(initial\ occupancy_{area_i}) = \alpha_1 + (\alpha_2 * \ historic_{area_i}) + error_{area_i} \tag{4.11}$$

$$z_{i,initial} \sim Bernoulli\ (initial.occupancy_i * \ extant_{i,initial}) \tag{4.8}$$

Here $area_i$ denotes the core area to which a given site *i* belongs. The coefficients α_1 and α_2 were given vague uniform priors (-20,20). We additionally included a core-area specific random effect term:

$$error_{area_i} \sim Normal\ (0, \tau)$$

$$\tau \sim Gamma\ (1,1)$$

For our citizen science dataset, all sites monitored in a year experienced the same false-positive and false-negative error rates, though we allowed these rates to vary between years. We used the same mean rate priors as those specified in the simulations, and annual rates were drawn as normally distributed random variables around these means. Standard deviations for each rate (on the logit scale) were assigned gamma(1,1) priors.

Species-specific colonisation and extinction rates were allowed to vary between core areas, but remained constant through time. We used a similar parameterisation as for the detection variables, with core-area specific colonisation and persistence rates drawn from normal distributions with gamma(1,1) standard deviation terms.

As in our simulations, we followed Miller et al. (2011) in using ‘confirmed’ observations to aid model convergence. We had no way to confirm *a posteriori* whether or not any observations made by the volunteers in the study was correct or false, therefore we used the associated estimates of population size to assign records as either confirmed or unconfirmed. We made the assumption that volunteers are less likely to misidentify a species (and thus create a false-positive record) when they reported large numbers of individuals than when they only saw a few individuals. To test this assumption, we ran two iterations of this model; one where all observations of population class four (the highest possible) were considered confirmed (hereafter the *FP-high* model), and a second that considered any observation with a population class greater than 1 (the smallest possible) as confirmed (hereafter *FP-low* model).

Data analysis

For each of the 12 species, we applied the FP-high and FP-low models, as well as a standard dynamic occupancy model considering only imperfect detection (hereafter *PT-model*; MacKenzie et al. 2003). We furthermore obtained unadjusted occupancy estimates from the raw data by assuming a species was absent from a surveyed site if it went undetected in all surveys in a given year, and present if it was detected at least once. All models were run in JAGS (Plummer 2003) using R2jags (Su & Yajima 2015) in R (R Core Team 2017). Code for the full model can be found in SI 4.5. For each model, we ran 3 chains of 150,000 iterations with a burn-in of 10,000, thinning the remaining samples by 1 in 30. Convergence was assessed using the Brooks-Gelman-Rubin statistic (Kéry & Schaub 2011).

For each model (FP-high, FP-low, PT-only, and unadjusted) we derived the relevant detection parameters and number of occupied sites for each core area under the different observation error scenarios. We report the number of occupied sites, rather than occupancy rates, as the number of monitored sites changed between years; for simplicity, we refer to this as occupancy except where this would lead to ambiguity. We also derived annual trends for each core area, which we define as the absolute change in occupancy between consecutive years (i.e. $\Psi_{t-1} - \Psi_t$). For both occupancy estimates and trends, we made pairwise comparisons between models to identify differences between observation-effect scenarios. Full details of the methods used to make these comparisons is given in SI 4.6.

RESULTS

Simulations

Our simulations demonstrated that our model is well able to estimate occupancy and detectability parameters under a range of realistic conditions. However, the results clearly show that true detection rates and false-positive rates can interact to induce bias in occupancy. When per-visit true-detection is high (0.85), even high false-positive rates (0.1) do not bias occupancy estimates (Fig. 4.1). With an intermediate detection probability of 0.55, such high false-positive rates lead to systematic overestimation of occupancy, and at low levels of true-detection (0.35), the situation is exacerbated and even false-positive rates of 0.05 can lead to extreme overestimation of occupancy (Fig. 4.1). This interaction between low detection and high false-positive rates can also lead to overestimation of colonisation, decreased precision of persistence, and underestimation of true detection (SI 4.7). However, for the range of true detection and false-positive error rates estimated for the species in our study, our model is able to accurately estimate occupancy. In most simulations, there was no evidence of temporal trends in occupancy bias (SI 4.7). There were two conditions when substantial temporal biases occasionally occurred (i.e. estimated occupancy rates diverged from true rates by more than 0.01 per year); firstly, when the highest false-positive and lowest detection rates occurred, and rarely when missing data were present and high pF combine with intermediate pT rates (SI 4.7).

When we applied the model to simulated datasets containing many missing years of data, model convergence became problematic: almost all models in simulation scenarios 1 and 2 converged, in contrast to scenarios 3 and 4 where non-convergence occurred in nearly 25% of simulations. Missing data decreased the precision of occupancy estimates, although the degree of bias remained the same relative to datasets without missing data.

Readers are directed to the interactive app (SI 4.7) where the effects of various parameter combinations, model types, and the presence of missing data on parameter estimation can be visualised.

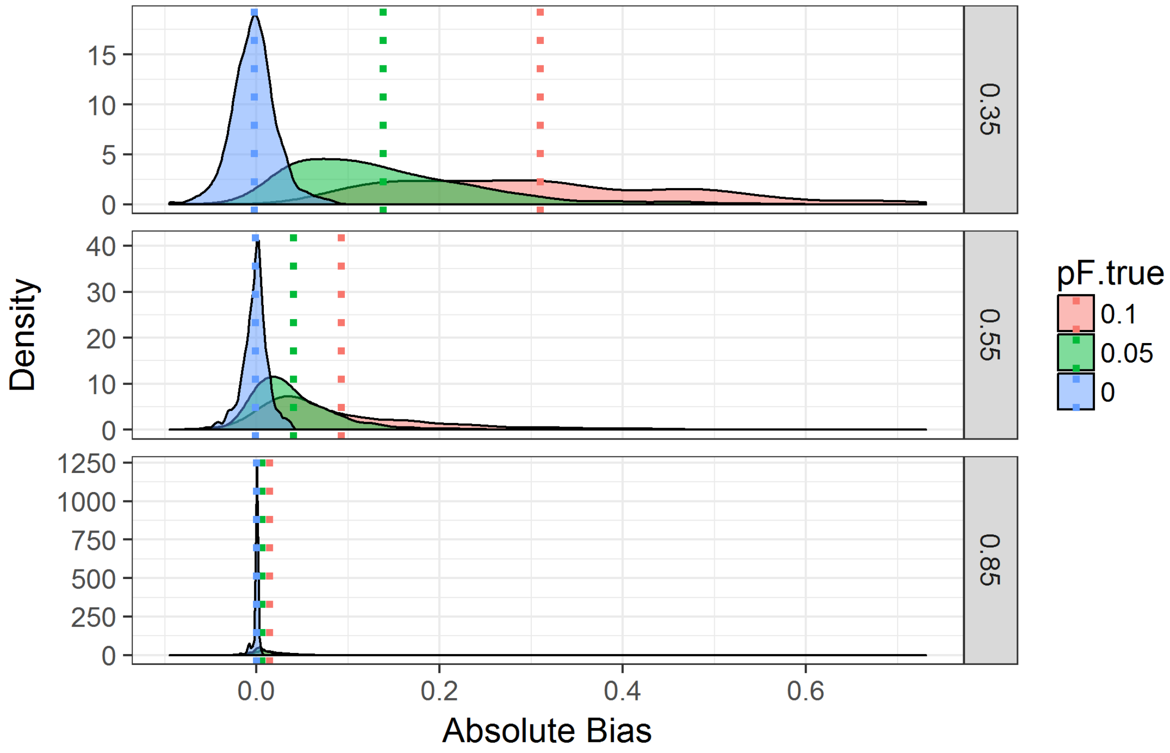


Figure 4.1- Density plot of absolute bias in occupancy estimates across all simulated datasets for simulation scenario 1. Top panel shows bias for datasets where detectability was high (0.85), middle panel shows results for simulations where detectability was 0.55, and bottom panel is for detectability of 0.35. Vertical bars represent the median bias for each group.

Case study: Volunteer monitoring

There was wide variation in true detection rates between species in our volunteer dataset, and for some species detection differed substantially between the different occupancy models (Fig. 4.2). Both FP models always produced higher estimates of true detection probability than the PT-only model. False-positive error rates were low for most species, although for four common species (*Bufo bufo*, *Ichthyosaura alpestris*, *Pelophylax esculentus* complex, *Rana temporaria*) at least one of the FP models estimated false-positive error rates exceeding 5%. Our simulation results suggest that the combinations of false-positive and true detection rates estimates for all species fell well within the range of parameter space for which occupancy estimates are unbiased.

For most species, both FP models produced comparable detection and false-positive estimates, although detection rates were more than 0.15 higher in the PF-low models for the two rarest species, *Lissotriton vulgaris* and *Hyla arborea*. However, these differences had little effect on the resulting occupancy estimates, which were comparable between these models (SI 4.8).

The differences in the observation effects accounted for in each of the occupancy models sometimes resulted in differing occupancy estimates (Figs 4.3-4.4, SI 4.8-4.9). For all species, a proportion of occupancy estimates differed significantly among models, although the magnitude of the differences was typically small (SI 4.8). When differences existed, PT-only models produced higher estimates than the FP-models. The unadjusted occupancy estimates were never higher than the PT-only model estimates, but were not consistently higher than those from the FP-models (SI 4.8). CI widths were 6.07 sites (sd: 4.17) for pT-only models, and 4.54 (3.27) and 5.45 (3.50) for FP-low and FP-high models, respectively (averaged across all species). However, in line with simulation results, we observed differences among species- with lower precision for species with higher false-positive error rates.

A similar pattern was seen in our measure of population trends. For all species, there were some occasions when annual changes in occupancy rates (our measure of trend) differed among the models (Table 4.1). However, the magnitude of such differences were small; in every comparison, the 95% CI included zero. There was therefore no strong evidence to suggest that occupancy trends, as measured here, differ between any of the models we considered.

Therefore, with some rare exceptions when occupancy differed slightly between models (see SI 4.8), the occupancy rates and trends did not differ significantly between models accounting for false-positive and false-negative observation errors, models accounting only for false-negatives, and the unadjusted observational data.

Table 4.1. The proportion of occupancy trends that are significantly different from 0, and the overall mean difference and 95% credible intervals of annual trends for pairwise model comparisons. Italicised row denotes column means. Bold values denote that 95% credible intervals do not overlap with 0. Full species names are provided in SI 4.4.

	Proportion significant						Mean difference in annual trend (95% CRI)					
Species	PT$_{only}$ - FP$_{high}$	PT$_{only}$ - naïve	FP$_{high}$ - naive	PT$_{only}$ - FP$_{low}$	FP$_{high}$ - FP$_{low}$	FP$_{low}$ - naïve	PT$_{only}$ - FP$_{high}$	PT$_{only}$ - naïve	FP$_{high}$ - naive	PT$_{only}$ - FP$_{low}$	FP$_{high}$ - FP$_{low}$	FP$_{low}$ - naïve
ALOB	0.00	0.19	0.29	0.01	0.00	0.31	0.01 (-3.00:2.00)	-0.14 (-5.25:2.00)	-0.15 (-5.25:1.67)	0.00 (-3.00:2.00)	-0.01 (-2.00:1.00)	-0.14 (-5.25:1.67)
BOVA	0.00	0.28	0.31	0.00	0.00	0.35	0.04 (-4.00:3.00)	0.05 (-5.50:5.00)	0.02 (-5.00:4.25)	0.08 (-4.00:3.00)	0.05 (-4.00:3.00)	-0.03 (-4.50:4.00)
BUBU	0.00	0.14	0.27	0.00	0.00	0.23	-0.13 (-7.00:5.00)	0.62 (-7.00:6.00)	0.74 (-5.50:5.00)	-0.10 (-7.00:5.00)	0.02 (-6.00:4.00)	0.72 (-5.75:5.50)
EPCA	0.00	0.11	0.24	0.01	0.00	0.24	-0.07 (-3.00:2.00)	-0.04 (-3.25:2.33)	0.03 (-2.67:2.00)	-0.06 (-3.00:2.00)	0.01 (-3.00:2.00)	0.02 (-2.75:2.00)
HYAR	0.00	0.13	0.27	0.00	0.00	0.41	0.08 (-2.00:2.00)	0.46 (-2.67:3.00)	0.37 (-2.50:2.50)	0.21 (-2.00:2.00)	0.12 (-2.00:1.00)	0.25 (-2.50:2.00)
ICAL	0.00	0.20	0.21	0.00	0.00	0.20	0.74 (-7.00:6.00)	1.12 (-7.67:8.00)	0.38 (-6.33:5.5)	0.50 (-7.00:6.00)	-0.24 (-7.00:4.00)	0.62 (-7.00:6.50)
LIHE	0.00	0.13	0.19	0.00	0.01	0.19	0.16 (-4.00:3.00)	0.11 (-5.33:3.50)	-0.05 (-5.00:3.00)	0.23 (-4.00:3.00)	0.07 (-4.00:3.00)	-0.12 (-5.00:3.00)
LIVU	0.00	0.09	0.15	0.00	0.00	0.23	-0.08 (-3.00:2.00)	-0.06 (-3.00:1.80)	0.02 (-2.50:1.20)	-0.08 (-3.00:2.00)	0.00 (-3.00:1.00)	0.02 (-2.25:1.20)
PEES complex	0.01	0.18	0.19	0.02	0.00	0.24	0.24 (-5.00:4.00)	0.76 (-10.33:5.50)	0.53 (-11.33:4.75)	0.34 (-5.00:4.00)	0.11 (-5.00:3.00)	0.42 (-11.33:4.50)
PERI	0.00	0.06	0.07	0.00	0.00	0.11	-0.06 (-4.00:2.00)	-0.10 (-3.80:2.00)	-0.04 (-3.00:2.00)	-0.06 (-4.00:2.00)	0.00 (-3.00:2.00)	-0.04 (-3.00:2.00)
RATE	0.01	0.18	0.23	0.01	0.01	0.19	0.13 (-8.00:6.00)	0.84 (-11.00:8.00)	0.70 (-8.33:6.25)	0.18 (-8.00:6.00)	0.05 (-7.00:5.00)	0.66 (-9.00:6.25)
TRCR	0.00	0.26	0.28	0.00	0.00	0.33	0.06 (-2.00:1.00)	0.14 (-5.00:1.67)	0.09 (-5.00:1.67)	0.08 (-2.00:1.00)	0.02 (-2.00:1.00)	0.06 (-5.00:1.25)
	0.00	*0.16*	*0.22*	*0.00*	*0.00*	*0.25*	*0.09*	*0.31*	*0.22*	*0.11*	*0.02*	*0.20*

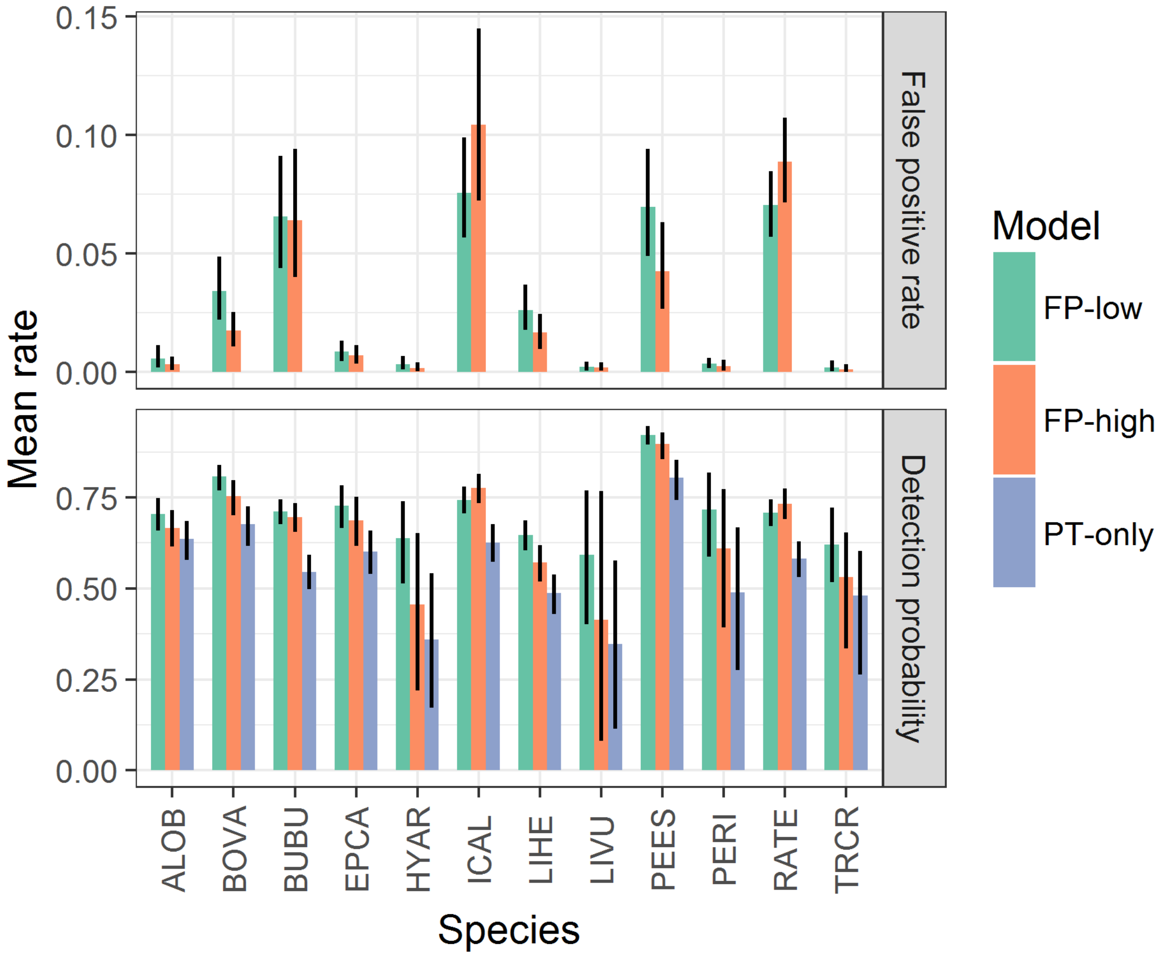


Figure 4.2- Detection probability and false-positive rates for the 12 species across 3 occupancy models. PEES denotes *P.esculentus* species complex.

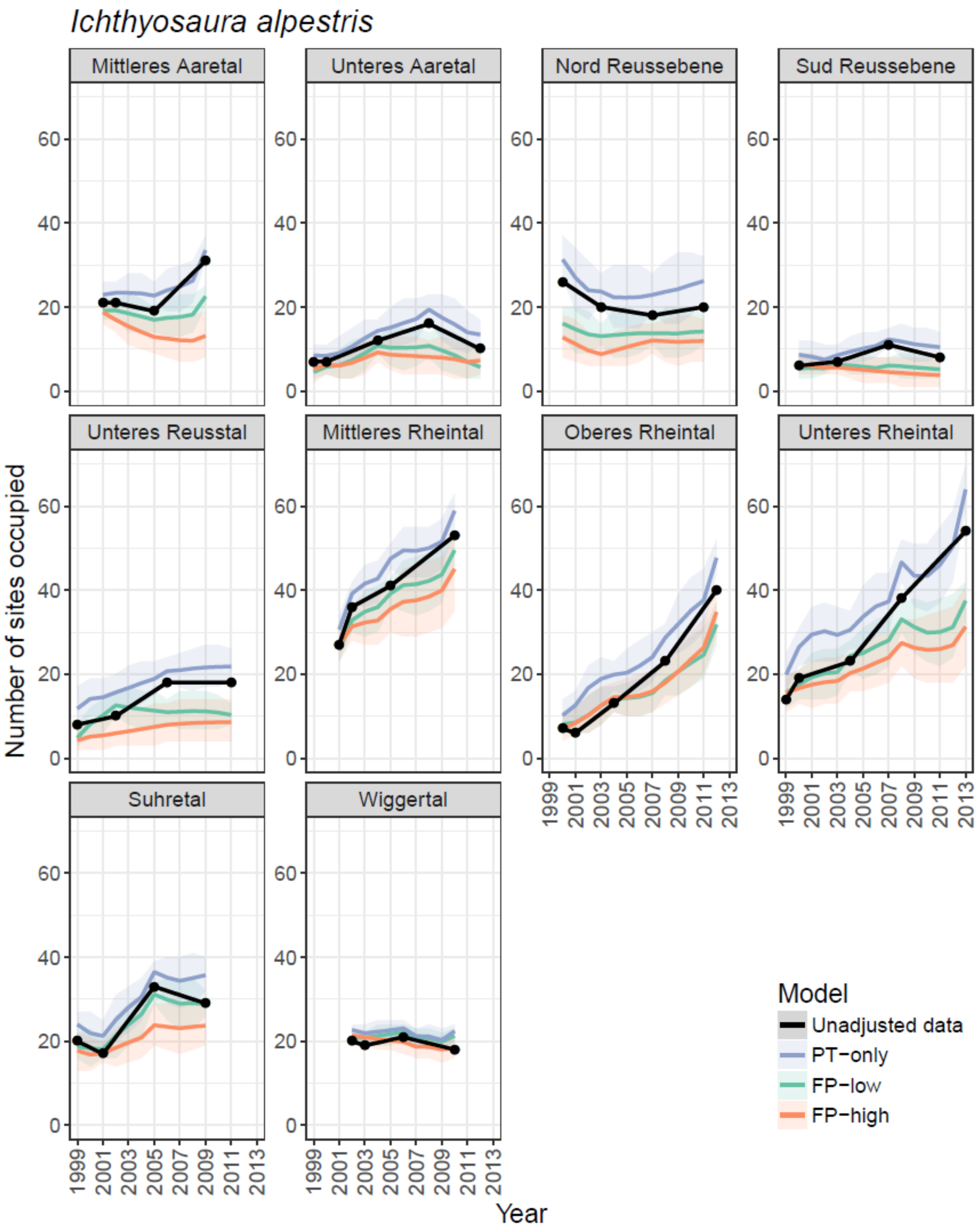


Figure 4.3- Number of occupied sites for a widespread species, *Ichthyosaura alpestris*, with a high estimated detection rate (pT) and false-positive rate (pF) over 10 core areas (panels).

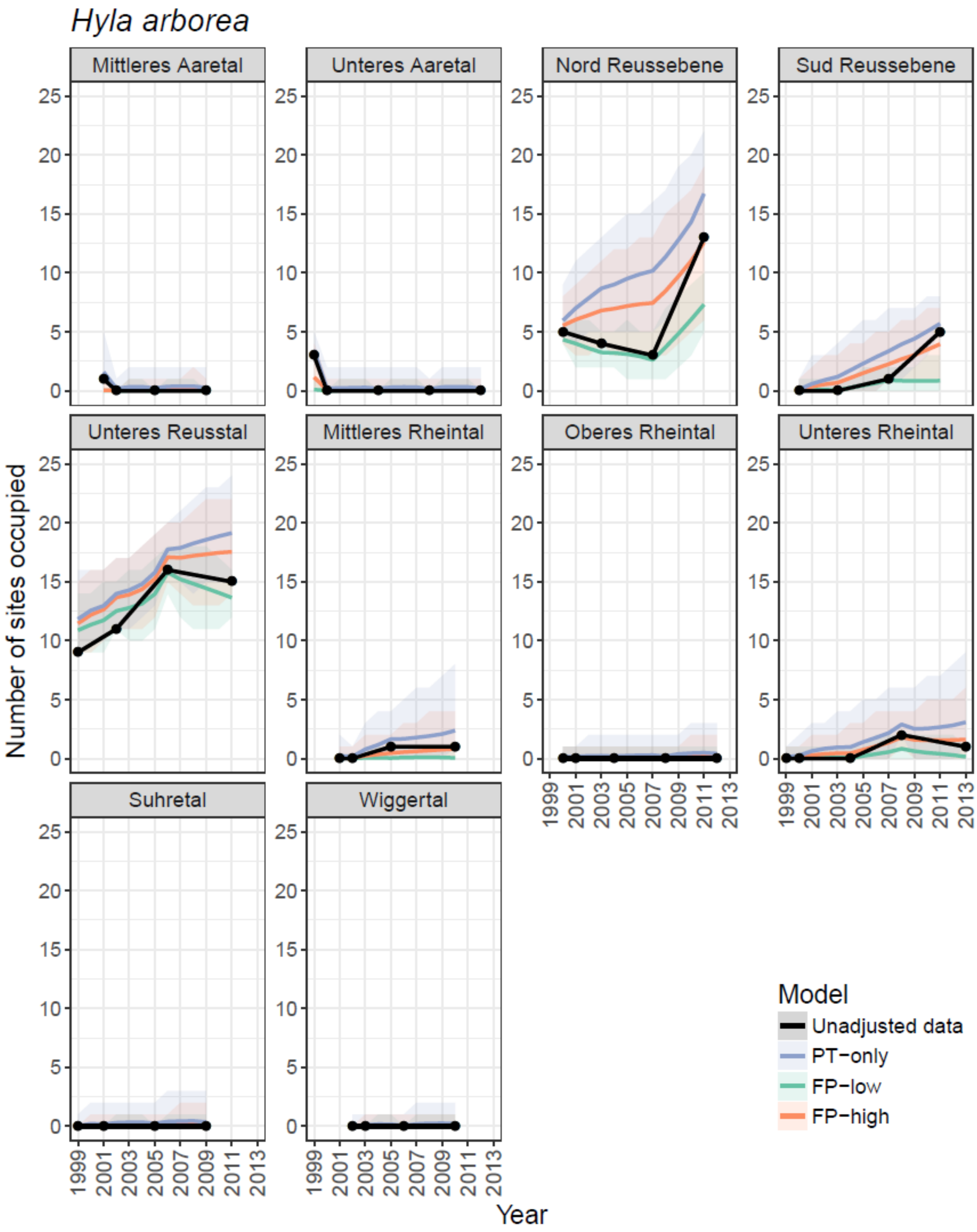


Figure 4.4. Number of occupied sites for a rare species, *Hyla arborea* over 10 core areas (panels).

DISCUSSION

Our model provides a method by which false-negative and false-positive error rates may be estimated and accounted for without the need for auxiliary data, and thus the quality of volunteer-collected occupancy datasets may be quantitatively assessed. We demonstrate that robust inference may be drawn from such datasets under a range of likely biological and sampling conditions, including when significant false-positive error rates exist. We have shown that as long as false-positive errors are modelled, their presence need not induce substantial occupancy bias unless detection probabilities are low and false-positives occur at a high rate. Our analysis of the quality of a long-term and large-scale volunteer amphibian monitoring program demonstrated that false-positive errors were rare for all but the most common species monitored. All species had sufficiently high detection probabilities that accurate estimates of occupancy could be derived regardless of the false-positive error rate.

A primary benefit of our model is in quantifying error rates from the dataset rather than relying upon auxiliary external data (such as expert opinion). We believe that detection probabilities and false-positive error rates represent two clear metrics that can be used to evaluate the quality of volunteer datasets (see equation 4.2). In our volunteer dataset we found that false-positive errors were rare for most species, but that high (>5%) error rates were possible - although detection probabilities were always sufficiently high to allow accurate occupancy estimation. False-positive error rates are most serious for rare species (Miller et al. 2011), and previous research on birds reported a rare-species bias (Farmer et al. 2012), whereby rarer species have higher false-positive rates than more common species. Encouragingly, our results show the opposite pattern, with rare species having only negligible error rates, and the more common species being those with higher rates. This suggests that volunteers in our study were cautious when reporting observations - if they were uncertain about the identity of a species, they appear more likely to conclude that the species was a common than a rare one. We suggest that this effect is likely to hold true in volunteer systems in which the suite of potential species is relatively small, and when volunteers are aware of the goals of the program, undertake some training in identification, and receive guidelines for reporting uncertain observations. In our case-study, dubious reports for rare species were double-checked and excluded from the dataset if considered implausible. Although it would be expected that such a process would act to reduce false-positive rates, so few records were removed in this process that a reanalysis of the data including these unreliable records had no effect on false-positive rates.

In our analysis of amphibian populations, we identified that failing to account for imperfect detection and/or false-positive errors sometimes led to quantitative differences in occupancy rates, but that the overall trends (annual changes in occupancy) were consistent between models. This suggests that the benefits of accounting for such errors may depend strongly on the goals of the monitoring program. For a conservation-manager trying to identify all sites that require conservation management, considering false-positives in monitoring data would help focus management at the appropriate locations. However, if the goal is long-term monitoring of population trends, our results suggest that the benefits of accounting for false-positive errors were slight- and program managers may decide that the results from standard occupancy modelling may be a sufficiently satisfactory method to achieve this goal.

All occupancy modelling is underpinned by the assumption that the true status of a site remains constant for all surveys within a season. In reality, this assumption of population closure may be violated - for example if individuals from one site temporarily use another location for foraging. In standard occupancy models, such a process would have two main effects; first, it would lead to the second site being wrongly classified as occupied when in fact there is no established population there. Secondly, at this second site, there would be only few positive detections within the survey history, which would lead to underestimation of detection rates and consequently further compound overestimation of the occupancy estimates. We see evidence of this in our results on tree-frogs (*H. arborea*); detection probabilities in our standard occupancy model (PT-only) are substantially lower than is typical for this species - which is easily detectable from calls (Cruickshank et al. 2016), and this led to inflated occupancy estimates (Fig. 4.4). Evidence exists that individuals frequently travel substantial distances within a breeding season in our study area (Angelone et al. 2011), suggesting that the closure assumption is violated, and that consequently there is a mismatch between the scale of monitoring (i.e. definition of a site) and scale of within-season movement for this species.

Standard occupancy models effectively model occurrences in a patch, and thus the presence of transient individuals leads to the site being classified as occupied. However, if one can ensure that only observations of populations that are established are considered as unambiguous (e.g. by recording signs of breeding), then the results of our false-positive model can instead effectively be interpreted as the presence of established populations- as rare observations of transient individuals would be classified as false-positive records. This distinction can be seen in Fig. 4.4; in Unteres Reusstal, tree-frogs have been established for many years and there is little evidence for large-scale movements (Angelone & Holderegger 2009). Consequently, the differences between FP and PT models are slight as there are few records of transients. In contrast, the species has recently been colonising Reussebene Nord (C. Bühler, *unpublished data*), and thus many observations of transient individuals led to large overestimates of occupancy by standard PT models. Our false-positive model suggests that the true rate of colonisation is substantially lower than would be inferred from standard models, as these rare observations of transient individuals at sites do not lead to classifications of such sites as occupied.

Allowing for false-positives in occupancy modelling requires the incorporation of some data-points representing reliable true observations. Past studies have utilised secondary datasets to obtain such data. In many cases (including ours) such extra data are not available, and alternative approaches will be necessary in order to classify some observations as 'true'. As species detectability is expected to increase with abundance (Tanadini & Schmidt 2011), we used population size class data associated with our observations to define observations of many individuals as true-positives. We applied models using different thresholds for a confirmed observation to test whether false-positives were more frequent in observations of few individuals. In two common species (*R. temporaria* and *I. alpestris*) false-positive rates were higher in the high-threshold model, which is suggestive that false-positive errors were made even when high abundances were reported. However, for all other species, the patterns observed were consistent with the idea that false-positive errors were only common in observations where few individuals were seen. Making and testing such hypotheses relating to the assumptions made in assigning observations as true is a good method by which false-positive models can be applied to existing datasets in which no observations can be considered true *a priori*.

Species monitoring studies have adapted to an understanding of the effects of imperfect detection, and are increasingly adjusting protocols to collect the necessary information to account for this issue (Kellner & Swihart 2014). In future studies, adapting protocols to allow a subset of observations to be confirmed as reliable is the only additional requirement necessary to allow false-positive errors to be simultaneously accounted for. This may be done by using a secondary detection method (Miller et al. 2011, 2013) or using experienced observers at a subset of sites (Chambert et al. 2015). In our models, non-identifiability was avoided even when as few as 2% of species records were treated as unambiguous, suggesting that only a small number of such observations are required. The consequences of false-positive errors are most extreme for rare species (Miller et al. 2011), which may also be hard to detect due to low abundances (Tanadini & Schmidt 2011). Our results suggest that false-positive errors only cause serious bias in combination with low detection probabilities. This suggests that modifying survey protocols to maximise species detection probability in a survey is likely to be the best way of ensuring accurate estimation of occupancy rates. However, modifications to minimise false-positive errors, such as informing volunteers as to how to report uncertain observations, requesting photographs of rare species report, or carrying out other data-quality assessments, may be more easily achieved; such efforts have previously been advocated by Miller et al (2015). Finally, we found that efforts to improve estimation of initial occupancy states was particularly important in our dynamic model; thus we recommend investing extra survey effort in the first year of multi-year monitoring, which would also lead to more accurate estimates of detection parameters.

One strength of the model is our construction of priors for the key parameters which avoids the need to make the restrictive hard constraints used in other models accounting for false-positives (Royle & Link 2006; Sutherland et al. 2013). Our weakly informative priors are based on realistic assumptions and avoid identifiability issues between key parameters (Royle & Link 2006; Miller et al. 2011), while allowing detectability rates to vary among years. This is considered essential in dynamic models to avoid bias (Miller et al. 2015). Our flexible approach produces accurate estimates of false-positive and occupancy rates, as well as detection probabilities.

To summarise, when setting up monitoring programs, we recommend survey designs that allow for an (ideally random) subset of surveys to be carried out with an unambiguous methodology; be that an accurate second survey method, or by sending surveyors known to be accurate and reliable. Investing extra survey effort in the first year of long-term programs will help to reduce identifiability issues and improve precision of parameter estimates, as will any feasible modifications to the survey protocols to improve detection probabilities. The final recommendation is that volunteers should be given identification training, and strong guidance as to how to report observations about which they are not entirely certain.

We have presented a general model to objectively evaluate the quality of monitoring data without the need for comparisons with expert data which may itself be flawed (Culverhouse et al. 2003; Austen et al. 2016). By applying it to a long-term volunteer dataset, we demonstrate that reliable measures of occupancy and trend can be derived even in the presence of false-positive errors. Occupancy rates were sometimes overestimated if false-positive error rates were unaccounted for, but population trends did not differ when false-positive errors were ignored. We recommend several simple modifications to sampling protocols which will allow the easy evaluation of occupancy rates and trends in future occupancy monitoring projects, even in the presence of false-positive errors. We therefore think that scientists and managers can use the model to fully access the information contained in data collected by citizen scientists.

ACKNOWLEDGEMENTS

This study was funded through a PhD scholarship from the Swiss Federal Office of the Environment (Forschungsvertrag 13.007PJ/M214-0617). We thank the Abteilung Landschaft und Gewässer of the canton of Aargau for providing permission to use the data from their amphibian monitoring program. We thank Arpat Ozgul for useful comments on the manuscript, and we are grateful to the many volunteers who did the field work for so many years and collected the data analysed in this study.

Literature Cited

Angelone, S. & Holderegger, R. (2009). Population genetics suggests effectiveness of habitat connectivity measures for the European tree frog in Switzerland. *Journal of Applied Ecology*, **46**, 879–887.

Angelone, S., Kienast, F. & Holderegger, R. (2011). Where movement happens: scale-dependent landscape effects on genetic differentiation in the European tree frog. *Ecography*, **34**, 714–722.

Austen, G.E., Bindemann, M., Griffiths, R.A. & Roberts, D.L. (2016). Species identification by experts and non-experts: comparing images from field guides. *Scientific Reports*, **6**, 33634.

Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Stuart-Smith, R.D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J.F., Pecl, G.T., Barrett, N. & Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, **173**, 144–154.

Burgman, M.A., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., Fidler, F., Rumpff, L. & Twardy, C. (2011). Expert status and performance. *PLoS ONE*, **6**, 1–7.

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

Crall, A.W., Newman, G.J., Stohlgren, T.J., Holfelder, K.A., Graham, J. & Waller, D.M. (2011). Assessing citizen science data quality: an invasive species case study. *Conservation Letters*, **4**, 433–442.

Cruickshank, S.S., Ozgul, A., Zumbach, S. & Schmidt, B.R. (2016). Quantifying population declines based on presence-only records for red-list assessments. *Conservation Biology*, **30**, 1112–1121.

Culverhouse, P.F., Williams, R., Reguera, B., Herry, V. & Gonzalez Gil, S. (2003). Do experts make mistakes? A comparison of human and machine labelling of dinoflagellates. *Marine Ecology Progress Series*, **247**, 17–25.

Dennis, E.B., Morgan, B.J.T., Brereton, T.M., Roy, D.B. & Fox, R. (2017). Using citizen science butterfly counts to predict species population trends. *Conservation Biology*, (In press) doi: 10.1111/cobi.12956.

Dickinson, J., 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.

Farmer, R.G., Leonard, M.L. & Horn, A.G. (2012). Observer effects and avian-call-count survey quality: Rare-species biases and overconfidence. *The Auk*, **129**, 76–86.

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.

Flory, C. (1999). Amphibien im Kanton Aargau. *Mitt. Aarg. Naturf. Ges*, **35**, 5–39.

Forrester, G., Baily, P., Conetta, D., Forrester, L., Kintzing, E. & Jarecki, L. (2015). Comparing monitoring data collected by volunteers and professionals shows that citizen scientists can detect long-term change on coral reefs. *Journal for Nature Conservation*, **24**, 1–9.

Foster-Smith, J. & Evans, S.M. (2003). The value of marine ecological data collected by volunteers. *Biological Conservation*, **113**, 199–213.

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

Kellner, K.F. & Swihart, R.K. (2014). Accounting for Imperfect Detection in Ecology: A Quantitative Review. *PLoS ONE*, **9**, e111436.

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

Kéry, M. & Schaub, M. (2011). *Bayesian population analysis using WinBUGS*. Academic Press.

Kéry, M. & Schmid, H. (2004). Monitoring programs need to take into account imperfect species detectability. *Basic and Applied Ecology*, **5**, 65–73.

Kéry, M. & Schmidt, B.R. (2008). Imperfect detection and its consequences for monitoring for conservation. *Community Ecology*, **9**, 207–216.

Lewandowski, E. & Specht, H. (2015). Influence of volunteer and project characteristics on data quality of biological surveys. *Conservation Biology*, **29**, 713–723.

MacKenzie, D.I., Nichols, J.D. & Hines, J. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, **84**, 2200–2207.

McDonald, T.L. (2003). Review of environmental monitoring methods: survey designs. *Environmental Monitoring and Assessment*, **85**, 277–292.

Meier, C. & Schelbert, B. (1999). Amphibienschutzkonzept Kanton Aargau. *Mitteilungen der aargauischen Naturforschenden Gesellschaft*, **35**, 41–70.

Miller, D.A.W., Bailey, L.L., Campbell Grant, E.H., McClintock, B.T., Weir, L.A. & Simons, T.R. (2015). Performance of species occurrence estimators when basic assumptions are not met: a test using field data where true occupancy status is known. *Methods in Ecology and Evolution*, **6**, 557–565.

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.

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

Nichols, J.D. (1992). Capture- Recapture Models: Using marked animals to study population dynamics. *Bioscience*, **42**, 94–102.

Plummer, M. (2003). JAGS: A program for analysis of bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, **124**.

Pocock, M.J.O., Roy, H.E., Preston, C.D. & Roy, D.B. (2015). The Biological Records Centre: a pioneer of citizen science. *Biological Journal of the Linnean Society*, **115**, 475–493.

Pollock, K.H., Nichols, J.D., Simons, T.R., Farnsworth, G.L., Bailey, L.L. & Sauer, J.R. (2002). Large scale wildlife monitoring studies: statistical methods for design and analysis. *Environmetrics*, **13**, 105–119.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ratnieks, F.L.W., Schrell, F., Sheppard, R.C., Brown, E., Bristow, O.E. & Garbuzov, M. (2016). Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers. *Methods in Ecology and Evolution*, **7**, 1226–1235.

Roth, T., Bühler, C. & Amrhein, V. (2016). Estimating Effects of Species Interactions on Populations of Endangered Species. *The American Naturalist*, **187**, 457–467.

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

Royle, J.A. & Kéry, M. (2007). A Bayesian state-space formulation of dynamic occupancy models. *Ecology*, **88**, 1813–23.

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

Schmeller, D.S., Henry, P.-Y., Julliard, R., Gruber, B., Clobert, J., Dziock, F., Lengyel, S., Nowicki, P., Déri, E., Budrys, E., Kull, T., Tali, K., Bauch, B., Settele, J., Van Swaay, C., Kobler, A., Babij, V., Papastergiadou, E. & Henle, K. (2009). Advantages of volunteer-based biodiversity monitoring in Europe. *Conservation Biology*, **23**, 307–16.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in ecology & evolution*, **24**, 467–71.

Stem, C., Margoluis, R., Salafsky, N. & Brown, M. (2005). Monitoring and evaluation in conservation: a review of trends and appraoches. *Conservation Biology*, **19**, 295–309.

van Strien, A.J., Pannekoek, J. & Gibbons, D.W. (2001). Indexing European bird population trends using results of national monitoring schemes: a trial of a new method. *Bird Study*, **48**, 200–213.

Su, Y.-S. & Yajima, M. (2015). R2jags: Using R to Run 'JAGS'. R package version 0.5-7. https://CRAN.R-project.org/package=R2jags

Sutherland, C., Elston, D.A. & Lambin, X. (2013). Accounting for false positive detection error induced by transient individuals. *Wildlife Research*, **40**, 490.

Tanadini, L.G. & Schmidt, B.R. (2011). Population Size Influences Amphibian Detection Probability: Implications for Biodiversity Monitoring Programs. *PLoS ONE*, **6**, e28244.

Tansley, A. (1904). The problems of Ecology. *New Phytologist*, **3**, 191–200.

SUPPLEMENTARY INFORMATION

SI 4.1: Description of Simulation study

For each scenario, we applied the model to simulated datasets spanning all combinations of the following parameter values:

Table SI 4.1.1 Parameter values used for the simulation study

Parameter	Value 1	Value 2	Value 3
Initial occupancy *psi1*	0.2	0.5	0.8
Colonisation rate *col*	0.02	0.075	0.2
Persistence rate *per*	0.5	0.75	0.9
True detectability *pT*	0.35	0.55	0.85
False-positive rate *pF*	0	0.05	0.1
Proportion of data considered 'confirmed' *prop.confirmed*	0.05	0.3	0.6

Therefore, for each scenario, we applied the model to 729 different datasets, spanning a range of plausible parameter values. Some parameter combinations were mathematically impossible to simulate (e.g. a combination of low true detectability and high false-positive results in a very low proportion of observations that are true detections; thus it is impossible to simulate a dataset with 60% of observations being fed to the model as 'true' detections).

In each simulation, we created datasets representing 15 years of data at 100 sites, with 3 visits per year. In order to assess the influence of missing data, we reduced datasets to the sparsity comparable to that of our real world datasets. Thus we simulated 15 years of data, but then converted all but 5 of these years to missing data- so that on average, sites were surveyed every 3 years.

In simulations where colonisation and extinction rates were allowed to vary, the mean rate was specified as given in the simulation run table above. Annual colonisation rates were randomly allocated either this value, or a rate 0.02 lower or 0.02 higher than the mean. Annual persistence rates were likewise allocated either the value specified in the simulation run, a value 0.05 lower, or 0.05 higher. For each simulation, we ran 3 chains of 40,000 iterations with a burn-in of 5,000, and thinned the remaining samples 1 in 10.

SI. 4.2: Number of sites existing in each core area

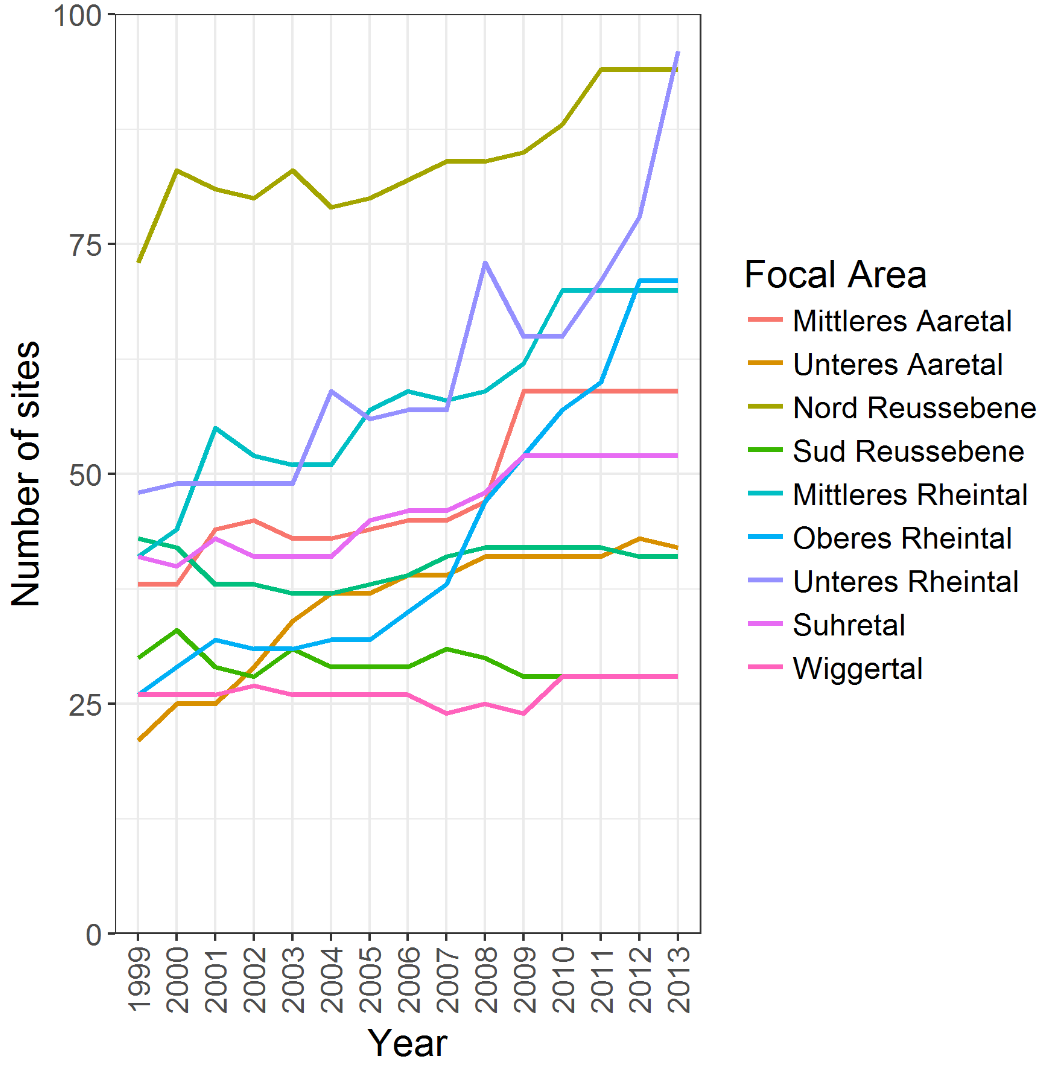


Figure SI 4.2 *The number of sites existing in each core area*

Table SI 4.3: Years in which each core area was surveyed

Core Area	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Aaretal Unteres	✓	✓				✓				✓				✓	
Aaretal Mittleres			✓	✓			✓				✓				
Reussebene Nord				✓	✓				✓				✓		
Reussebene Sud				✓	✓				✓				✓		
Reusstal Unteres	✓			✓				✓					✓		
Rheintal Mittleres			✓	✓			✓					✓			
Rheintal Oberes		✓	✓			✓				✓				✓	
Rheintal Unteres	✓	✓				✓				✓					✓
Suhretal	✓		✓				✓				✓				
Wiggertal				✓	✓			✓				✓			

Table SI 4.4: Species specific population size classes and species codes

Population counts were classified into 4 population size classes based on the number of individuals seen the observer. The table below gives for each species the number of individuals necessary to warrant inclusion in these size classes. Two thresholds were used to create 'confirmed' datasets- population class 4 only (high-threshold false-positive model), or population class 2 and above (low-threshold false-positive model)

Species name	Species Code	Population Class 1	Population Class 2	Population Class 3	Population Class 4
Triturus cristatus, *Lissotriton vulgaris,* *Lissotriton helveticus,* *Ichthyosaura alpestris*	TRCR LIVU LIHE ICAL	1-3	4-10	11-40	> 40
Bombina variegata, *Epidalea calamita,* *Pelophylax ridibundus* *Pelophylax esculentus* complex,	BOVA EPCA PERI PEES	1-5	6-30	31-100	> 100
Rana temporaria	RATE	1-5 adults, 1-40 spawn clumps	6-50 adults, 40-100 spawn clumps	51-200 adults, 100-400 span clumps	> 200 adults, > 400 spawn clumps
Bufo bufo	BUBU	1-5 adults, 1-20 spawn chains	6-50 adults, 20-100 spawn chains	51-200 adults, 100-400 spawn chains	> 200 adults, > 400 spawn chains
Alytes obstetricans, *Hyla arborea*	ALOB HYAR	1-5	6-20	21-60	> 60

Supplementary Information 4.5: JAGS code for the full false-positive dynamic occupancy model

```
model {
#### process priors  ###
mean.survival ~ dbeta(2,1)                                    #mean dynamic rates
mean.colonisation ~ dbeta(1,2)
surv.transformed <- log(mean.survival/(1-mean.survival))      #transform to logit scale
col.transformed <- log(mean.colonisation/(1-mean.colonisation))
surv.error ~ dgamma(1,1)                                      #ensure error can have wide tails.
col.error ~ dgamma(1,1)
surv.precision <- pow(surv.error,-2)                          # convert sd to precision
col.precision <- pow(col.error,-2)

for(a in 1:narea){
survival[a]     ~  dnorm(surv.transformed,surv.precision)     #area specific survival rate
colonisation[a]     ~      dnorm(col.transformed,col.precision)   #area specific colonisation rate
col[a] <- 1/ (1+exp(-colonisation[a]))                        #rates on probability scale
surv[a] <-  1/ (1+exp(-survival[a]))
psi.error[a] ~ dnorm (0,precision)                            #correctly specifies the precision for
                                                              the core-area random effect of
                                                              initial occupancy
}

alpha1     ~    dunif(-20,20)                                 # probablility of a site being
                                                              occupied in first year given no
                                                              historic records
alpha2     ~        dunif(-20,20)                             # effect of historic records on
                                                              occupancy in year 1
tau        ~    dgamma(1,1)                                   # specifies sd of the core-area
                                                              random effect of initial occupancy
precision <- pow(tau,-2)

### Detection Priors ###

pT.mean ~ dunif(0,1)                                          # true detection probability
pF.mean~ dbeta(1,2                                            # false positive error rate (for either
                                                              population class 4 (high-threshold
                                                              model) or 2,3,4 (low-threshold
                                                              model)
pT.transformed <- log(pT.mean/(1-pT.mean))                    # transform to logit scale
pF.transformed <- log(pF.mean/(1-pF.mean))
pT.error ~ dgamma(1,1)                                        # ensure this can have wide tails.
pF.error ~ dgamma(1,1)
pT.precision <- pow(pT.error,-2)
pF.precision <- pow(pF.error,-2)
r11 ~ dunif(0,1)
                                                              ##probability of detecting
                                                              'confirmed' populations

for (y in 1:nyear){
ppT[y] ~ dnorm( pT.transformed , pT.precision)
ppF[y] ~ dnorm (pF.transformed , pF.precision)
pT[y] <- 1/ (1+exp(-ppT[y]))                                  #annual true detection rate
pF[y] <- 1/ (1+exp(-ppF[y]))                                  #annual false positive detection rate
}
```

```
#######MODEL#######

####  Process model ####

 for(i in 1:nsite){

logit(init.occ[i]) <-alpha1+(alpha2*historic[area[i]])+ psi.error[area[i]]
#common intercept and slope within a core area, but with error term to allow for between site
variation.

 z[i,start[area[i]]] ~ dbern(init.occ[i]*existing[i,start[area[i]]])
#occupancy state in the first year a site was surveyed

            for(t in start[area[i]]:(end[area[i]]-1)){

    occs[i,t+1]<- ((z[i,t]* (surv[area[i]])))  + ((1-z[i,t])*((col[area[i]])))
#area specific annual survival and extinction
  }
}

###  Observation Model   #####

for(i in 1:nsite){
     for (t in 1:nyear){
             p[i,t] <-detectable[i,t]* ((z[i,t]*pT[t])+((1-z[i,t])*pF[t]))
            #detection of any population (allowing false positives)
     r[i,t] <- detectable[i,t]*z[i,t]*r11
            #detection rate of confirmed populations (i.e size class 4 (high-threshold) or 2,3,4 (low-
            threshold)

        for(j in 1:nrep){
        y[i,j,t] ~ dbern (p[i,t])
        w[i,j,t] ~ dbern (r[i,t])
            }
        }
        }

  ###  derived parameters ####

  for (l in 1:narea){
for(t in start[l]:end[l]){
            n.occ[l,t] <- sum(z[carea[l]:(carea[l+1]-1),t])
            #number of occupied sites in each core area and year
              prop.occ[l,t] <- n.occ[l,t]/sum(existence[carea[l]:(carea[l+1]-1),t])
            #proportion of occupied sites in each core area and year
  }

 for(r in start[l]:(end[l]-1)){

 prop.trend[l,r]<-prop.occ[l,r+1]-prop.occ[l,r]
            # annual changes in the proportion of occupied sites

 n.trend[l,r]<-n.occ[l,r+1]-n.occ[l,r]
            # annual changes in the number of occupied sites
 }
 }
 }
```

Supplementary Information 4.6: Full description of methods used to compare occupancy estimates and annual trends.

Description of how trend and occupancy estimates were calculated

We performed comparisons between pairs of models, to identify the differences between models in terms of a) differences in estimated occupancy rates between years, and b) differences in estimated population trends (measured as the change in occupancy from one year to the next). We performed pairwise comparisons between 4 sources:

1. A dynamic occupancy model accounting for imperfect detection only. (PT-Model)
2. A dynamic occupancy model accounting for imperfect detection, and false-positive errors, using only the highest population class as 'confirmed' observations (FP-high; see main text for explanation)
3. A dynamic occupancy model accounting for imperfect detection, and false-positive errors, using all observations not of the lowest population class as 'confirmed' observations (FP-low).
4. Raw observations- whether or not any positive observations for a site were made in a given year.

Sources 1-3 produce a distribution of occupancy in each year- i.e. a mean and a measure of uncertainty surrounding this mean. We sought to incorporate this uncertainty in our comparisons. Source 4 has no associated uncertainty, as the assumption is made that the data are perfect (i.e. the species would have been observed at least once if present), the comparisons between this source and the other models was carried out slightly differently (see information below).

Occupancy

Our models provide for each year and core area, an occupancy measure with associated uncertainty (Fig SI 4.4.1). For sources 1-3, these estimates were available for each core area for every year between the year the area was first surveyed, and the year that the core area was last surveyed. For the raw data, each core area was surveyed on average 4.3 times (SI 4.2), leading to 43 core area/year combinations for which raw occupancy were available. We restricted occupancy analyses to these 43 core area/year combinations to allow reasonable comparisons between all data sources.

For each of these 43 combinations, we took the posterior distributions of occupancy estimates for the two models under comparison (Fig SI 4.6.2, which equates to the occupancy estimates for year 4 in Fig SI 4.6.1). By subtracting one distribution from the other, we obtain a distribution of the differences in occupancy estimates between models at this given point (Fig SI 4.6.3). We used two metrics to judge the differences between models. The mean of this distribution gives the average difference in occupancy estimates between the two models during this time point. By calculating the 95% credible interval of this distribution, we can judge whether or not this difference is significant at this time point. We took the mean difference between each of the 43 core area/year combinations as a measure of the difference between occupancy estimates between models, and also calculated the proportion of the 43 combinations for which these differences were significantly different from zero (SI 4.8).

For comparisons between sources 1-3 and source 4, the process was slightly different, as the raw data has no surrounding uncertainty. Thus the comparison for any given point appears as in Fig SI 4.6.4, with a point estimate (from the raw data) being subtracted from a distribution (from one of the occupancy models). This again produces a distribution of the differences, which is treated as described above.

Trends comparisons

Our method of comparing occupancy trends is similar to the process for comparing occupancy estimates. Here we define trend as the difference in estimated occupancy rate between year t and t+1. For sources 1-3 we simply calculated this parameter by subtracting the occupancy distribution for each model at time t from that of time t+1. For the raw data, we had to use a slightly different method. As occupancy rates were only available for years when an area was surveyed (e.g. years 1,4,6 and 10 in our example), we assumed that occupancy changed at a constant rate between these 'observed years', and used the gradient of this linear change between observed years as a measure of annual population trend (Fig SI 4.6.5). After this first step, we followed the procedure described above, to assess differences between the models in terms of population trend.

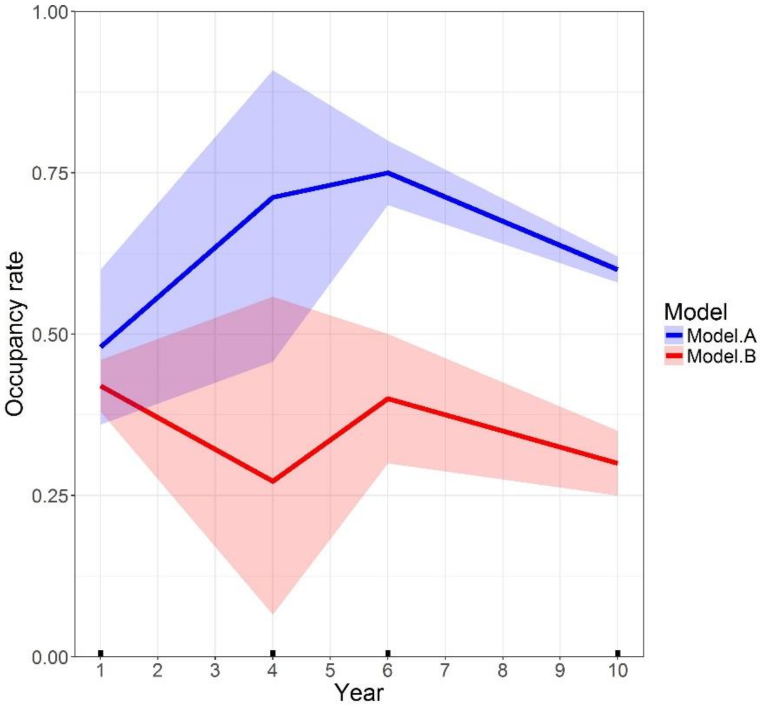


Figure SI 4.6.1. Two occupancy models under comparison. Occupancy estimates with associated uncertainty distributions were estimates for each year. Dash marks at the x-axis represent the years for which raw data were available for this area (here, years 1,4,6 and 10). Thus these are the years for which occupancy estimate comparisons were made.

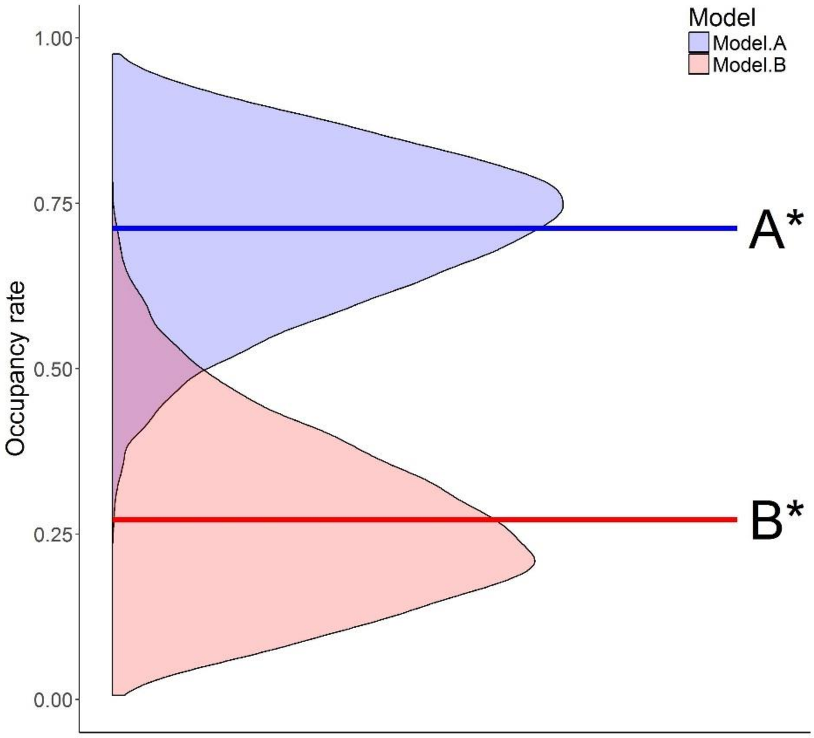


Figure SI 4.6.2- Occupancy estimates for Models A and B relating to year 4 in Figure 1. Each model has a mean value (A* and B*), and a distribution of estimates around this mean.

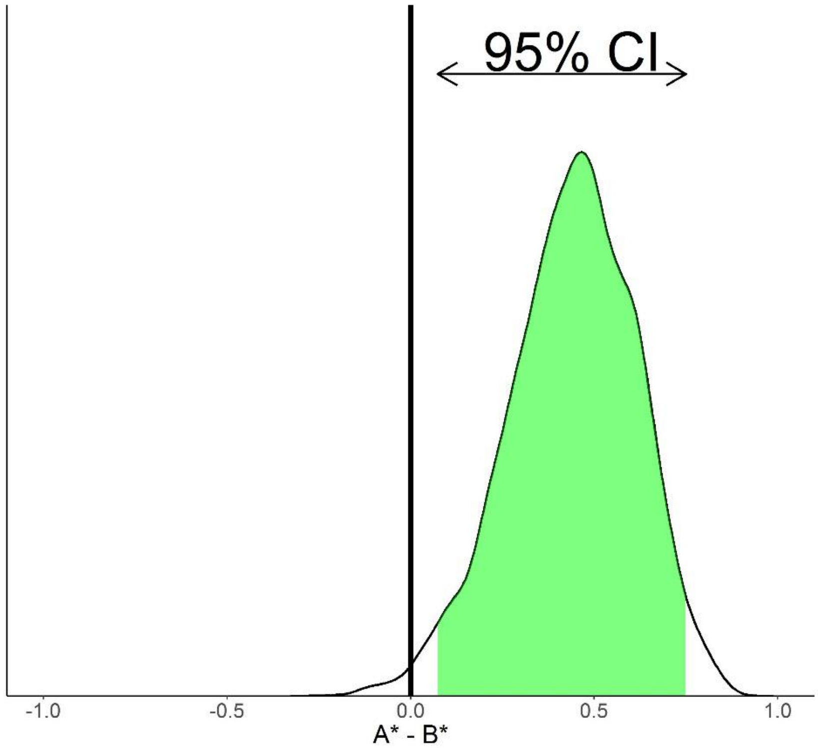


Figure SI 4.6.3-Distribution of the differences in occupancy estimates between Models A and B in year 4. Red shading represents the 95% credible interval of this differences. Here, the 95%CI does not encompass zero, thus the occupancy estimates for the two models are significantly different for year 4.

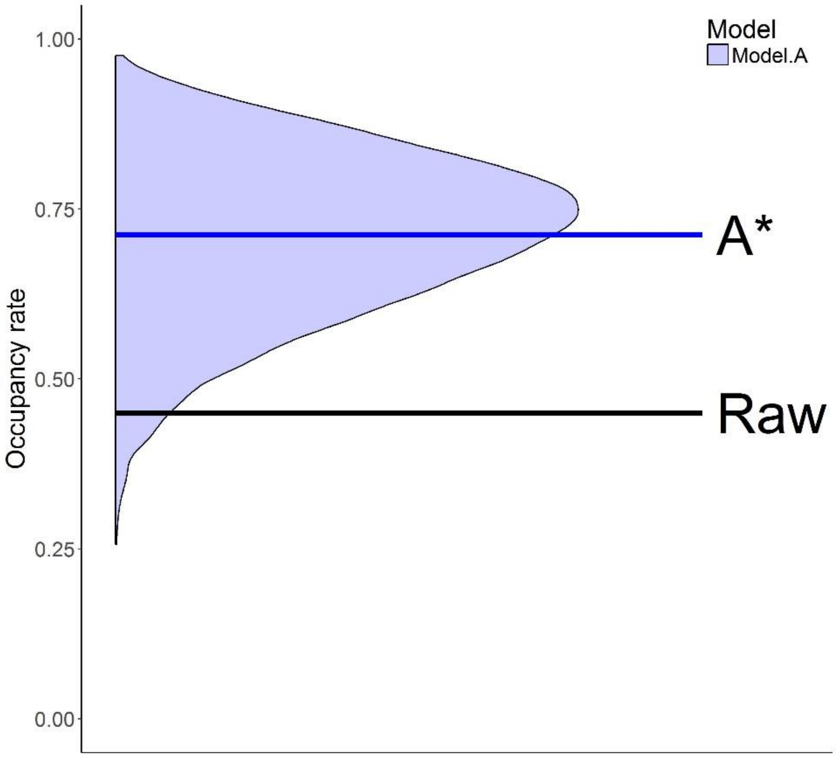


Figure SI 4.6.4- Comparison of occupancy estimates between Model A and the raw data. The raw data represents a point estimate with no uncertainty, so here the comparison is made by subtracting this point estimate from the distribution produced by Model A

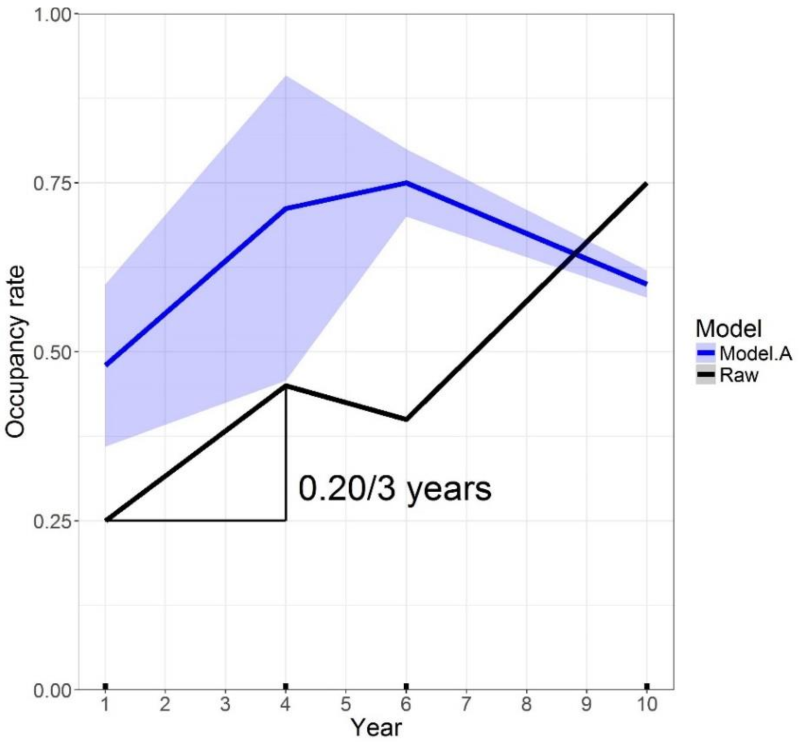


Figure SI 4.6.5- Comparison of population trend between Model A and the raw data. Annual trends for the raw data are calculated as the gradient between observed occupancy estimates. Here the first 3 trend estimates for the raw data are 0.067

Supplementary Information 4.7: App for visualising and exploring simulation results.

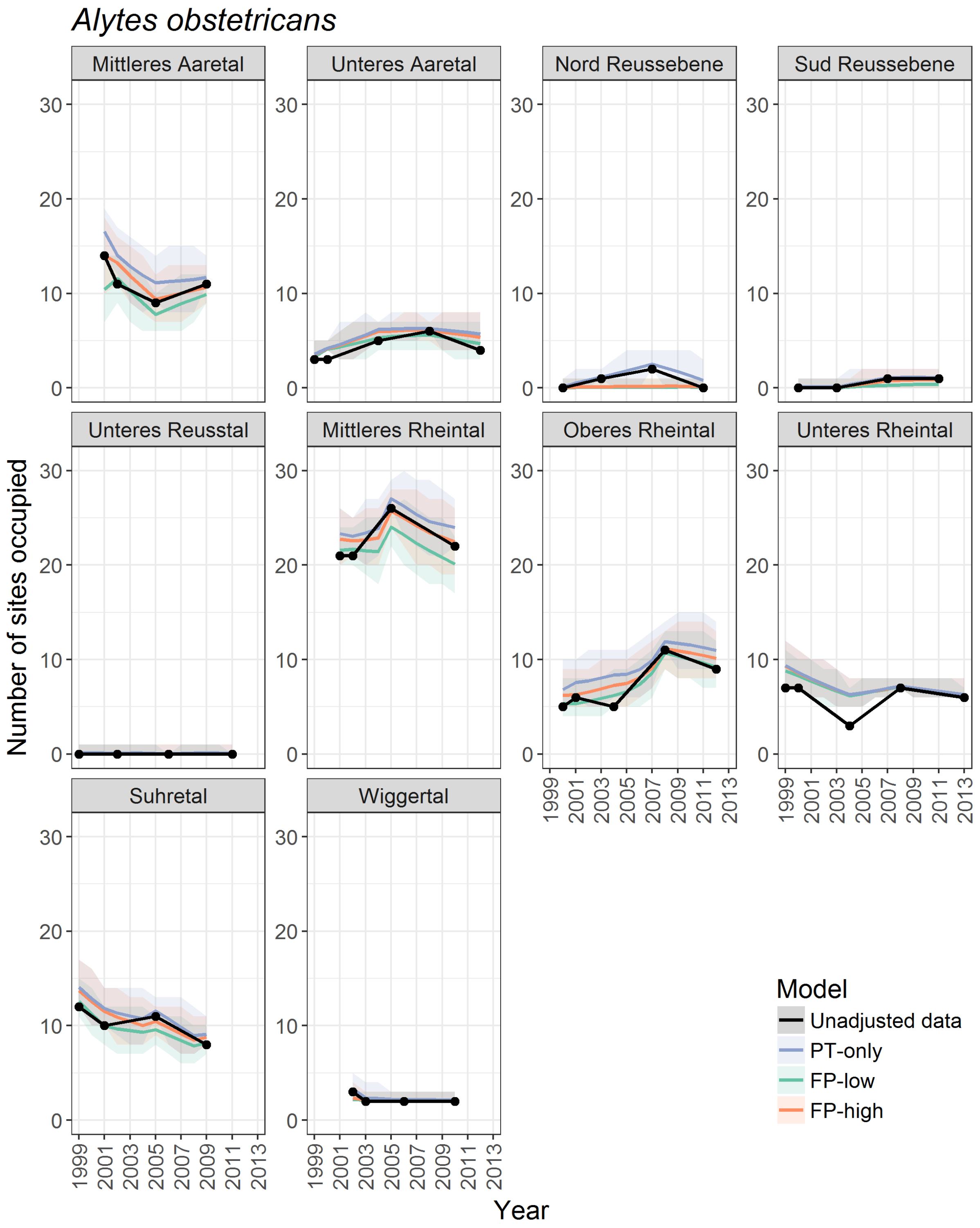
The Shiny app is currently available in a GitHub depository and can be accessed by running the following code in Rstudio (requires packages “shiny”, “ggplot2”, and “dplyr”)

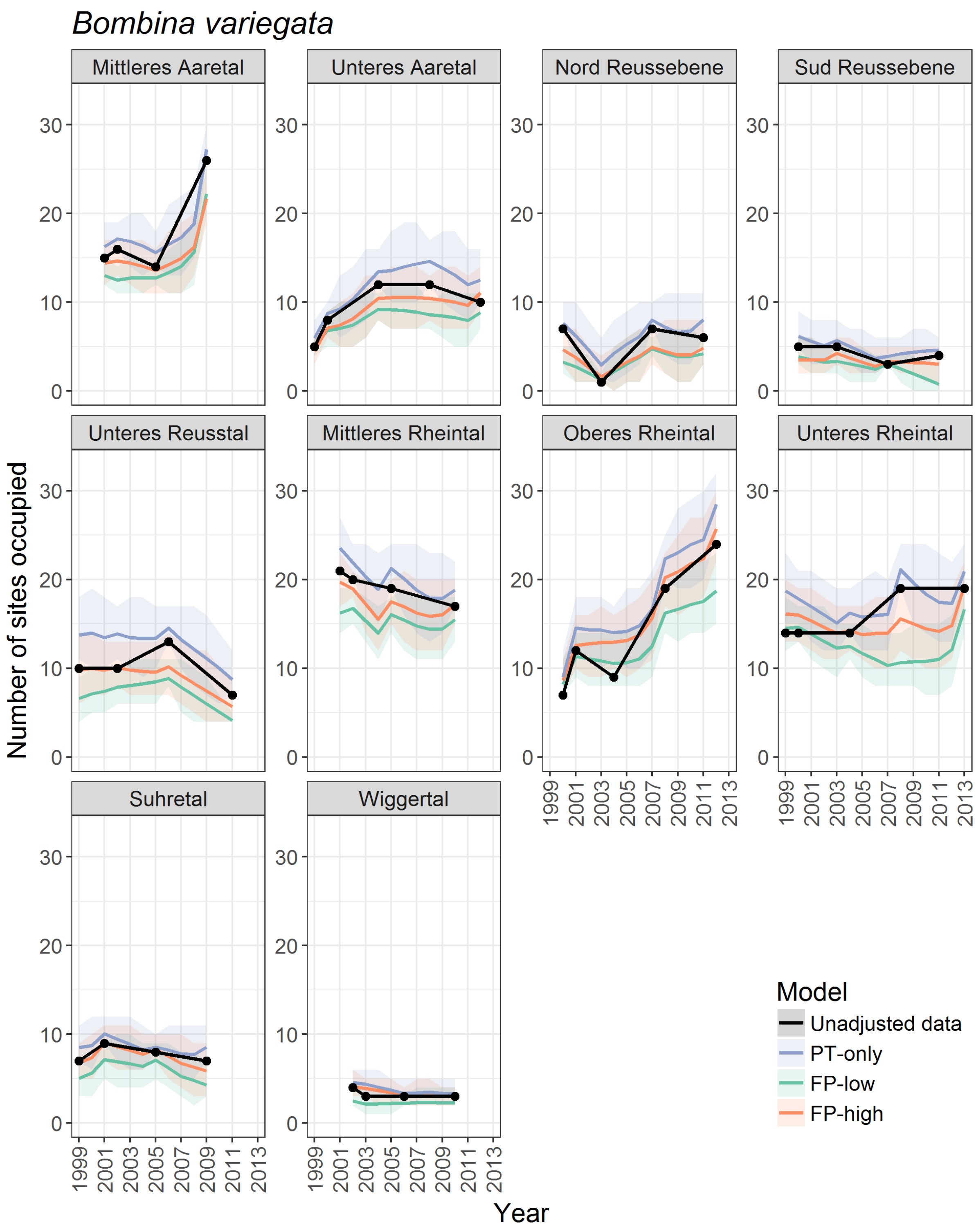
```
runGitHub( "Aargau-demonstation", "sscruickshank”)
```

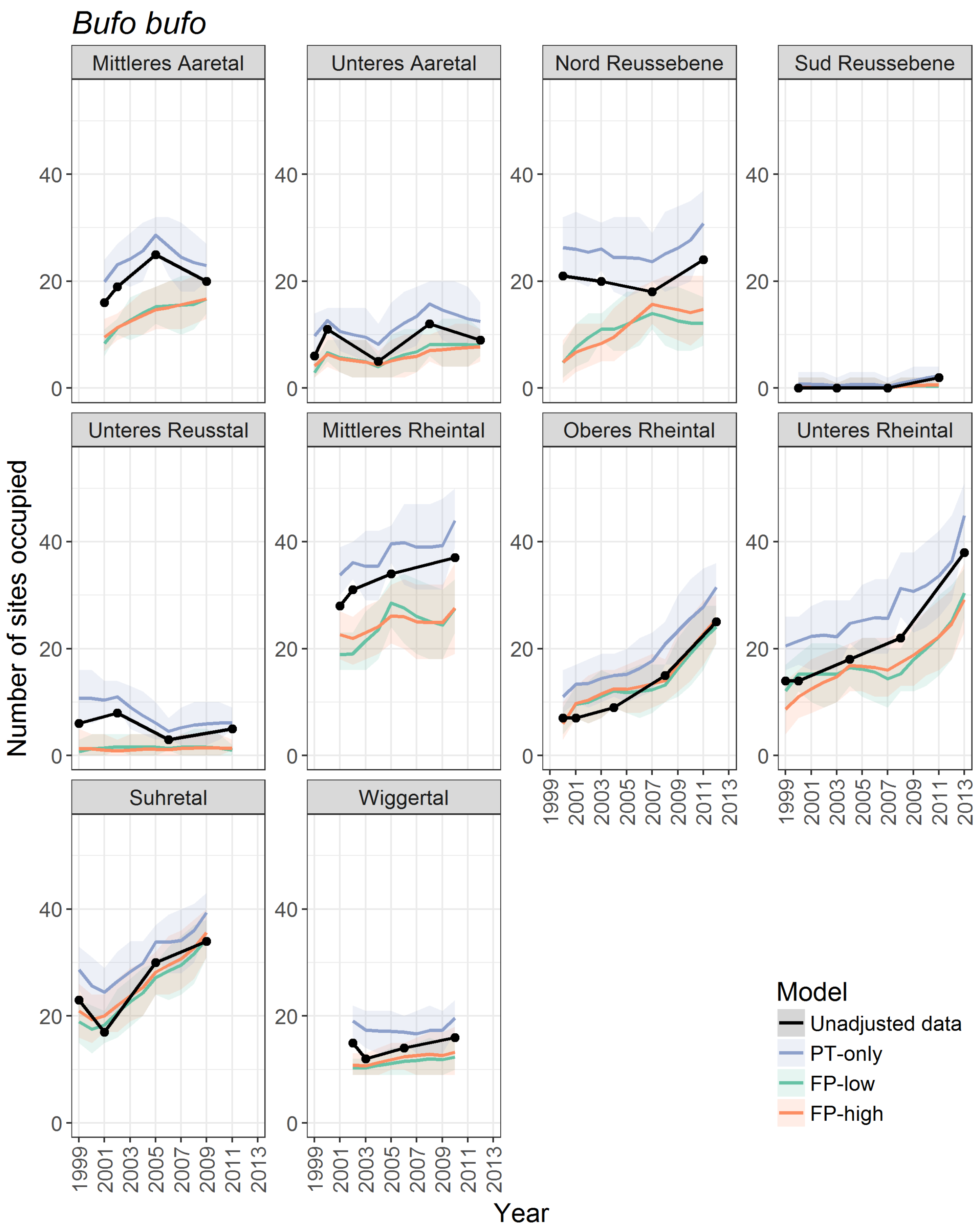
Table SI 4.8. Proportion of annual occupancy comparisons significantly different from 0, and overall mean and 95% credible intervals of differences in the number of occupied sites for pairwise model comparisons. Italicised row denotes column means. Bold value denote that 95% credible intervals do not overlap with 0. Species names are provided in SI 4.4

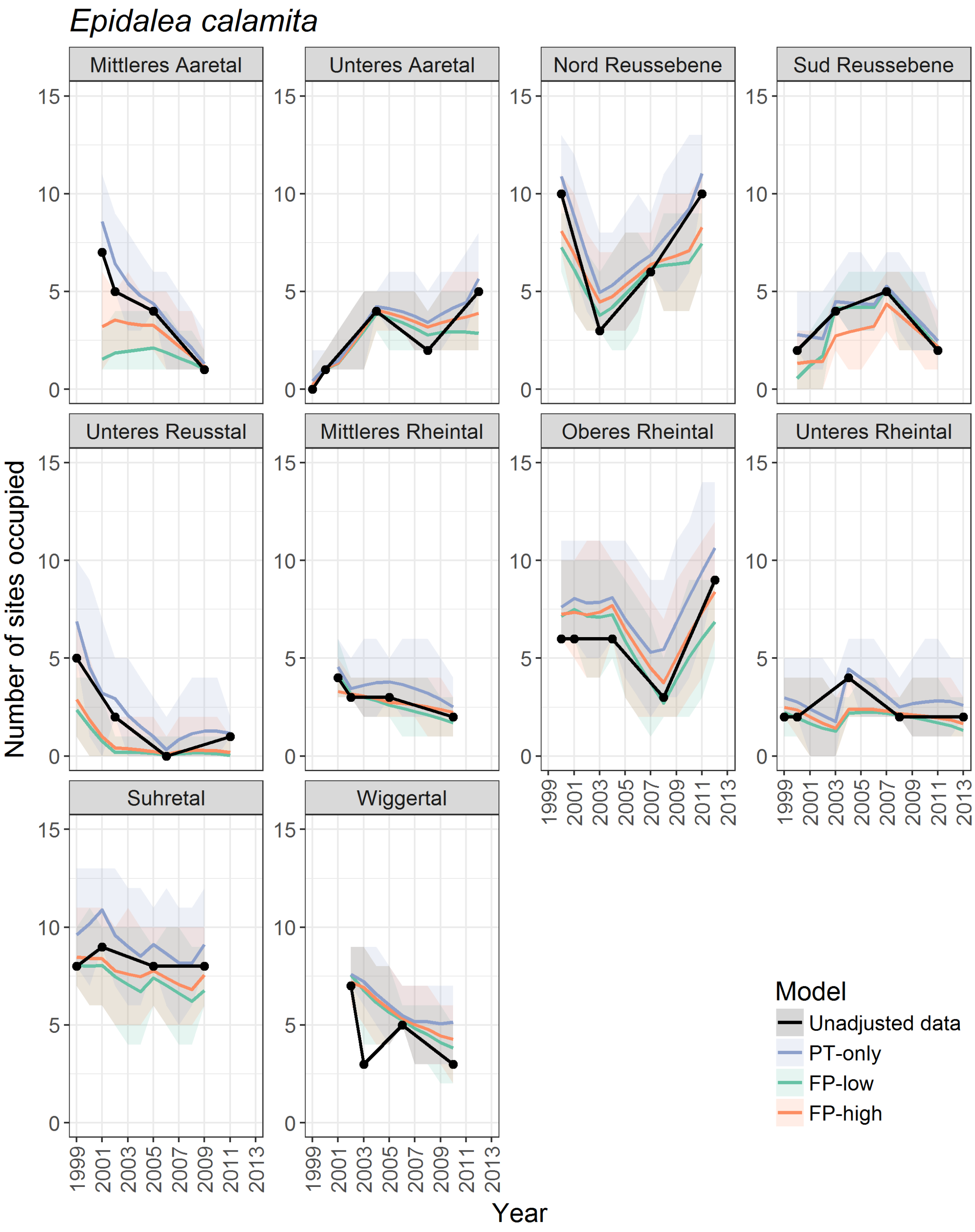
Species	Proportion significant						Mean difference in number of occupied sites trend (95% CRI)					
	PT_{only} - FP_{high}	PT_{only} - naïve	FP_{high} - naïve	PT_{only} - FP_{low}	FP_{high} - FP_{low}	FP_{low} - naïve	PT_{only} - FP_{high}	PT_{only} - naïve	FP_{high} - naïve	PT_{only} - FP_{low}	FP_{high} - FP_{low}	FP_{low} - naïve
ALOB	0.02	0.09	0.07	0.05	0.00	0.12	0.55 (-2 : 3)	1.09 (0 : 3)	0.54 (-2 : 3)	1.18 (-1 : 4)	0.63 (-2 : 3)	-0.09 (-3 : 2)
BOVA	0.05	0.16	0.09	0.49	0.02	0.44	2.25 (-1 : 6)	1.89 (0 : 5)	-0.36 (-4 : 2)	3.89 (0 : 8)	1.64 (-2 : 5)	-2 (-7 : 1)
BUBU	0.72	0.74	0.51	0.79	0.00	0.60	8.18 (0 : 17)	4.51 (0 : 9)	-3.67 (-15 : 2)	8.45 (0 : 18)	0.27 (-6 : 4)	-3.94 (-14 : 1)
EPCA	0.02	0.02	0.09	0.19	0.00	0.26	1.19 (-2 : 4)	1.05 (0 : 3)	-0.14 (-3 : 2)	1.54 (-1 : 5)	0.34 (-2 : 2)	-0.48 (-4 : 1)
HYAR	0.02	0.05	0.02	0.21	0.00	0.16	0.64 (-2 : 3)	1.03 (0 : 4)	0.38 (-2 : 3)	1.42 (0 : 6)	0.78 (-1 : 4)	-0.4 (-5 : 0)
ICAL	0.70	0.65	0.56	0.65	0.14	0.47	8.95 (0 : 19)	4.32 (0 : 9)	-4.64 (-20 : 2)	6.94 (0 : 14)	-2.02 (-11 : 3)	-2.62 (-12 : 3)
LIHE	0.05	0.37	0.00	0.56	0.05	0.28	2.83 (-2 : 8)	3.11 (0 : 7)	0.28 (-4 : 3)	4.64 (0 : 10)	1.81 (-2 : 6)	-1.52 (-7 : 1)
LIVU	0.02	0.02	0.05	0.00	0.02	0.00	0.51 (-3 : 3)	1.04 (0 : 4)	0.53 (-1 : 3)	1.06 (-1 : 4)	0.56 (-1 : 3)	-0.02 (-1 : 1)
PEES complex	0.33	0.42	0.30	0.58	0.02	0.58	3.66 (-1 : 10)	2.81 (0 : 7)	-0.85 (-9 : 4)	5.04 (0 : 12)	1.38 (-3 : 5)	-2.23 (-10 : 3)
PERI	0.00	0.02	0.00	0.07	0.00	0.09	0.79 (-2 : 3)	1.01 (0 : 3)	0.22 (-2 : 2)	1.22 (-1 : 4)	0.43 (-1 : 2)	-0.21 (-2 : 1)
RATE	0.81	0.77	0.49	0.70	0.09	0.49	**8.91** **(1 : 20)**	**4.93** **(1 : 9)**	-3.98 (-16 : 2)	**7.6** **(1 : 16)**	-1.32 (-9 : 3)	-2.67 (-11 : 2)
TRCR	0.00	0.00	0.00	0.05	0.00	0.05	0.36 (-2 : 2)	0.54 (0 : 2)	0.18 (-1 : 1)	0.77 (-1 : 3)	0.41 (-1 : 2)	-0.23 (-3 : 0)
	0.23	*0.28*	*0.18*	*0.36*	*0.03*	*0.29*	*0.09*	*0.31*	*0.22*	*0.11*	*0.02*	*0.20*

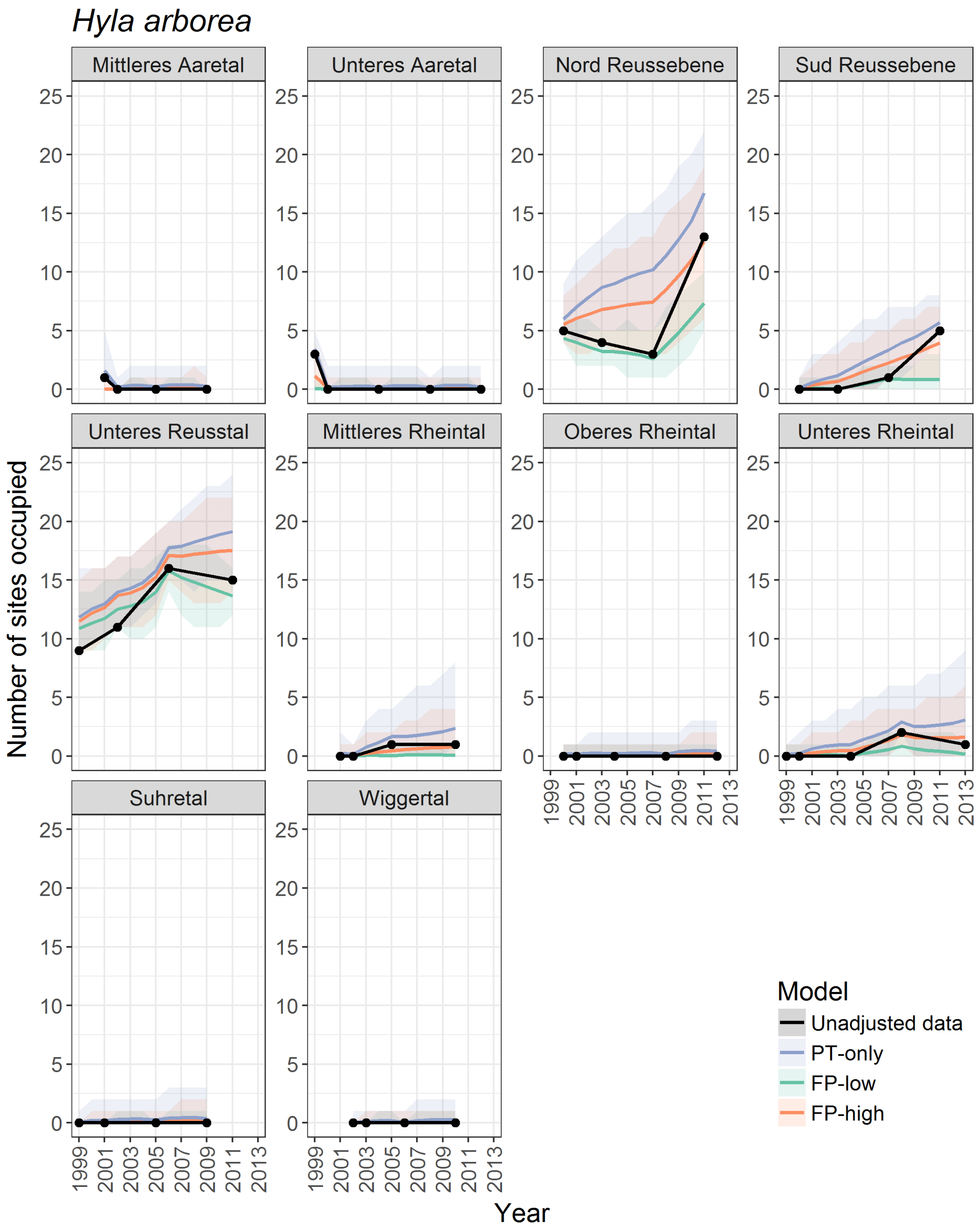
Supplementary Information 4.9: Graphical representation of population trends for the 3 occupancy models and the raw data for all 12 species in each core area

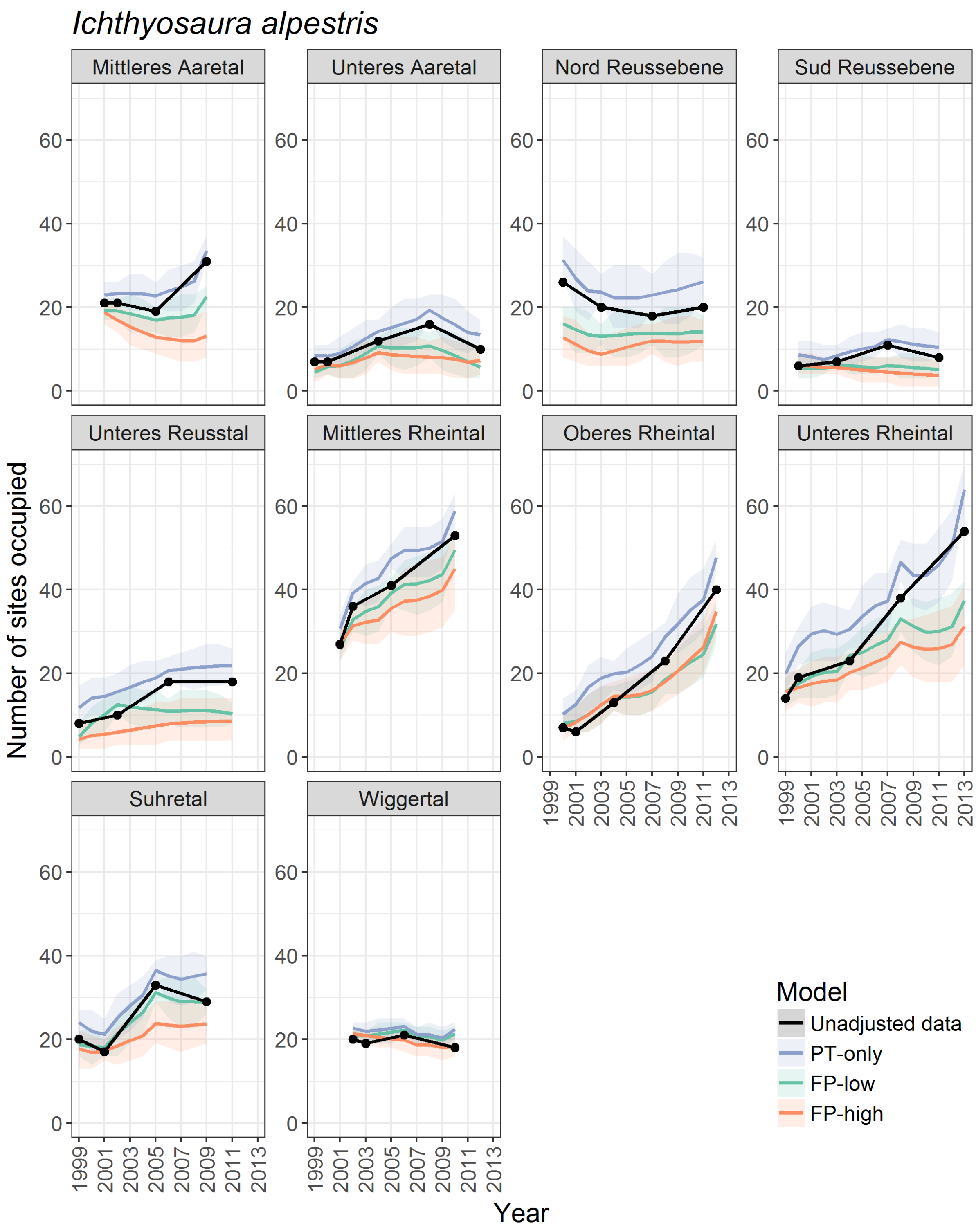


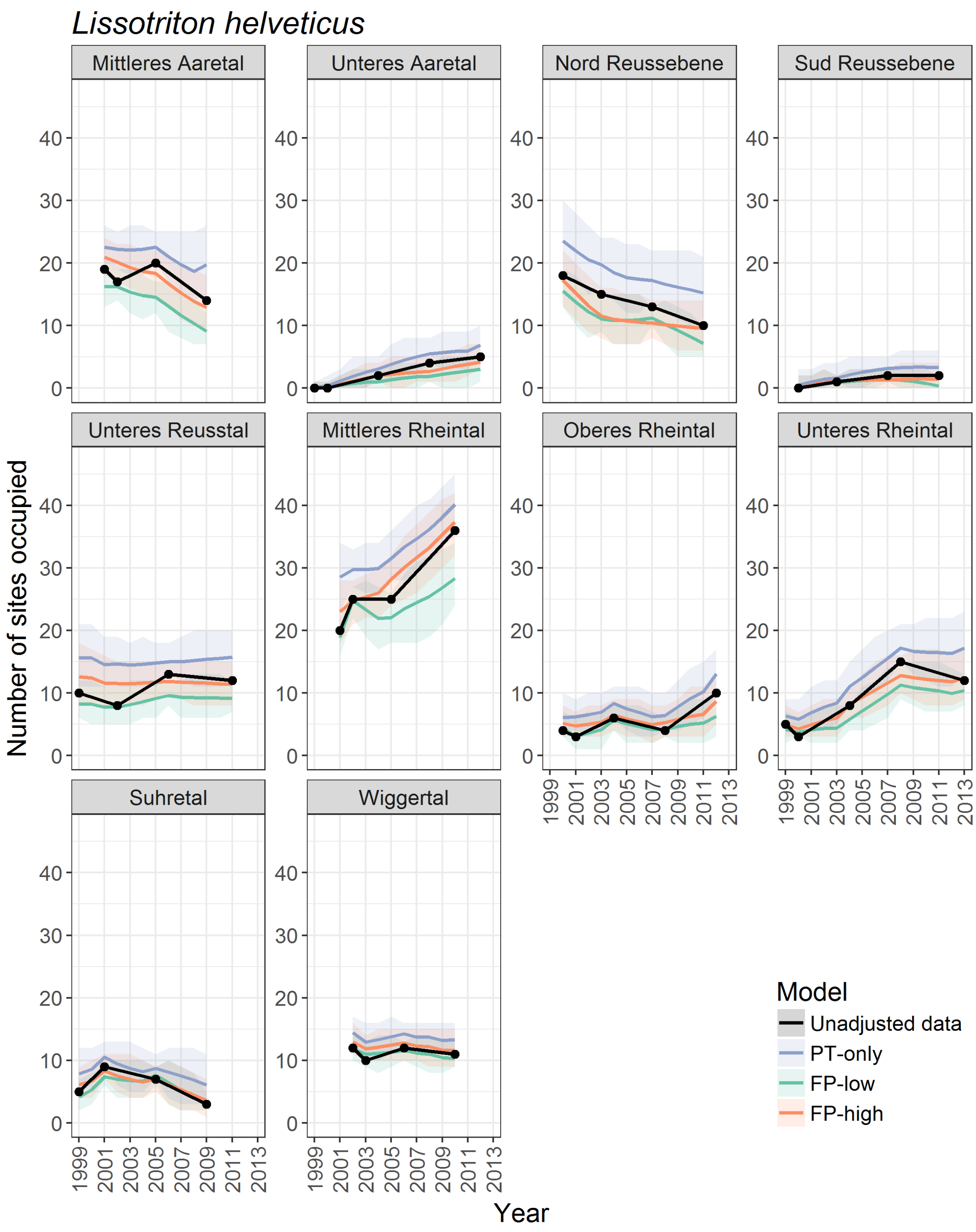

Bombina variegata
Mittleres Aaretal
Unteres Aaretal
Nord Reussebene
Sud Reussebene
Unteres Reusstal
Mittleres Rheintal
Oberes Rheintal
Unteres Rheintal
Suhretal
Wiggertal
Number of sites occupied
Year
0
10
20
30
1999
2001
2003
2005
2007
2009
2011
2013
Model
Unadjusted data
PT-only
FP-low
FP-high

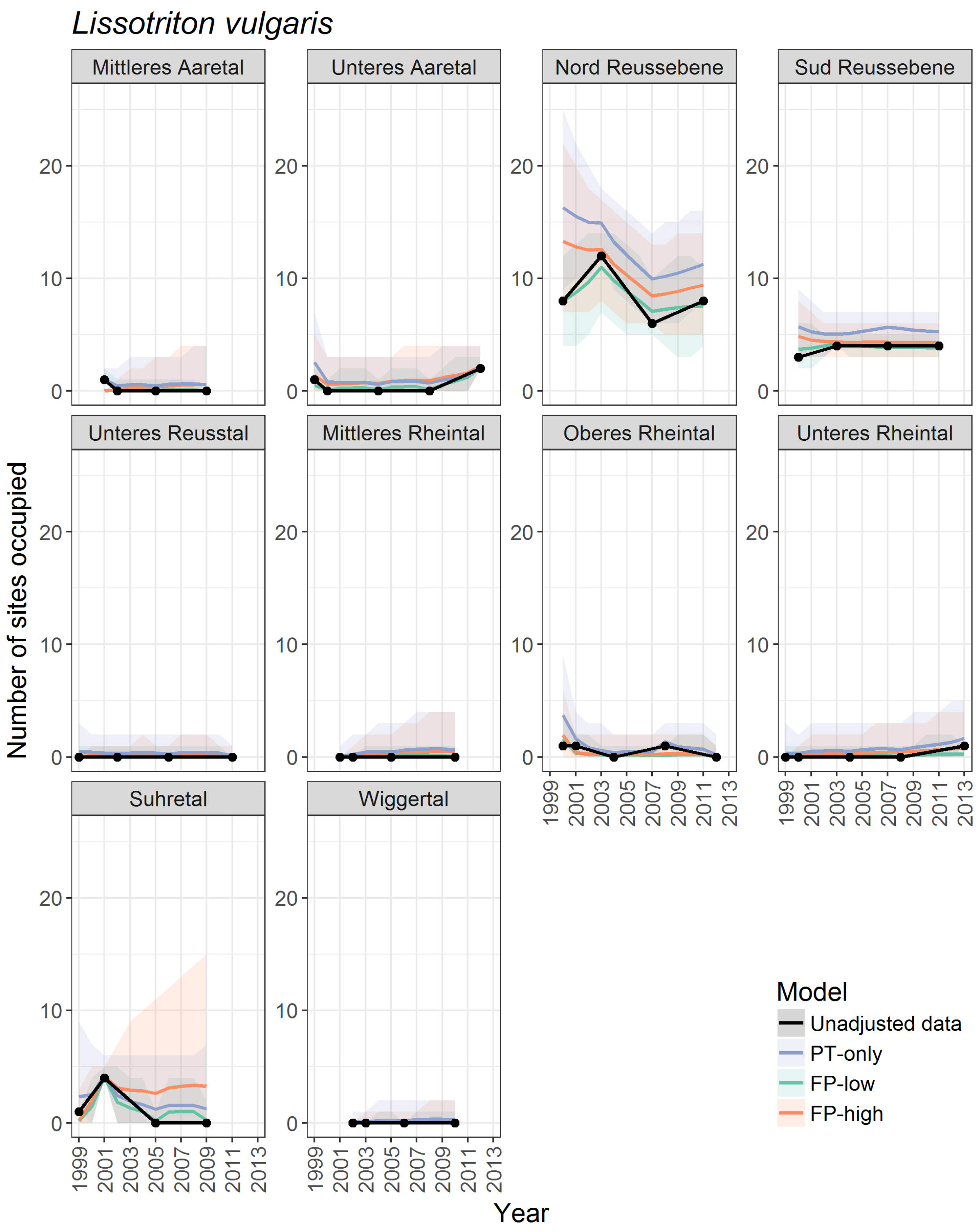

Bufo bufo
Mittleres Aaretal
Unteres Aaretal
Nord Reussebene
Sud Reussebene
Unteres Reusstal
Mittleres Rheintal
Oberes Rheintal
Unteres Rheintal
Suhretal
Wiggertal
Number of sites occupied
0
20
40
Year
1999
2001
2003
2005
2007
2009
2011
2013
Model
Unadjusted data
PT-only
FP-low
FP-high

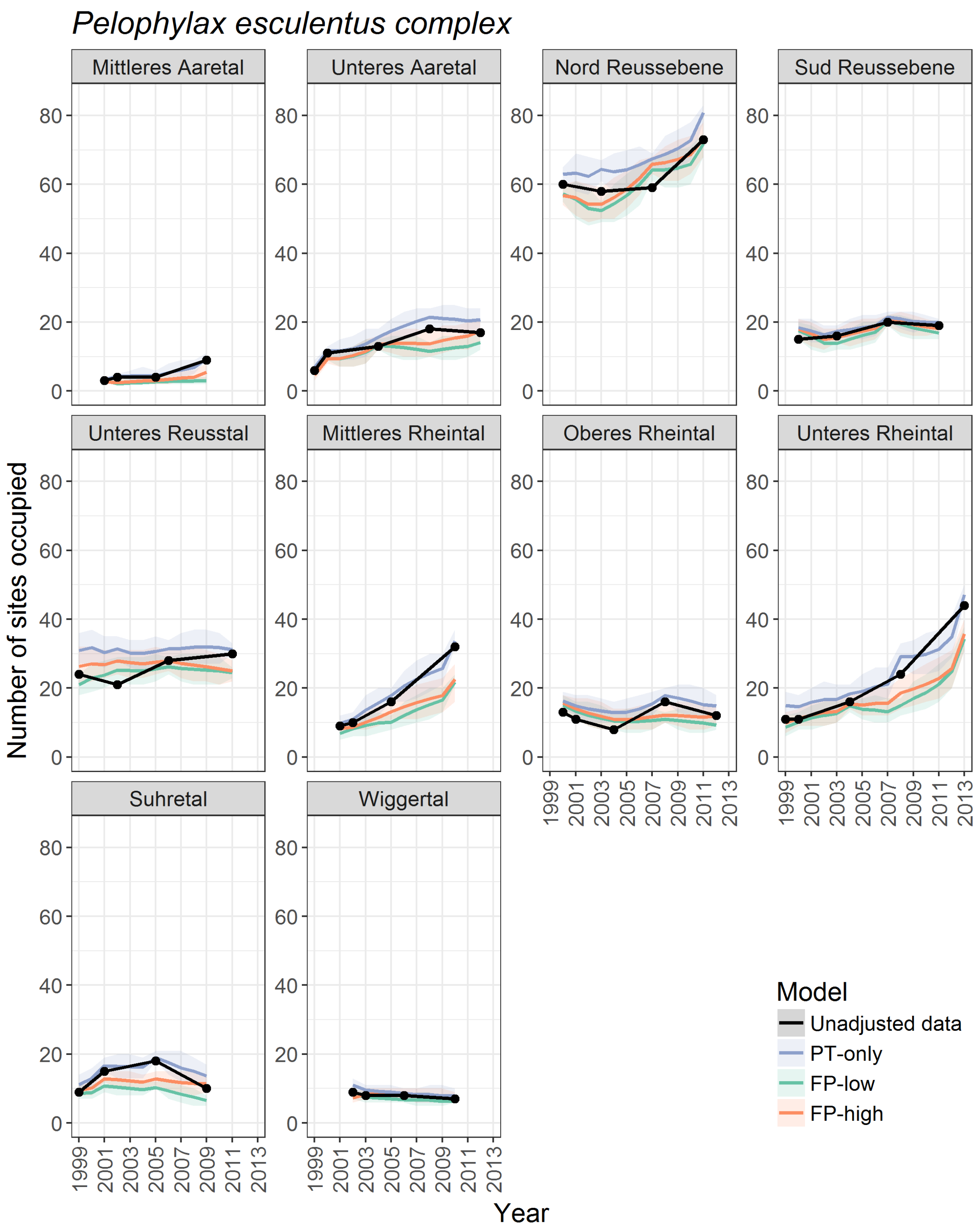

Epidalea calamita
Mittleres Aaretal
Unteres Aaretal
Nord Reussebene
Sud Reussebene
Unteres Reusstal
Mittleres Rheintal
Oberes Rheintal
Unteres Rheintal
Suhretal
Wiggertal
Number of sites occupied
Year
0
5
10
15
1999
2001
2003
2005
2007
2009
2011
2013
Model
Unadjusted data
PT-only
FP-low
FP-high

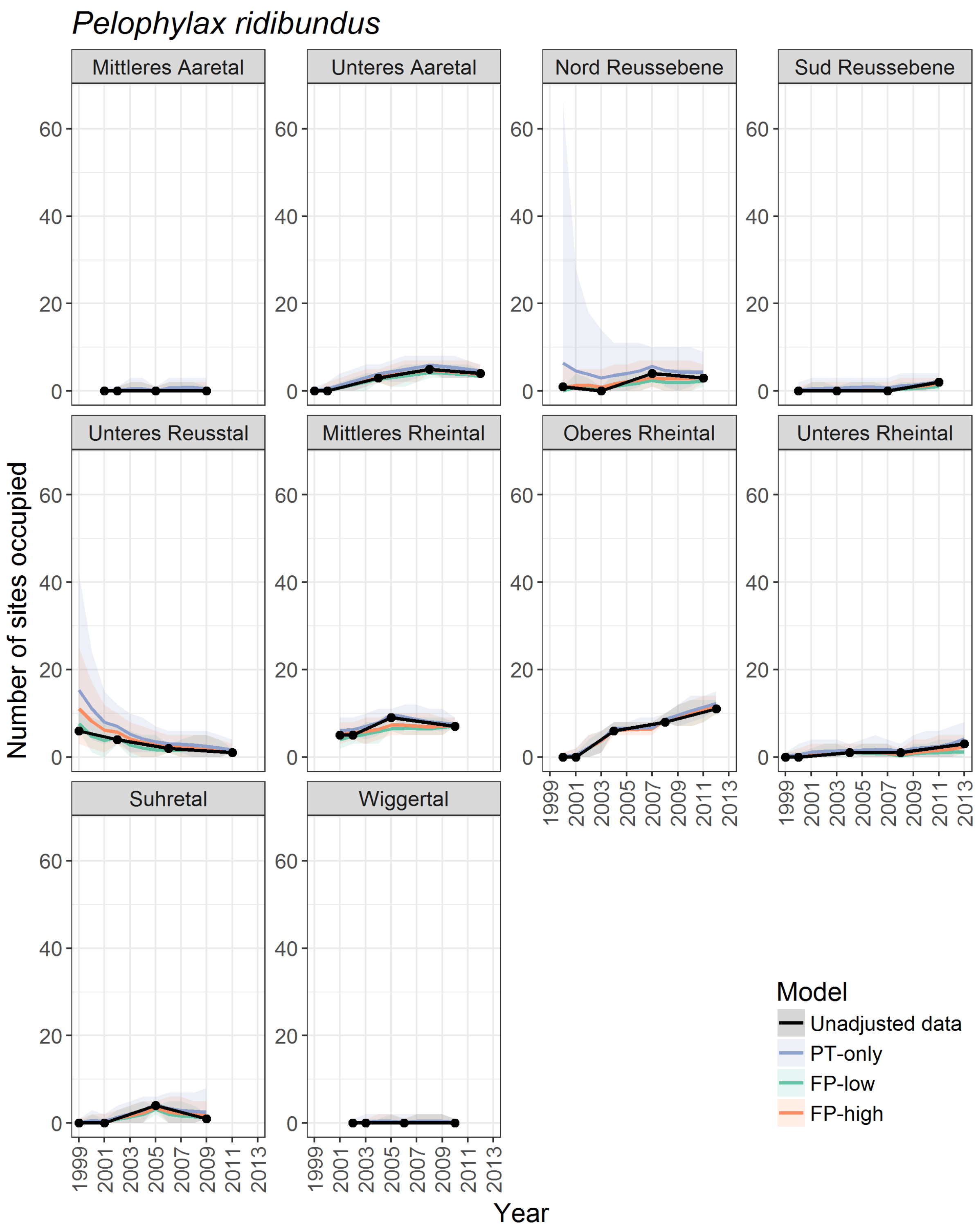

Hyla arborea
Mittleres Aaretal
Unteres Aaretal
Nord Reussebene
Sud Reussebene
Unteres Reusstal
Mittleres Rheintal
Oberes Rheintal
Unteres Rheintal
Suhretal
Wiggertal
Number of sites occupied
0
5
10
15
20
25
1999
2001
2003
2005
2007
2009
2011
2013
Year
Model
Unadjusted data
PT-only
FP-low
FP-high

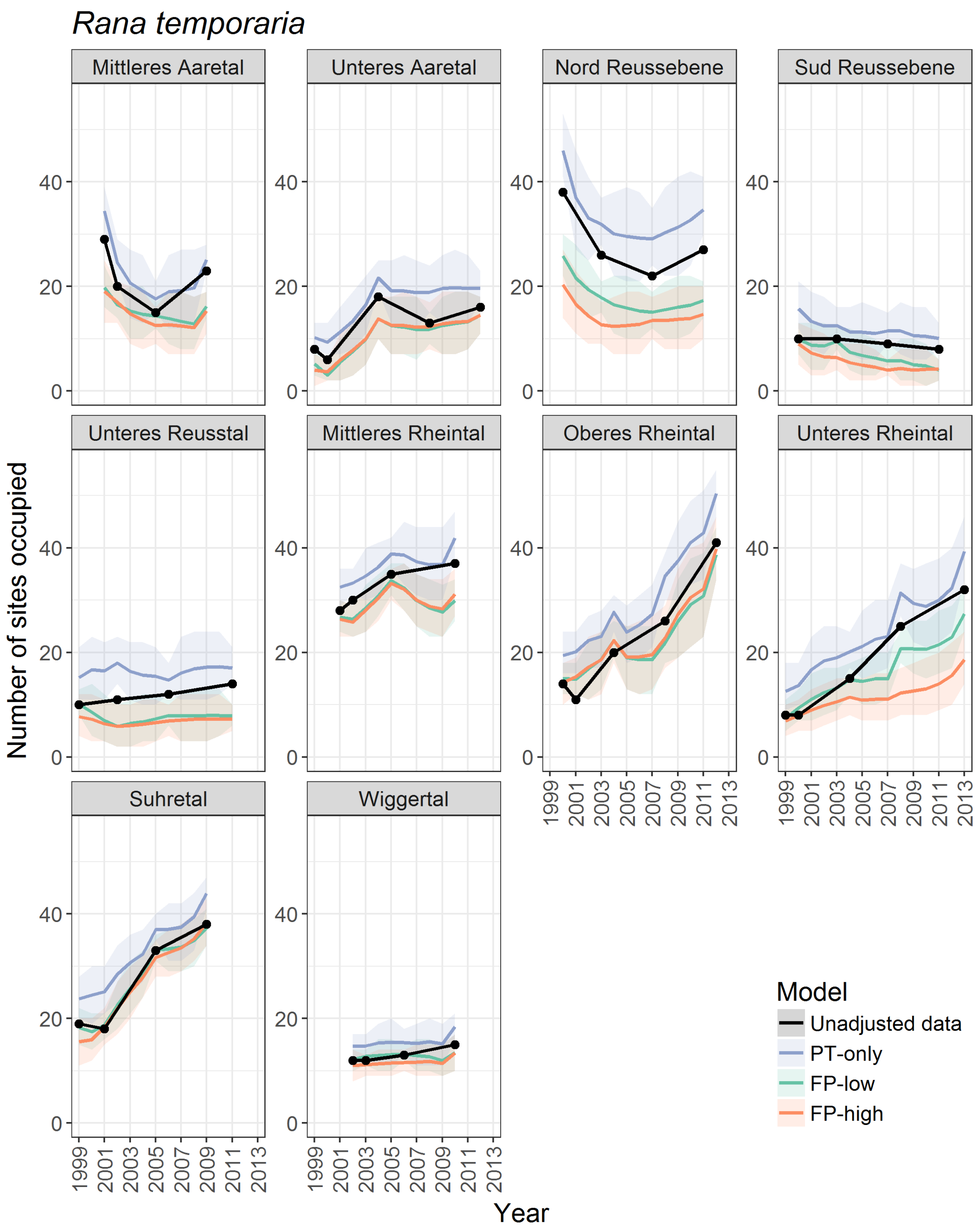

Ichthyosaura alpestris
Mittleres Aaretal
Unteres Aaretal
Nord Reussebene
Sud Reussebene
Unteres Reusstal
Mittleres Rheintal
Oberes Rheintal
Unteres Rheintal
Suhretal
Wiggertal
Number of sites occupied
0
20
40
60
Year
1999
2001
2003
2005
2007
2009
2011
2013
Model
Unadjusted data
PT-only
FP-low
FP-high

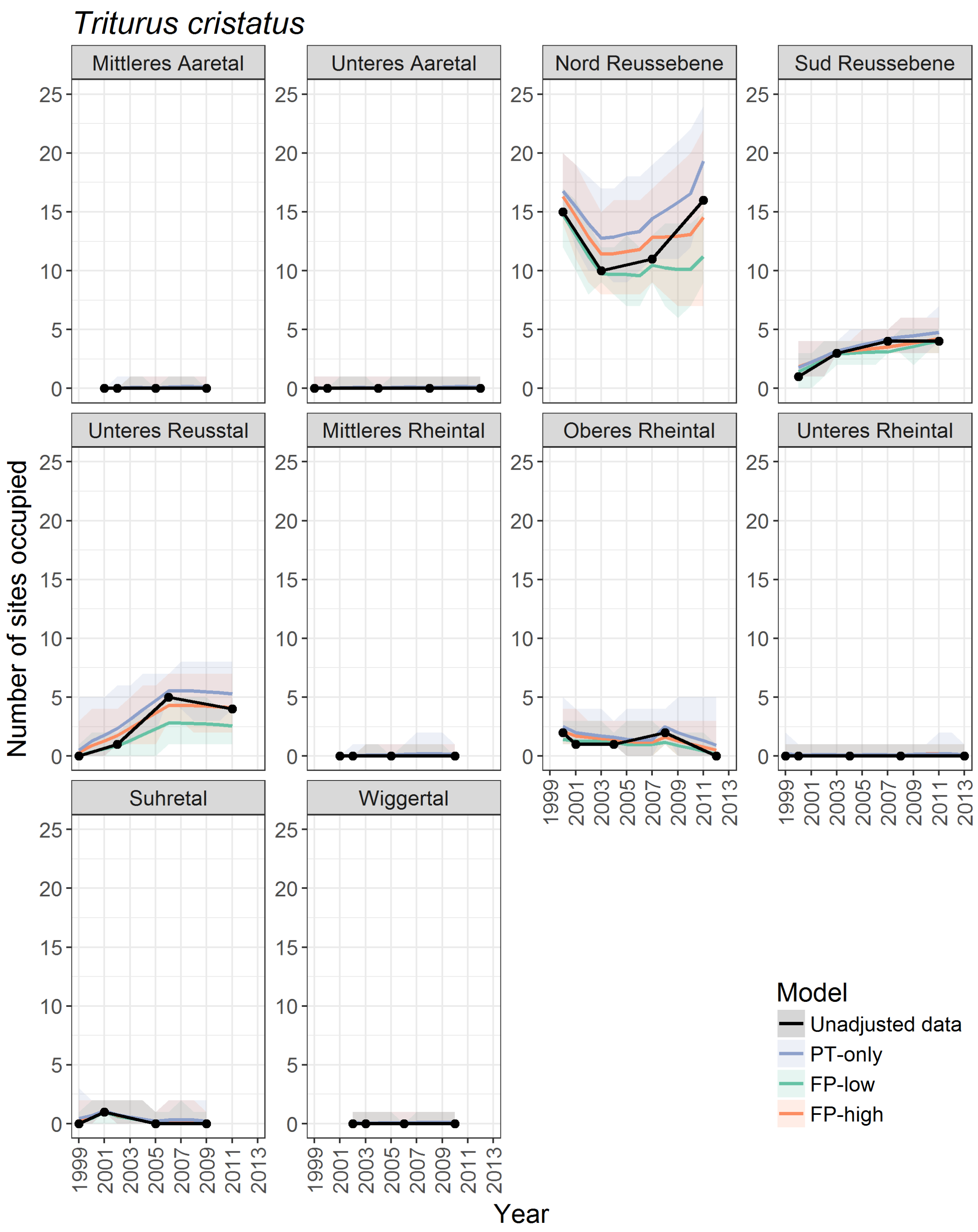

Lissotriton helveticus
Mittleres Aaretal
Unteres Aaretal
Nord Reussebene
Sud Reussebene
Unteres Reusstal
Mittleres Rheintal
Oberes Rheintal
Unteres Rheintal
Suhretal
Wiggertal
Number of sites occupied
0
10
20
30
40
Year
1999
2001
2003
2005
2007
2009
2011
2013
Model
Unadjusted data
PT-only
FP-low
FP-high


Lissotriton vulgaris
Mittleres Aaretal
Unteres Aaretal
Nord Reussebene
Sud Reussebene
Unteres Reusstal
Mittleres Rheintal
Oberes Rheintal
Unteres Rheintal
Suhretal
Wiggertal
Number of sites occupied
0
10
20
Year
1999
2001
2003
2005
2007
2009
2011
2013
Model
Unadjusted data
PT-only
FP-low
FP-high


Pelophylax esculentus complex
Mittleres Aaretal
Unteres Aaretal
Nord Reussebene
Sud Reussebene
Unteres Reusstal
Mittleres Rheintal
Oberes Rheintal
Unteres Rheintal
Suhretal
Wiggertal
Number of sites occupied
0
20
40
60
80
Year
1999
2001
2003
2005
2007
2009
2011
2013
Model
Unadjusted data
PT-only
FP-low
FP-high


Pelophylax ridibundus
Mittleres Aaretal
Unteres Aaretal
Nord Reussebene
Sud Reussebene
Unteres Reusstal
Mittleres Rheintal
Oberes Rheintal
Unteres Rheintal
Suhretal
Wiggertal
Number of sites occupied
0
20
40
60
1999
2001
2003
2005
2007
2009
2011
2013
Year
Model
Unadjusted data
PT-only
FP-low
FP-high


Rana temporaria
Mittleres Aaretal
Unteres Aaretal
Nord Reussebene
Sud Reussebene
Unteres Reusstal
Mittleres Rheintal
Oberes Rheintal
Unteres Rheintal
Suhretal
Wiggertal
Number of sites occupied
0
20
40
Year
1999
2001
2003
2005
2007
2009
2011
2013
Model
Unadjusted data
PT-only
FP-low
FP-high


Triturus cristatus
Mittleres Aaretal
Unteres Aaretal
Nord Reussebene
Sud Reussebene
Unteres Reusstal
Mittleres Rheintal
Oberes Rheintal
Unteres Rheintal
Suhretal
Wiggertal
Number of sites occupied
0
5
10
15
20
25
Year
1999
2001
2003
2005
2007
2009
2011
2013
Model
Unadjusted data
PT-only
FP-low
FP-high

CHAPTER FIVE

Availability for detection and its impact on abundance estimation using open N-mixture models

Sam S. Cruickshank [1,*], Arpat Ozgul [1], Benedikt R. Schmidt [1,2]

Manuscript in preparation

---

[1] *Department of Evolutionary Biology and Environmental Studies, University of Zürich, Zürich, Switzerland*
[2] *Info Fauna karch, Neuchâtel, Switzerland*
*sam.cruickshank@ieu.uzh.ch

ABSTRACT

1. The accurate monitoring of population abundance is fundamental in order to identify declines and implement appropriate management. Population sizes across a large number of sites can be assessed by analysing repeated counts using N-mixture models. However, meeting the assumptions of the modelling framework can be challenging in natural populations, and issues have been identified with analysing populations when detection probabilities are low.

2. As amphibian species often have low detection probabilities, and individuals do not remain within the breeding pond throughout the breeding season when counts are made (and are therefore not always available for detection), it seems likely that determining population size accurately using count-based methods may be problematic. However, the extent to which these issues inhibit our ability to obtain accurate estimates of population size is unknown.

3. In this study, we use simulated data to explore the consequences of different types of reduced availability for detection and imperfect detection rates upon the ability of both mechanistic and phenomenological open N-mixture models to estimate abundance accurately across years. We use two forms of open N-mixture models to assess situations where a constant proportion of the population are available for detection in each survey, and develop a new formulation of the model to account for availability which varies in a predictable manner over the course of repeated surveys.

4. We found that when both detection probabilities and the proportion of the population available for sampling were high, abundances could be reliably estimated, however when either of these parameters dropped below 50%, highly biased abundance estimates could arise, particularly when populations were strongly declining and when initial abundances were low. In contrast, when patterns in availability can be modelled mechanistically, reliable estimates of abundance can be obtained, even when availability levels are low.

Keywords:

Availability, N-mixture, Dail-Madsen model, abundance estimation, imperfect detection

INTRODUCTION

Estimation of population size is a fundamental task for conservation ecologists, as accurate measures are necessary in order to identify those populations under threat and evaluate the effectiveness of management (Legg & Nagy 2006; Pereira & Cooper 2006; Magurran *et al.* 2010). Population size is a useful state variable as declines can be detected with greater sensitivity through changing abundances than by assessing coarse measures such as occupancy (MacKenzie & Nichols 2004; Joseph *et al.* 2006; Dorazio 2007; IUCN Standards and Petitions Subcommittee 2017). However, making complete censuses of most natural populations is impossible, and there are many processes that may confound attempts to assess population size accurately (Pollock *et al.* 2004; Mazerolle *et al.* 2007; Schmidt & Pellet 2009).

One of the main challenges in estimating abundance is the issue of imperfect detection (Royle & Nichols 2003; Schmidt & Pellet 2009; Clement *et al.* 2017). As individuals are never detected with certainty, expected counts E(C) must therefore be considered to represent only a proportion of the true population size N (Nichols 1992):

$$E(C) = N * p \tag{5.1}$$

Where p represents detection probabilities. Obtaining accurate measures of population size therefore requires not only counts, but also information from which p can be estimated (Royle 2004). Detection probabilities can be estimated from mark-recapture datasets; however the time-consuming and expensive nature of such monitoring (Yoccoz *et al.* 2001) mean that this approach is generally unfeasible at large scales. Alternatively, the development of N-mixture modelling (Royle 2004) provides a means by which detection probabilities, and thus abundance, can be estimated from relatively cheap count data across many sites.

A second phenomenon that may also influence our ability to accurately estimate abundance is that of reduced availability; when not all individuals are present at a site when counts are made. As approaches that estimate abundance from counts assume that all individuals are equally detectable in all surveys, issues arise when a proportion of individuals are not present at a site during a survey

(Chandler *et al.* 2011). This could arise if there is migration into or out of the population between surveys (O’Donnell *et al.* 2015), or if only a proportion of the area used by the population is surveyed (Royle & Dorazio 2006). In cases such as this, it is necessary to modify equation 5.1 to reflect our understanding that an individual can only be counted if it is both available for detection (a) and then detected given that it was available (p):

$$E(C) = N * p * a \tag{5.2}$$

We refer to this product of availability and detection probability as the *realised detection rate* (τ). Here, availability is the proportion of the total population that is present in the study area during a survey; it can therefore be considered a measure of site closure (Diefenbach *et al.* 2007; Schmidt *et al.* 2013). By acknowledging that availability may not equal one, it is important to recognise that abundance estimates relate not to the total number of individuals present in a survey, but rather to the superpopulation, representing all individuals that make use of a site over the period in which repeat surveys take place (Schmidt *et al.* 2013). However, as N-mixture modelling works on the assumption that all individuals are equally detectable in all surveys (Royle 2004), imperfect availability can be seen as a case of extreme heterogeneity in detection, and so will likely impact our ability to accurately determine true abundances (Nichols *et al.* 2009; Wilson *et al.* 2014).

Monitoring is often planned to try and minimise the possibility for population closure violations and maximise the proportion of the population available for sampling (MacKenzie & Royle 2005). In cases where repeat counts are made concurrently or at short intervals, it may be possible that no migration occurs (MacKenzie & Royle 2005). However, in many situations movement into or out of a site are unavoidable. This issue is particularly relevant to amphibian populations, where there is often no feasible monitoring design that can ensure population closure. For many species, making counts at ponds during the breeding season is the best opportunity to sample the entire population, as individuals may be dispersed throughout the terrestrial environment outside of this period (Alford & Richards 1999; Semlitsch & Bodie 2003). However, for species with extended breeding seasons, there may be substantial turnover of individuals throughout this season (Wagner *et al.* 2011). Intensive studies of breeding population reveal that individuals frequently enter and exit breeding ponds in an unpredictable manner (Bertram *et al.* 1996; Given 2002). Indeed, Grafe and Meuche (2005) intensively studied male tree frog attendance in a pond over an entire season and found not only that availability was low- at most 57% of males were present during a given night- but also that availability varied substantially from day to day. Under such circumstances, it is implausible to assume that populations are closed, thus we must acknowledge the presence of reduced availability when estimating abundance and strive to understand its consequences.

Reduced availability represents a violation of the assumption of closure that underpins the N-mixture modelling framework. We therefore set out to examine the impacts of reduced availability upon the robustness of population size estimates derived from count data, and to identify the conditions under which reliable abundance estimates can be derived. In this study we focus on two different ways in which availability may be reduced. We initially consider the situation where availability for detection may be reduced to an unknown extent, but remains constant throughout time. This situation could arise when repeated counts are made at a breeding pond experiencing high turnover- resulting, for example, from individuals entering the pond briefly before leaving once they have reproduced (e.g. Wagner *et al.* 2011). The resulting counts may remain relatively steady, but with maximum counts substantially lower than the true number of individuals that made use of the pond. Reflecting the fact that in real survey data it is typically not possible to disentangle the influence of detection and availability independently, we refer to this situation as that of *unpredictable availability*, as it is only possible to estimate the compound parameter *realised detection* (equation 5.2).

In our second situation, we examine the scenario where availability is believed to be less than one, and is expected to vary between surveys following a predictable pattern. We consider here the case study of amphibian egg masses, which are commonly used to estimate the size of the breeding female population (Crouch & Paton 2000; Grant *et al.* 2005; Salvidio 2009). Females (through their proxy of egg-mass counts) are unavailable at the start of the breeding season, but availability is expected to increase through the season as egg-clutches are laid. Towards the end of the season, counts of egg-masses should approach the total abundance of the female breeding population (Fig. 5.1). We refer to this scenario as that of *variable availability*, and propose a new formulation of open N-mixture model which explicitly models this pattern in availability.

In this study, we use simulations to examine the influence of reduced availability for detection upon our ability to accurately estimate population abundances using both phenomenological and mechanistic open N-mixture models (Dail & Madsen 2011; Hostetler & Chandler 2015). Our goal was to evaluate the conditions under which abundance can be reliably derived. We first explored the performance of the standard Dail-Madsen (Dail & Madsen 2011) and exponential growth (Hostetler & Chandler 2015) models in estimating abundance from a range of biological datasets generated under a range of availability and detection probability combinations. We present a novel formulation of open N-mixture model which explicitly models the availability process, and finally explore the performance of this model under a range of demographic scenarios.

METHODS

N-mixture and Dail-Madsen Models

Open N-mixture models (Dail & Madsen 2011) represent an extension of the single-season N-mixture model of Royle (2004), in which a population model describes changes in abundance between primary periods. The model estimates abundance $N_{i,t}$ in each of *i* sites and *t* time periods using $y_{i,k,t}$ repeated counts. Abundances in the first time period are modelled as a Poisson process where λ is an estimated parameter describing the mean abundance across sites:

$$N_{i,1} \sim Poisson\ (\lambda) \qquad (5.3)$$

The counts in each year are traditionally described as a binomial trial conditional on abundance, where the success probability p represents the detection probability- however to account for reduced availability, we instead use realised detection as the success probability:

$$y_{i,t,k} \sim Binomial\ (N_{i,t}\ *\ \tau) \qquad (5.4)$$

In this study we evaluated two forms of open N-mixture model; 1) the exponential growth model of Hostetler & Chandler (2015), which models population growth between years as a function of the intrinsic population growth rate; and 2) the standard Dail-Madsen (DM) model (Dail & Madsen 2011) a mechanistic model whereby abundances are estimated as a function of survival and recruitment rates. In the exponential formulation of Hostetler & Chandler (2015), changes in abundance between years are modelled as a function of *r*, the intrinsic population growth rate:

$$N_{i,t+1} \sim Poisson\ (N_{i,t} \ * \ e^{r}) \tag{5.5}$$

In contrast, in the Dail-Madsen model survival and recruitment rates are estimated, and changes in population size are derived as a combination of new recruits and survival from previous years:

$$N_{i,t+1} = S_{i,t} + G_{i,t} \tag{5.6}$$

$$S_{i,t} \sim Bernoulli\ (surv \ * \ N_{i,t}) \tag{5.7}$$

$$G_{i,t} \sim Poisson\ (rec \ * \ N_{i,t}) \tag{5.8}$$

Where $S_{i,t}$ represents the number of individuals from time t which survive into time t+1 with survival probability *surv*, and $G_{i,t}$ denotes the number of individuals recruited into the population between t and t+1, which is a function of the per-capita recruitment rate *rec*, and the number of individuals present in time t.

Therefore three parameters are estimated in the exponential model (τ, λ, r) and four in the Dail-Madsen model ($\tau, \lambda, surv, rec$). We kept all parameters constant throughout our simulations, therefore these two formulations can perfectly describe the same pattern of abundance as $r = rec * (1-surv)$.

Egg-mass Count Model

In order to account for the expected accumulation of egg masses throughout a season in our hypothetical dataset, we modified the observation sub-model of the exponential open N-mixture model described above. Abundances are modelled as in the exponential growth model (Equations 5.2 and 5.5). However, the relationship between these true abundances and the $y_{i,j,k}$ repeated counts must be adapted to reflect our expectation that availability increases over the course of the survey season. In our example case study of egg counts, we use the expectation that the number of egg masses present at the start of the season will be very low, will increase over the course of the season and tend towards the number of breeding females. This is analogous to the logistic growth function in which abundances N_t at time t are a function of initial population size N_0 at time 0, the population carrying capacity K, and the population growth rate r:

$$N_t = \frac{K}{1+\left(\frac{K-N_0}{N_0}\right) e^{-rt}} \tag{5.9}$$

If we replace the carrying capacity K with the true abundance at a site $N_{i,t}$, and the population growth rate r with the rate at which egg masses accumulate (i.e. the rate at which availability increases; *acc.rate*), then this equation can be used to describe the counts $y_{i,j,t}$ at a site (Fig 5.1). Equation 5.9 is conditioned on the initial abundance N_0, which we can replace with y_0, the count at time 0, which represents the start of the breeding season. However, as the logistic growth function will not allow growth from an initial count of 0, we actually replace y0 with the value 1. The final modification involves replacing the parameter t. In our modified equation, t represents the time elapsed since the start of the breeding season. We replace this value with two components. Firstly we have a matrix $\text{timediff}_{i,j,t}$ which gives, within each year, the date since sampling started. However, the time since sampling started is not the same as the start of the breeding season (which may in fact differ among sites), thus to this matrix we add one final parameter, $\text{lag}_{i,t}$, which offsets the timediff data such that the first observed count need not take value 1. The lag is estimated from the data and is specified as:

$$lag_{i,t} \sim Poisson\ (timelag) \tag{5.10}$$

For the purposes of our simulations, we give timelag a uniform (0,10) distribution. With these modifications, the logistic growth function of Equation 9 can now be used to relate the abundance at time t to the observed counts:

$$avail.eggs_{i,t,k} = \frac{N_{i,t}}{1+\left(N_{i,t}-1\right) e^{-acc.rate*\left(lag_{i,t}+\ timediff_{i,t,k}\right)}} \tag{5.11}$$

$$y_{i,t,k} \sim Poisson\ (avail.eggs_{i,t,k}) \tag{5.12}$$

This is described graphically in Figure 5.1. We note that this description of the observation process is incompatible with counts of zero at sites with non-zero abundances (as this is inconsistent with our simplification of setting y_0 to equal 1), therefore any such observations should not be included in the data.

We do not explicitly account for imperfect detection using this formulation of the model, as we envisage it being applied to data in which the detection rate for egg clusters is close to 1 (e.g. Grant *et al.* 2005). However, we note that incorporating the effects of imperfect detection would be possible by replacing Equation 5.12 with a modified version of equation 5.3 in which $N_{i,t}$ is replaced by $avail.eggs_{i,t,j}$ rounded to the nearest integer.

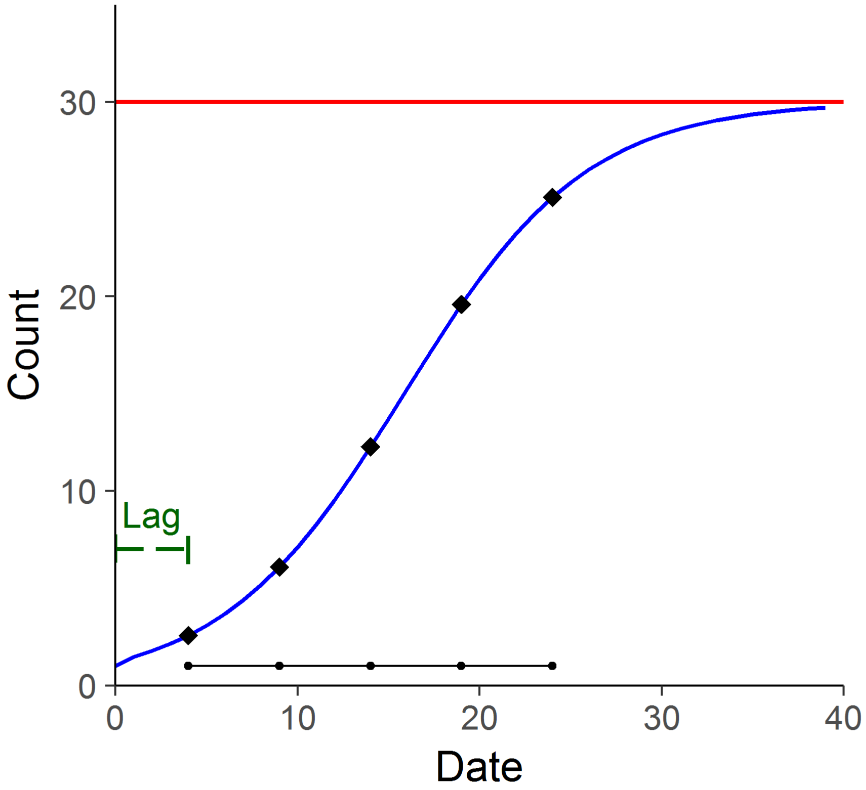


Figure 5.1. Schematic describing the expected pattern of counts as a result of increased availability (Equation 11). The abundance at the site ($N_{i,t}$) is represented by the red horizontal line. Successive counts are made (here at 5 day intervals),and the respective counts $y_{i,j,t}$ (black diamonds) are used to estimate the egg accumulation rate (*acc.rate*, blue line). Within the dataset, the time difference between surveys is known, (black horizontal line), however the time between the hypothetical start of the survey (y-axis) and the first survey (here at x=4) is unknown, and is estimated from the data as the parameter *lag*.

Simulations

We evaluated the performance of all models by applying them to a series of datasets in which the underlying population dynamics were identical. We simulated abundance datasets representing abundances across 50 sites for a 10 year-period. With initial abundances of 5, 10, 20 and 50 individuals per site, we simulated population growth with intrinsic growth rates ranging from strongly declining (r = -0.3) to strongly growing (r = 0.3). We generated these growth rates mechanistically using the same survival rate and varying recruitment rates (Table 5.1). We therefore generated 20 abundance datasets in which model performance was evaluated.

To explore the effects of unpredictable availability, from each of these datasets we generated 3 repeat counts per site and season using 25 values of realised detection rate, which we derived by combining a range of plausible availability and detection probability parameters (Table 5.2). We therefore applied the exponential growth and DM model to 500 different count datasets. For each scenario, we assessed model convergence, absolute error (estimated value - true value) in the key parameters (r, lambda, realised detection, and surv/rec as appropriate). To assess bias in abundance for each scenario, we calculated mean absolute error to identify cases when abundance was systematically over- or under-estimated, as well as relative root mean squared error (relative RMSE). RMSE is calculated as $\sqrt{\sum(N_{estimated} - N_{true})^2/500}$ and is used as general metric of model performance which is highly sensitive to outliers (see Chai & Draxler 2014; Bellier *et al.* 2016; Veech *et al.* 2016). We calculated relative RMSE by dividing RMSE by the mean abundance across each simulated dataset; therefore this metric represents the magnitude of the mean error in abundance estimates as a proportion of the true mean abundance.

In both models, we gave τ a uniform (0,1) prior and lambda a half-normal (0,0.001) prior. For the exponential growth model, r was given a normal (0,0.01) prior. It has been suggested that survival and recruitment may be non-identifiable in this model without the inclusion of extra information on these parameters (for example from CMR data; [Bellier *et al.* 2016; Zipkin *et al.* 2017]). We tested this idea by applying the Dail-Madsen model with two sets of priors. In the first, we used vague uniform (0,1) priors for survival and recruitment. In the second, we restricted the uniform priors on survival and recruitment to span 0.15 either side of their true simulated value. This simulates the kind of prior information that may be available from CMR data. Note that for recruitment of 0 the prior range was set to (0,0.3).

We used the same abundance datasets to assess the egg count model. In this case, we simulated a sequence of 5 counts at 7 day intervals, using Equation 5.11 with an accumulation rate of 0.2. The lag for each site and year was drawn from a Poisson distribution with expected value of 5. The egg accumulation rate was assigned a uniform (0,0.3) prior, *timelag* a uniform (0,10), and r and lambda were assigned normal (0,0.10) and half-normal (0,0.001) priors respectively.

All models were run in JAGS (Plummer 2003) using package *jagsUI* (Kellner 2016) through R (R Core Team 2017). We ran 3 chains with 5000 samples for adaptation and burn-in, followed by 50,000 iterations which we then thinned 1 in 10. Convergence was assessed using the Gelman-Rubin statistic (Gelman *et al.* 2004). Model convergence is an important parameter as non-convergence is indicative that the estimated parameters may be unstable and thus that inference should not be drawn from the model (Cowles & Carlin 1996).

Table 5.1. Combinations of survival and per capita recruitment used to simulate populations ranging from strongly growing to strongly declining.

Simulation description	Overall population growth rate	Demographic rates	
	r	*recruitment*	*survival*
Strong decline	-0.3	0.0	0.7
Weak decline	-0.1	0.2	0.7
Stable	0	0.3	0.7
Weak growth	0.1	0.4	0.7
Strong growth	0.3	0.6	0.7

Table 5.2. Realised detection probabilities used in simulations. Values are a product of detection probability (columns) and availability for detection (rows)

	0.2	**0.35**	**0.55**	**0.75**	**0.95**
0.2	0.04	0.07	0.11	0.15	0.19
0.4	0.08	0.14	0.22	0.30	0.38
0.6	0.12	0.21	0.33	0.45	0.57
0.8	0.16	0.28	0.44	0.60	0.76
1.0	0.20	0.35	0.55	0.75	0.95

RESULTS

Unpredictable Availability

Model convergence was good for the exponential growth models- the model converged for 97.8% of simulation scenarios. In contrast, convergence was more problematic for the Dail-Madsen model, particularly when non-informative priors on survival and recruitment were used. In such situations, the model converged in only 59% of scenarios, whereas with informative priors performance improved and convergence was reached in 90% of scenarios. Consequently, we report here only the results from models with informative priors, and only results from simulations that converged. We note that although convergence would eventually have been reached in these models (in further tests, convergence was reached after 2,000,000 iterations), errors in abundance were extremely high, and thus non-convergence in this situation can be considered analogous to poor model performance. In all cases, convergence failures were associated with datasets with low realised detection rates in combination with high λ and strong population growth.

Overall, exponential growth and DM models estimated abundances accurately across most simulated scenarios. However, for both models, relative RMSE was strongly related to realised detection rates (Fig 5.2, Fig SI 5.1). For exponential growth models, relative RMSE was 0.658 when realised detection was below 0.5, but only 0.096 above this threshold. The relative magnitude of abundance error was much lower in DM models, where relative RMSE was 0.237 for simulations where $\tau<0.5$, but only 0.083 above this threshold. In both models, abundance error was more extreme when low realised detection rates were found in combination with low initial abundance and strong population declines. In such situations, abundances were commonly estimated with error exceeding 100% (Fig 5.2, Fig SI 5.1), particularly in the exponential growth model. In terms of absolute bias in abundance estimates, across simulated scenarios exponential models tended towards systematic overestimation (mean absolute error 7.32 ± 11.0), whereas the DM models were unbiased (mean absolute error -0.19 ± 5.8).

Population growth rates tended to be underestimated in the exponential model (mean absolute error across scenarios: -0.021 ± 0.026). In simulations where populations were strongly growing or declining, population growth rates were more strongly underestimated at all levels of realised detection (Fig SI 5.2). Precision was generally high, with mean 95% CRI spanning 0.057 ±0.036. Precision increased at higher initial population sizes and with stronger population growth. In contrast, population growth rates were well estimated in DM models (mean bias 0.000 ± 0.010; Fig SI 5.3), although error was elevated in simulations with low initial population sizes.

Initial population sizes were strongly overestimated by the exponential growth model across scenarios where realised detection rates were low (Fig SI 5.4). In contrast, at low realised detection rates, DM models produced a combination of over- and under-estimates of lambda, however the overall magnitude of errors was substantially lower (Fig SI 5.5). For both models, precision was lower with reduced realised detection; overall the DM model produced smaller 95% CRI (mean 9.6 ± 10.1) than the exponential growth model (mean 13.8 ± 13.6).

In both models, realised detection rates were poorly estimated at intermediate detection rates, but estimated well or with slight underestimation at extremely high or low rates (Figs SI 5.6 and SI 5.7). Bias was much stronger in the DM model (mean across scenarios: -0.22 ± 0.21) than in the exponential model (-0.05 ± 0.04). Precision, however was similar between models, with 95% CRI spanning 0.07 and 0.08 in the exponential and DM models, respectively.

Survival was held constant in all simulations. In scenarios with no recruitment, survival was consistently underestimated by the DM model, whereas when recruitment occurred, a combination of over- and under-estimates of survival were produced (Fig SI 5.8). With the lowest realised detection rates, 95% CRI spanned nearly the entirety of the 0.3 prior distribution, whereas at higher levels of realised detection, precision was increased (Fig SI 5.9). Similarly, recruitment was consistently overestimated in those scenarios where no recruitment was simulated ($r = -0.3$) and otherwise a combination of over-and under-estimates were produced in different simulation scenarios (Fig SI 5.10). Precision in recruitment followed an identical pattern to that of survival, with very low precision at low realised detection rates.

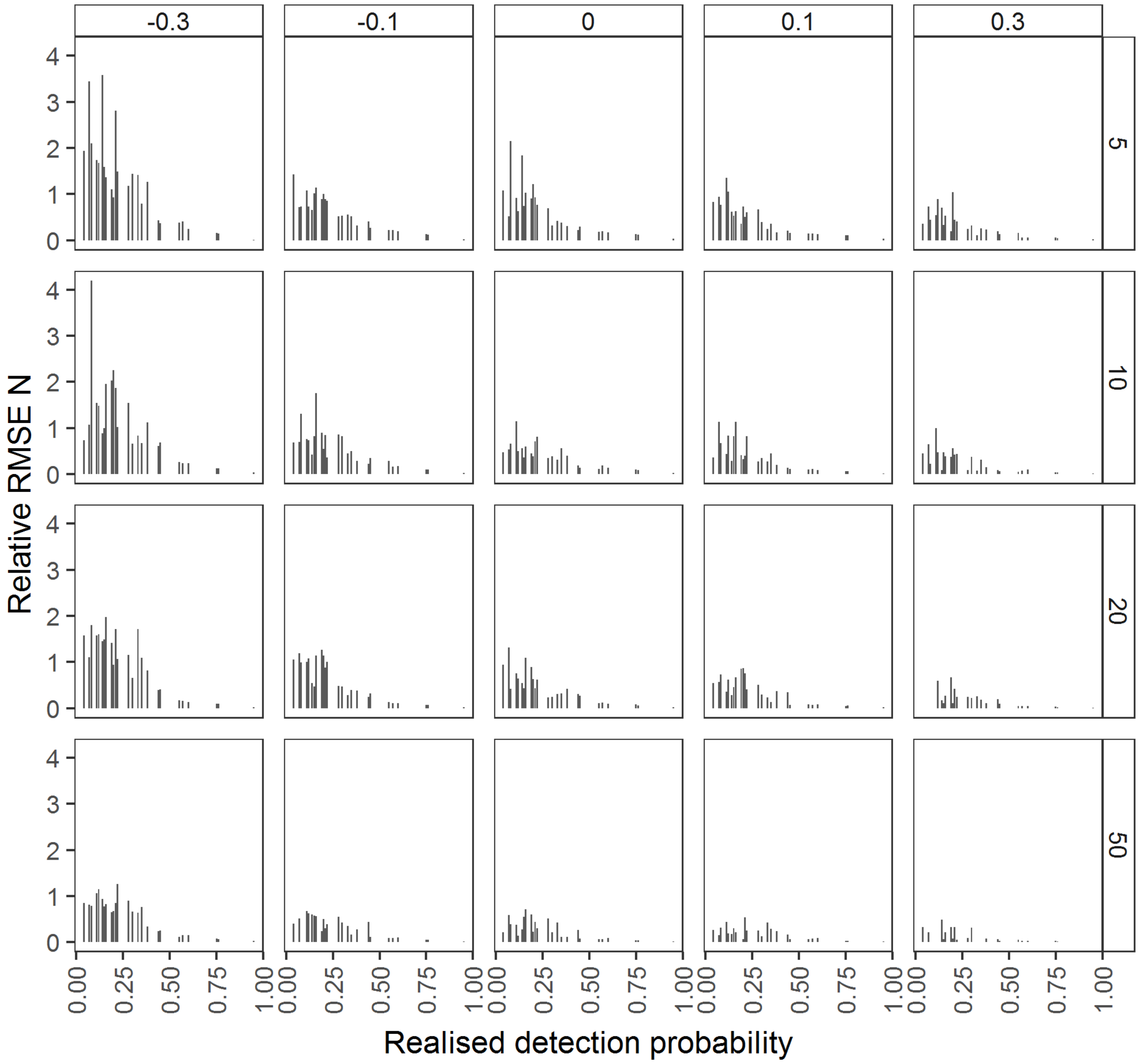


Figure 5.2. Root mean squared error in abundance for datasets analysed using the standard exponential growth model. Panels show increasing population growth rate from left to right, and increased initial population size from top to bottom.

Variable Availability

In the simulated datasets, maximum counts were on average 4.77 lower than true abundances, albeit with wide variation both within and between scenarios (overall sd: 28.17). Therefore availability for detection was substantially less than 1 across the simulated datasets. The deficit between true abundance and maximum counts was greater in scenarios with larger initial population size.

The egg count model performed well under all simulated values of population growth rate and initial abundance (Fig 5.3), and there were no cases where convergence was not reached. Abundances were estimated well (Fig 5.4), with a mean relative RMSE across all simulations of 0.159 ± 0.070. The poorest estimation occurred when population sizes were low and strongly declining; in this scenario mean abundance error reached nearly 35% of total true abundance. Mean absolute error averaged across simulations was -0.009 ± 1.16, suggesting that there was no systematic over- or under-estimation of abundance between simulation scenarios.

Population growth rates were well estimated (mean error across simulations: 0.026 ± 0.225), although with strongly declining populations when initial population sizes were small, growth rates could be underestimated and have low precision (Fig SI 5.11). Lambda was estimated accurately in most scenarios, although when populations were exhibiting strong growth or declines, this parameter could be underestimated (Fig SI 5.12). 95% CRI increased with increased lambda, but were generally small (2.87 ± 1.4; Fig SI 5.12). Finally, the egg accumulation rate was poorly estimated when initial population size was low and populations were strongly declining, but in all other scenarios this parameter was accurately and precisely estimated (Fig SI 5.13)

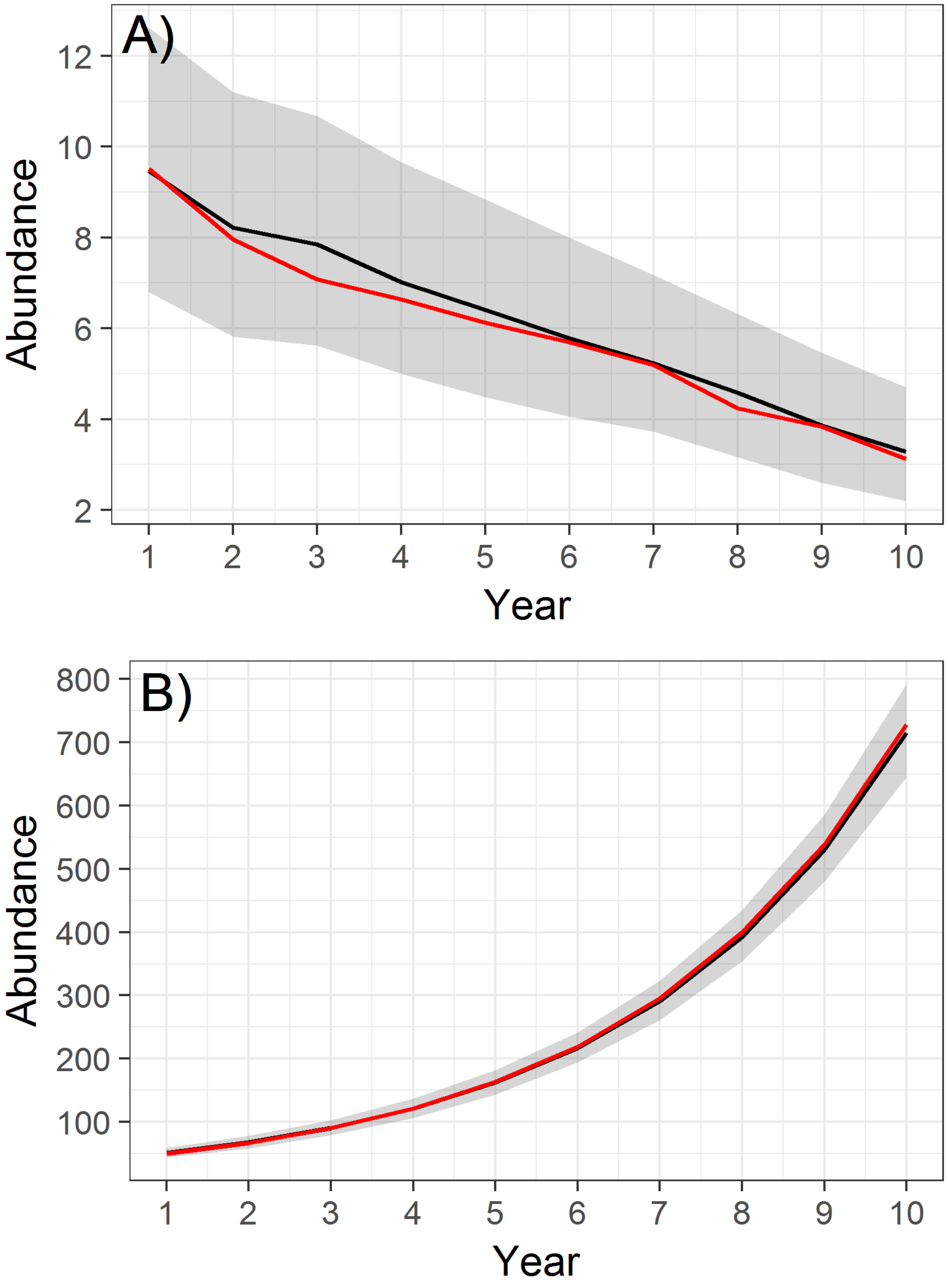


Figure 5.3. True population trajectories (red) and egg-count model estimates (black with grey ribbons representing 95% CRI) for two datasets under conditions of A). λ= 5, r= -0.1, B) λ= 50, r= 0.3

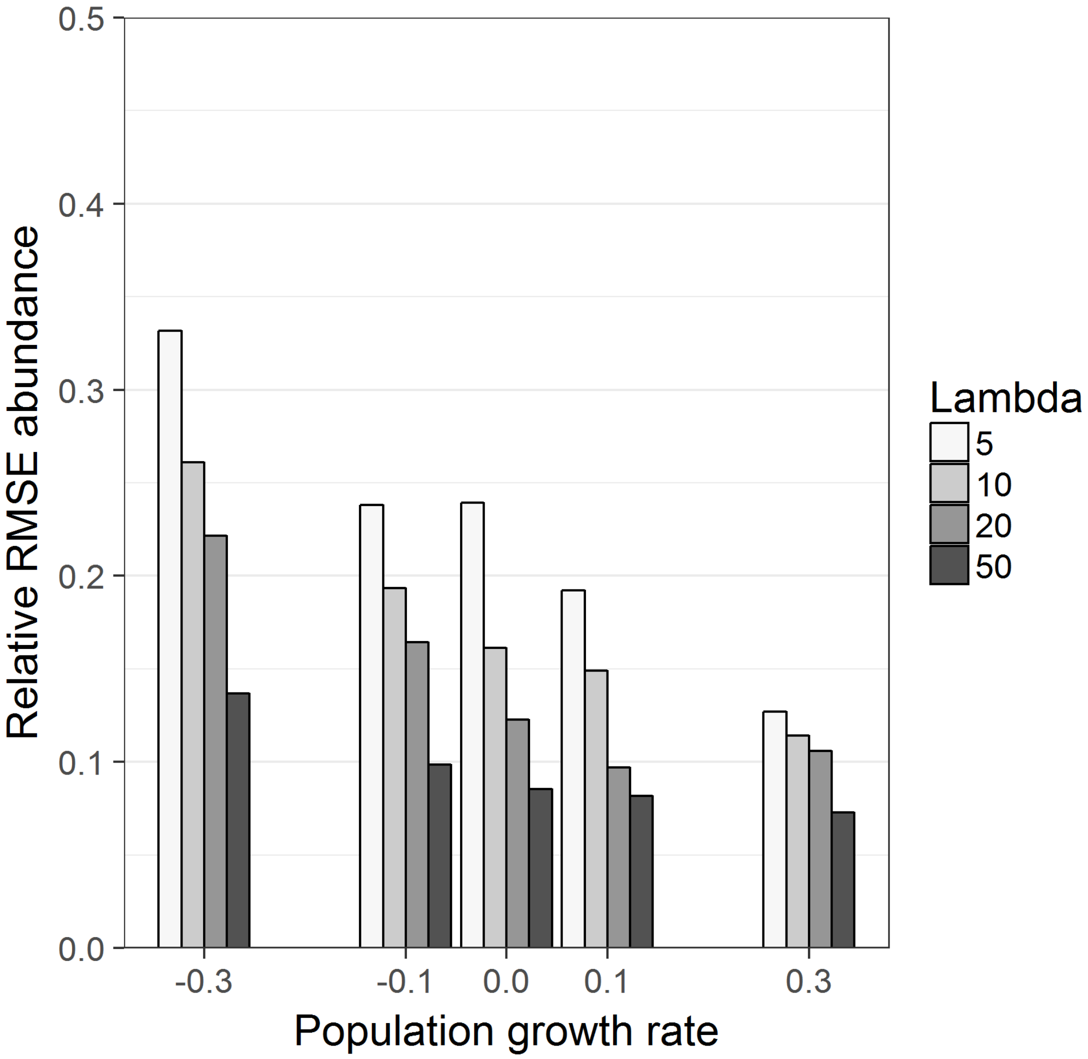


Figure 5.4. Root square mean error in abundance for datasets analysed using the egg count model.

DISCUSSION

It has long been recognised that low detection probabilities can pose challenges for N-mixture models and may leads to unrealistically high abundance estimates (Royle 2004; Joseph *et al.* 2009; Kéry & Royle 2015; Veech *et al.* 2016). Given that many amphibian species are cryptic and individuals are frequently hard to detect (Schmidt & Pellet 2009), this suggests that estimating population sizes may be challenging even in cases where the closure assumption is truly met. In this study we demonstrate that this issue is exacerbated when only a subset of the total population are available for detection in any given survey. Reduced availability interacts with detection probabilities to influence the realised detection rate; thus if either of these are low it becomes almost impossible to ensure that abundance estimates are reliable. We found that mechanistic population models such as the Dail-Madsen are better able to deal with reduced availability and provide more reliable estimates of abundance - provided that prior information on demographic rates are available. This suggests that some combination of count data and mark-recapture at a subset of sites might be an optimal solution to monitoring populations in which availability cannot be assumed to be perfect (Zipkin *et al.* 2017).

If count data relate to a population in which availability is not equal to one, then the resulting detection estimates will necessarily be lower than found in similar studies of the species in which population closure is assured. However, the consequences of reduced availability also influence how we interpret the resulting abundance estimates. Rather than assessing the number of individuals present in a pond, N-mixture models will produce estimates of the size of the superpopulation of individuals making use of the pond over the course of successive surveys (Schmidt *et al.* 2013), which may be substantially higher. Relatively stable counts can therefore easily give the impression of a population being small, whereas in reality only a proportion of the much larger superpopulation is exposed to sampling in any given survey. Indeed, individual based monitoring frequently reveals true abundances can be substantially higher than those estimated from counts alone (Pellet *et al.* 2007; Schmidt & Pellet 2009). Correctly interpreting unexpectedly high abundance estimates can therefore be a difficult task; if detection probabilities are high and there is evidence that closure violations are minimal, then the inference is likely robust. However, if there are reasons to believe that either of these parameters is low, then it may be more likely that the abundance represents ‘freak’ estimates we report for low realised detection rates, and which have previously been discussed in the context of single-season N-mixture models (Royle 2004; Kéry & Royle 2015). Given this potential conflict, in cases where there is a reason to expect reduced availability, it seems wise to design monitoring protocols to maximise detection probabilities such that realised detection rates may still remain high.

In the face of reduced availability, both open N-mixture model variants were able to accurately estimate abundances as long as realised detection rates were high (above approximately 0.50). Below this, both produced biased estimates, particularly when initial population sizes were low and when population were strongly declining. However, the potential for extreme bias was much greater in the exponential model of Hostetler & Chandler (2015) than in the Dail-Madsen model (Dail & Madsen 2011) (Fig. 5.2 and Fig. SI 5.1) in which population growth was decomposed into survival and recruitment rates. Under these conditions, the exponential model systematically overestimated true abundances, whereas the Dail-Madsen model showed no systematic bias. This suggests that the Dail-Madsen model is better able to analyse data in which availability may be reduced; however it is important to note that in our simulations we had to use informative priors in order for the models to converge and produce reliable estimates. Several authors have previously suggested that extra information in the form of informative priors (Zipkin *et al.* 2014; Morris *et al.* 2015) or by incorporating data on demographic rates from other sources such as CMR data (Bellier *et al.* 2016; Zipkin *et al.* 2017) may be necessary in order to improve the accuracy of the DM-model. The failure of so many models with uninformative priors suggests that such information is indeed essential, particularly when realised detection rates are low. In our Dail-Madsen simulations, priors were set such that the true value sat at the centre of the prior distribution. The high accuracy and low precision of our demographic rate estimates appear therefore to reflect the strong importance of these priors, and cautions that the count data likely do not contain sufficient information to estimate these rates. We therefore recommend the collection of extra information (e.g. mark-recapture at a subset of sites) if population sizes are to be evaluated using the Dail-Madsen model when perfect availability cannot be assumed.

Careful study design can minimise, but in most cases cannot eliminate, reduced availability or violations of closure arising from species phenology (Diefenbach *et al.* 2007; McClure *et al.* 2011). Although assessing population availability is often a time-consuming and costly process (Diefenbach *et al.* 2007) and is therefore not feasible in many monitoring applications, we have shown that issues relating to reduced availability can be mitigated if expected patterns in availability can be incorporated mechanistically into the observation process. In our egg-count example we were able to accurately assess population sizes even though availability was almost always less than one and was very low in most counts. Such approaches have been used with success in analysing counts of avian and marine mammal populations (Mordecai *et al.* 2011; Borchers *et al.* 2013), and although patterns in amphibian availability may typically be highly variable and impossible to reliably predict (e.g. Grafe & Meuche

2005), we have demonstrated that this method could nonetheless be beneficially applied to some forms of amphibian monitoring count data.

Our egg-count model represents a simplified example in which availability varies predictably, but it can readily be extended to incorporate more complex scenarios. For example, having a single accumulation rate to describe the increase in counts over time will likely be problematic if there is large variation in abundance between sites. By incorporating covariates or a random effect on this parameter, the model should be sufficiently flexible to accurately estimate true abundances when counts are highly over-dispersed. Similarly, given the well documented effects of climate-change on species phenology (Parmesan 2006), availability for detection may differ between years (McClure *et al.* 2011), thus for analysis of long-term datasets it may be beneficial to also allow for annual differences in egg-accumulation rates, in order to account for differing availability between years as a consequence of changing phenology. Although we decided not to explicitly build imperfect detection explicitly into this model as we envisage it being applied to count data where detection probabilities are high, our model could readily be extended to accommodate detection probabilities explicitly. Finally, in this study we chose to explore the behaviour of open N-mixture models in populations exhibiting exponential growth as this allowed us to assess performance on the same simulated datasets as for the mechanistic and phenomenological open N-mixture models. However, our model can be easily modified to include other forms of population dynamics; indeed although we do not present results here, the model can also accurately model density-dependent population growth.

Previous research has identified that low detection probabilities can cause problems when estimating abundances within closed populations. We demonstrate that these problems are also present in open N-mixture models, and that biased inference can be exacerbated when populations are not uniformly available for detection. We argue that this issue is particularly important for the analysis of amphibian populations, as detection probabilities are frequently low and assumptions of complete availability cannot typically be made. We demonstrate that when population availability is close to one and detection probabilities are high, abundance estimates from N-mixture models can be relied upon, however if either parameter is below 50%, abundances may become highly biased. In cases where expected patterns in availability can be explicitly modelled, we have shown that abundances can be reliably inferred even when availability is often frequently low. In cases where availability is unknown, mechanistic models such as the Dail-Madsen represent the best method for minimising bias, although the accuracy of this model is likely dependent upon the availability of prior information on demographic rates. The best approach for monitoring large-scale trends in abundance may therefore be a combination of count-based monitoring informed by mark-recapture at a subset of locations. The guidance we provide here will help to minimise the consequences of violating the closure assumption underlying abundance modelling, however study design cannot remove the influence of species phenology and so some violations of modelling assumptions are always possible. Given the potentially serious bias that can arise from ignoring these effects, we therefore emphasise that potential violations of population closure are not ignored by those designing monitoring studies, and that effort be invested into understanding potential patterns such that availability need not bias abundance estimates.

Literature Cited

Alford, R.A. & Richards, S. (1999). Global amphibian declines: a problem in applied ecology. *Annual Review of Ecology and Systematics*, **30**, 133–165.

Bellier, E., Kéry, M. & Schaub, M. (2016). Simulation-based assessment of dynamic N -mixture models in the presence of density dependence and environmental stochasticity. *Methods in Ecology and Evolution*, **7**, 1029–1040.

Bertram, S., Berrill, M. & Nol, E. (1996). Male Mating Success and Variation in Chorus Attendance within and among Breeding Seasons in the Gray Treefrog (Hyla versicolor). *Copeia*, **3**, 729–734.

Borchers, D.L., Zucchini, W., Heide-Jørgensen, M.P., Cañadas, A. & Langrock, R. (2013). Using Hidden Markov Models to Deal with Availability Bias on Line Transect Surveys. *Biometrics*, **69**, 703–713.

Chai, T. & Draxler, R.R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, **7**, 1247–1250.

Chandler, R.B., Royle, J.A. & King, D.I. (2011). Inference about density and temporary emigration in unmarked populations. *Ecology*, **92**, 1429–1435.

Clement, M.J., Converse, S.J. & Royle, J.A. (2017). Accounting for imperfect detection of groups and individuals when estimating abundance. *Ecology and Evolution*, **7**, 7304–7310.

Cowles, M.K. & Carlin, B.P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, **91**, 883–904.

Crouch, W.B. & Paton, P.W.C. (2000). Using Egg-mass Counts to Monitor wood frog populations. *Wildlife Society Bulletin*, **28**, 895–901.

Dail, D. & Madsen, L. (2011). Models for Estimating Abundance from Repeated Counts of an Open Metapopulation. *Biometrics*, **67**, 577–587.

Diefenbach, D.R., Marshall, M.R., Mattice, J.A. & Brauning, D.W. (2007). Incorporating Availability for Detection in Estimates of Bird Abundance. *The Auk*, **124**, 96–106.

Dorazio, R.M. (2007). On the choice of statistical models for estimating occurrence and extinction from animal surveys. *Ecology*, **88**, 2773–2782.

Gelman, A., Carlin, J.P., Stern, H.S. & Rubin, D.S. (2004). *Bayesian Data Analysis*. CRC/ Chapman & Hall, Boca Raton, FL.

Given, M.F. (2002). Interrelationships among Calling Effort, Growth Rate, and Chorus Tenure in Bufo fowleri. *Copeia*, **2002**, 979–987.

Grafe, T.U. & Meuche, I. (2005). Chorus tenure and estimates of population size of male European tree frogs Hyla arborea: implications for conservation. *Amphibia-Reptilia*, **26**, 437–444.

Grant, E.H.C., Jung, R.E., Nichols, J.D. & Hines, J.E. (2005). Double-observer approach to estimating egg mass abundance of pond-breeding amphibians. *Wetlands Ecology and Management*, **13**, 305–320.

Hostetler, J.A. & Chandler, R.B. (2015). Improved state-space models for inference about spatial and temporal variation in abundance from count data. *Ecology*, **96**, 1713–1723.

IUCN Standards and Petitions Subcommittee. (2017). *Guidelines for Using the IUCN Red List Categories and Criteria*. Prepared by the Standards and Petitions Subcommittee. Available from: http://www.iucnredlist.org/documents/RedListGuidelines.pdf.

Joseph, L.N., Elkin, C., Martin, T.G. & Possingham, H.P. (2009). Modeling abundance using N - mixture models: the importance of considering ecological mechanisms. *Ecological Applications*, **19**, 631–642.

Joseph, L.N., Field, S., Wilcox, C. & Possingham, H.P. (2006). Presence-absence versus abundance data for monitoring threatened species. *Conservation Biology*, **20**, 1679–1687.

Kellner, K.F. (2016). *jagsUI: A Wrapper Around 'rjags' to Streamline 'JAGS' Analyses*. R package version 1.4.4. Available from: https://CRAN.R-project.org/package=jagsUI.

Kéry, M. & Royle, J.A. (2015). *Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS: Volume 1: Prelude and Static Models*. Academic Press.

Legg, C.J. & Nagy, L. (2006). Why most conservation monitoring is, but need not be, a waste of time. *Journal of Environmental Management*, **78**, 194–199.

MacKenzie, D.I. & Nichols, J.D. (2004). Occupancy as a surrogate for abundance estimation. *Animal Biodiversity and Conservation*, **1**, 461–467.

MacKenzie, D.I. & Royle, J.A. (2005). Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology*, **42**, 1105–1114.

Magurran, A.E., Baillie, S.R., Buckland, S.T., Dick, J.M., Elston, D.A., Scott, E.M., Smith, R.I., Somerfield, P.J. & Watt, A.D. (2010). Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in Ecology & Evolution*, **25**, 574–582.

Mazerolle, M.J., Bailey, L.L., Kendall, W.L., Andrew Royle, J., Converse, S.J. & Nichols, J.D. (2007). Making great leaps forward: accounting for detectability in herpetological field studies. *Journal of Herpetology*, **41**, 672–689.

McClure, C.J.W., Burkett-Cadena, N.D., Ligon, R.A. & Hill, G.E. (2011). Actual or Perceived Abundance? Interpreting Annual Survey Data in the Face of Changing Phenologies. *The Condor*, **113**, 490–500.

Mordecai, R.S., Mattsson, B.J., Tzilkowski, C.J. & Cooper, R.J. (2011). Addressing challenges when studying mobile or episodic species: hierarchical Bayes estimation of occupancy and use. *Journal of Applied Ecology*, **48**, 56–66.

Morris, W.K., Vesk, P.A., McCarthy, M.A., Bunyavejchewin, S. & Baker, P.J. (2015). The neglected tool in the Bayesian ecologist’s shed: A case study testing informative priors’ effect on model accuracy. *Ecology and Evolution*, **5**, 102–108.

Nichols, J.D. (1992). Capture- Recapture Models: Using marked animals to study population dynamics. *Bioscience*, **42**, 94–102.

Nichols, J.D., Thomas, L. & Conn, P.B. (2009). Inferences About Landbird Abundance from Count Data: Recent Advances and Future Directions. *Modeling Demographic Processes in Marked Populations* (eds D.L. Thomson, E.G. Cooch & M.J. Conroy), pp. 201–235. Springer US, Boston, MA.

O’Donnell, K.M., Thompson, F.R. & Semlitsch, R.D. (2015). Partitioning detectability components in populations subject to within-season temporary emigration using binomial mixture models. *PLoS ONE*, **10**, 1–18.

Parmesan, C. (2006). Ecological and Evolutionary Responses to Recent Climate Change. *Annual of Ecology, Evolution and Systematics*, **37**, 637–669.

Pellet, J., Helfer, V. & Yannic, G. (2007). Estimating population size in the European tree frog (Hyla arborea) using individual recognition and chorus counts. *Amphibia-Reptilia*, **28**, 287–294.

Pereira, H.M. & Cooper, H.D. (2006). Towards the global monitoring of biodiversity change. *Trends in Ecology & Evolution*, **21**, 123–129.

Plummer, M. (2003). JAGS: A program for analysis of bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, **124**.

Pollock, K.H., Marsh, H., Bailey, L.L., Farnsworth, G.L., Simons, T.R. & Alldredge, M.W. (2004). Separating components of detection probability in abundance estimation: an overview with diverse examples. *Sampling Rare or Elusive Species: Concepts, Designs, and Techniques for Estimating Population Parameters* (ed W.L. Thompson), pp. 43–58. Island Press.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Royle, J.A. (2004). N-Mixture Models for Estimating Population Size from Spatially Replicated Counts. *Biometrics*, **60**, 108–115.

Royle, J.A. & Dorazio, R.M. (2006). Hierarchical models of animal abundance and occurrence. *Journal of Agricultural, Biological, and Environmental Statistics*, **11**, 249–263.

Royle, J.A. & Nichols, J.D. (2003). Estimating abundance from repeated presence-absence data or point counts. *Ecology*, **84**, 777–790.

Salvidio, S. (2009). Detecting amphibian population cycles: The importance of appropriate statistical analyses. *Biological Conservation*, **142**, 455–461.

Schmidt, J.H., McIntyre, C.L. & MacCluskie, M.C. (2013). Accounting for incomplete detection: What are we estimating and how might it affect long-term passerine monitoring programs? *Biological Conservation*, **160**, 130–139.

Schmidt, B.R. & Pellet, J. (2009). Quantifying abundance: counts, detection probabilities, and estimates. *Amphibian Ecology and Conservation*, 465–480.

Semlitsch, R.D. & Bodie, J.R. (2003). Biological Criteria for Buffer Zones around Wetlands and Riparian Habitats for Amphibians and Reptiles. *Conservation Biology*, **17**, 1219–1228.

Veech, J.A., Ott, J.R. & Troy, J.R. (2016). Intrinsic heterogeneity in detection probability and its effect on N -mixture models. *Methods in Ecology and Evolution*, **7**, 1019–1028.

Wagner, N., Pellet, J., Lötters, S., Schmidt, B.R. & Schmitt, T. (2011). The superpopulation approach for estimating the population size of ‘prolonged’ breeding amphibians: Examples from Europe. *Amphibia-Reptilia*, **32**, 323–332.

Wilson, T.L., Schmidt, J.H., Thompson, W.L. & Phillips, L.M. (2014). Using double-observer aerial surveys to monitor nesting bald eagles in Alaska: Are all nests available for detection? *Journal of Wildlife Management*, **78**, 1096–1103.

Yoccoz, N.G., Nichols, J.D. & Boulinier, T. (2001). Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution*, **16**, 446–453.

Zipkin, E.F., Rossman, S., Yackulic, C.B., Wiens, J.D., Thorson, J.T., Davis, R.J. & Grant, E.H.C. (2017). Integrating count and detection-nondetection data to model population dynamics. *Ecology*, **98**, 1640–1650.

Zipkin, E.F., Sillett, T.S., Grant, E.H.C., Chandler, R.B. & Royle, J.A. (2014). Inferences about population dynamics from count data using multistate models: a comparison to capture-recapture approaches. *Ecology and Evolution*, **4**, 417–426.

SUPPLEMENTARY INFORMATION

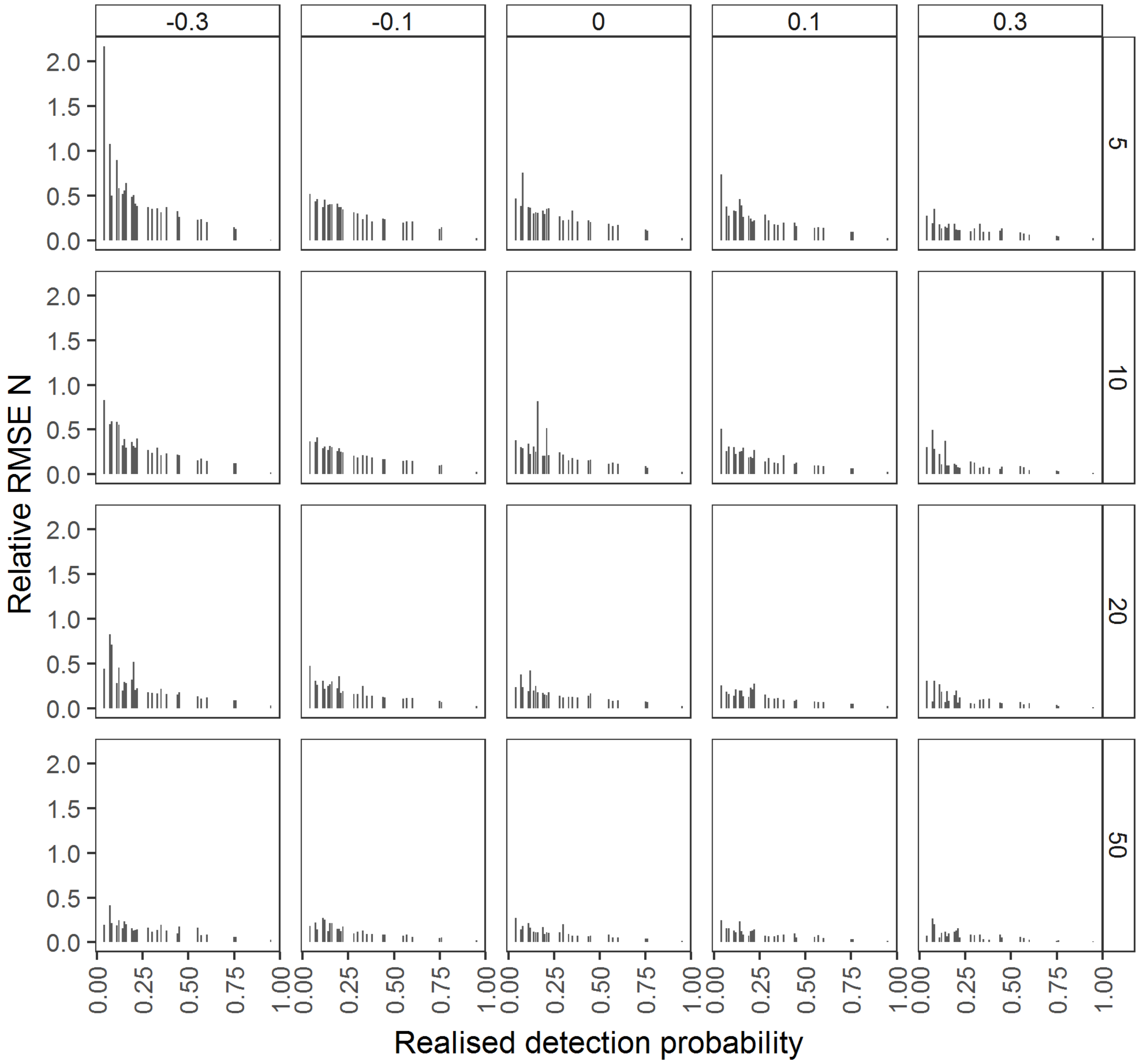


Figure SI 1. root mean squared error in abundance for simulated datasets analysed under the Dail-Madsen model. Panels represent simulations with population growth rate (columns: left to right), and increasing initial population size (rows: top to bottom)

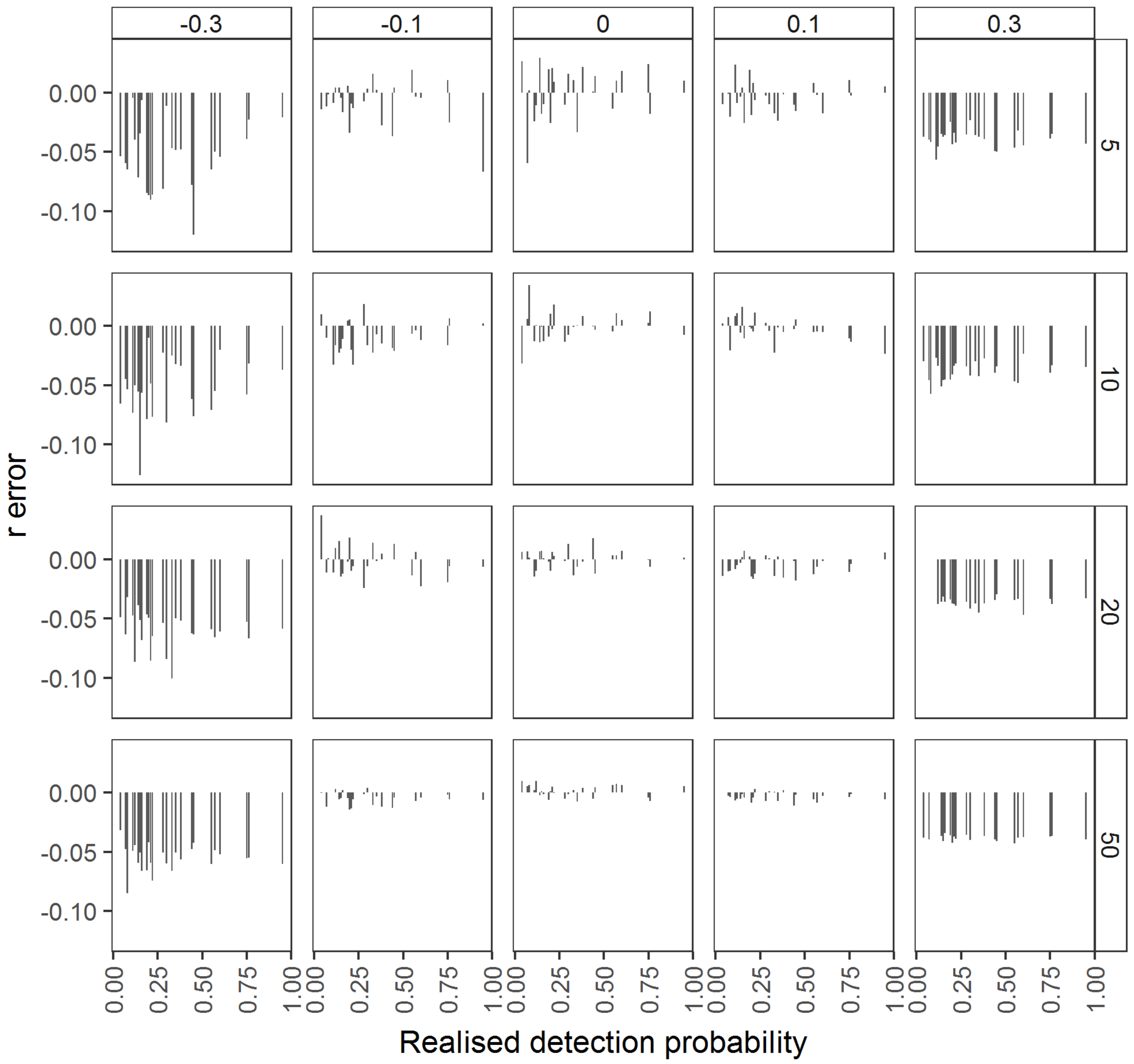


Figure SI 2. absolute error (r_{est}- r_{true}) in the population growth rate for simulated datasets analysed under the exponential growth model. Panels represent simulations with population growth rate (columns: left to right), and increasing initial population size (rows: top to bottom)

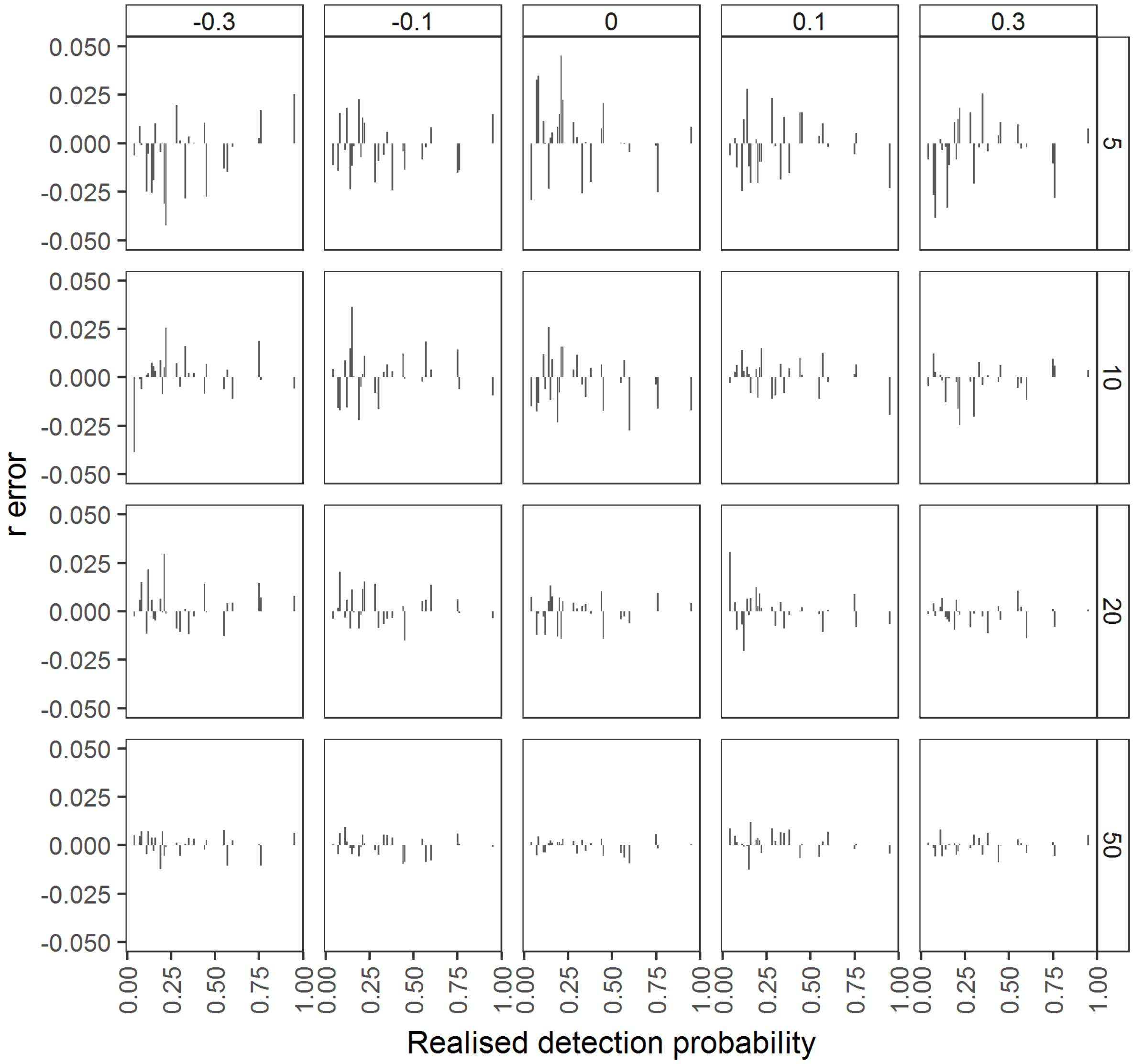


Figure SI 3. Absolute error (r_{est}- r_{true}) in the population growth rate for simulated datasets analysed under the Dail-Madsen model. Panels represent simulations with population growth rate (columns: left to right), and increasing initial population size (rows: top to bottom)

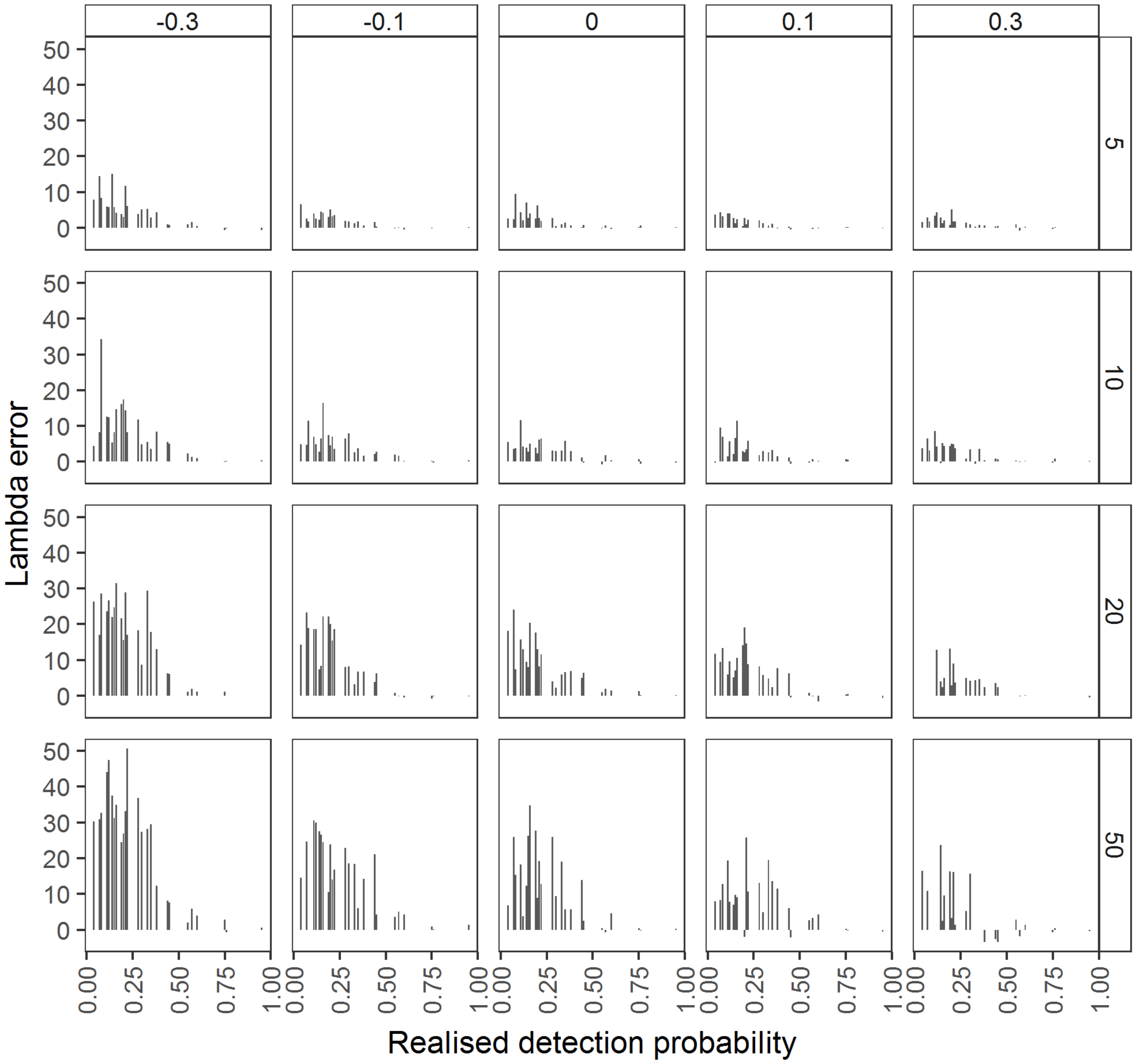


Figure SI 4. Absolute error (λ_{est}- λ_{true}) error in initial population size for simulated datasets analysed under the exponential growth model. Panels represent simulations with population growth rate (columns: left to right), and increasing initial population size (rows: top to bottom)

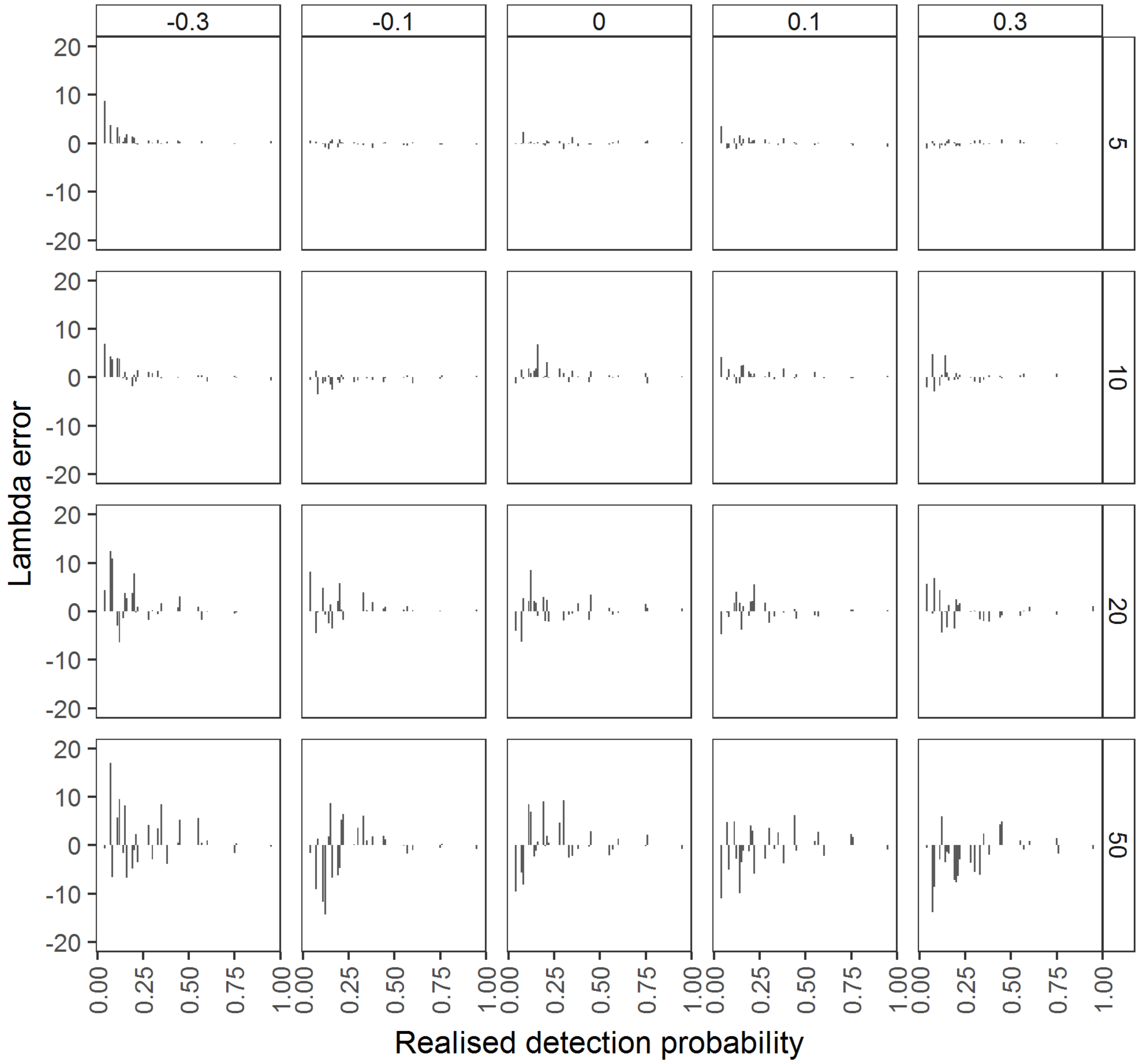


Figure SI 5. Absolute error (λ_{est}- λ_{true}) error in initial population size for simulated datasets analysed under the Dail-Madsen model. Panels represent simulations with population growth rate (columns: left to right), and increasing initial population size (rows: top to bottom)

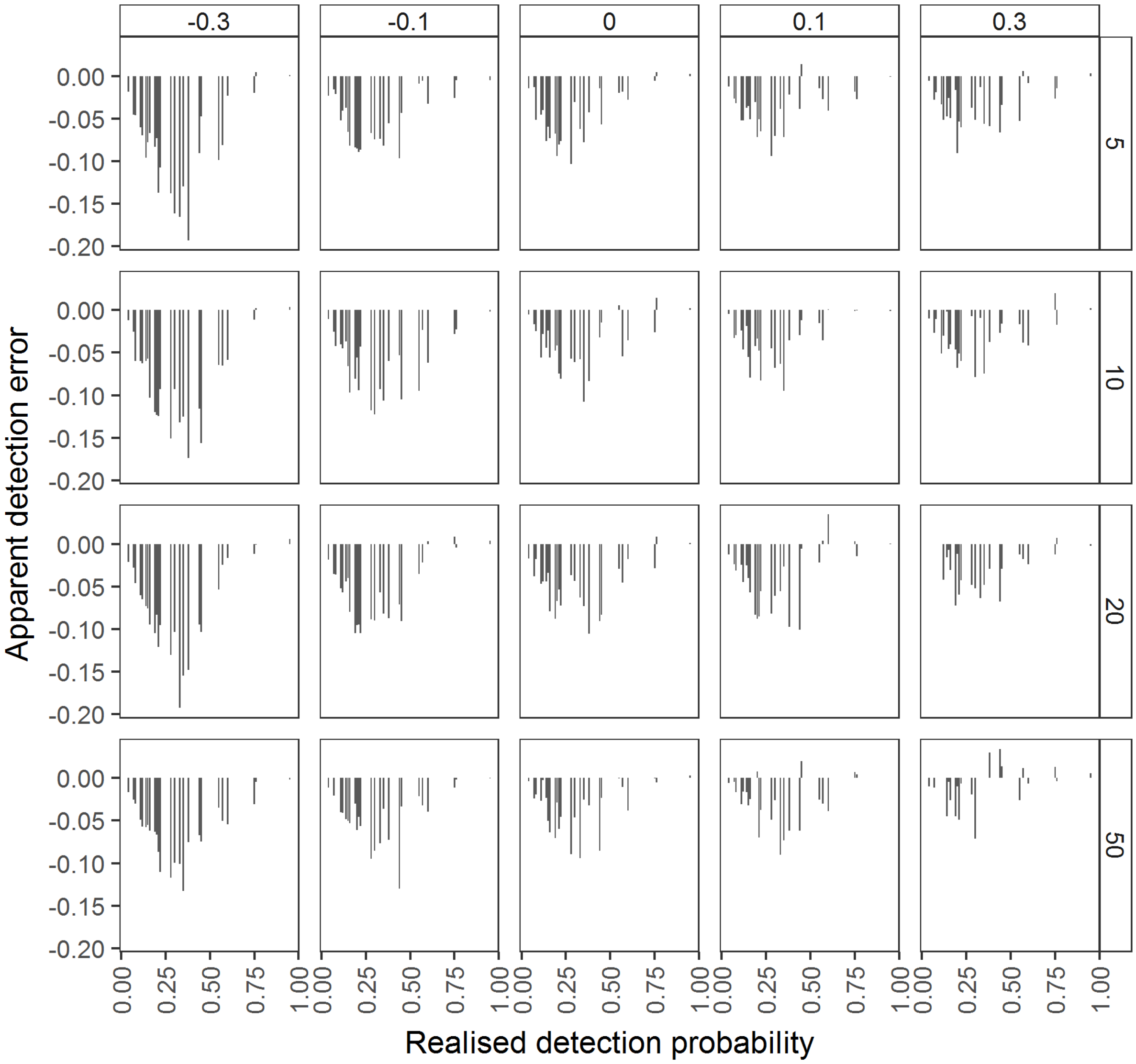


Figure SI 6. Absolute error ($\tau_{est} - \tau_{true}$) in realised detection for simulated datasets analysed under the exponential growth model. Panels represent simulations with population growth rate (columns: left to right), and increasing initial population size (rows: top to bottom)

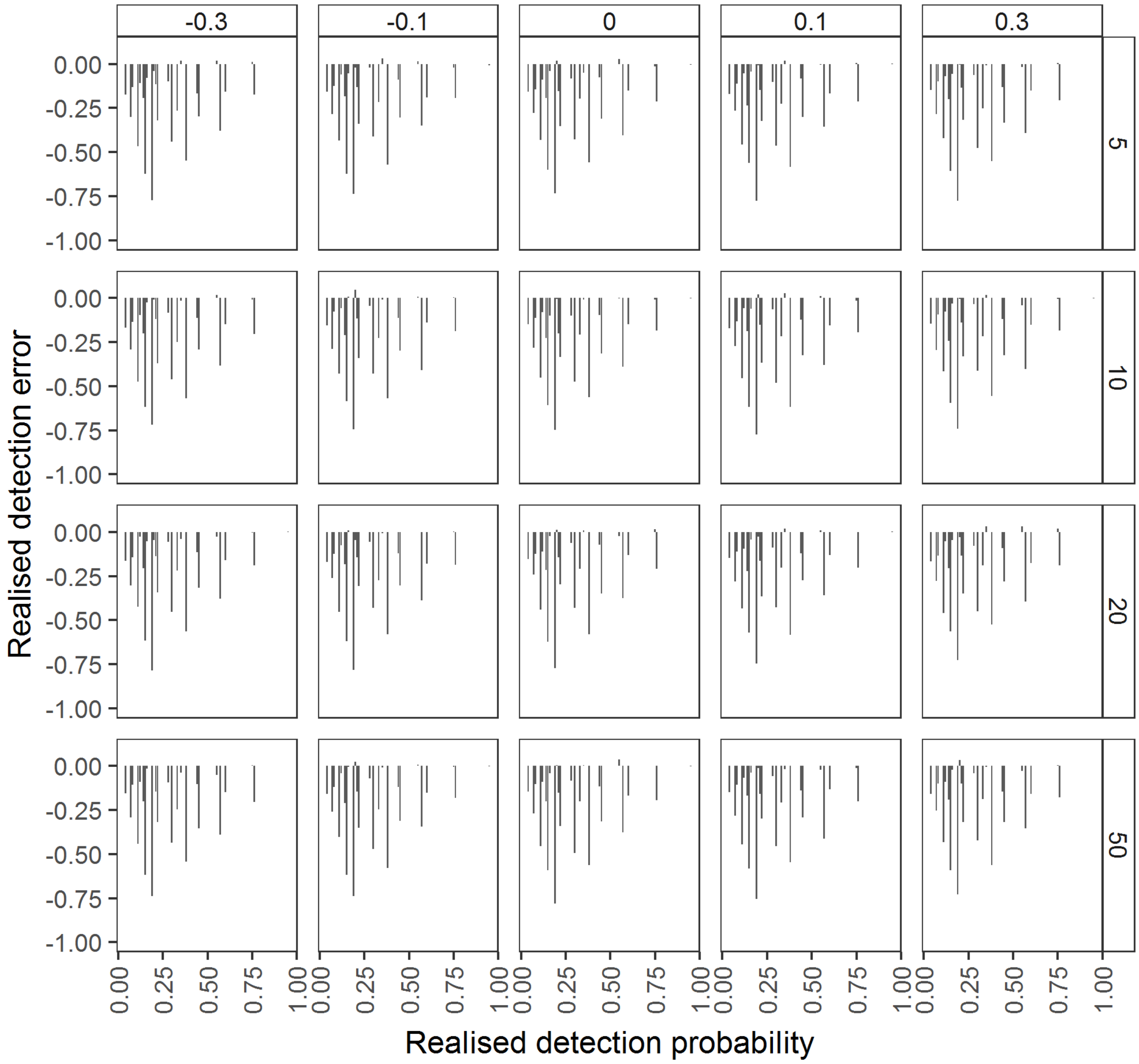


Figure SI 7. Absolute error ($\tau_{est} - \tau_{true}$) in realised detection for simulated datasets analysed under the Dail-Madsen model. Panels represent simulations with population growth rate (columns: left to right), and increasing initial population size (rows: top to bottom)

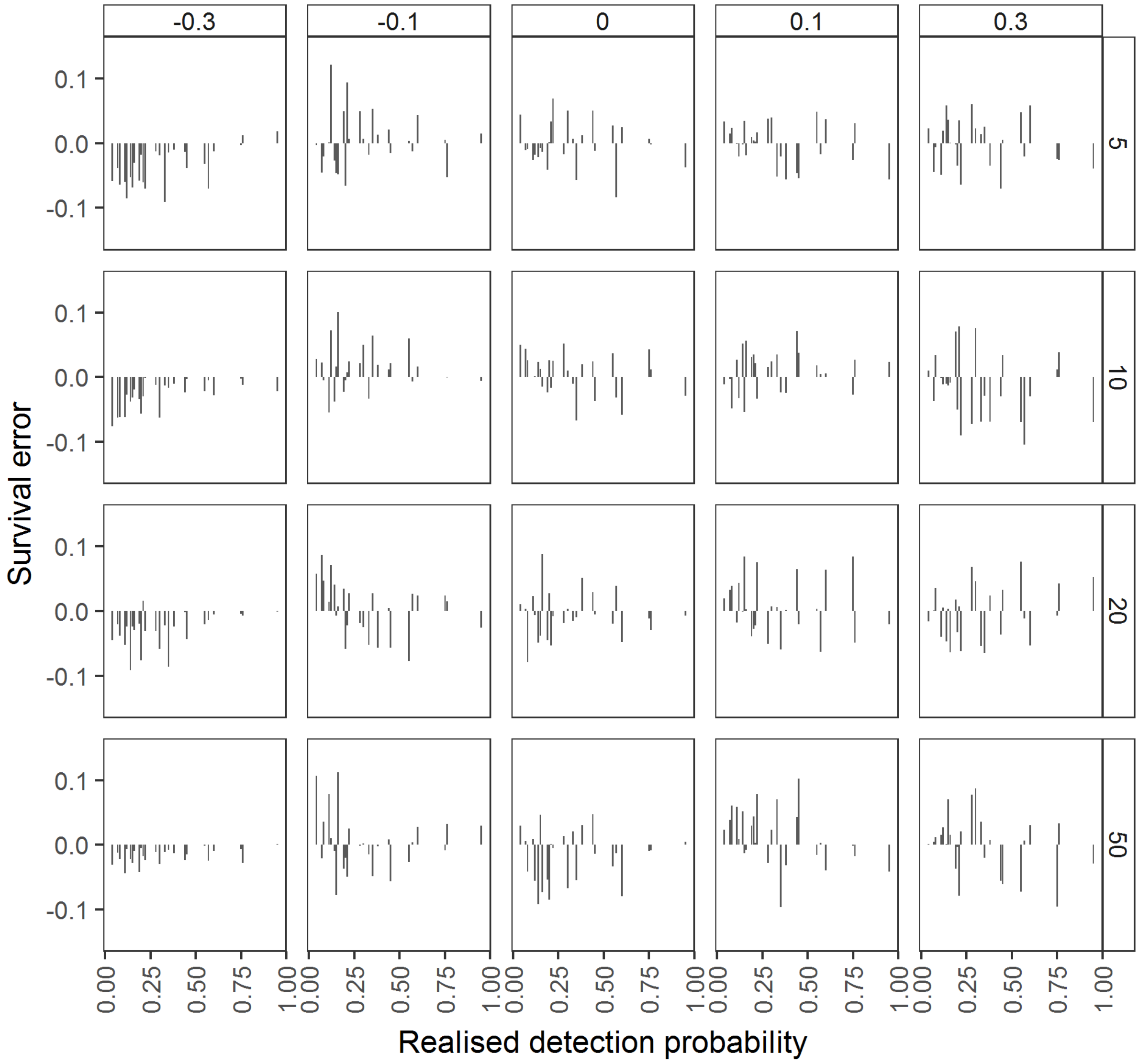


Figure SI 8. absolute error ($surv_{est}$-$surv_{true}$) error in survival probability for simulated datasets analysed under the Dail-Madsen model. Panels represent simulations with population growth rate (left to right), and increasing initial population size (top to bottom). Note that for all simulations the true survival rate was 0.7

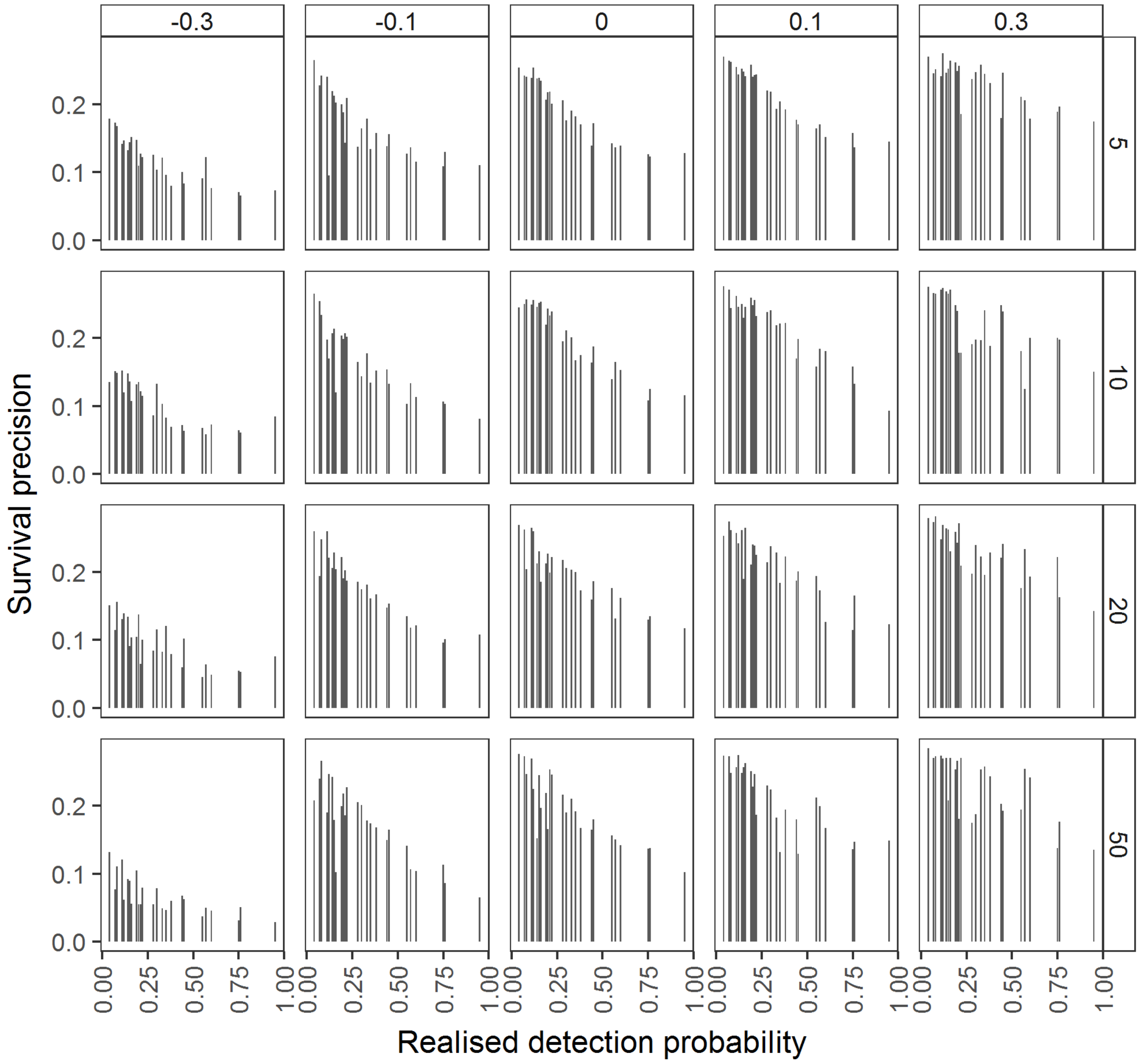


Figure SI 9. Survival probability precision (width of 95% CRI) for simulated datasets analysed under the Dail-Madsen model. Panels represent simulations with population growth rate (left to right), and increasing initial population size (top to bottom)

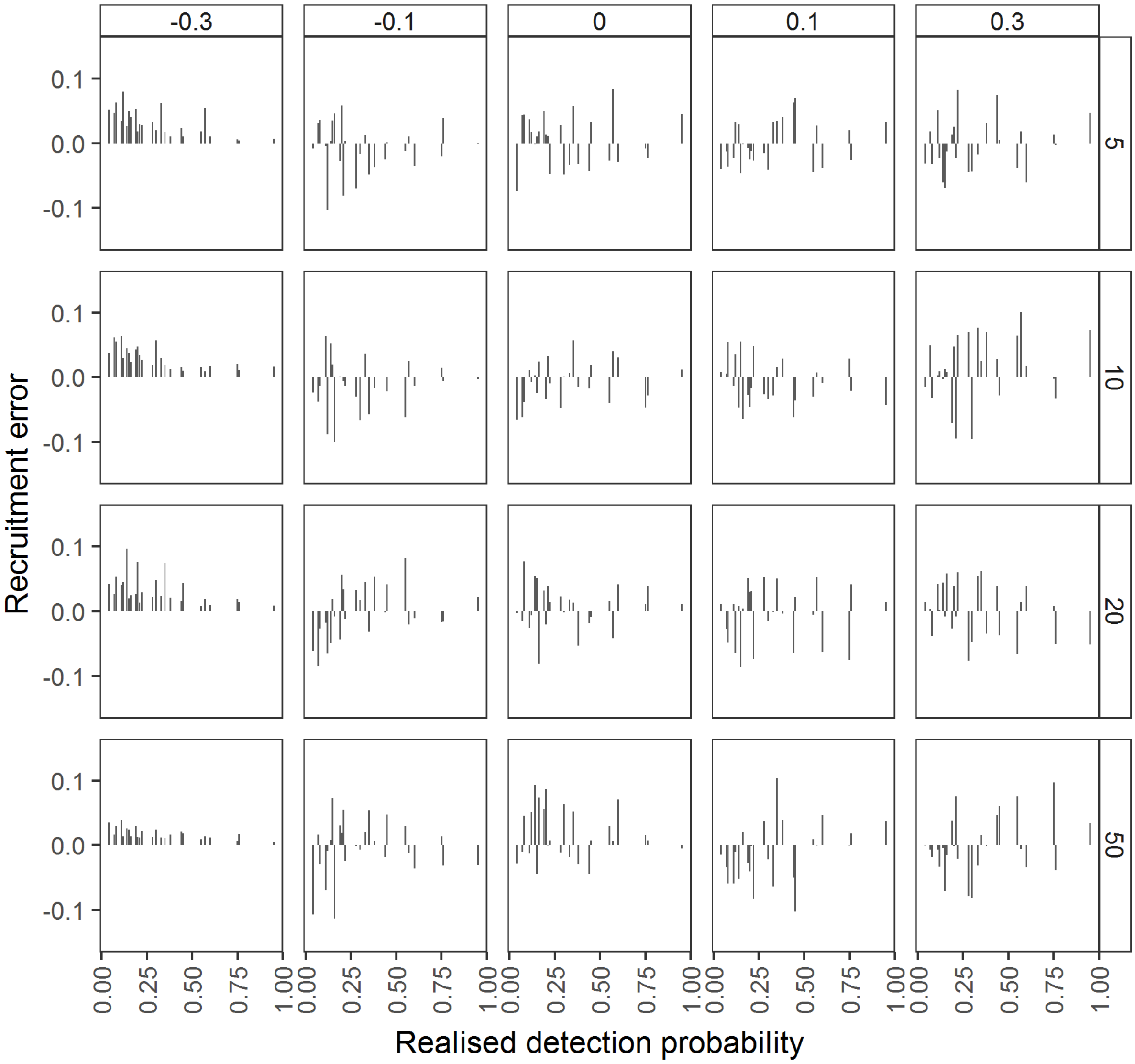


Figure SI 10. Absolute error (rec_{est}-rec_{true}) in per-capita recruitment rates for simulated datasets analysed under the Dail-Madsen model. Panels represent increasing initial population size (rows; top to bottom), and increasing population growth rate (columns; left to right). As survival rate is constant in all simulations, simulations in the left column had 0 recruitment, and recruitment increases up to the rightmost column in which the true value of per-capita recruitment is 0.6.

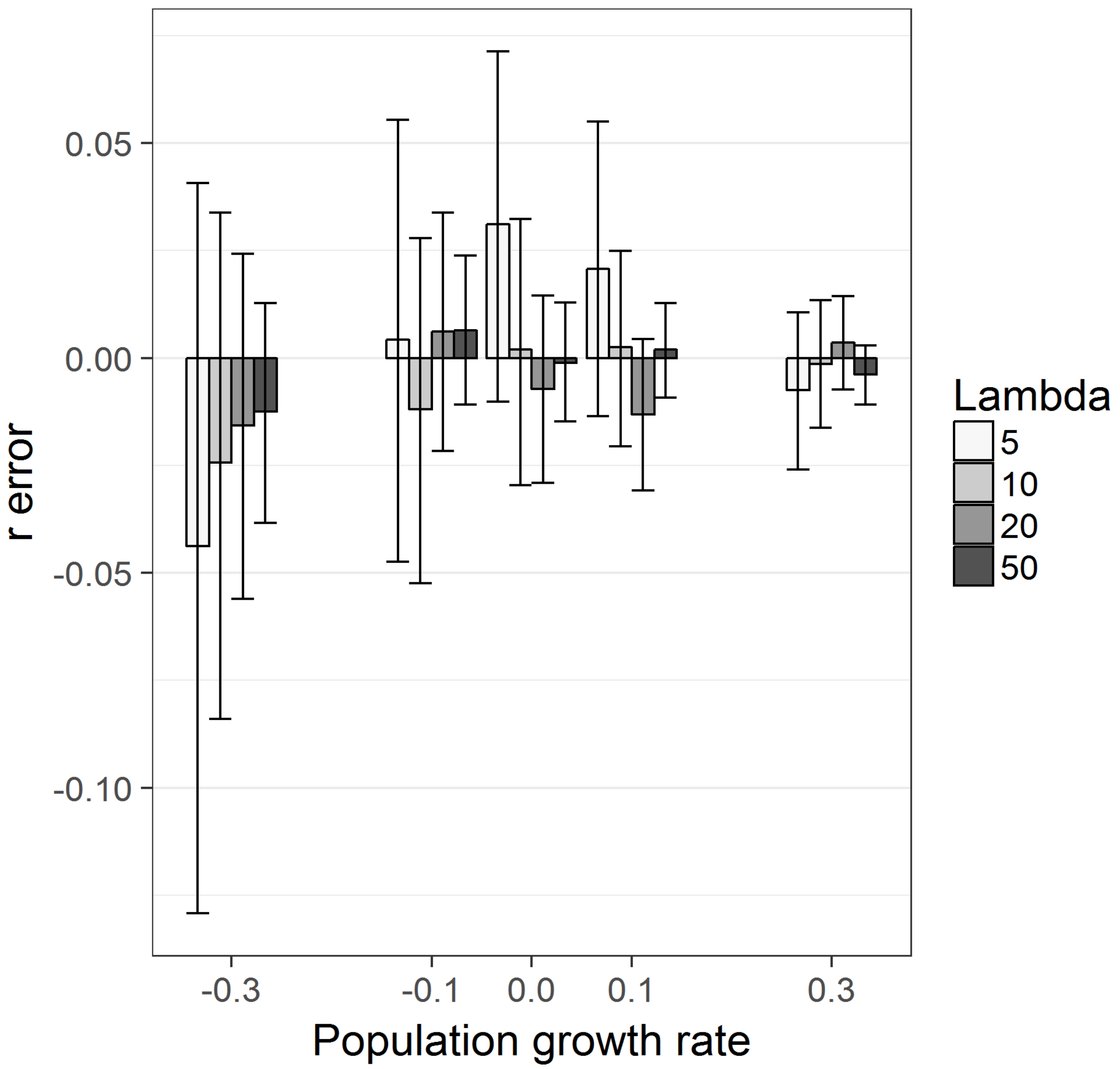


Figure SI 11. Absolute error (r_{est}-r_{true}) in the population growth rate with 95% CRI for datasets analysed using the egg count model.

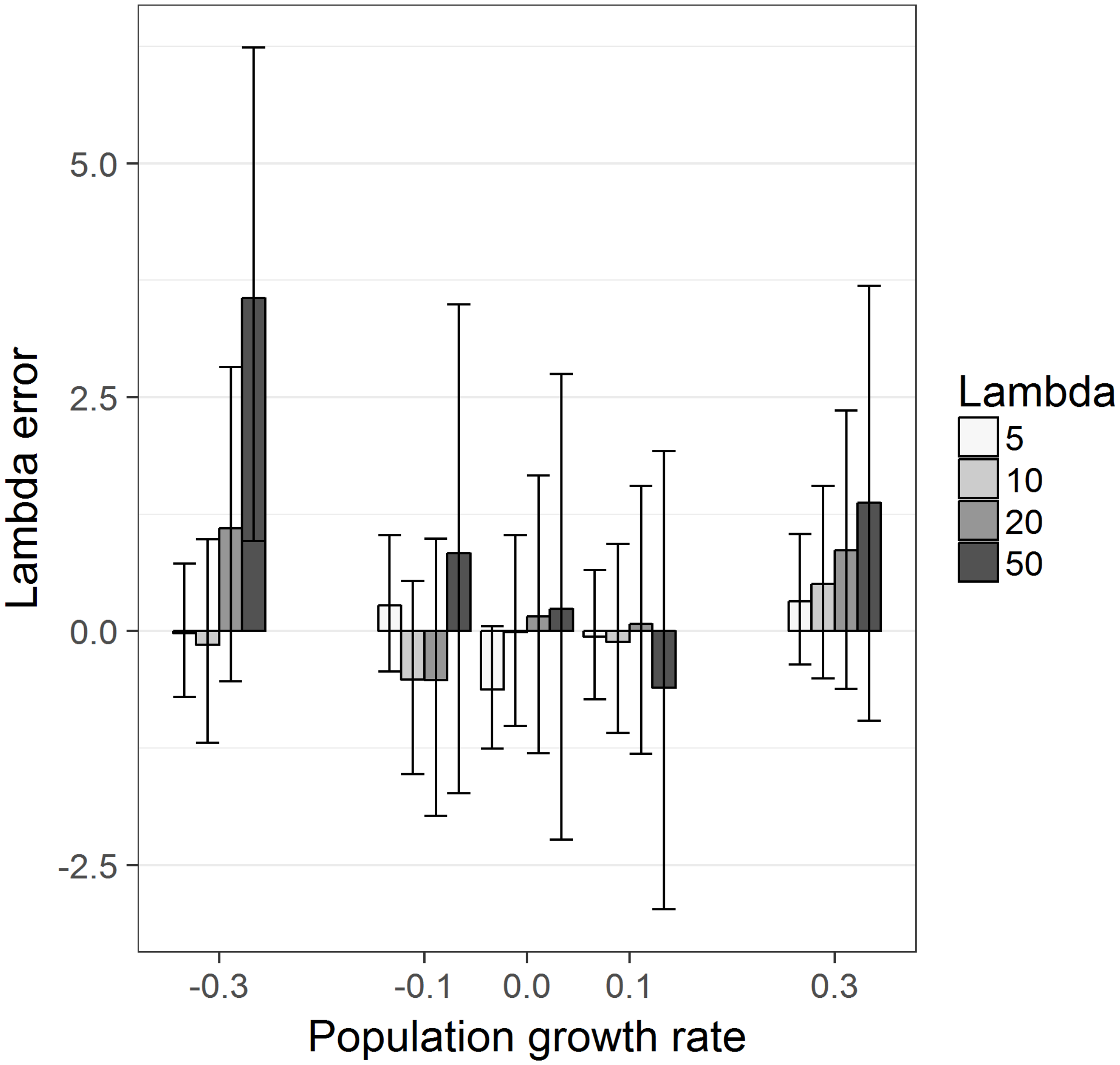


Figure SI 12. Absolute error (λ_{est}- λ_{true}) in initial population size, with 95% CRI for datasets analysed using the egg count model.

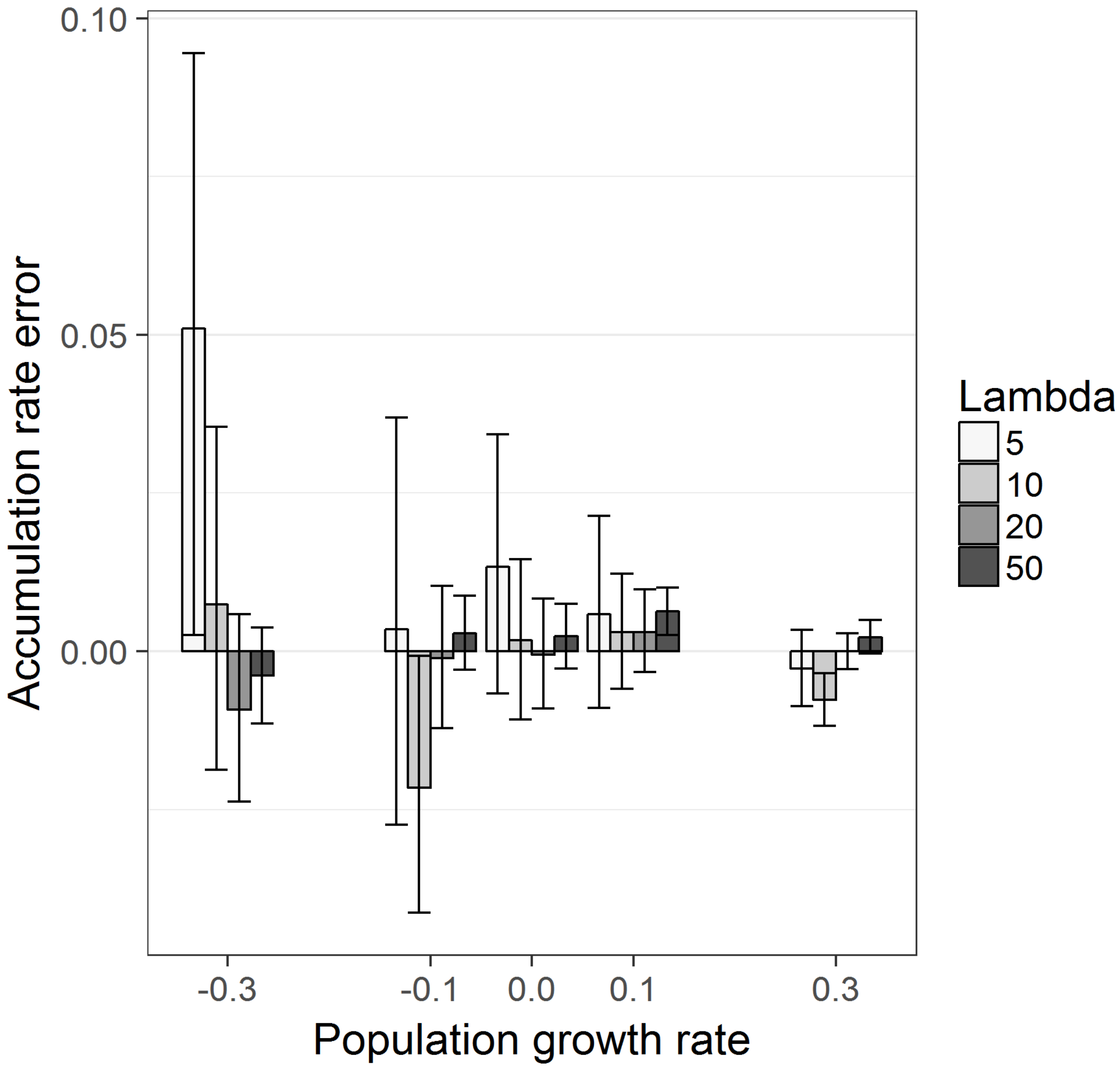


Figure SI 13. Absolute error (acc.rate$_{est}$-acc.rate$_{true}$) in egg accumulation rate and 95% CRI for datasets analysed using the egg count model.

CHAPTER SIX

Error rates and variation between observers are reduced with the use of photographic matching software for capture-recapture studies

Sam S. Cruickshank [1,*], Benedikt R. Schmidt [1,2]

Published as:

Cruickshank, S.S. and Schmidt, B.R. (2017) *Error rates and variation between observers are reduced with the use of photographic matching software for capture-recapture studies*. **Amphibia-Reptilia** 38(3): 315-325

---

[1] *Department of Evolutionary Biology and Environmental Studies, University of Zürich, Zürich, Switzerland*
[2] *Info Fauna karch, Neuchâtel, Switzerland*
*sam.cruickshank@ieu.uzh.ch

ABSTRACT

Photographic capture-mark-recapture (CMR) permits individual recognition whilst avoiding many of the concerns involved with marking animals. However, the construction of capture histories from photographs is a time-consuming process. Furthermore, matching accuracy is determined based on subjective judgements of the person carrying out the matching, which can lead to errors in the resulting datasets- particularly in long-term projects where multiple observers match images. We asked 63 volunteers to carry out two photographic-matching exercises using a database of known individuals of the yellow-bellied toad (*Bombina variegata*). From these exercises we quantified the matching accuracy of volunteers in terms of false-acceptance and false-rejection rates. Not only were error rates greatly reduced with the use of photographic-matching software, but variation in error rates among volunteers was also lowered. Furthermore, the use of matching software led to substantial increases in matching speeds and a 87% reduction in the false-rejection rate. As even small error rates have the potential to bias CMR analyses, these results suggest that computer software could substantially reduce errors in CMR datasets. The time-savings and reduction in variance among observers suggest that such methods could be particularly beneficial in long-term CMR projects where a large number of images may be matched by multiple observers.

Keywords :

computer-aided pattern recognition; individual identification; mark-recapture; misidentification; natural markings; photo-identification; Wild-ID

INTRODUCTION

Over recent decades, developments in capture-mark-recapture (CMR) methods have been incredibly influential in our understanding of the demography and dynamics of natural populations (Lebreton et al. 1992). The ability to track individuals over long time periods has been fundamental to increasing our understanding of population demography and ecology (Clutton-Brock and Sheldon 2010). However, there are welfare and ethical concerns associated with marking animals (May 2004), and tags may be lost from individuals (Grant 2008; Campbell *et al.* 2009). Furthermore, the process of marking may impact survival (Silvy et al. 2012), return rates (Parris and McCarthy 2001), movement (Schmidt and Schwarzkopf 2010), or behaviour (Murray and Fuller 2000). All of these effects have the potential to induce serious bias in parameters of interest such as survival (Hastings et al. 2008; Morrison et al. 2011) and population size estimates (Stevick and Palsbøll 2001; Yoshizaki et al. 2009).

For suitable species, many of these issues can be mitigated through the use of natural, rather than invasive artificial marks. In species that exhibit distinctive features such as unique colour patterning, these natural marks can be 'captured' through photography; images can later be compared visually to an existing database to determine individual identities and construct capture histories (Bolger et al. 2012). In relation to invasive marks such as PIT tagging or toe clipping, welfare concerns from using photography are typically reduced (Narayan et al. 2011), and it may even be possible to identify individuals without the need to capture them (e.g. Arzoumanian 2005; Morrison and Bolger 2014; Halloran et al. 2015). Furthermore, given that photographic CMR is cheaper than purchasing tags, and thus time is the major factor limiting the number of individuals that can be tracked, it is unsurprising that photographic CMR is becoming an increasingly popular method of tracking individuals.

However, the use of natural markings to identify individuals also poses some uncertainties. Traditional identifiers used in CMR studies typically allow quick, easy, and unambiguous identification of individuals upon recapture; through for example unique codes in the case of PIT tags and bird rings, or distinctive colour combinations in visible implant elastomers. In contrast, identifying recaptures through their natural markings involves much more subjectivity- a good understanding of the variation within the population is often necessary in order to determine whether two images depict the same individual. Substantial variation may exist between observers in their ability to assess whether an image represents a recapture or a new individual (Scott 1978, Sears et al. 1990, Treilibs et al. 2016). This can be an issue particularly in long-term CMR projects where the task of image matching may be performed by numerous observers, resulting in varying error rates throughout the course of the study. Characterising the frequency of such errors and ensuring that they remain at an acceptably low level should therefore be an essential first step when setting up new photographic CMR databases (Bolger et al. 2012, Sacchi et al. 2016). There is also a need to examine inter-observer variability in matching ability and implement methods that minimise these differences (Marshall and Pierce 2012), as failing to account for such variation can lead to biased inference when analysing CMR data with unmodelled heterogeneity (Link 2003, Burt et al. 2014).

Errors in assigning capture histories arise if the markings fail to meet either of two assumptions; firstly, markings must be sufficiently different to permit individual recognition, and secondly, markings must remain constant through time so as to allow re-identifications (Bolger et al. 2012). A failure to correctly identify recaptures, for example as a result of tag loss, or a failure to recognise that two images represent the same individual, leads to a false rejection. In contrast, the erroneous acceptance that two different individuals are the same, for example through a misreading of unique marker ID, or through wrongly thinking two photographs represent the same animal, leads to a false acceptance. These rates will be unique to each study species and identification method, and thus pilot studies should be carried out to assess these rates and validate that an appropriate method is being used and is able to generate unbiased capture histories (Sacchi et al. 2016).

Computer software to aid photographic matching represents a method by which error rates may be reduced. These programs typically calculate similarly scores between images and filter out implausible comparisons. As the number of comparisons required in photographic identification increases exponentially with the number of images added to a database (Arntzen et al. 2004, Sacchi et al. 2010), software can drastically reduce this substantial logistical burden (Kelly 2001). Some software fully automates the matching process (Arzoumanian 2005; Town et al. 2013), in which case error rates depend solely on the performance of the algorithms underpinning the software. However most still require human input to confirm or reject the most likely matches suggested by the algorithms (Bolger et al. 2012; Merkle and Fortin 2013; Hiby et al. 2013; Moya et al. 2015), and thus error rates represent a combination of both human and software error.

Validation studies of computer-aided matching software have to date focussed on comparisons between computer-aided matching and traditional ('by eye') matching with a single observer (Halloran et al. 2015, Morrison et al. 2016). Some studies have used multiple observers but did not quantify among-observer variability (Bolger et al 2012, Dala-Corte et al. 2016). Here, we use a standardised dataset with a large number of volunteers to assess whether variability among observers may be influenced by the uptake of computer-aided matching software. To do so, we recruited volunteers to assess whether variation in error rates among participants differed when photographic-matching was carried out with and without the use of Wild-ID, a freely available photographic-identification software. Participants carried out two matching exercises using a database containing photographs of known individuals of the yellow-bellied toad (*Bombina variegata*). We used this information to assess relationships between error rates, matching method and the time taken to complete each exercise, and compared variation in the error rates among volunteers to assess whether computer software has the potential to lead to more consistent results in long-term CMR studies.

MATERIALS AND METHODS

Field data collection

We collected test photographs of the yellow-bellied toad (*Bombina variegata*) as part of a larger long-term capture-recapture study (Hasen, 2011; Jordan, 2012; Brandt, 2015) . Adults of this species have unique ventral markings which become fixed shortly after metamorphosis and remain stable throughout adult life, making them highly suitable for photographic CMR (Vörös et al. 2007; Gollmann and Gollmann 2011; Cayuela et al. 2016). We caught individuals by hand from a wild population near Schwyz in central Switzerland (47˚00' N, 8˚36' E, approximately 450m a.s.l) during May 2014. Photographing the animals involved placing them in a petri dish with a lid containing a foam insert.

Inverting the petri dish therefore revealed the ventral pattern, which was then photographed using a 8.1 Megapixel Sony DSC-W150 camera. We illuminated the petri dish with a Olight R40 Seeker torch on low power setting fixed approximately 30cm from the petri-dish at an inclination of 60˚ so as to avoid reflectance glare from the petri-dish. We mounted both the camera and an external light source on tripods so as to standardise lighting patterns, minimise glare, and ensure a standard focal distance for all photographs (Fig. 6.1). Multiple images of each animal were taken; as individuals move around within the petri dish during this process, this resulted in several non-identical images of each individual. Handling times were approximately 20 seconds per animal, and individuals were released at their original capture location once photographs had been taken.

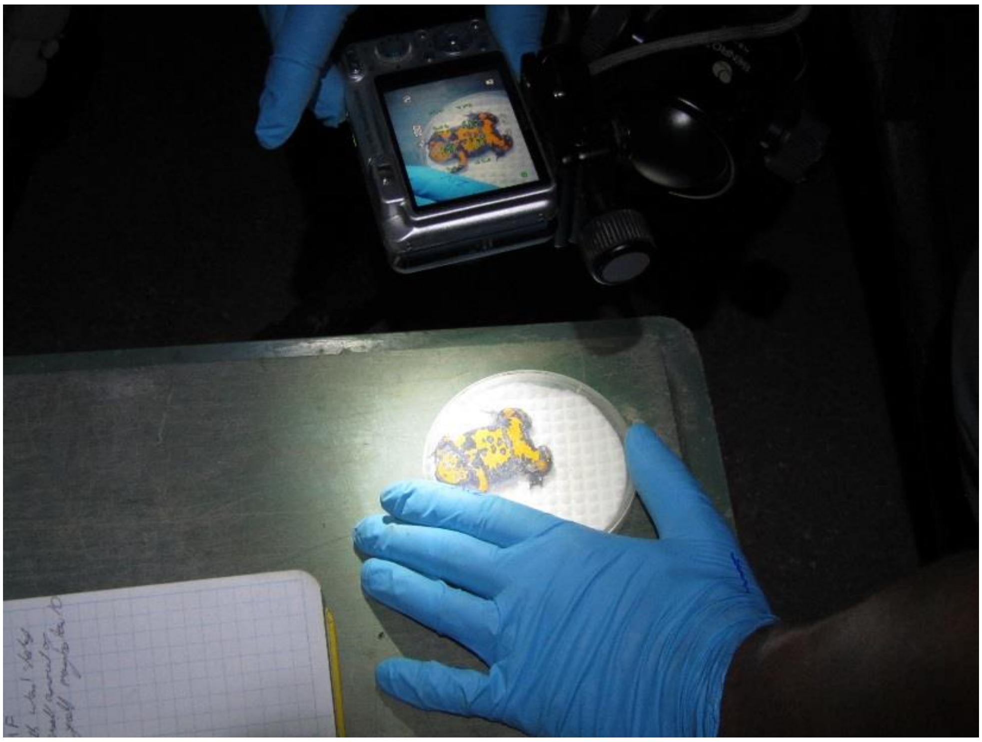


Figure 6.1. Procedure for taking photographs. Individuals were inverted in a petri dish and photographed with a tripod-mounted camera illuminated with an external light source.

Selection of photographs

To create the test dataset used for the photo-matching exercise, we selected 100 colour photographs that were considered to be of high quality (i.e. in focus and with no camera glare). We included photographs of 80 individuals; for 60 of these individuals a single image was included, and for the remaining 20 individuals two photographs from the same capture event were included, selected specifically so that both images showed the individual in a different position (i.e. the photographs did not look identical) (Fig. 6.2). All images were cropped to maximise the area of the frame filled by the ventral pattern without removing any features that could be useful for pattern matching (as recommended by Bolger et al. 2012).

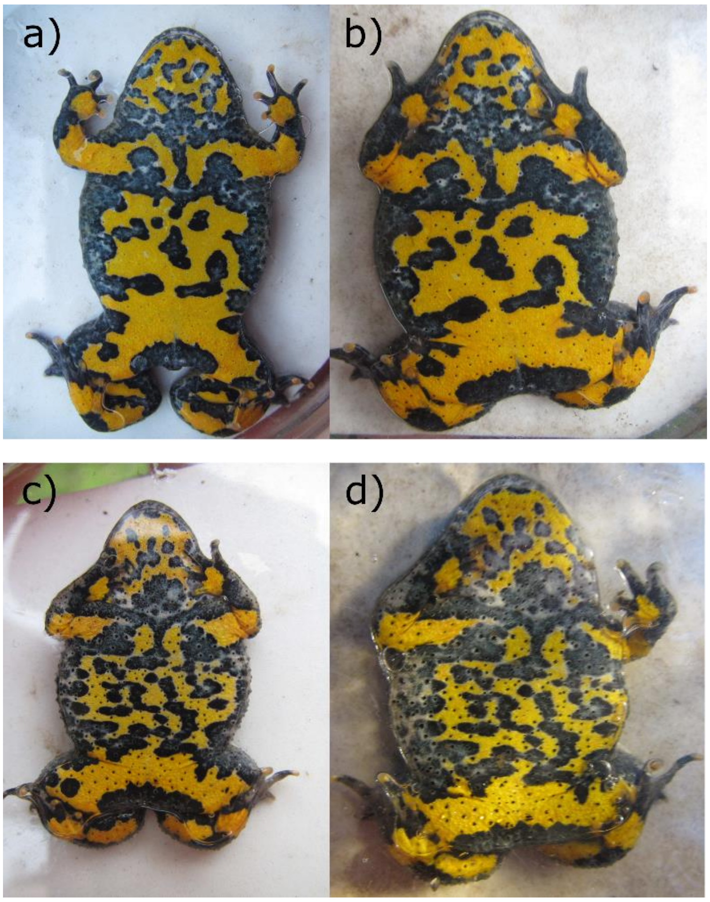


Figure 6.2. Ventral patterns from 2 Yellow-bellied toad individuals. a) and b) represent one individual, c) and d) are of a second individual

Volunteer matching exercises

In task one, 63 volunteers were asked to identify matches by eye using a paper copy of the photographic dataset. Volunteers were recruited through personal and professional networks and comprised a mix of scientists with and without experience of working with amphibians, students, and non-scientists. Participants were provided with the set of 100 8x6 cm images printed onto cards alongside unique codes. They were told that some individuals were depicted in two photographs within the dataset, and instructed to take as much time as they felt necessary in order to identify the number of matching pairs present. Volunteers reported the time they spent on the exercise, the number of matches identified, and the codes of the matching individuals. These unique codes allowed later identification of which true pairs were missed, and of any erroneous pairings.

In task two, the volunteers were asked to complete the same exercise using the Wild-ID software (Bolger et al. 2012). We chose Wild-ID as this is a freely available software designed to be applicable to any species with distinctive patterning. Furthermore, the program requires relatively little pre-processing of images prior to matching, which we consider an important quality in any software that could be applied for processing large photographic datasets. This pre-processing was not included in the time recorded by volunteers to complete the photographic matching exercise.

Participants were provided with an electronic folder containing the photographic database and a copy of the Wild-ID software (Bolger et al. 2012). This program uses a Scale Invariant Feature Transform (SIFT) algorithm (Lowe 2004) to identify distinctive features of each image. The algorithm then makes pairwise comparisons between all images, comparing these features irrespective of scale, orientation and colour, in order to produce pairwise similarity scores. Using these scores, the program presents to the user up to 20 best potential matches for each photograph in decreasing order of similarity. The user then views these top matches and confirms or rejects the suggestions. The output file from the program lists details of all pairings confirmed by the user, as well similarity scores and the rank of the matched photograph within the potential matches as measured using the SIFT algorithm.

In order to avoid our results being biased by a ‘learning’ effect- i.e. remembering patterns belonging to images with a match from the first exercise, and using this information in the second exercise, several steps were taken:

(1) The order in which participants carried out the exercises was randomly assigned;
(2) Volunteers completed their second exercise a minimum of one week after the first;
(3) Volunteers were informed that the two exercises used two different photographic databases
(4) All photographs were assigned different identification codes in each exercise.

Comparison of matching methods

From these data, we calculated the False Acceptance (FAR) and False Rejection Rates (FRR) (Jain 2007; Bendik et al. 2013) for each volunteer for both matching methods. False acceptance rates were calculated as the number of false matches made divided by the total number of pairwise comparisons between non-matching images (4930); FRR as the number of true matches missed divided by the total number of matching pairs (20). Using error rates from all volunteers for both exercises, we used GLM with these rates (FAR and FRR) as response variables to test for differences among the two different matching methods (by eye or with Wild-ID), and to assess whether lower rates resulted from participants spending more time on an exercise. As each volunteer produced two sets of error rates (one for each exercise), we also included a categorical predictor variable "order" specifying whether the error rate for a given method derived from the first or second exercise that the volunteer attempted. This allowed us to test for the 'learning effect' described above. As FRR were overdispersed, we used quasibinomial GLM in analysis of this rate.

To examine variation between observers for each exercise, we used permutation tests to compare variances in error rates between the matching methods. 100,000 unconstrained randomisations were performed in which the observed error rates were assigned randomly to one of the two matching methods. Variance ratios were calculated for each permutation and these used to create a null distribution. The null hypothesis that variances did not differ between methods could then be rejected if the observed variance ratio lay in the upper 5% of this distribution (one tailed-test, as variance ratio is constrained to be >1). Significance was assessed at the $\alpha=0.05$ level in all analyses.

Performance of Wild-ID

To evaluate the performance of the Wild-ID program, we examined all pairwise similarity scores calculated by the algorithm to assess whether matches could be confirmed on the basis of a similarity score above a certain threshold. In order to get a better idea of the ability of individuals using this software to easily identify matching images, we also examined the rankings of confirmed images from a larger photographic dataset for this species. This dataset (S. Cruickshank, unpubl. data) contained

7499 images and 2999 confirmed matches, and was constructed by three of the volunteers in this study using Wild-ID.

RESULTS

Comparison of matching methods

Participants spent significantly longer when matching photographs by eye (mean: 71.02 minutes, range: 28-200) than when using Wild-ID (mean: 47.98, range: 17-115) (GLM: $t_{72} = -4.88$, $P < 0.001$; Fig. 6.3). Time spent also varied as an interaction between the method used for matching and the order in which the tasks were completed (GLM: $t_{72} = 0.410$, $P < 0.05$). Volunteers spent longer matching photographs by eye than when using Wild-ID, and for a given method, the volunteers completed the task faster if it was the second task that they completed- i.e. there was evidence of a learning effect. Thus for the first exercise attempted, volunteers using Wild-ID completed the exercise on average 10.73 minutes faster than those matching by eye, whereas for the second exercise, this average time difference increased to 30.03 minutes.

False acceptances were rare; across 623,700 comparisons made by the volunteers in this study, only 12 resulted in false matches. This represented an overall FAR of 0.00002. False acceptance rates did not differ between the two different exercises, and were not influenced by the order in which the exercises were attempted, or with the time spent on the exercise (binomial GLM, all $P > 0.05$). Only one photograph was falsely matched to another on multiple occasions, however never to the same photograph twice. Thus, it appears there were no confusing photo pairs which were falsely matched multiple times in different exercises. Variation in FAR did not differ significantly between assessment methods (permutation test; $P = 0.489$).

Volunteers frequently failed to identify matching image pairs, predominantly in the eye-matching exercise. This led to a FRR of 0.131 ± 0.134 (mean ± SD) during the eye-matching exercise, and a significantly lower rate of 0.017 ± 0.055 (mean ± SD) with the use of Wild-ID (quasibinomial GLM: $t_{123} = -2.71$, $P < 0.01$, dispersion parameter 2.76; Fig. 6.4). Additionally, error rates were lower when more time was spent on an exercise (quasibinomial GLM: $t_{123} = -0.025$, $P < 0.01$; Fig. 6.5), however there was no effect of the order in which exercises were attempted. Variation in FRR among participants was significantly reduced when volunteers were allowed to use the Wild-ID software (permutation test, $P < 0.001$).

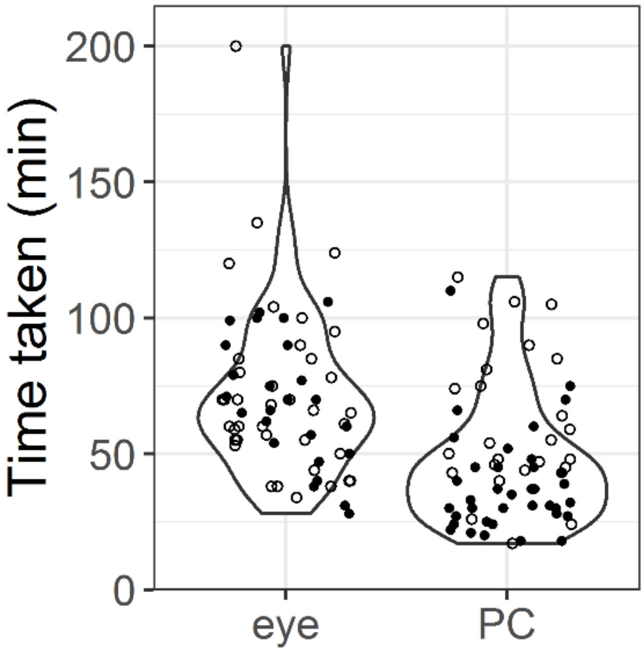


Figure 6.3. Violin plot showing the distribution of time taken by volunteers to complete the photographic matching exercise with and without the use of Wild-ID software. Open points denote that the datapoint represents the first exercise a participant completed, with closed dots representing the second exercise. The violin represents the density of the data across the range of time taken by volunteers to complete the exercises.

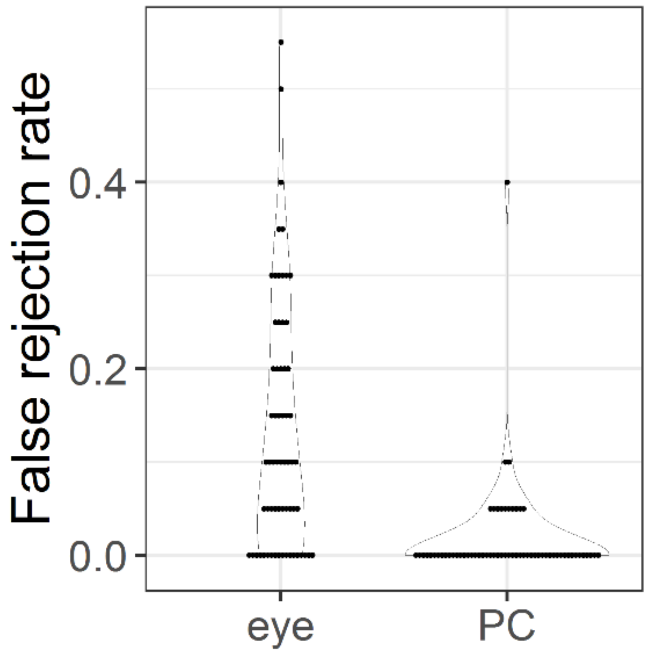


Figure 6.4. Violin plot showing false rejection rates for participants completing the photo-matching exercise by eye and with the use of Wild-ID software. The violin represents the density of the data across the range of reported false rejection rates.

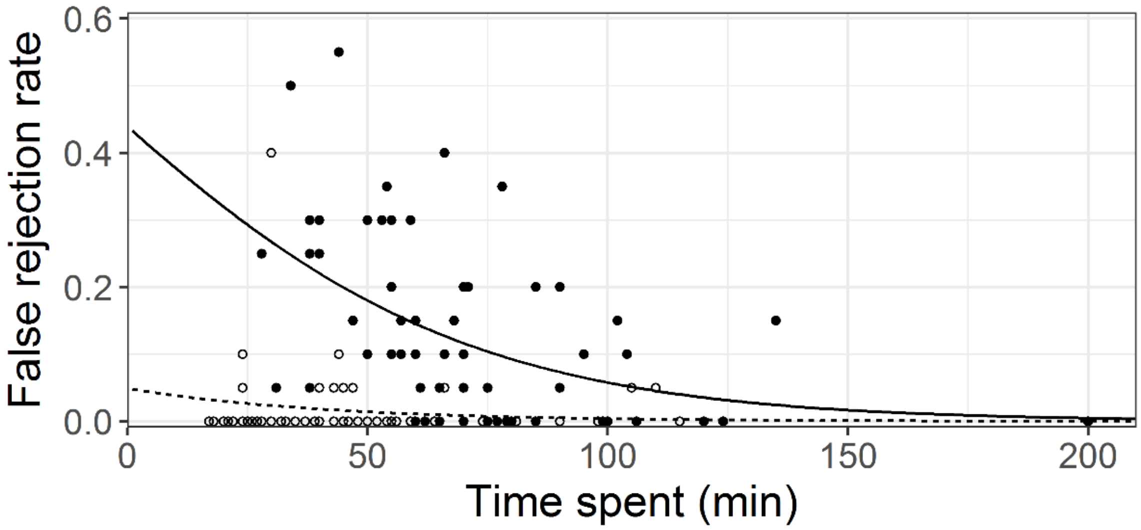


Figure 6.5. Relationships between the time spent on matching exercises and false rejection rates for the two different matching methods. Lines represent predicted rates from the FAR GLM (see results). Open points and the dotted line refer to the error rates for Wild-ID, closed points and full line relates to the task completed by eye.

Performance of Wild-ID

The Wild-ID algorithm for similarity scoring was very effective on the test dataset- for all true photograph pairs, the correct corresponding image was the first presented to the user for review. Similarity scores were often low, even for matching images; the lowest-scoring image pair had a calculated similarity score of only 3.15%. However, all pairs with a score above this were true matches, and all those below this threshold were non-matching pairs.

In our separate, larger dataset, of all the matches identified through consideration of the 20 top potential images ranked by WILD-ID, 91.6% of identified photo pairs were presented as the best candidate match. As such, if the top image was the only potential match considered, this would lead to a FRR of 0.084. As this dataset consisted of photographs collected and analysed over several years by several observers, it was not possible to be entirely certain that all true matches were identified. However, as only a negligible number of additional matches were found by consideration of lower ranked images (Fig. 6.6), yet we nonetheless always considered the top 20 ranked images, we can be confident that the majority, if not all, true matches were identified by this point. After considering the top two candidate pairs, over 95% of matching photograph pairs were correctly identified (Fig. 6.6).

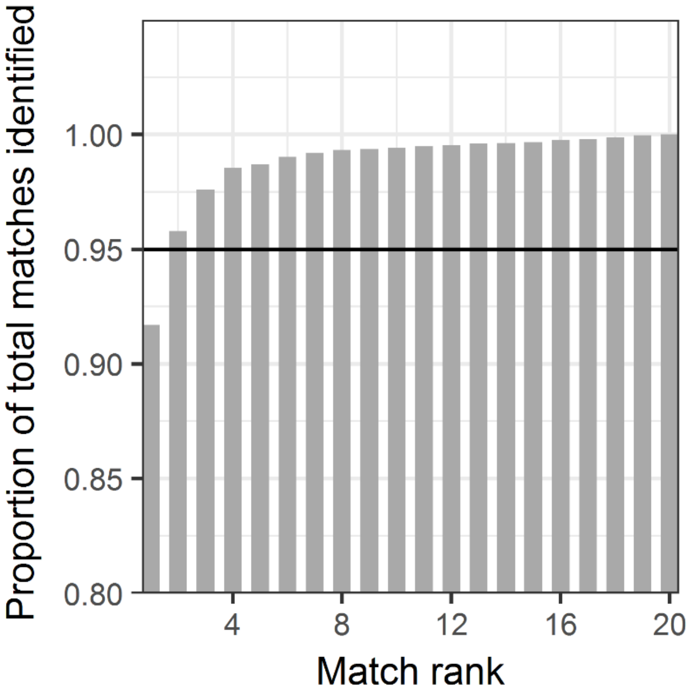


Figure 6.6. The cumulative proportion of matches found by match rankings for a large photographic dataset (7499 images, 2999 matches). The black horizontal line represents identification of 95% of true matches.

DISCUSSION

One of the main limitations of photographic CMR is the difficulty in unambiguously determining whether two photographs represent the same individual. Error rates will differ with matching methods and study species- indeed, assessing the frequency of these errors is an essential step in ensuring that photographic CMR data are reliable. Our results add to a growing body of evidence suggesting that computer-aided photographic matching can lead to substantially reduced error rates, and requires less time (Bolger et al. 2012; Treilibs et al. 2016, Dala-Corte et al. 2016). We also provide the first evidence that using such systems can also lead to reduced variation in error rates among observers, in addition to a reduction in error rates themselves. We believe that this result likely reflects a general rule for such software and thus believe that photographic identification software could be of particular benefit in minimising errors in CMR studies that involve multiple observers carrying out photographic matching.

The possibility of false rejections is considered a major disadvantage in photographic CMR (Yoshizaki 2007; Link et al. 2010), as even small false rejection rates have been shown to induce substantial bias in demographic parameter estimates (Creel et al. 2003; McDonald et al. 2003; Lukacs and Burnham 2005). The fact that we recorded high FRR when observers matched images by eyes, even in our relatively small photographic database, emphasises that this can be a real limitation for such studies. We collected these data to validate methods for a larger study involving approximately 7500 images- a size not uncommon in CMR studies. Had this larger dataset been matched exclusively by eye, our median rates suggest that 300 recaptures would have been overlooked, compared to none through the use of Wild-ID at median FRR rates.

Errors likely occur when fatigue sets in during the matching process, leading to a reduction in performance (Sears et al. 1990). By reducing the effective number of comparisons that need to be made, photo-ID software benefits users by allowing more photographs to be processed during the period in which the observer is still operating at peak concentration levels. Furthermore, as the most implausible comparisons are effectively filtered out by computer software, positive feedback (in the form of identifying correct matches), could act to increase the time a user spends matching before attention spans wane. The 87% reduction in false rejection rates with the use of Wild-ID suggest that this filtering approach can be extremely effective in ensuring high and consistent matching accuracies.

We tested for the existence of FAR as such errors lead to several capture histories being merged into one, which can severely bias survival and population size estimates (Pradel et al. 1997). Although most photographic CMR studies have not considered this issue (although see Kenyon et al. 2009; Morrison et al. 2011; Bolger et al. 2012), the extremely low rates we observed in our test exercise appear to be the norm rather than the exception. Most published estimates appear very low (0.00-0.001) irrespective of the taxa under study or software used (e.g. Sherley et al. 2010; Morrison et al. 2011; Bolger et al. 2012; Drechsler et al. 2015). Indeed, the simulations of Morrison et al (2001) suggest that even the highest reported FAR to date (0.025; Kelly 2001) is unlikely to induce sufficient bias for concern in survival analyses.

We found that variation in error rates among observers was reduced with the use of computer software. Marshall and Pierce (2012) identified differences in observer ability as a major issue in photographic mark-recapture studies, and it has long been recognised that changes in staff over the course of mark-recapture studies can lead to loss of identification experience and therefore induce errors in CMR datasets (Scott 1978). Our findings suggest that photo-matching software represents an appropriate tool to deal with these issues and reduce potential biases induced by human subjectivity. With the use of an appropriate software system, relationships between user experience and error rates (Schofield et al. 2007; Huffard et al. 2008; Waite and Mellish 2009) need not impact data quality. Our study examined only one study system, however the results of our study are in concordance with others that have examined differences in error rates through the use of photographic software, and thus we believe that it is reasonable to expect that reduced variance among observers is a pattern likely to hold in other systems. Ultimately, however, CMR projects should always carry out pilot studies to evaluate the error rates for the chosen matching method. If the study is designed to be a long-term project, it would be prudent to evaluate variation in errors among several observers as a proxy for potential turnover in observers throughout the course of the study.

Photo-matching software is designed primarily in order to ease the logistical burden that matching photographs represents. Increased matching speeds have been reported in other comparisons between manual and computer-aided matching (Elgue et al. 2014; Halloran et al. 2015; Dala-Corte et al. 2016); indeed, the improvement in matching speed that we report here represents one of the more modest time savings reported. The fact that this time reduction comes associated with a reduction in error rates is likely a consequence of the effectiveness of the software algorithm in accurately comparing photographs. Given that one of the major costs of photographic CMR is associated with the time required to complete matching, this finding is valuable and may mean that photographic approaches can now be applied to studies where this logistical burden would otherwise have been too great.

The use of Wild-ID software in this study system appears highly successful. True matches were always presented as the most likely candidate in our test dataset, and examination of our much larger dataset suggested that over 95% of true matches would be identified if only the top two best candidate matches were to be considered. The similarity scores for all true matches were above a given threshold, and all non-matches below this threshold, suggesting the possibility of discriminating matches solely based on similarity scores without the need to visually confirm images. We take these findings as confirmation that the pattern matching algorithm underlying Wild-ID is highly accurate for our study species.

However, the potential benefits of computer software to any photographic study rely on several conditions being met. Of primary importance is that the species of interest has sufficiently unique markings that remain constant through time, and are thus suitable for photographic CMR. This condition will preclude many species from being studied using photographic CMR, leaving no choice but to utilise artificial invasive marks. Secondly, it is essential that the photographs taken are of high quality; the ability of software to reliably compare images is greatly reduced when photo quality is poor (Kelly 2001), and even simple efforts to standardise photographs greatly improve performance (Bendik et al. 2013). Finally, different matching algorithms underpin different programs, thus careful selection of a software that is able to register and compare the features of interest in a given species is key. Although the generic pattern-matching software we apply here has been shown to be appropriate in many different taxa (e.g. Cross et al. 2014; Halloran et al. 2015; Dala-Corte et al. 2016), there are situations where more time-intensive programs (e.g. Gamble et al. 2007; Hiby et al. 2013) may be more appropriate.

The application of photographic-matching software to mark-recapture studies can lead to a number of direct benefits in terms of dataset accuracy and the time savings. The use of software can lead to reductions not only in the absolute error rates when matching, but just as importantly can also reduce variability in matching abilities among observers. These findings suggest that photographic CMR software can mitigate some of the potential downsides to using CMR with natural markings (Arntzen et al. 2004), and thus be of benefit to many CMR applications. We highlight that photo-matching software could be particularly beneficial when applied in long-term studies where the time required to match images would otherwise be prohibitively expensive. As long-term studies often involve several observers completing image matching over the course of a study, the use of computer software would furthermore benefit such projects by decreasing the influence of individual subjectivity, which in turn would lead to increased accuracy in the resulting datasets. Ultimately, however, the utility of any photographic CMR system should be tested using pilot studies to quantify the major possible error rates and evaluate the suitability of different methods for the project.

ACKNOWLEDGEMENTS

We thank all the volunteers who gave their time to participate in the photograph matching exercises, as well as to H. Brandt for assisting with the capture and photographing of the toads in this study, and A. Jansen van Rensburg for providing constructive feedback on an earlier draft of this manuscript. The Laboratorium der Urkantone approved the use and treatment of the toads in the study (Bewilligung Nr. 12-009). The project was funded through a PhD scholarship from the Swiss Federal Office of the Environment (Forschungsvertrag 13.007PJ/M214-0617).

Literature Cited

Arntzen, J.W., Goudie, I., Halley, J. & Jehle, R. (2004). Cost comparison of marking techniques in long-term population studies: PIT-tags versus pattern maps. *Amphibia-Reptilia*, 305–315.

Arzoumanian, Z. (2005). An astronomical pattern-matching algorithm for computer-aided identification of whale sharks Rhincodon typus. *Journal of Applied Ecology*, 999–1011.

Bendik, N.F., Morrison, T.A., Gluesenkamp, A.G., Sanders, M.S. & O'Donnell, L.J. (2013). Computer-assisted photo identification outperforms visible implant elastomers in an endangered salamander, *Eurycea tonkawae*. *PloS one*, **8**, e59424.

Bolger, D.T., Morrison, T.A., Vance, B., Lee, D. & Farid, H. (2012). A computer-assisted system for photographic mark-recapture analysis. *Methods in Ecology and Evolution*, **3**, 813–822.

Brandt, H. (2015): Dispersal in a Metapopulation of the Yellow-Bellied Toad : Does Conservation Action Work? MSc Thesis. University of Zürich, Zürich

Burt, M.L., Borchers, D.L., Jenkins, K.J. & Marques, T.A. (2014). Using mark-recapture distance sampling methods on line transect surveys. *Methods in Ecology and Evolution*, **5**, 1180–1191.

Campbell, T.S., Irvin, P., Campbell, K.R., Hoffmann, K., Dykes, M.E., Harding, A.J. & Johnson, S.A. (2009). Evaluation of a new technique for marking anurans. *Applied Herpetology*, **6**: 247–256.

Cayuela, H., Arsovski, D., Thirion, J.-M., Bonnaire, E., Pichenot, J., Boitaud, S., Brison, A.-L., Miaud, C., Joly, P., Besnard, A. (2016): Contrasting patterns of environmental fluctuation contribute to divergent life histories among amphibian populations. *Ecology*, **97**, 980–991.

Clutton-Brock, T. & Sheldon, B.C. (2010). Individuals and populations: the role of long-term, individual-based studies of animals in ecology and evolutionary biology. *Trends in Ecology and Evolution*, **25**, 562–73.

Creel, S., Spong, G., Sands, J.L., Rotella, J., Zeigle, J., Joe, L., Murphy, K.M. & Smith, D. (2003). Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. *Molecular Ecology*, **12**, 2003–2009.

Cross, M.D., Lipps, G.J., Sapak, J.M., Tobin, E.J. & Root, K. V. (2014). Pattern-recognition software as a supplemental method of identifying individual eastern box turtles (*Terrapene c. carolina*). *Herpetological Review*, **45**, 584–586.

Dala-Corte, R.B., Moschetta, J.B. & Becker, F.G. (2016). Photo-identification as a technique for recognition of individual fish: a test with the freshwater armored catfish Rineloricaria aequalicuspis Reis & Cardoso, 2001 (Siluriformes: Loricariidae). *Neotropical Ichthyology*, **14**, e150074.

Drechsler, A., Helling, T., Steinfartz, S. (2015): Genetic fingerprinting proves cross-correlated automatic photo-identification of individuals as highly efficient in large capture-mark-recapture studies. *Ecology and Evolution*. **5**, 141–151.

Elgue, E., Pereira, G., Achaval-coppes, F. & Maneyro, R. (2014). Validity of photo-identification technique to analyze natural markings in *Melanophryniscus montevidensis* (Anura : Bufonidae). *Phyllomedusa*, **13**, 59–66.

Gamble, L., Ravela, S. & McGarigal, K. (2007). Multi-scale features for identifying individuals in large biological databases: an application of pattern recognition technology to the marbled salamander Ambystoma opacum. *Journal of Applied Ecology*, **45**, 170–180.

Gollmann, G. & Gollmann, B. (2011). Ontogenetic change of colour pattern in Bombina variegata: implications for individual identification. *Herpetology Notes*, **4**, 333–335.

Grant, E.H.C. (2008). Visual Implant Elastomer Mark Retention through Metamorphosis in Amphibian Larvae. *The Journal of Wildlife Management*, **72**, 1247–1252.

Halloran, K.M., Murdoch, J.D. & Becker, M.S. (2015). Applying computer-aided photo-identification to messy datasets: a case study of Thornicroft's giraffe (*Giraffa camelopardalis thornicrofti*). *African Journal of Ecology*, **53**, 147–155.

Hasen, D. (2011): Bestandsuntersuchung von Gelbbauchunken-Populationen im Talkessel Schwyz-Ingenbohl mittels Impressum. BSc Thesis. Zurcher Hochschule Fur Angewandte Wissenschaften, Zurich

Hastings, K., Hiby, L. & Small, R. (2008). Evaluation of a computer-assisted photograph-matching system to monitor naturally marked harbor seals at Tugidak Island, Alaska. *Journal of Mammalogy*, **89**, 1201–1211.

Hiby, L., Paterson, W.D., Redman, P., Watkins, J., Twiss, S.D. & Pomeroy, P. (2013). Analysis of photo-id data allowing for missed matches and individuals identified from opposite. *Methods in Ecology and Evolution*, **4**, 252–259.

Huffard, C.L., Caldwell, R.L., DeLoach, N., Gentry, D.W., Humann, P., MacDonald, B., Moore, B., Ross, R., Uno, T. & Wong, S. (2008). Individually unique body color patterns in octopus (*Wunderpus photogenicus*) allow for photoidentification. *PLoS ONE*, **3**, e3732.

Jain, A. (2007). Technology: biometric recognition. *Nature*, **449**, 38–40.

Jordan, M. (2012): Untersuchung von Populationsgrösse und Dispersionsverhalten der Gelbbauchunken-Population im Talkessel Schwyz-Ingenbohl. BSc Thesis. Zurcher Hochschule Fur Angewandte Wissenschaften, Zurich

Kelly, M.J. (2001). Computer-aided photograph matching in studies using individual identification: an example from Serengeti cheetahs. *Journal of Mammalogy*, **82**, 440–449.

Kenyon, N., Phillott, A. & Alford, R.A. (2009). Evaluation of the photographic identification method (PIM) as a tool to identify adult Litoria genimaculata (Anura: Hylidae). *Herpetological Conservation and Biology*, **4**, 403–410.

Lebreton, J., Burnham, K.P., Clobert, J., Anderson, D.R., David, R., Monographs, S.E. & Mar, N. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs*, **62**, 67–118.

Link, W.A. (2003): Nonidentifiability of Population Size from Capture-Recapture Data with Heterogeneous Detection Probabilities. *Biometrics* , **59**, 1123–1130.

Link, W., Yoshizaki, J., Bailey, L.L. & Pollock, K.H. (2010). Uncovering a latent multinomial: analysis of mark–recapture data with misidentification. *Biometrics*, 178–185.

Lowe, D.G. (2004). Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision*, **60**, 91–110.

Lukacs, P.M. & Burnham, K.P. (2005). Review of capture-recapture methods applicable to noninvasive genetic sampling. *Molecular ecology*, **14**, 3909–19.

Marshall, A. & Pierce, S. (2012). The use and abuse of photographic identification in sharks and rays. *Journal of fish biology*, 1361–1379.

May, R.M. (2004). Ethics and Amphibians. *Nature*, **431**, 403.

McDonald, T.L., Amstrup, S.C. & Manly, B.F.J. (2003). Tag loss can bias Jolly-Seber capture-recapture estimates. *Wildlife Society Bulletin*, **31**, 814–822.

Merkle, J.A. & Fortin, D. (2013). Likelihood-based photograph identification: Application with photographs of free-ranging bison. *Wildlife Society Bulletin*, **38**, 196–204.

Morrison, T.A. & Bolger, D.T. (2014). Connectivity and bottlenecks in a migratory wildebeest Connochaetes taurinus population. *Oryx*, **48**, 1–9.

Morrison, T.A., Yoshizaki, J., Nichols, J.D. & Bolger, D.T. (2011). Estimating survival in photographic capture-recapture studies: overcoming misidentification error. *Methods in Ecology and Evolution*, **2**, 454–463.

Moya, Ó., Mansilla, P.-L., Madrazo, S., Igual, J.-M., Rotger, A., Romano, A. & Tavecchia, G. (2015). APHIS: A new software for photo-matching in ecological studies. *Ecological Informatics*, **27**, 64–70.

Murray, D.L. & Fuller, M.R. (2000). A Critical Review of the Effects of Marking on the Biology of Vertebrates. *Research Techniques in Animal Ecology: Controversies and Consequences* (eds L. Boitani & T.K. Fuller), pp. 15–64. Colombia University Press, New York.

Narayan, E.J., Molinia, F.C., Kindermann, C., Cockrem, J.F. & Hero, J.M. (2011). Urinary corticosterone responses to capture and toe-clipping in the cane toad (Rhinella marina) indicate that toe-clipping is a stressor for amphibians. *General and Comparative Endocrinology*, **174**, 238–245.

Parris, K.M., McCarthy, M.A. (2001): Identifying effects of toe clipping on anuran return rates: the importance of statistical power. *Amphibia-Reptilia.* **22**, 275–289.

Pradel, R., Hines, J.E., Lebreton, J.-D. & Nichols, J.D. (1997). Capture-Recapture Survival Models Taking Account of Transients. *Biometrics*, **53**, 60–72.

Sacchi, R., Scali, S., Pellitteri-Rosa, D. (2010): Photographic identification in reptiles: a matter of scales. *Amphibia-Reptilia* **31**: 489–502.

Sacchi, R., Scali, S., Mangiacotti, M., Sannolo, M., Zuffi, M.A.L. (2016): Digital identification and analysis. In: Reptile Ecology and Conservation. A Handbook of Techniques, p. 59–72. Dodd, C.K., Ed., Oxford University Press

Schmidt, K. & Schwarzkopf, L. (2010). Visible implant elastomer tagging and toe-clipping: Effects of marking on locomotor performance of frogs and skinks. *Herpetological Journal,* **20**, 99–105.

Sherley, R.B., Burghardt, T., Barham, P.J., Campbell, N., Cuthill, I.C. (2010): Spotting the difference: Towards fully-automated population monitoring of African penguins Spheniscus demersus. Endangered Species Research. **11**: 101–111.

Schofield, M.R., Barker, R.J. & MacKenzie, D.I. (2007). Flexible hierarchical mark-recapture modeling for open populations using WinBUGS. *Environmental and Ecological Statistics*, **16**, 369–387.

Scott, D.K. (1978). Identification of individual Bewick's swans by bill patterns. *Animal marking: recognition marking of animals in research* (ed B. Stonehouse), pp. 160–168. Macmillan, London.

Sears, R., Williamson, J.M., Wenzel, F.W., Berube, M., Gendron, D. & Jones, P. (1990). Photographic Identification of the Blue Whale (Balaenoptera musculus) in the Gulf of St. Lawrence, Canada. *Report of the International Whaling Commission*, **Special Is**, 335–342.

Silvy, N.J., Lopez, R.R. & Peterson, M.J. (2012). Techniques for marking wildlife. *The wildlife techniques manual.* (ed N.J. Silvy), pp. 230–257. Johns Hopkins University Press.

Stevick, P. & Palsbøll, P. (2001). Errors in identification using natural markings: rates, sources, and effects on capture recapture estimates of abundance. *Canadian Journal of Fisheries and Aquatic Science*, **58**, 1861–1870.

Town, C., Marshall, A. & Sethasathien, N. (2013). Manta Matcher: automated photographic identification of manta rays using keypoint features. *Ecology and Evolution,* **3**, 1902–1914.

Treilibs, C.E., Pavey, C.R., Hutchinson, M.N. & Bull, C.M. (2016). Photographic identification of individuals of a free-ranging, small terrestrial vertebrate. *Ecology and Evolution*, **6**, 800–809.

Vörös, J., Szalay, F., Barabás, L. (2007): A new method for quantitative pattern analysis applied to two European Bombina species. *Herpetological. Journal*. **17**, 97–103.

Waite, J.N. & Mellish, J.A.E. (2009). Inter- and intra-researcher variation in measurement of morphometrics on Steller sea lions (Eumetopias jubatus). *Polar Biology*, **32**, 1221–1225.

Yoshizaki, J. (2007). *Use of natural tags in closed population capture-recapture studies: modeling misidentification*. PhD dissertation. North Carolina State University.

Yoshizaki, J., Pollock, K.H., Brownie, C. & Webster, R. (2009). Modeling misidentification errors in capture-recapture studies using photographic identification of evolving marks. *Ecology,* **90**, 3–9.

CHAPTER SEVEN

General Discussion

CHAPTER SEVEN

GENERAL DISCUSSION

In this thesis I set out the explore some of the challenges faced when analysing species monitoring and distribution data. More specifically, I examined the extent to which various forms of observation error may lead to biased inference, and presented methods by which the influence of such errors can be addressed during data analysis. I focussed on the analysis of occupancy data to describe species distributions, count data to assess population sizes, and errors in identifying individuals when constructing capture histories for demographic analyses. Data collection in each of these cases present different pathways by which measurement error may be introduced, however common to all is the issue of imperfect detection. As individuals, populations or species are never detected with 100% certainty, imperfect detection causes false-negative records in all forms of monitoring data. Accounting for measurement errors such as imperfect detection typically requires that repeat measurements are made of the parameter of interest. The differences between these repeat measurements can then be used to estimate error rates, and thereby adjust accordingly to produce unbiased estimates.

Although the use of occupancy modelling and other statistical adjustments that account for detectability have grown strongly over recent decades, there has been some backlash against these practises, which one author has notably termed “statistical machismo” (McGill 2012). These arguments take two main forms: a) that if detectability is important, we must throw away old datasets which do not contain sufficient information to estimate detection rates, which would then hamper both conservation and research, and b) that the extra data that must be collected means that fewer sites can be monitored than if detectability is ignored (Banks-Leite *et al.* 2014). Both concerns are valid, yet we do not feel that they are sufficiently strong arguments to ignore detection probabilities. Firstly, although extra data is needed to estimate detection, if surveys are carried out by multiple observers then getting each observer to record data independently will generate information allowing detection to be considered with little to no extra investment (Nichols *et al.* 2000). Alternatively, spatial rather than temporal replication can provide the same data without the need for a site to be visited on multiple occasions (Srivathsa *et al.* 2017). Although the value of old data is indeed reduced when detectability cannot be assumed, such data are not useless. As we highlighted in chapter two, the problem with such data is that we cannot reliably establish locations where a species was not present in the past. As such, we can only estimate species trends by revisiting sites which were known to be occupied in the past. Any resulting measure of change can only account for the loss of species from sites, and not also for colonisation of new areas. That is, turnover in a metapopulation cannot be distinguished from a decline. Therefore, in revisitation studies, what we estimate is the worst-case scenario; the maximum possible extent of species decline. However, we demonstrated that if we fail to account for detection probabilities in contemporary data, then declines are overestimated even more strongly. For nearly half of the species we examined, ignoring imperfect detection would result in population declines being overestimated to the extent that the species would be classified within a higher extinction threat category under IUCN Red-List guidelines. As the Red List is commonly used to prioritise species for conservation management, this simple example represents a strong counter-argument to claims that accounting for detection is not worthwhile. The increasing tendency for observation databases to record non-detections in addition to species observations is a very promising sign as it means that future studies will be able to account for detection probabilities at multiple points in time and thus be able to estimate true rates of population change.

Recording non-detections is an essential first step towards more accurately being able to monitor changes through time, however interpreting what multiple non-detections represent is not so straightforward. A single observation at a site is typically sufficient to classify it as occupied, yet a non-detection can arise if the site is unoccupied, or if individuals remain undetected at an occupied site. In most cases, multiple non-detections are therefore required in order to confidently classify a site as unoccupied, yet determining the exact amount of effort that should be invested can be a challenge. In chapter three, we used data on detectability and prevalence for Swiss reptile species to explore the differences between two frameworks that provide guidance on survey effort requirements. Information on both detection probabilities and expected species prevalence are necessary in order to evaluate if monitoring has sufficient power to infer absence, which is problematic as for most species there is no natural spatial scale at which prevalence should be assessed. Our analysis also concluded that for rare species, which are frequently those of most interest to conservation biologists, it will normally not be possible to conduct a sufficient number of surveys to become confident that a site is truly unoccupied. This conclusion relates to structured monitoring programmes in which specific sites are monitored at designated occasions using standardised methods, and often at high costs (Kamp *et al.* 2016). However, unstructured surveying, in which participants record species observations opportunistically, provides a possible solution to this challenge. Citizen science has grown considerably in recent years (Silvertown 2009) and there has been proliferation of readily accessible platforms by which anyone can submit species observations (e.g. iNaturalist 2007; Sullivan *et al.* 2009). If such systems are constructed in such a way that non-detections of a species can be reported (e.g. by asking volunteers to report all observed species from a list), then obtaining sufficient survey effort to reliably infer species absences may become much more feasible. We should note that the analysis of unstructured monitoring data is complicated by the fact that spatial biases in reporting are a major feature in these datasets and are not trivial to deal with. However methods to account for these issues are undergoing a period of intense development (Hochachka *et al.* 2012; van Strien *et al.* 2013; Bird *et al.* 2014; Isaac *et al.* 2014; Swanson *et al.* 2016; Dennis *et al.* 2017b,a; Johnston *et al.* 2017), and it seems likely that unstructured mass-participation monitoring will become an increasingly influential means of monitoring species in the future.

Although the potential of citizen-science for species monitoring is high, there are some concerns which may be limiting a more widespread use of volunteers for data collection. In chapter four, we addressed the oft-cited (though rarely tested) concern that volunteer data quality are low (Foster-Smith & Evans 2003; Bird *et al.* 2014). There are two main forms of error that can arise in occupancy data; false-negatives, which arise as a result of imperfect detection, and false-positives, which may result from an observer believing wrongly that they have observed a species that is truly absent. We developed a dynamic occupancy model that simultaneously accounts for both forms of error, thereby allowing the quality of occupancy datasets to be evaluated by directly estimating error rates. In our case study of volunteer monitoring, we found that only the most common species had non-negligible rates of false-positive observations, and that the presence of these records had no substantial influence on population trends. Assuming that volunteers are given good guidance and training, we believe that concerns over volunteer data quality are overstated in most cases.

Monitoring must take place at an appropriate scale to ensure the results can be relied upon. In amphibian monitoring, most studies have focussed on ponds or wetlands as spatially discrete units. It is appealing to consider that individuals in apparently isolated patches within a landscape constitute discrete population units. As our understanding of movement behaviour has advanced, it has been recognised that such assumptions rarely hold true (Smith & Green 2005; Pittman *et al.* 2014), yet monitoring of many species remains focussed at this scale. This can be problematic when species spend a significant proportion of time outside of the area in which monitoring takes place; as this can readily lead to false-negative records within the occupied patch, or to false-positive records in neighbouring patches. In chapter five we saw that standard occupancy models overestimated tree frog occupancy rates in study regions where distances between adjacent sites are smaller than the distances that individuals are known to regularly traverse. We were able to account for this using our false-positive model which classifies transient individuals as false-positives, yet in standard occupancy models this will result in sites which do not support a resident population being wrongly classified as occupied (Sutherland *et al.* 2013). Issues of temporary emigration from a site being monitored can also have a strong effect when abundances are estimated from count data, as we showed in chapter six. If individuals are temporarily absent from a site, or if a site only covers a proportion of the habitat area used by a population, then only a proportion of the total population will be available for detection during a survey. Unless detection probabilities are high, we found that this can easily result in highly biased estimates of population size. In cases of temporary emigration, where phenology is likely to lead to predictable changes in availability, we found that it may be possible to still obtain accurate estimates of population size if the underlying phenology can be explicitly modelled. In situations when patterns in availability cannot be explicitly described, bias might be avoided by decreasing the time between successive surveys (so as to minimise the possibility of temporary emigration) or to increase the size of the spatial units surveyed.

The tracking of individuals over long-time periods has yielded many insights into species distributions and behaviour. Studies are increasingly turning towards less invasive means of monitoring individuals than 'traditional' animal marking methods such as PIT-tags or toe-clipping, which may negatively impact the welfare of animals (McCarthy & Parris 2004; Gauthier-Clerc *et al.* 2004). Photographic monitoring, often using camera-traps, is frequently used to estimate the distribution and density of species. We found that error rates by researchers matching images collected from such studies may be high. This suggests that there may be a trade-off in individual based monitoring; although photographs reduce welfare concerns, higher error rates may arise when identifying individuals through markings which may not always be unambiguously identifiable. The impact of wrongly constructed capture histories can be strong (Stevick & Palsbøll 2001; Yoshizaki *et al.* 2009), thus it is encouraging that we found that the use of generic freeware photo-identification software can have such strong effects in reducing error rates when identifying individuals from images. Increasingly, monitoring studies are using genetic data; either to monitor individuals in a non-invasive manner (Lukacs & Burnham 2005), or by identifying species presence using environmental DNA (Bohmann *et al.* 2014; Thomsen & Willerslev 2015). Issues of species detection remain relevant in such data (will the individual pass through this hair trap and leave a sample? Is the species' DNA present in water sampled for eDNA?), however, there is an additional level of complexity which must also be considered. Less-than-perfect accuracy in diagnostic genetic analyses means that false-positive and false-negative errors are possible even for samples in which the DNA has been 'captured'. Effort has been put into the development of models to account for this extra level of uncertainty (Knapp *et al.* 2009; Wright *et al.* 2009; Guillera-Arroita *et al.* 2017). Nevertheless, it is recognised that a great number of factors influence the amount of DNA present in environmental samples (Goldberg *et al.* 2016), and so the extent to which this form of sampling will be able to reliably replace more traditional monitoring, remains unclear.

Although I have sought to demonstrate the possibilities for dealing with measurement error in monitoring data, the real key to effective monitoring is in carefully determining the goals of a program at the planning stage (Yoccoz *et al.* 2001). By designing survey protocols carefully, the magnitude of observation errors can be reduced. Ultimately this represents a more cost-effective means to improve the utility of monitoring data than adjusting for errors statistically after data have been collected. I showed that with higher detection probabilities, the absence of a species at sites can be reliably inferred with feasible amounts of effort; that false-positive errors in observational data can be dealt with statistically as long as detection probabilities are not low, and that abundance estimation can be reliable in surveys with imperfect availability, provided that detection rates are high. Careful survey design, taking into account knowledge of the natural history of the target species, can therefore sidestep many of the difficulties that may confound effective monitoring, as well as ensuring that the data collected are actually informative and useful for management. Monitoring a species to extinction is not the goal of conservation biologists, yet too frequently surveying occurs without any consideration as to whether the data will be useful in informing management decisions. Programs should therefore be planned with explicit consideration as to how the data can be incorporated into adaptive management (McCarthy & Possingham 2007; Aceves-Bueno *et al.* 2015; Canessa *et al.* 2016) and thus improve conservation outcomes.

Literature Cited

Aceves-Bueno, E., Adeleye, A.S., Bradley, D., Tyler Brandt, W., Callery, P., Feraud, M., Garner, K.L., Gentry, R., Huang, Y., McCullough, I., Pearlman, I., Sutherland, S.A., Wilkinson, W., Yang, Y., Zink, T., Anderson, S.E. & Tague, C. (2015). Citizen Science as an Approach for Overcoming Insufficient Monitoring and Inadequate Stakeholder Buy-in in Adaptive Management: Criteria and Evidence. *Ecosystems*, **18**, 493–506.

Banks-Leite, C., Pardini, R., Boscolo, D., Cassano, C.R., Püttker, T., Barros, C.S. & Barlow, J. (2014). Assessing the utility of statistical adjustments for imperfect detection in tropical conservation science. *Journal of Applied Ecology*, **51**, 849–859.

Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Stuart-Smith, R.D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J.F., Pecl, G.T., Barrett, N. & Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, **173**, 144–154.

Bohmann, K., Evans, A., Gilbert, M.T.P., Carvalho, G.R., Creer, S., Knapp, M., Yu, D.W. & de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology and Evolution*, **29**, 358–367.

Canessa, S., Guillera-Arroita, G., Lahoz-Monfort, J.J., Southwell, D.M., Armstrong, D.P., Chadès, I., Lacy, R.C. & Converse, S.J. (2016). Adaptive management for improving species conservation across the captive-wild spectrum. *Biological Conservation*, **199**, 123–131.

Dennis, E.B., Morgan, B.J.T., Brereton, T.M., Roy, D.B. & Fox, R. (2017a). Using citizen science butterfly counts to predict species population trends. *Conservation Biology*, (In press) doi: 10.1111/cobi.12956.

Dennis, E.B., Morgan, B.J.T., Freeman, S.N., Ridout, M.S., Brereton, T.M., Fox, R., Powney, G.D. & Roy, D.B. (2017b). Efficient occupancy model-fitting for extensive citizen-science data. *Plos One*, **12**, e0174433.

Foster-Smith, J. & Evans, S.M. (2003). The value of marine ecological data collected by volunteers. *Biological Conservation*, **113**, 199–213.

Gauthier-Clerc, M., Gendner, J.-P., Ribic, C.A., Fraser, W.R., Woehler, E.J., Descamps, S., Gilly, C., Le Bohec, C. & Le Maho, Y. (2004). Long-term effects of flipper bands on penguins. *Proceedings of the Royal Society B: Biological Sciences*, **271**, S423–S426.

Goldberg, C.S., Turner, C.R., Deiner, K., Klymus, K.E., Thomsen, P.F., Murphy, M.A., Spear, S.F., McKee, A., Oyler-McCance, S.J., Cornman, R.S., Laramie, M.B., Mahon, A.R., Lance, R.F., Pilliod, D.S., Strickler, K.M., Waits, L.P., Fremier, A.K., Takahara, T., Herder, J.E. & Taberlet, P. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, **7**, 1299–1307.

Guillera-Arroita, G., Lahoz-Monfort, J.J., van Rooyen, A.R., Weeks, A.R. & Tingley, R. (2017). Dealing with false-positive and false-negative errors about species occurrence at multiple levels. *Methods in Ecology and Evolution*, **38**, 42–49.

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.

iNaturalist. (2007). iNaturalist Research-grade Observations. Occurrence Dataset. URL https://doi.org/10.15468/ab3s5x [accessed 5 October 2017]

Isaac, N.J.B., van Strien, A.J., August, T. a., de Zeeuw, M.P. & Roy, D.B. (2014). Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, **5**, 1052–1060.

Johnston, A., Fink, D., Hochachka, W.M. & Kelling, S. (2017). Estimates of observer expertise improve species distributions from citizen science data. *Methods in Ecology and Evolution*, **38**, 42–49.

Kamp, J., Oppel, S., Heldbjerg, H., Nyegaard, T. & Donald, P.F. (2016). Unstructured citizen science data fail to detect long-term population declines of common birds in Denmark (B. Schröder, Ed.). *Diversity and Distributions*, **22**, 1024–1035.

Knapp, S.M., Craig, B.A. & Waits, L.P. (2009). Incorporating Genotyping Error Into Non-Invasive DNA-Based Mark–Recapture Population Estimates. *Journal of Wildlife Management*, **73**, 598–604.

Lukacs, P.M. & Burnham, K.P. (2005). Review of capture-recapture methods applicable to noninvasive genetic sampling. *Molecular ecology,* **14**, 3909–19.

McCarthy, M.A. & Parris, K.M. (2004). Clarifying the effect of toe clipping on frogs with Bayesian statistics. *Journal of Applied Ecology*, **41**, 780–786.

McCarthy, M.A. & Possingham, H.P. (2007). Active adaptive management for conservation. *Conservation Biology,* **21**, 956–963.

McGill, B. (2012). Statistical Machismo. *Dynamic Ecology Blog*. URL https://dynamicecology.wordpress.com/2012/09/11/statistical-machismo/

Nichols, J.D., Hines, J.E., Sauer, J.R., Fallon, F.W., Fallon, J.E. & Heglund, P.J. (2000). A double-observer aproach for estimating detection probability and abundance from point counts. *The Auk*, **117**, 393–408.

Pittman, S.E., Osbourn, M.S. & Semlitsch, R.D. (2014). Movement ecology of amphibians: A missing component for understanding population declines. *Biological Conservation*, **169**, 44–53.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology and Evolution*, **24**, 467–71.

Smith, M.A. & Green, D.M. (2005). Dispersal and the metapopulation paradigm in amphibian ecology and conservation : are all amphibian populations metapopulations ? *Ecography*, **28**, 110–128.

Srivathsa, A., Puri, M., Kumar, N.S., Jathanna, D. & Karanth, K.U. (2017). Substituting space for time: Empirical evaluation of spatial replication as a surrogate for temporal replication in occupancy modelling. *Journal of Applied Ecology*, **38**, 42–49.

Stevick, P. & Palsbøll, P. (2001). Errors in identification using natural markings: rates, sources, and effects on capture recapture estimates of abundance. *Canadian Journal of Fisheries and Aquatic Science*, **58**, 1861–1870.

van Strien, A.J., van Swaay, C.A.M. & Termaat, T. (2013). Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology,* **50**, 1450–1458.

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.

Sutherland, C., Elston, D.A. & Lambin, X. (2013). Accounting for false positive detection error induced by transient individuals. *Wildlife Research*, **40**(6), 490–498.

Swanson, A., Kosmala, M., Lintott, C. & Packer, C. (2016). A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conservation Biology*, **30**, 520–531.

Thomsen, P.F. & Willerslev, E. (2015). Environmental DNA - An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, **183**, 4–18.

Wright, J.A., Barker, R.J., Schofield, M.R., Frantz, A.C., Byrom, A.E. & Gleeson, D.M. (2009). Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples. *Biometrics*, **65**, 833–840.

Yoccoz, N.G., Nichols, J.D. & Boulinier, T. (2001). Monitoring of biological diversity in space and time. *Trends in Ecology and Evolution*, **16**, 446–453.

Yoshizaki, J., Pollock, K.H., Brownie, C. & Webster, R. (2009). Modeling misidentification errors in capture-recapture studies using photographic identification of evolving marks. *Ecology*, **90**, 3–9.

Acknowledgements

First and foremost, I would like to thank my supervisor, Beni Schmidt, as without his guidance and patient encouragement this whole thing would have fallen apart long ago. Thank you for innumerable discussions, for giving me a lot of flexibility in my work, for getting me involved in some really interesting collaborations, and for being very good at making sure that there is always money available. Thanks also to Arpat Ozgul for his supervision and for embedding me into his research group - the last few years would have been much less enjoyable without the PopEcol group! Lukas Keller and Rolf Holderegger make up the rest of my PhD committee and I am very grateful for useful discussions and for spotting potential issues on the horizon long before I would have.

Much of the research presented here is based on data collected by a great number of people, the vast majority of whom I will never meet. I thank them all, and to everyone at KARCH for collating and granting me access to the great Swiss datasets upon which this thesis is built. Particular thanks go to Thierry Bohnenstengel at KARCH and Simon Capt at CSCF who dealt with my specific requests for data. In addition to my co-authors and those people acknowledged in the relevant chapters, I would like to thank Jérôme Pellet, Thomas Hertach, Michel Schlitner, Thomas Röösli and Hugo Cayuela for sharing their knowledge and experience of the various field-sites in which I have worked over the last few years.

Although it didn’t make it into the final thesis, most of my time over the last 4 years has involved thinking about yellow-bellied toads in Schwyz. This research is a long-term endeavour which started long before I even knew of the existence of either the yellow-bellied toad, or Schwyz itself. I’m indebted to Daniel Hasen and Martin Jordan for starting the mark-recapture data-collection, and Hannelore Brandt for continuing the study and showing me the ropes. Over the last few years, I have had a lot of assistance in the field; thank you to my field assistants Sarah Bänziger, Matteo Brezzi and Joaquim Golay. I am immensely grateful to the many people who were persuaded to leave the comfort of their homes to spend all night wading through thigh-deep mud (often in thunderstorms) in order to

take photographs of toads! I'm even more surprised that many of these people weren't put off by the experience, and later volunteered to spend hours staring at the photographs with nothing but the promise of chocolate. I thank you all for your willingness to procrastinate for my benefit. Fending off sleep in the early hours while driving back from Schwyz was sometimes a challenge, so I am exceptionally grateful to Martin Wehrle at Tierpark Goldau for providing with accommodation close to the field site.

I was attracted to apply for this PhD because it presented an opportunity not only to broaden my experience as a field-biologist, but also to head into the lab and tackle some genetics. I very rapidly realised that this was **not** my thing, and so I'm glad that there were people around who either helped me avoid messing things up too badly, or just took over entirely. My gratitude goes to Glauco Camenisch, Ursina Tobler, Martina Schenkel, Jasmin Winkler, and Trent Garner and Lola Brookes at IOZ in London. Ramon Müller took on what was going to be 'chapter 7' for his Master's thesis, which allowed me to avoid most of that lab-work while still getting the benefits of a season's fieldwork catching Midwife toad tadpoles. Thanks also to Stiftung Claraz and Stiftung Artenschutz who both provided funds for labwork.

A number of administrative challenges very quickly arise when you move to a new country just after they have voted to tighten their immigration rules; thus I quickly learned the value of the admin staff at the institute. Thanks to Ursina Tobler, Jacqui Moser and Lilian Dutoit for performing financial magic at times, and Michel Nakano and Cornelia Carnal for the ridiculous amount of time spent sorting out field-cars over the last few years. Isabel Schöchli and Florence Zufferey provided some essential support, which included helping me filling out what I was informed to be the most bizarre accident insurance claim the institute has ever had to deal with. I hope I wasn't too much of a pain

The road to a PhD is often rough and seemed at times never-ending. I thank all of the PopEcol group, past and present, for their conversation and distractions along the journey. I realised that I was in good company when my officemates each started trying to claim the title of "the most horrible person in the office"! I'm especially grateful to Tina Cornioley, Beni Schmidt, Richard Baxter, Alex van Rensburg and Julia Hatzl for commenting on sections of this thesis. Particular thanks to Julia and Beni for translating my summary into German. To all my friends in Zürich , thanks for all the good times!

I’ll inevitably forget some of you, but particular thanks to Beni, Rich, Janina, Nina, Alex, Reini, Glauco, Irina, Annina, Katie, Gian-Marco, Enrica and Oli. Particular thanks to Alex, Katie, Gian-Marco and Meg, my past and present flatmates, who have supported/cajoled/given me a kick up the arse to get this work done, as conditions dictated. Alex deserves particular note for talking me out of quitting, for doing all the genomic work which I don’t understand, and for diligently injecting me with rabies shots after an unfortunate encounter with a hyena. Elsewhere in the world, many others kept me sane through skype chats and by providing shelter and distraction; particularly Aidan and Hannah, Cat, Kirstie, Helen and Eliott, Joe, Tom, and Paul. Thanks to my family for supporting me through the tough times in the last few years, and of course to Julia, for her support and for always being there. And for making sure that I didn’t forget to eat while writing up.

My computer would like to thank the communities of Stack Overflow and the JAGS forum, without whom it would have taken much more of a kicking, and would certainly have been thrown off the office balcony at least once.

Finally I wish to thank Sam, to whom this thesis is dedicated. The friendly (but always unspoken) competition between us undoubtedly helped nudge me onto the academic path, and it’s not an exaggeration to say that without his friendship I would be a very different person today. You left tremors in the world that will not fade, and I am saddened that you will never get the chance to finally call me “The Doctor”.

Sam Stephen CRUICKSHANK

Date of Birth: 12th August 1988
Nationality: British

Education

01.2014 - 11.2017 **PhD in Ecology** (submission in 2017)

University of Zürich

Thesis title: Dealing with uncertainty in Amphibian and Reptile population monitoring for conservation
Supervisors: Dr. B. Schmidt and Prof. A. Ozgul

09.2012 - 09.2013 **MSc Ecology, Evolution and Conservation (Distinction)**

Imperial College London- Silwood Park

Thesis title: Should the Leopard frog change its spots: An examination of factors associated with parasitism in the Northern Leopard Frog (*Rana* p*ipiens*)
Supervisors: Prof. T. Garner and Dr. D. Orme

09.2007 - 06.2011 **BSc Natural Sciences (Class 1 Honours)**

University of Bath

Dissertation title: Spatio-temporal invertebrate community dynamics in response to the management of *Calluna* heathland
Supervisors: Dr. B. Smith and Dr. M. O'Connell

07.2004 - 06.2006 Godalming College.
A-levels: Biology, Mathematics, Chemistry
AS-levels: Economics, Geography, Critical Thinking

Publications

Cruickshank, S.S., Schmidt, B.R. (2017) Error rates and variation between observers are reduced with the use of photographic matching software for capture-recapture studies. **Amphibia-Reptilia** 38 (3): 315-325

Cruickshank, *S.S.*, Ozgul, A., Zumbach, S., Schmidt, B.R. (2016) Quantifying population declines based on presence-only records for red-list assessments. **Conservation Biology** 30:1112-1121.

Campbell Grant, E.H., Miller, D.A.W., Schmidt, B.R., Adams, M.J., Amburgey, S.M., Chambert, T., Cruickshank, S.S., Fisher, R.N.,[…], Muths, E. (2016) Quantitative evidence for the effects of multiple drivers on continental-scale amphibian declines. **Scientific Reports** 6:25625

Canessa, S. Bozutto, C. Grant, E., Cruickshank, S.S., Fisher, M.,[…], Schmidt, B.R. (in press) Decision making for mitigating wildlife diseases: from theory to practice for an emerging fungal pathogen of amphibians. **Journal of Applied Ecology**

Cruickshank, S.S., Bühler, C., Schmidt, B.R. (in review) The impact of false-positive and false-negative observations on occupancy and trends derived from volunteer monitoring data.

Miller, D.A.W., Grant, E.H.C.,[…], Cruickshank, S.S.,[…],Sigafuss, B.R. (in review) Quantifying Climate Sensitivity and Climate Driven Change in North American Amphibian Communities.