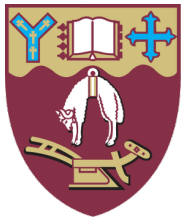
Improving the reliability of citizen science data

Julie Mugford

A thesis
submitted in partial fulfilment
of the requirements for the degree
of
Doctor of Philosophy
in
Computational Applied Mathematics



Department of Mathematics and Statistics
University of Canterbury
New Zealand
2021

Abstract

Citizen science is a growing movement that enables volunteers to help scientists collect and analyse information. Citizen science can solve problems that may, through ordinary methods, be intractable. In the context of ecology this is extremely important, as global biodiversity is in sharp decline, and there are currently not enough resources for traditional ecological monitoring to meet the current monitoring demands. However, there are some factors that make citizen science problematic, and raise concerns about reliability. Individual citizen scientists may vary in ability, there is the opportunity for persistent bias in the data, and the level of participant guidance given varies widely between citizen science projects.

In this thesis we use two contrasting mathematical methodologies: A Bayesian approach, and stochastic processes (particularly, random walk theory), to quantify and improve the reliability of citizen science data. We apply our methods to iNaturalist NZ, a citizen science project that provides participants with an online community to share and classify various observations of biota when and how they choose.

Our Bayesian approach combines individual citizen scientists' classifications for an image into a likely final classification. This approach improves on the common, but simple, majority vote method by estimating and utilising a measure of each participant's ability to correctly classify an image. This approach optimises the citizen scientists' classification efforts while also ensuring a desired level of final classification certainty.

Our stochastic process approach models the stochastic nature of citizen scientists and their observations. We use random walk theory to model a citizen scientist encountering and sharing observations of a given species and scale this up to simulate multiple years of iNaturalist NZ observation data. We use the simulated data to test the ability to specify a statistical model that differentiates between temporal changes in the number of species observations due to variation in observer behaviour, versus ecological changes in the species abundance. Without sufficient metadata about observer behaviour it is difficult to specify an appropriate statistical model. Observer metadata may be explicitly collected, or information could be inferred from the observation data. For example, we show that it is well supported by the iNaturalist NZ data that the probability of an observer sharing an observation decays as they share more observations during a walk.

The methods developed in this thesis are applicable to a wide range of citizen science projects. Our methods are able to take large, possibly unreliable datasets, and both quantify the reliability and improve the usability of the data.

Acknowledgements

First and foremost, I would like to thank my supervisors Alex James, Mike Plank, Elena Moltchanova, Andrea Byrom, and Jon Sullivan. Your ongoing support, generosity with your time, encouragement, and words of wisdom made my PhD experience extremely positive. I would particularly like to acknowledge my main supervisor, Alex James. Your relentless positivity, practical approach to mathematical modelling, and what seems like an ability to know a suitable method for every real-world problem was a large part in my motivation for doing a PhD under your supervision. I can safely say, I'm very pleased with that decision. I'm extremely grateful for all the time and learning I have received from you throughout my PhD.

My PhD was co-funded by Canterbury University, Mannaki Whenua, and Te Pūnaha Matatini. I'm thankful for the support and funding from all three of these organisations and the wide range of perspectives they contributed to this work. Te Pūnaha Matatini enriched my PhD experience beyond anything I could have imagined. In particular, the Te Pūnaha Matatini Whānau was an absolute highlight of my PhD experience. The opportunities within the Whānau have greatly grown my confidence, and the friendships I made, I will cherish for life. The school of Mathematics and Statistics at Canterbury University has been a massive part of my education, from my undergraduate years, to honours year, and now my PhD. I would like to acknowledge the amazing administration staff, academic staff, and fellow students that have made this place an absolute joy to spend so many years. I would particularly like to thank my officemate, Fareeda Begum. I'm so grateful that I got to share so much of this experience with you and I look forward to a life long friendship with you.

I would like to thank my parents for all their support and raising me in a way that installed a strong work ethic, which I've relied on deeply in recent months. To my brother, wider family, friends, and Ministry of Transport colleagues, thank you for your continued support and encouragement. All your words of support have kept the fire burning throughout this journey. I would like to thank Daniel Guppy who generously proof-read this entire thesis and provided valuable feedback. To Nikki Edwards and Cor Story, your introduction to endurance triathlon, and support has not only served as a great escape from PhD life, but, it has taught me that with self belief, determination, and a smile, anything is possible!

Julie Mugford

"What you do makes a difference, and you have to decide what kind of difference you want to make."
Jane Goodall

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Deputy Vice-Chancellor's Office
Postgraduate Research Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 3b has been published in Ecological Informatics
https://doi.org/10.1016/j.ecoinf.2021.101313

Please detail the nature and extent (%) of contribution by the candidate:

Julie Mugford was the lead author of this work. The work was carried out solely by her with guidance from her supervision team.

Certification by Co-authors:
If there is more than one co-author then a single co-author can sign on behalf of all
The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the Doctoral candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Alex James* Signature: *Alex James* Date: *23 May 2021*

Contents

Nomenclature

Chapter 3a and 3b: Image classification

α	Classification accuracy
$\hat{\theta}$	Estimated user accuracy
$\hat{t}_i$	Classification of image i
ρ_i	Prevalence of class i images
θ	User accuracy
$a^{(k)}$, $b^{(k)}$	Prior parameters for user k accuracy
I	Total number of images
i	Counter for number of images
I_C	Number of classification stage images
I_T	Number of testing stage images
J	Total number of classes
j	Counter for number of classes
K	Total number of users
k	Counter for number of users
$m^{(k)}$	Number of images user k identifies
n	Number of identifications per image
t_i	True identity of image i
$x^{(k)}$	Identification from user k
$y^{(k)}$	Number of images user k correctly identifies

Chapter 4a: Mechanistic model

δx	Individual step size in random walk
Ω	Area of field
τ	Time between random walk steps

D Diffusion coefficient

L Length of observer walk

n Number of individuals

p Probability an individual in encountered and shared to iNaturalist NZ

p_{enc} Probability an individual in encountered

p_{shared} Probability an observation is shared to iNaturalist NZ

R_H Individual's home range radius

R_P Observer's perceptive radius

V_W Observer walking speed

Chapter 4b: Multi year simulations

α Type 1 error rate

β Type 2 error rate

ϕ Dispersion parameter

K Number of observers

n Number of individuals

p Probability of an observation

t Year counter

W Number of walks per observer per year

w Walk counter

W_t Total number of walks by all users in year t

x_t Individual observations in year t

$x_t^{(w)}$ Individual observations on walk w, and in year t

Chapter 4c: Variation in observer behaviour

Δ_i AIC difference

$\lambda^{(i,j)}$ Decay parameter for observer i on walk j

a, b	Beta distribution parameters
i	Observer counter
j	Walk counter
k, θ	Gamma distribution parameters
n	Number of individuals
p	Probability of an observation
$y^{(i,j)}$	Number of observations by observer i on walk j

Chapter 1

Introduction

Citizen science.
Engaging the curious.
Helping scientists.

Citizen science is a growing movement that enables volunteers to help scientists collect and analyse information (Bonney et al., 2014; Gura, 2013; Cohn, 2008a; Tulloch et al., 2013; Bonney et al., 2009; Baker, 2016; Pocock et al., 2017). Participants are engaged in authentic research experiences at various stages in the scientific process (Dickinson et al., 2012). Citizen scientists offer a free source of labour, skills, and computing power (Silvertown, 2009; Cohn, 2008a) that enables researchers to tackle otherwise intractable, laborious, or costly research problems (Gura, 2013; Tulloch et al., 2013; Franzoni and Sauermann, 2014). Project goals range from answering very specific scientific questions, to broad and unguided data gathering and analysis, to purely providing outreach and education to the public (Wiggins and Crowston, 2011; Gura, 2013; Miller-Rushing, 2017; Silvertown, 2009). Typically, citizen science projects involve participants either (1) completing recognition, classification, or problem-solving tasks that require human competencies; or (2) contributing data or observations according to a protocol (Wiggins and Crowston, 2011). Throughout this thesis we refer to the former as classification-based citizen science projects, and the latter as observation-based citizen science projects. Some citizen science projects include both a classification phase and an observation phase.

In general, citizen science projects and protocols are designed for anyone to follow (Silvertown, 2009). However, project complexity and level of guidance may vary greatly across projects (Pocock et al., 2017). For example, classification-based citizen science projects hosted by Zooniverse provide participants with tutorials and guides (Simpson et al., 2014), whereas iNaturalist provides participants with an online community to share and classify various observations of biota when and how they choose. Furthermore, in the classification phase of iNaturalist, no explicit training is given, thus it is unlikely that all classifications are accurate.

There has been ongoing research about how to ensure the quality and certainty of citizen science data, in response to scepticism about data quality (Bird et al., 2014). In this thesis, techniques are explored to make use of citizen science datasets generated from projects with minimal protocols or guidance, where it was previously uncertain that any conclusions could be drawn from the data.

Citizen science has grown since its inception, and rapidly in recent years. Citizen science concepts date back to the 1880s, when the American Ornithologists' Union collected and analysed citizens' data regarding bird strikes into lighthouses (Droege, 2007). In the 1990s the term "citizen science" emerged from the Cornell Lab of Ornithology (Baker, 2016) and the first peer-reviewed article involving citizen science was published (Follett and Strezov, 2015). Today, for example, citizen science projects include: classifying the morphological features of galaxies, deciphering fragments of

pre-modern and medieval Jewish texts, and sharing and classifying observations of biota. There has been an explosion of citizen science projects since the early 2000s (Bonney et al., 2014, 2009; Baker, 2016; Pocock et al., 2017). For example, in 2007 the citizen science hosting platform, Zooniverse launched their first citizen science project - Galaxy Zoo (Raddick et al., 2009). As of 2021, almost 15 years since the first project, on Zooniverse there are almost 80 active projects, 180 paused projects, and just over 50 finished projects. Citizen science has proliferated in ecology (Pocock et al., 2017; Dickinson et al., 2010). Pocock et al. (2017) undertook a comprehensive search for ecology and environment-based citizen science projects since the 1940s and found an almost 10% increase in new projects annually since the early 2000s'. Much of this recent growth is due to advancements and increased availability of technology, such as internet, cameras, and GPS availability on mobile phones (Silvertown, 2009; Bonney et al., 2014; Gura, 2013).

There is no doubt that citizen science has contributed significantly to scientific knowledge (Reed et al., 2013; Bonney et al., 2014). For example, Bautista-Puig et al. (2019) found over 5000 peer-reviewed articles that involve citizen science on Web of Science. The citizen science project Foldit that turned protein folding into a game led to the discovery of a desired protein structure before professional AIDS researchers working on the same problem (Khatib et al., 2011). eBird data - a bird watching citizen science project - has formed one of the largest biodiversity datasets in the world (Wiggins, 2011).

However, the amount of data gathered from participants, and their backgrounds and motivation, can vary greatly among citizen science projects. Typically, citizen data are gathered by a small number of participants (Sauermann and Franzoni, 2015a; Eveleigh et al., 2014). Participants' backgrounds and skills also may vary greatly. For example, some participants may be experts in a field, whereas others are preschool students, prisoners, or alzheimer's patients (Gura, 2013; Baker, 2016). Such differences in volunteers' skills contribute to the perceived low quality of citizen science data (Bird et al., 2014; Dickinson et al., 2010). Participants' motivations are also a factor (Rotman et al., 2012; Reed et al., 2013; Raddick et al., 2009). They may include: interest in the subject area, desire to contribute to science, the community and social cohesion created with other participants, opportunity to use their skills/knowledge, that the project is easy and/or enjoyable to do, and opportunities for education (Rotman et al., 2012; Reed et al., 2013; Raddick et al., 2009).

Monitoring changes in species' populations is crucial to identifying extinction risk (Mace and Lande, 1991), evaluating the performance of conservation efforts (McKinley et al., 2017), monitoring ecological responses to climate change, and reporting

against national and international targets (Butchart et al., 2010). However, traditional species monitoring by trained experts and in standardised forms is resource intensive and costly. Further, these traditional ecological monitoring schemes typically do not deliver long-term data or at regional scales, making it difficult to monitor biodiversity change across space and time (Hudson et al., 2014; Schmeller et al., 2009). Thus, citizen science can often be the only practical way to answer various ecological questions (Tulloch et al., 2013; Miller-Rushing et al., 2012).

There have been insights into the impact of climate change using data from citizen science projects which include the presence, abundance, demography, and phenology of organisms (Dickinson et al., 2012). For example, Hickling et al. (2006) used citizen science data to detect the movement north, and to higher elevations, of a wide range of taxa during a period of regional climate warming. Citizen scientists have the ability to frequently monitor large areas, and citizen science projects are becoming an important tool in early detection and tracking of invasive species (Dickinson et al., 2012). For example, a large network of citizen scientists was used to track the spread of a house finch eye disease in the USA (Dhondt et al., 1998). The scale of citizen science data makes it a particularly useful tool for studying effects on biodiversity of habitat loss and fragmentation (Dickinson et al., 2010). For example, Zuckerberg and Porter (2010) used New York State's two Breeding Bird Atlases (1980–1985 and 2000–2005) to investigate the response of 25 forest birds to a range of forest cover loss and land fragmentation.

In spite of the useful contributions that citizen science has made to scientific knowledge, especially in ecology, there has been considerable debate about the utility of such data in a scientific framework (Bird et al., 2014). Detractors often raise concerns over the precision of citizen science data due to variability among the participants, and the opportunity for bias in the data. For example, detractors cite under-detection of species or non-random distribution of citizen scientists' efforts (Crall et al., 2011) as reasons to be sceptical about the contribution that citizen science can make in ecology and conservation. Data collection protocols specified by the project, and the ability to apply post-hoc statistical manipulation to the data, can influence the types of research questions citizen science data may be able to answer (Bird et al., 2014; Crall et al., 2011; Wiggins et al., 2011). Typically, species observation citizen science data is recorded over time and space and may include only presence data, or presence and absence data (Bird et al., 2014).

iNaturalist is an example of a citizen science project that gathers presence-only ecological data. This type of data collection protocol improves accessibility of the project according to varying levels of participants' motivation, commitment, and skills (Bird

et al., 2014). However, a lack of information on species' absence may restrict what questions can be answered by the data, and what type of analyses are possible (Pearce and Boyce, 2006). Furthermore, variation in participants' sampling efforts spatially, temporally, and across taxa is often not known and difficult to infer, making it challenging to isolate true ecological trends from variability in participants' efforts and abilities. Conversely, ecological citizen science projects with more extensive protocols that gather presence-absence data do exist. Data from these projects can more readily be utilised to answer questions relating to spatial and/or temporal distributions of species. However, there is still potential for error, such as the inability to distinguish imperfect detection from true absence (Bird et al., 2014).

A focus of citizen science research has been on developing techniques to address uncertainty and variability in citizen science data. Within classification projects, there has been research on improving the quality of final classifications, based on a consensus of the individual votes. This has included: estimating participant accuracy based on proxies for past performance (e.g. similarity of an individual's decision to those of others (Kurvers et al., 2019)), weighting individual votes based on a measure of participant confidence (See et al., 2013); estimating participant classification ability from images with known ground truth and then weighting votes accordingly (See et al., 2015); and using Bayesian modelling to re-evaluate the quality of the consensus label following each new vote (Siddharthan et al., 2016). There has also been research in the ability to identify the experts in the crowd when there is a small pool of experts and limited time for the identification process (Katsagounos et al., 2021). For observation-based citizen science, particularly in ecology, much research has focused on methods to detect signals of species abundance changes using noisy data. For example, Isaac et al. (2014) simulated noisy data to test how 11 different statistical trend estimation methods performed at reliably detecting a temporal trend of a focal species' population.

New Zealand is an interesting case study for understanding and improving the utility of ecological citizen science data. The long period of isolation of New Zealand, slow evolution, and lack of native terrestrial mammals has resulted in a unique and high level of endemic biodiversity (Diamond, 1990). Native species, their genetic diversity, and the habitats and ecosystems that support them are greatly important to New Zealand and its citizens (Department of Conservation and Ministry for the Environment, 2000). However, since settlement of humans in New Zealand, biodiversity has been in decline, and now many of the unique endemic taxa are at risk of extinction or have gone extinct (Russell et al., 2015). This is largely driven by loss and disruption of natural areas and ecosystems, and the introduction of plant and animal pests (Department of Conservation and Ministry for the Environment, 2000). It is not surprising that ecological conservation is of great importance in New Zealand. This is

evident through the strong public support for conservation groups (Russell et al., 2015) and ambitious national policy statements. For example, Predator Free 2050 aims to eliminate eight introduced mammalian predators from New Zealand by 2050 (Russell et al., 2015), and the 2018 - 2028 Kiwi Recovery Plan outlines the goal to reach 100,000 kiwi by 2030 by growing all kiwi species by at least 2% per year (Germano et al., 2018).

However, biodiversity monitoring is expensive and only feasible for a very limited number of taxa, spatial locations, and time periods. It is unrealistic to assume that conservation managers in New Zealand will have the resources to monitor all threatened taxa or pests that must be controlled or eliminated. Hence (and especially given New Zealand's strong national identity with conservation), there is a need for data and resources from wider sources, and citizen science is a good candidate. Citizen science data provides crucial extra information for evidence-based decision-making when conservationists are faced with the above prioritisation decisions.

Citizen science, and particularly iNaturalist NZ (the New Zealand specific branch of the global iNaturalist network), thus has the potential to play a vital role in New Zealand's conservation management and bio-security efforts. In the first 7 years of iNaturalist NZ (2013 - 2019) there were almost 800,000 observations, shared by just over 15,500 observers, and almost 6,000 participants that helped classify observations. The observations cover almost 16,000 species and include plants, animals, fungi, chromista, and protozoans. When the iNaturalist community agree on the identity of an observation it becomes 'research grade'. Of the almost 800,000 observations on iNaturalist NZ more than half (450,000) are currently considered research grade. However, because iNaturalist is a completely curiosity-driven citizen science platform with minimal guidance in both the classification and observation phases, there is uncertainty about the usefulness of the data it contains, including data that are considered research grade. Throughout this thesis the iNaturalist citizen science platform is used as the motivation for developing methods to improve the usefulness of citizen science data, and the iNaturalist NZ dataset is used as a case study to demonstrate these methods in practice.

This thesis is in two parts based on the classification and observation phases of the iNaturalist citizen science platform. In the two parts of this thesis we use two distinct and contrasting mathematical methodologies. First, we consider the classification phase of iNaturalist. Currently, the iNaturalist classification phase provides minimal training for the participants, collects no information about a participant's classification ability, and combines multiple classifications for each observation with a simple majority vote. Although this is a simple method and easy to implement, it is not the most efficient use of valuable citizen scientist resources, and it leaves open the question: how trustworthy are the classifications? In Chapter 3a we show how a

Bayesian approach could be used to collate participant classifications, by developing a measure of a participant's ability to correctly classify an observation. In Chapter 3b we apply a Bayesian classification approach to the first five years of the iNaturalist NZ dataset. This method optimises citizen scientists' classification efforts while also ensuring a desired level of final classification certainty.

In the second part of this thesis, we assume the iNaturalist research grade identities are accurate, and consider how these observations could contribute to knowledge of temporal abundance changes in the taxa recorded. Given the curiosity-driven nature of sharing observations to iNaturalist, and the minimal record of observer effort, this is not a straightforward exercise. In Chapter 4a and Chapter 4b we use stochastic processes (particularly, random walk theory) to build a simulation of the iNaturalist NZ dataset that includes many aspects of expected noise due to observer behaviour. The simulation is used to test the ability to specify a statistical model to differentiate between changes in species abundances due to variation in observer effort, versus ecological changes in species abundances. For simplicity, we initially assume that observer behaviour is constant over time and across observers. However, in Chapter 4c we revisit this assumption and use model fitting techniques to explore how observer behaviour may vary.

This thesis develops and contributes methods to improve the reliability and usability of citizen science projects that have minimal participant guidance or protocols. Although the methods developed were motivated by iNaturalist and applied to the iNaturalist NZ dataset, they are applicable to any classification and observation-based citizen science project or platform with a similar participant protocol. Furthermore, given iNaturalist has minimal participant guidance, these new methods are easily applicable to citizen science projects with tighter protocols. The latter would have more information to input into the models developed, and therefore the results would be more reliable. For example, there may be information on a participant's ability to correctly classify an image, or a temporal measure of observer effort, that could be incorporated into the models developed here.

Thesis outline

Chapter 2 (iNaturalist NZ). The iNaturalist NZ citizen science project is introduced. We cover the history of iNaturalist NZ, how it works, and contributions it has made to scientific knowledge in New Zealand. We present some key insights about the scale and state of the current dataset.

Chapter 3a (Including user accuracy in classification based citizen science projects). We outline a Bayesian approach which uses individual classifications for an image to produce a likely final classification. The chapter begins by exploring how we would include participant ability in the final classification decision if it was known. We then consider how we would estimate participant classification ability if the observations' true identities were known. To finish, we outline a Bayesian method to estimate participant ability, and utilise this in the final classification decision when there are no observations with a known true identity.

Chapter 3b (Citizen science decisions: A Bayesian approach optimises effort). We apply the Bayesian method that is outlined in Chapter 3a to the first 5 years of iNaturalist NZ data. This method optimises the citizen scientists' classification efforts while also ensuring a desired level of final classification certainty.

Chapter 4a (A mechanistic model of a citizen scientist encountering individuals).We use random walk theory to describe a mechanistic model of a single citizen scientist encountering a given taxon and sharing observations to iNaturalist NZ. For simplicity we assume that the citizen scientists' behaviour is constant over time. We find that the resulting distribution of the number of observations shared per day by the observer is binomially distributed.

Chapter 4b (Extracting the change signal from noisy ecological citizen science data). We introduce a generalised linear model that may be used to detect if there is a significant annual trend in a given taxon abundance from the iNaturalist NZ data. This model takes into account that the annual number of observations for a taxon may vary due to changes in the underlying taxon abundance, and due to variation in the citizen scientist behaviour. We scale up the single observer model in Chapter 4b to a multi-year simulation of participants' observation behaviours. In the simulated data there is no annual change in the taxon abundance, and we simulate some of the noise we would expect to see in the iNaturalist NZ dataset due to observer behaviour. The simulation is used to test the ability to specify a statistical model to differentiate between changes in species observations due to variation in observer effort, versus ecological changes in species abundance.

Chapter 4c (Modelling variation in citizen scientist's observation sharing behaviours). In Chapter 4a and Chapter 4b we assumed that observer behaviour was consistent across different observers (inter-observer) and did not change for each observer over time (intra-observer). In this chapter we revisit this assumption by using maximum-likelihood methods to estimate the parameters for three candidate models that model variation in inter and intra-observer behaviour. We find that the model that allows observer effort to decay over time is the most supported model.

Chapter 5 (Final Discussion). Discusses the insights and contributions this thesis has made to understanding and improving the usefulness of citizen science data.

Thesis publications

A paper presenting the work in Chapter 3b has been published by Ecological Informatics (Mugford et al., 2021).

Chapter 2

iNaturalist NZ — Mātaki Taiao

iNaturalist.
Users share nature photos,
and vote on labels.

Introduction

iNaturalist NZ - Mātaki Taiao - (https://inaturalist.nz/) is a New Zealand citizen science project that is building a living record of biodiversity in New Zealand. Their vision is to improve the visibility of biodiversity changes in New Zealand, by providing an online platform where nature watchers can share their observations of biota (New Zealand bio-recording network trust, 2020). Scientists, and environmental managers will be able to use their data to monitor changes in biodiversity, and anyone can use their breadth of records to learn about New Zealand's natural history. Beyond sharing observations of biota, iNaturalist NZ participants can comment on and identify other participants' observations.

The history of iNaturalist NZ

iNaturalist NZ is run by the New Zealand Bio-Recording Network Trust (NZBRN), a charitable trust dedicated to bio-recording. NZBRN began in 2005 with a grant from the New Zealand Terrestrial and Freshwater Biodiversity Information System programme (TFBIS) to prototype a community nature observation system for New Zealand. They adapted a Swedish nature observation system, Artportalen, to create a New Zealand version. The New Zealand version gathered several hundred thousand observations of mostly birds, fungi, and plants. The observations came from a small devoted group of participants, and data from the New Zealand Garden Bird Survey.

In 2010 the original NZBRN system was becoming outdated as it was far from the standard of modern internet technology. An upgrade was required and more funding was received from the New Zealand Government's TFBIS programme. After a review of systems, iNaturalist, an open source community nature observation system from the USA, stood out. iNaturalist had many attractive features but in particular it was user-friendly and made full use of modern social media to build an online community.

In August 2012 the New Zealand chapter of iNaturalist was launched and joined the growing network of countries, e.g. USA, Canada, Mexico, and Colombia, that use the iNaturalist platform. At the end of June 2020 the iNaturalist network had over 48 million observations of almost 300,000 species observed by over 1.4 million observers. iNaturalist NZ made up 1,043522 of these observations, 18,041 of the species, and 20,455 of the observers.

How iNaturalist NZ works

iNaturalist NZ participants mostly contribute in two ways: they share observations of biota, or they help identify observations that other participants have shared. A partic-

ipant is able to share an observation to iNaturalist NZ by using either the iNaturalist NZ website or their mobile application. When a participant shares an observation they are prompted to include a photo or sound clip, but there is the option to share an observation with no media. Participants are able to fill in details about their observation, e.g. the species name, date, and location of the observation. The mobile application automatically fills in the date, time, and location field, but these can be manually overridden by the user. If the participant does not know some of the details they are able to leave those fields as unknown. This is particularly common for the species name field as many observers use iNaturalist NZ to learn the identity of taxa they have seen.

Once a participant has shared an observation it is visible to all of the iNaturalist NZ members and all of the public. Other participants are able to suggest an identification for the observation. Observations on iNaturalist NZ are assigned to one of three quality grades, "casual", "needs ID", or "research", depending on data quality and how many participants have identified the observation and the level of consensus between the identifiers. Below are the definitions of the quality grades.

Verifiable observation: A verifiable observation has to pass the following data quality assessments: have a date, be geo-referenced, have a photo or sound clip, and not be a captive or cultivated organism. Verifiable observations are labelled "needs ID" until they either attain research grade status or are voted to causal grade because the iNaturalist NZ community vote that at least one of the required data quality assessments are not met.

Casual grade: Casual grade observations are observations that are considered non-verifiable.

"Needs ID" grade: "Needs ID" grade observations are verifiable observations that need further identifications to reach a majority vote on a taxon.

Research grade: Research grade observations are verifiable observations where the majority ($> 50\%$) of identifiers have agreed at a species level or lower on the taxon. However this comes with the caveat that an observation may become research with a majority vote at the genus level if the observation has been flagged as "it's as good as it can be".

iNaturalist NZ participants are able to identify observations by either agreeing with the previous participants' identification or by suggesting another identification from the database of taxa. Identifications can be made at any level of taxonomic rank. Once an observation is considered research grade, further identifications may still be contributed to the observation and the classification or quality grade of the observation could change.

Computer vision

Computer vision was added to the iNaturalist platform throughout the second half of 2017. The computer vision algorithm is trained on historical research grade iNaturalist observations and provides identifiers with an automated list of taxon suggestions. Computer vision was first integrated into the iNaturalist iOS application on June 29, 2017. However, it was not fully integrated into the Android application and web observation uploader until September 2017. Since 2017 there have been multiple updates to the computer vision model.

Access to iNaturalist NZ data

The iNaturalist NZ data can be accessed and downloaded in multiple ways. The iNaturalist NZ website has a tool to explore shared observations, and also download observations to csv or kml. All research grade observations are shared weekly with the Global Biodiversity Information Facility (GBIF, https://www.gbif.org), the global repository of biodiversity data. iNaturalist NZ is the second largest contributor of New Zealand observations to GBIF (New Zealand bio-recording network trust, 2020), after eBird (a bird watching citizen science project). There are also multiple R packages that use the iNaturalist API to download data, e.g. 'iNatTools' (https://github.com/pjhanly/iNatTools). Throughout the sharing of iNaturalist data, every photo, observation, and audio file uploaded to iNaturalist is owned by the uploading participant. Each participant has complete control over the copyright or Creative Commons license attached to each file and observation.

Datasets used within this thesis

In this thesis two snapshots of the iNaturalist NZ dataset are used due to the timing of the work. In Chapter 3a, and 3b the first five years of iNaturalist NZ data from August 2012 - July 2017 are used. In Chapter 4b, and 4c seven years of iNaturalist NZ data from January 2013 - December 2019 are used. In 3a, and 3b we required data on both iNaturalist NZ observations and identifications, whereas, in Chapter 4b, and 4c we only required observation data. Therefore, in this data section we only present facts about the first five years of identification data, but facts about the observation data from January 2013 to December 2019. Where applicable, throughout this thesis we present further relevant facts and summaries about the iNaturalist NZ dataset.

Throughout this thesis we only consider observations that were contributed after the project shifted to using the iNaturalist platform, i.e. observations contributed from August 2012.

The facts and figures

Participation has been growing on iNaturalist NZ since it's launch in August 2012. Participation comes in two forms: contributing observations, or contributing identifications. To align with the remainder of the thesis we summarise observation data from the first complete seven years of iNaturalist NZ data (January 2013 - December 2019), and identification data from the first five years of iNaturalist NZ (August 2012 - July 2017).

Observers and observations

From January 2013 to December 2019 there has been an increase in both the number of observers on the iNaturalist NZ platform, and the number of observations shared, Figure 2.1. Between January 2013 and December 2019, 15,699 observers shared 798,020 observations to iNaturalist NZ. The rate that new observers have been participating on iNaturalist NZ, and the rate that new observations have been contributed to iNaturalist NZ during the first seven years has been increasing, Table 2.1. For example, in 2013 on average 2 new observers contributed to iNaturalist NZ per day, and 92 new observations were added to iNaturalist NZ per day. By 2019 these rates had leaped up to an average of 13 new observers contributing to iNaturalist NZ per day, and 695 new observations being added to iNaturalist NZ daily.

Year	Average new daily observers	Average new daily observations
2013	2	92
2014	2	136
2015	6	189
2016	4	312
2017	7	315
2018	9	448
2019	13	695

Table 2.1: **The rate that new observers have been joining iNaturalist NZ and that observations have been added to iNaturalist NZ has grown since 2013.** Average new daily observers and observations for the first complete 7 years of iNaturalist NZ data.

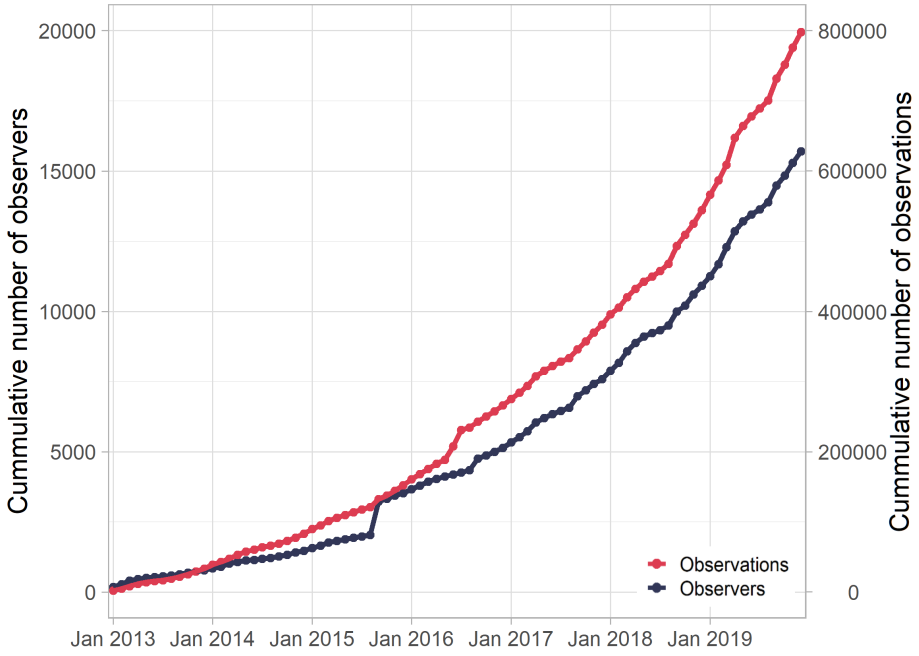


Figure 2.1: **The number of observers and observations on iNaturalist NZ has been increasing steadily since January 2013.** Cumulative number of observers and observations on iNaturalist NZ between January 2013 and December 2019. 15,699 observers shared 798,020 observations to iNaturalist NZ between January 2013 and December 2019.

Quality grades

As described in the introduction to this chapter, observations are assigned to one of three quality grades: casual, "needs ID", or research. From the almost 800,000 observations shared to iNaturalist NZ between January 2013 and December 2019 at the time of downloading the data (April 2020), 20% (165,100) were casual grade, 23% (183,404) were "needs ID" grade, and 57% (449,516) were research grade.

Taxa

From the 449,516 observations that have been classified to research grade, 11,407 different taxa have been identified. iNaturalist NZ observations are categorised into 13 'iconic' taxa based on the contributed identifications. Table 2.2 shows: the number of research grade observations per iconic taxon, the number of observers that have shared a research grade observation within each iconic taxon, and the number of unique taxa within each iconic taxon with a research grade observation. Some iconic taxa are nested in lower taxonomic ranks than others (e.g. animals, insects) and observations are assigned to the lowest matching iconic taxon. The plant iconic taxon is the largest on all

three metrics: 55% of research grade observations are of plant taxa, 50% of observers have shared a research grade plant observation, and 45% of the 11,407 taxa with research grade observations are a plant taxon. On the other hand, 44% of observers have shared a research grade bird observation, but only 3.6% of the research grade taxa are bird taxa.

Iconic taxa	**Observations**		**Observers**		**Unique taxa**	
	proportion	(value)	proportion	(value)	proportion	(value)
Protozoans	0.09%	(424)	1.7%	(174)	0.4%	(48)
Amphibians	0.1%	(613)	2.7%	(281)	0.1%	(8)
Chromista	0.6%	(2590)	3.0%	(318)	0.7%	(82)
Reptiles	0.7%	(2900)	7.8%	(816)	0.6%	(63)
Ray finned fish	1.0%	(4782)	5.9%	(615)	2.2%	(255)
Mammals	1.6%	(7255)	11.9%	(1248)	0.7%	(76)
Arachnids	2.5%	(10995)	13.7%	(1441)	2.3%	(262)
Animal	2.8%	(12370)	10.0%	(1048)	4.2%	(477)
Mollusks	3.9%	(17615)	9.8%	(1034)	6.3%	(719)
Fungi	4.9%	(21841)	18.2%	(1912)	13.0%	(1478)
Insects	12.1%	(54295)	37.7%	(3954)	20.9%	(2384)
Birds	14.4%	(64924)	44.0%	(4615)	3.6%	(411)
Plants	55.4%	(248833)	50.4%	(5289)	44.9%	(5126)
Other	0.02%	(75)	0.4%	(46)	0.16%	(18)
All observations	**100%**	**(449512)**	**-**	**(10499)**	**100%**	**(11407)**

Table 2.2: **Over half of the iNaturalist NZ research grade observations are of a plant taxa.** This table just contains information about research grade observations. The plant iconic taxa is the most common in terms of number of observations (55.4% of all observations), number of observers that have shared an observation within the iconic taxa (50.4% of observers), and number of individual taxa within the iconic taxa (44.9% of taxa).

On average 134 new taxa were observed on iNaturalist NZ every month between January 2013 and December 2019, Figure 2.2. However, the number of new taxa observations per month was higher in the earlier years of iNaturalist NZ. In 2013 there were on average 232 new taxa observed every month, whereas, in 2019 the average number

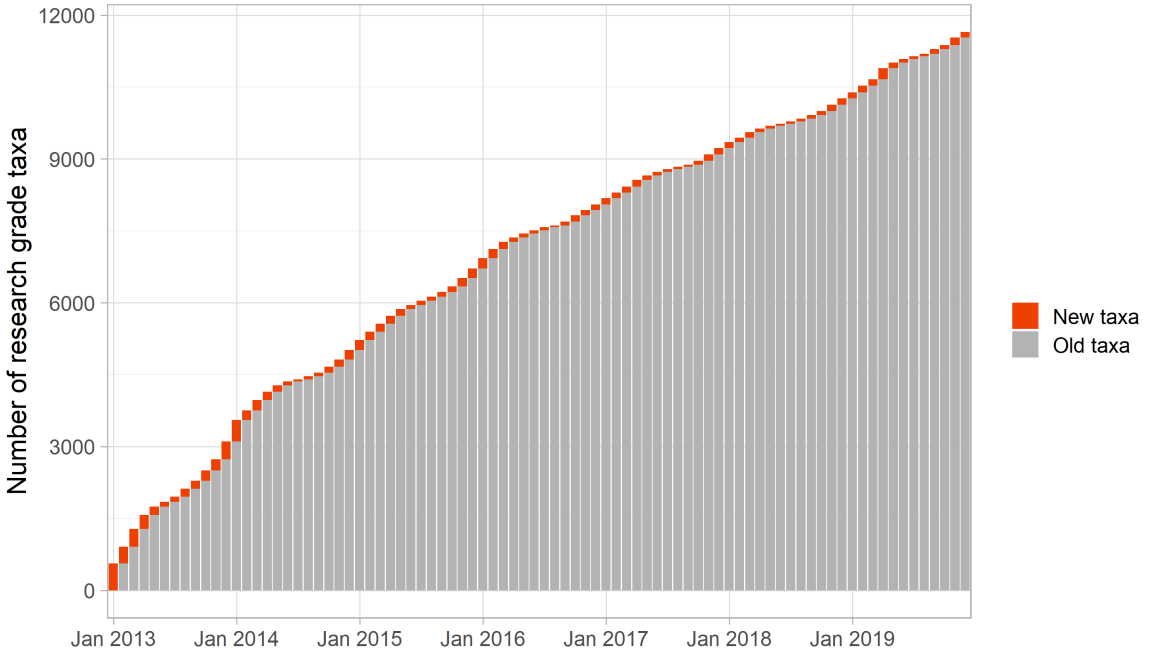


Figure 2.2: **New taxa are continuously being observed and classified to research grade on iNaturalist NZ.** Every month between January 2013 and December 2019, at least one new taxa has been classified to research grade on iNaturalist NZ. Between this period 11,407 different taxa have had a research grade observation.

of new taxa observed every month was 115. Some of the first taxa added to iNaturalist NZ were the Auckland tree weta (*Hemideina thoracica*), Pukeko (*Australasian swamphen*) and White basket fungus (*Ileodictyon cibarium*). In the first month on iNaturalist NZ a lot of the more abundant native species were observed, for example, Tui (*Prosthemadera novaeseelandiae*) and Kererū (*Hemiphaga novaeseelandiae*). However, it was not until the second month of iNaturalist NZ that a house sparrow (*Passer domesticus*) was observed.

Observer contributions

In citizen science projects, and iNaturalist NZ, it is common for a small number of contributors to make up a large share of the contributions (Sauermann and Franzoni, 2015a), iNaturalist NZ is no exception. From the 15,699 observers that have contributed at least one observation to iNaturalist NZ, the top 10 observers have shared 22% (175,064) of the observations. 75% of observers have shared fewer than 10 observations, and 35% of observers have only shared 1 observation. Furthermore, the top 10% of observers have shared 90% of the observations, the next 10% of observers, shared 5% of the observations, and the remaining 80% of observers shared the remaining 5%

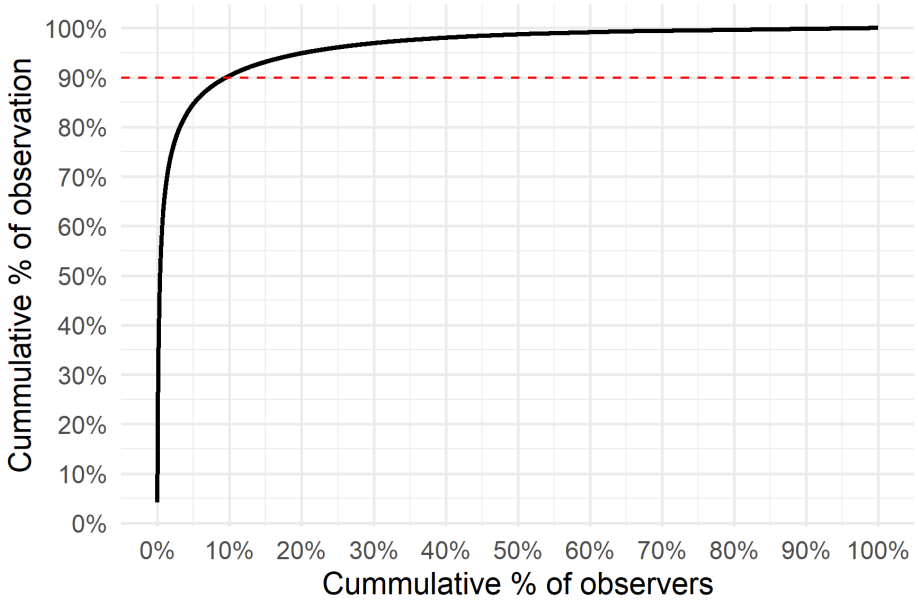


Figure 2.3: **A small number of observers share the majority of the iNaturalist NZ observations.** The top 10% of observers shared 90% of the observations between January 2013 and December 2019. The next 10% of observers, shared 5% of the observations, and the remaining 80% of observers shared the remaining 5% of observations.

of observations, Figure 2.3.

Observation locations

Figure 2.4 shows a density plot of observation locations throughout NZ. Note, for mapping ease we did not plot the observations from the Chatham Islands. The majority of observations are near the major urban areas within New Zealand (Auckland, Wellington, Christchurch, Dunedin).

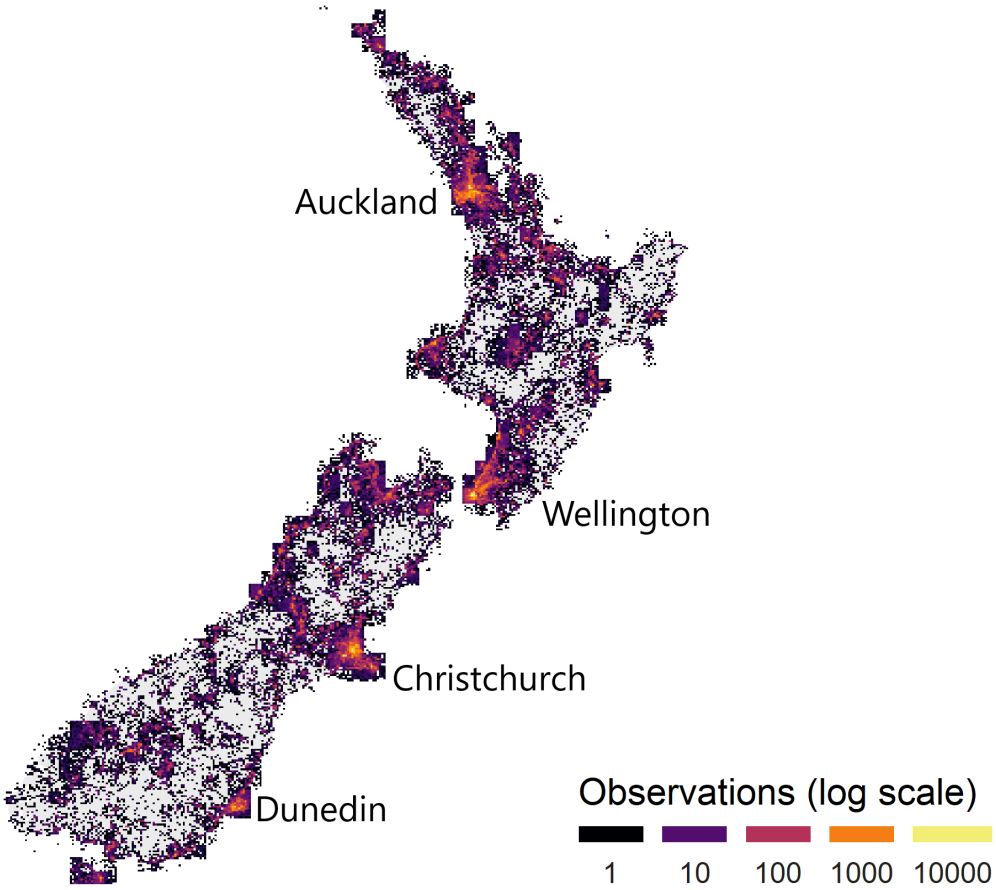


Figure 2.4: **The majority of iNaturalist NZ observations are recorded near major cities in New Zealand.** Density plot of all observations shared to iNaturalist NZ between January 2013 and December 2019.

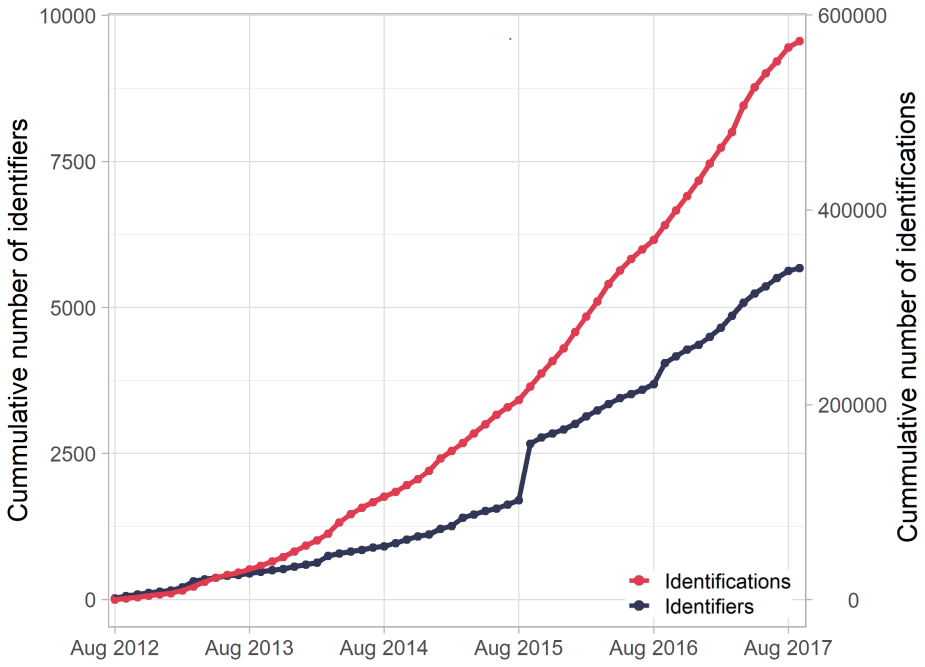


Figure 2.5: **The number of identifiers and identifications on iNaturalist NZ increased steadily between August 2012 and July 2017.** Cumulative number of identifiers and identifications on iNaturalist NZ between August 2012 and July 2017. 5,673 identifiers contributed 573,523 identifications to iNaturalist NZ between August 2012 and July 2017.

Identifiers and identifications

Between August 2012 and July 2017, there were 259,770 identifications shared to iNaturalist NZ that received 573,523 identifications by 5,673 identifiers. Figure 2.5 shows how the number of identifications and identifiers have increased over this time period. The first identification for the vast majority (90%) of observations is by the participant that shared the observation.

Identifications per observations

The majority of observations (83%) require two or three identifications to become research grade, Figure 2.6. 73% of the observations in the casual grade have had one observation and 74% of observations in the "needs ID" grade have had one or two identifications. In all the quality grades there are some observations that have had more than 3 identifications.

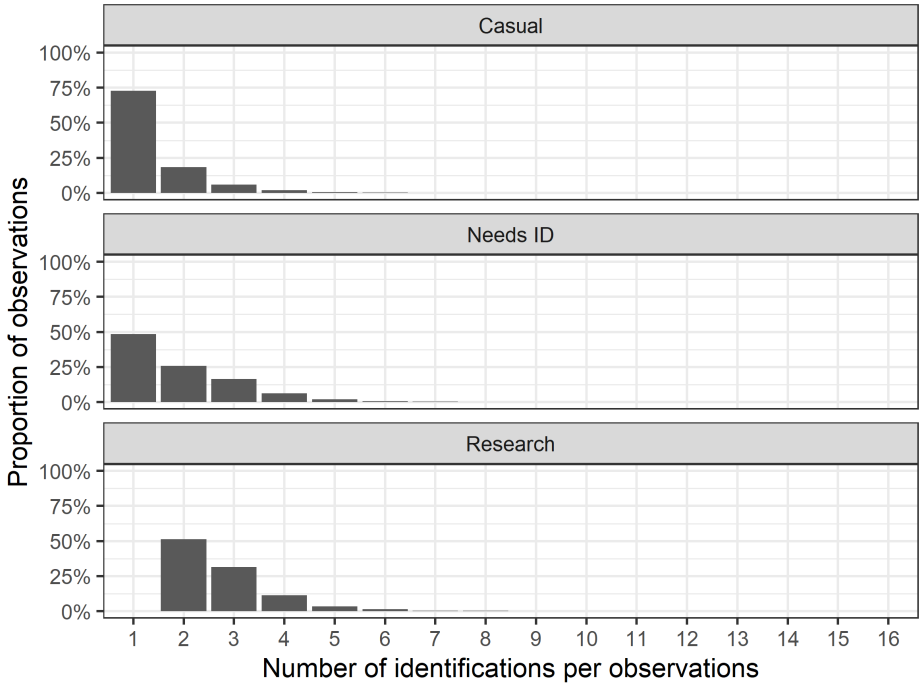


Figure 2.6: **Most observations take 2-3 identifications to reach research grade.** Most observations in casual grade have had one identification. Most observations in "needs ID" grade have had 1-2 identifications. However, in all quality grades there are a number of observations with more than three identifications.

Identifier contributions

Similar to the observations, a small proportion of the identifiers do the bulk of the identifications. From the 5,673 identifiers that have contributed at least one identification to iNaturalist NZ, the top 10 identifiers have done 37% (213,003) of the identifications. 76% of identifiers have shared fewer than 10 identifications, and 36% of identifiers have only shared 1 identification. Furthermore, the top 10% of identifiers have shared 95% of the identifications, Figure 2.7.

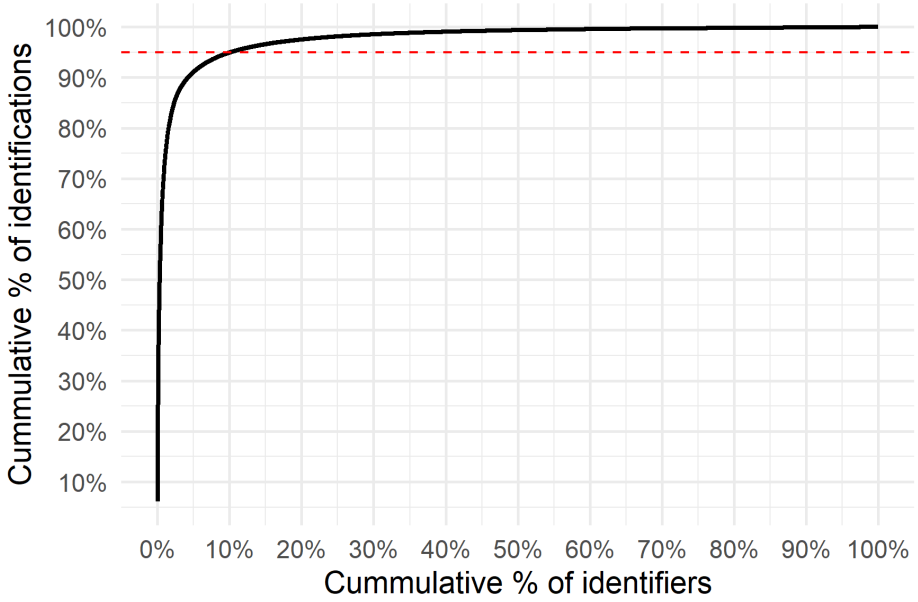


Figure 2.7: **A small number of identifiers contribute the majority of the iNaturalist NZ identifications.** The top 10% of identifiers shared 95% of the identifications between August 2012 and July 2017.

Contributions to science and society

Notable discoveries

iNaturalist NZ is home to a large number of observations of a large range of taxa. Furthermore, some exceptionally important observations of new taxa discoveries or rediscoveries in New Zealand have been shared to iNaturalist NZ. For example, in April 2016 a new species of large mushroom, *Asproinocybe daleyae* was first observed on iNaturalist NZ (Lebel et al., 2020). In May 2016 the first New Zealand incursion of the weed great willowherb, *Epilobium hirsutum*, was discovered in North Canterbury and reported on iNaturalist NZ. Biosecurity New Zealand was notified about this observation and it triggered a biosercurity response from the Ministry of Primary Industries. This provides a great example of how the iNaturalist NZ community can both alert the authorities to new biosecurity incursions and also provide valuable widespread surveillance during an eradication attempt. On the iNaturalist NZ website there is a maintained project of all the important discoveries and rediscoveries shared on iNaturalist NZ (https://inaturalist.nz/projects/new-zealand-discoveries).

Events

iNaturalist NZ has hosted and collaborated on many projects. For example, in 2019 NZBRN collaborated with the Christchurch City Council to enter the Christchurch district into the global City Nature Challenge via an iNaturalist NZ project. In the four days of the 2019 challenge, 323 people in Christchurch made 17,592 observations of 2,360 species, as identified by 458 identifiers. iNaturalist NZ collaborated with the annual New Zealand Garden Bird Survey, operated by Manaaki Whenua (Landcare Research). The Garden Bird Survey runs for a week and encourages people to spend an hour in their garden counting all of the birds they see or hear. NZBRN helps to promote the event and encourages people to confirm their identifications by uploading photos or sound recordings to iNaturalist NZ. Another important and long running collaboration is The Great Kererū Count, and below we explore this in more detail.

The Great Kererū Count

Kererū *Hemiphaga novaeseelandiae* are an iconic New Zealand forest pigeon and the only pigeon endemic to New Zealand. The Great Kererū Count is an annual citizen science project that runs for up to 2 weeks during spring in New Zealand. Participants are asked to share sightings of kererū through either chance encounters or timed surveys. The Great Kererū Count uses iNaturalist NZ as a platform to share and gather observations. The Great Kererū Count has been running since 2014 and aims are to

encourage public participation in a meaningful science project, as well as gather scientific data about the nationwide distribution and population of kererū to help promote healthy and abundant populations of kererū.

Kererū are the most observed species on iNaturalist NZ and in the 7 years from 2013 - 2019 there have been 8404 research grade kererū observations by 1886 observers. Over half of these observations have occurred during one of the 6 Great Kererū Counts. Table 2.3 shows the breakdown of kererū observations on iNaturalist NZ by year, and Great Kererū Count, versus other times of the year.

Year	Great Kererū Count observations	Other Kererū observations	Total Kererū observations	Great Kererū Count observers	Other Kererū observers	Total Kererū observers
2013	NA	127	127	NA	45	45
2014	34	221	255	17	70	77
2015	694	517	1211	312	218	496
2016	515	702	1217	182	178	332
2017	864	768	1632	194	263	413
2018	1318	927	2245	222	340	510
2019	948	769	1717	201	354	511
Total	**4373**	**4031**	**8404**	**971**	**1121**	**1886**

Table 2.3: **Over half of the kererū observations on iNaturalist NZ have occured during a Great Kererū Count.** In the 7 years of iNaturalist NZ data from 2013 - 2019 there have been 8404 kererū observations by 1886 observers. The first Great Kererū Count took place in 2014.

The Great Kererū Count as an iNaturalist NZ observer recruiter

During the 6 years that the Great Kererū Count has been running 971 observers have shared a kererū observation to iNaturalist NZ. 733 (75%) of these observers were first time iNaturalist NZ observers. 359 (37%) of these observers contributed multiple observations to iNaturalist NZ. 135 (14%) of these observers contributed an observation to iNaturalist NZ outside of a Great Kererū Count period and 90 (9%) of the Great Kererū Count observers contributed an observation of something other than a kererū. This flow of how some Great Kererū Count observers continue to be active iNaturalist NZ observers is illustrated in Figure 2.8.

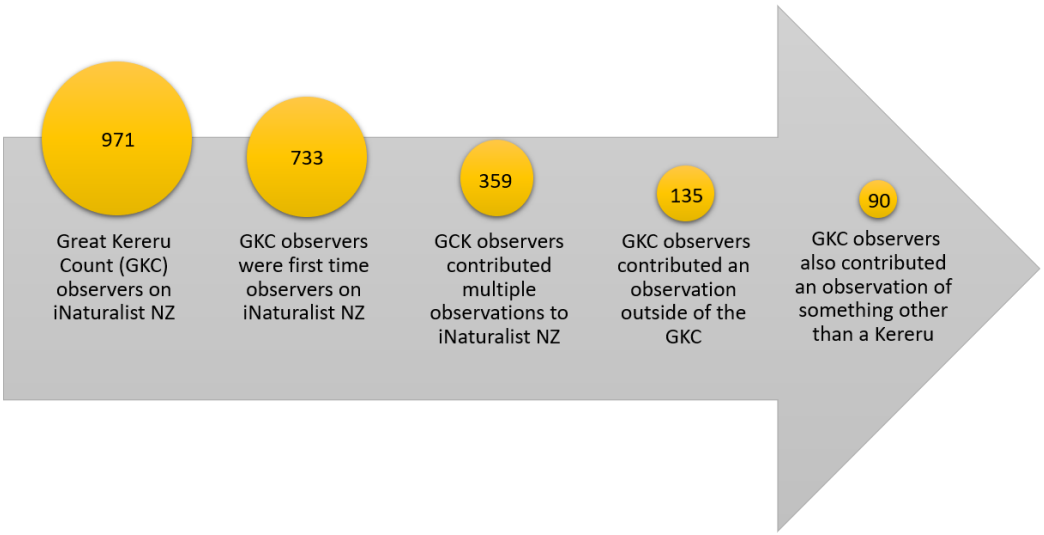


Figure 2.8: **Approximately 9% of Great Kererū Count observers that use iNatuaralist for the first time during the Great Kererū Count continue to use iNaturalist NZ to share observations of other taxa.** This diagram shows how many of the 971 Great Kererū Counter observers on iNaturalist NZ are new recruits to iNaturalist NZ and if they contribute to iNaturalist NZ outside of the Great Kererū Count.

Scientific uses of iNaturalist NZ data

As well as providing a community for participants to share and learn about nature there are many scientific questions iNaturalist NZ may be able to answer. iNaturalist NZ contains three broad types of data of value to researchers. Firstly, there is the extensive database of species occupancy data with iNaturalist NZ (now the largest online source of data on species distributions for many New Zealand taxa). Secondly, there is the large number of annotations on observations, including when different life stages are active, what is flowering when, and which species interact. Thirdly, iNaturalist NZ often provides the only photos on the internet of rarer and more obscure New Zealand native species.

iNaturalist NZ observation data is being used in a variety of research projects. For example, data from iNaturalist NZ have been used by a Victoria University based project on lizards in New Zealand urban areas, and an Otago University based project on the threatened NZ mistletoe, *Tupeia antarctica*. A collaboration between the University of Canterbury and Scion has also been using identified images of pest insects on iNaturalist to train better image recognition technology for border biosecurity (New Zealand bio-recording network trust, 2020).

iNaturalist NZ has been an important collaborator in the response to myrtle rust

since it arrived in New Zealand in 2017. Myrtle rust is an invasive pathogen that attacks trees in the Myrtaceeae family, which includes *Eucalyptus* and a number of native trees. The iNaturalist NZ API was used to directly enter myrtle rust observations into iNaturalist NZ from the myrtle rust reporter application that was built in 2017. Further to this, in 2019 iNaturalist NZ observation data of susceptible native New Zealand trees was shared with Manaaki Whenua - Landcare Research. They combined this data with other plant datasets and climate datasets to make projections about which forests in New Zealand are most vulnerable to being badly affected by myrtle rust.

Final remarks

Throughout this chapter we have provided an overview of the iNaturalist NZ citizen science project, the breadth of the data on iNaturalist NZ, and some highlights of the contributions the project is making to science and society in New Zealand. Throughout the years that iNaturalist NZ has been operating, it has seen steady growth in the number of participants and taxa observed. In the remainder of this thesis we explore methods to improve the scientific usefulness of this expansive dataset. We begin by exploring how the identification process may be optimised with a Bayesian approach (Chapters 3a and 3b). We then investigate if the iNaturalist NZ dataset can be used to reliably detect biological changes in species abundances (Chapters 4a, and 4b). Finally, we explore the different types of observation behaviour that may be at play within iNaturalist NZ, Chapter 4c. Throughout these subsequent chapters, further analysis of iNaturalist NZ data is provided as it becomes relevant.

Chapter 3a

Including user accuracy in classification based citizen science projects

Classifications,
With user accuracy,
beat a simple vote.

Abstract

Image classification citizen science projects are very common and an effective way of classifying a large number of unknown images. Citizen science projects so far often rely on the "wisdom of the crowd" through majority vote methods to produce accurate classifications and assume all volunteer citizen scientists have equal ability.

Initially we assume we have perfect knowledge of user accuracy, and we include this knowledge in the collective classification process with Bayes' formula. We find that including user accuracy in the classification decision process improves classification accuracy compared to a majority vote method.

We introduce a two-stage classification method, where we initially estimate user accuracies in a testing stage and then use Bayes' theorem to include this information in the classification process. In the testing stage the user classifies some images with known ground truth. We show there is a trade-off on classification accuracy of distributing user responses between the testing stage and the classification stage.

We explore the independent Bayesian classifier combination method. This methods estimates both user accuracy and includes this information in the collective classification process when there are no images with known ground truth.

We show the value of including an estimate of user accuracy in a citizen science collective classification processes. However, obtaining an estimate of user accuracy is often not simple.

Introduction

Image classification citizen science projects are very common and an effective way of classifying a large number of unknown images. Some common examples are where participants identify galaxy types on Zoouniverse projects, land types on Geowiki, or species on iNaturalist NZ. Although computer image recognition has advanced a long way from Google's breakthrough in 2012 of training their AI to recognise cats in YouTube videos, there is still a need for image classification by humans. This is because computers may miss interesting features, and for many citizen science projects the image recognition algorithms are not adequate. Citizen science and computer image recognition may also complement each other. For example, citizen science image classification projects produce valuable training datasets for computer image recognition algorithms, whilst, computer image recognition algorithms are being integrated into classification based citizen science projects to suggest possible classifications to users. For example, in 2017, iNaturalist integrated a computer image recognition algorithm into their observation sharing platforms, which, where possible, suggests the species of the uploaded images. Training data for the algorithm was taken from iNaturalist images for species that have at least 20 research grade observations.

The citizen science image classification process, shown in Figure 3a.1, normally involves users viewing images and indicating what they believe the identity of the images are. Multiple users identify each image resulting in the citizen science project building up a database of user identifications per image. It is a fundamental feature of classification based citizen science projects to make classifications based on multiple user responses rather than the knowledge of an expert. Many classification based citizen science projects use majority vote to combine individual user responses into a collective classification. This has the potential to introduce errors due to variations in users ability to identify species or events (Bird et al., 2014). However, group judgements have long been noted to be able to be more accurate than individual decisions. This concept dates back to the Condorcet Jury Theorem (Condorcet, 1785). The Condorcet Jury Theorem states that the probability of a group making the correct decision using majority vote tends to 100% as the size of the group increases provided each individual's probability of being correct, θ_i, is greater than 50%. Similarly, the probability of the group making a correct decision tends towards 0% if $\theta_i < 0.5$. Owen et al. (1989) generalised the Condorcet Jury Theorem to: if the mean probability, $\bar{\theta}$, of the group is greater than 50%, then as the size of the group increases the probability of making the correct decision will tend towards 100%. The improved accuracy of judgements from the wisdom of the crowd effect has been shown in many simple examples. For example, Treynor (1987) asked 56 students to estimate the number of jelly beans in a jar and found that the average guess of the group, 871, was very close to the actual amount,

850, and better than 98% of individual guesses.

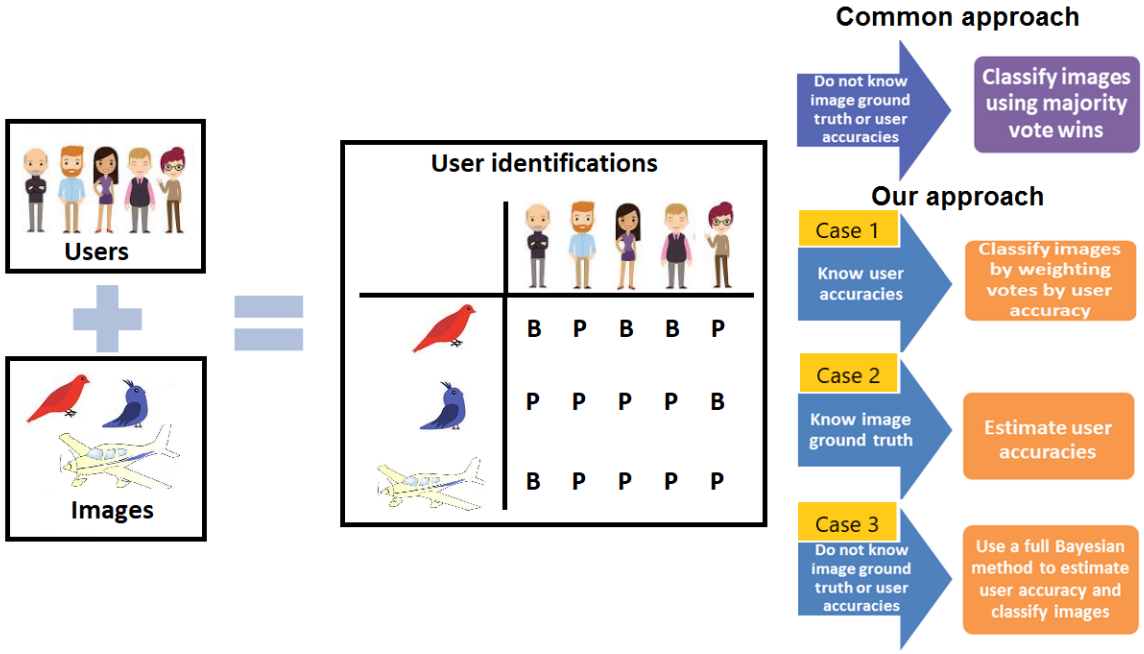


Figure 3a.1: **Citizen science users view images and indicate what they believe the identity of the image is. These individual images are then combined to collectively decide the image identities.** P = plane, B = bird. Commonly collective decisions are made using majority vote. Our approach is broken down into three cases as outlined in the image.

Despite the simplicity and improved accuracy of using majority vote to collate user identifications there are concerns with the method. In particular, each users' identifications are equally weighted even though there are often a wide range of user abilities. For example, iNaturalist NZ uses majority vote to make image classifications and we know there are a range of identifier profiles on iNaturalist NZ, e.g. expert ecologists, enthusiast nature watchers, school children, and tourists. It is unlikely that all these types of users have the same accuracy and therefore we may want to weight their responses unequally. Other approaches to enhance the quality of collective classifications have included acquiring information on users' confidence in their vote and then using this information to weight the votes by placing more emphasis on the votes with high confidence (See et al., 2013). This method is problematic as some users overestimate their abilities while other users underestimate their abilities (Kruger and Dunning, 1999). An improvement of this basic confidence weighted method may be using the "surprisingly popular vote" approach that makes a collective classification based on the class that is more popular than predicted by the users (Prelec et al., 2017). More recent approaches have involved using estimates of users' accuracies to reduce the weighting of the least accurate users and magnify the weighting of the most accurate users (Foody

et al., 2018).

Throughout this chapter *accuracy* can be used in many ways. Therefore, for clarity, we define the following stricter definitions of accuracy:

- *User accuracy*, denoted as θ, is the probability that a user correctly identifies an image.
- *Classification accuracy*, denoted as α, is the proportion of images correctly classified by the group of users.

Other easily confused terms are *identifications* and *classifications*. For this chapter we make the following distinctions:

- *Identification* refers to an individual user's response of the identity of an image.
- *Classification* refers to a judgement based on a collection of individual user identifications.

It is common for citizen science projects that utilise user accuracies in the collective classification decisions to incorporate a user testing stage in the project. Testing stages normally involve the user identifying a range of images that have a known ground truth and therefore provide the researches with a dataset to estimate user accuracy. In the absence of images with known ground truth, a ground reference data set or validation points are often produced by experts in the field (Foody et al., 2013; Gengler and Bogaert, 2016). Many approaches have been used to estimate user accuracy from the user identifications and the corresponding ground truth or expert opinions. For example, a simple method is to calculate the proportion of images the user correctly identified (Foody et al., 2013). However, this simple approach is problematic as it is common to obtain a 0 or 1 estimate for user accuracies when samples are small. An improvement is to use a Bayesian estimator that generalises this simple proportion of correct identifications to include prior information about the users' accuracy and therefore avoid the problematic case of estimating user accuracy as 0 or 1. More advanced methods to estimate user accuracy include latent class analysis and Bayesian statistics. Latent class models describe the relationship between observed variables, user identifications, and latent variables, e.g. the true class of an image. If the model is an adequate fit the parameters may be used to indicate the quality of identifications made by each user. For example, Foody et al. (2018) fitted a latent class model to user identifications of land cover types and used the model parameters as a measure of user accuracy. Bayesian statistical methods may also be used to estimate user accuracy and this is the technique we will focus on in this chapter. For example, Gengler and Bogaert (2016) used validation points obtained from an expert to update the prior probability distributions of user accuracies to obtain a posterior probability distribution for user

accuracy.

In this chapter we work through three different cases, shown in Figure 3a.1, to build up two additional methods to majority vote to collectively classify images.

Case 1

In case 1 we assume we have perfect knowledge of all user accuracies and use Bayes' theorem to include this information in the collective classification process. We initially start by assuming all users have the same accuracy. We compare the classification accuracy from classifying images with Bayes' theorem against the classification accuracy of using majority vote. As expected, including user accuracy in the classification decision improves classification accuracy compared to majority vote. This motivates us to explore how we would use Bayes' theorem to include user accuracy in the classification decision in the case of multiple classes of users, e.g. amateurs and experts, and the case of all users having different user accuracies.

Case 2

Motivated by the improvements in classification accuracy by including user accuracy in classification decisions we explore how we would estimate user accuracy in a testing stage where image ground truth is known. We introduce a two-stage classification method where we initially estimate user accuracies in a testing stage and then using Bayes' theorem (as in case 1) we use these user accuracy estimations to classify images in a classification stage. We explore the trade-off on classification accuracy of distributing user responses between the testing and classification stages. We highlight the risk on classification accuracy of reducing the number of identifications per image in the classification stage.

Case 3

Motivated by the trade-off results from case 2, and with no testing stage on iNaturalist NZ, we begin building this case by reviewing previous literature on estimating user accuracies in citizen science projects in the absence of a testing stage or any images with known ground truth. We focus on a Bayesian classifier combination method and outline a suitable Gibbs sampling algorithm to estimate the model that best combines our image classification methods from case 1 and user accuracy estimation methods from case 2.

Bayesian background

Bayesian statistics is an approach to data analysis and parameter estimation based on Bayes' theorem. Bayesian methods are based on the idea that before any experiments are run or any data is gathered there is usually some information about the unknown parameter and therefore a prior belief about the possible parameter value. Typically a Bayesian approach involves: (1) capturing available knowledge about a given parameter via the prior distribution; (2) determining the likelihood function using the information about the parameters available in the observed data; (3) combining both the prior distribution and the likelihood function using Bayes' theorem to obtain the posterior distribution. The posterior distribution balances prior knowledge with observed data and reflects one's updated knowledge. The posterior distribution is used to conduct inferences.

Prior probability distributions express the knowledge of the parameter distribution before the experiment. Prior distributions are often based on previous experiments or on basic knowledge of the parameter space, e.g. an accuracy parameter has to lie between 0 and 1. If no information is known, uninformative prior distributions may be used, e.g. a uniform distribution for a bounded range such as a probability. The use of prior distributions in Bayesian statistics is a key difference from frequentist statistics and is often the point of debate between the two philosophies due to prior distributions being subjective. Different statisticians may use different prior distributions, and therefore sensitivity analysis of the posterior distribution to different prior distributions is important.

A prior distribution and likelihood function are said to be conjugate when the resulting posterior distribution is the same type of distribution as the prior distribution. There are many conjugate models. One of relevance to this chapter is the Beta-Binomial model where a Beta prior distribution and Binomial likelihood model result in a Beta posterior probability distribution. Many real-world Bayesian models are too complex to produce analytical and tractable posterior distributions due to difficult multi-dimensional integrals. Historically Bayesian statistics was largely restricted to conjugate models for computational simplicity. However, due to recent advances in computing power and techniques such as Markov chain Monte Carlo (a method of calculating numerical approximations of multi-dimensional integrals) most posterior distributions can now be estimated numerically.

Case 1: Collective classification decisions with known user accuracy

In this case we outline a simple example of users classifying images of bird and planes. Initially we assume:

- Users are independent.
- There are only two image classes: birds, and planes.
- We have perfect knowledge of every user's accuracy.
- User accuracies are the same for both image classes.
- All images are of equal difficulty to identify.
- User accuracy does not change over the number of identifications the user has done.

For each image we use Bayes' formula to calculate the posterior probability that the image belongs to each class given the user identifications for the image and class prevalences. An image is classified as the class with the highest posterior probability. In the case that both classes have equal posterior probabilities we will classify the image as a bird. Initially we assume all users have the same user accuracy. We compare the classification accuracy from using Bayes' formula to using a simple majority vote classification rule for a range of user identifications per image and user accuracies. We then extend our analysis to the case that users are either amateurs or experts and then generalise the analysis to all users having a unique identification accuracy.

Common user accuracy

First, we assume all users have the same user accuracy, θ. We investigate how the expected classification accuracy changes as more users identify images or user accuracy varies. We also consider the optimal user accuracy and number of identifications per image given a trade-off between these two inputs.

Probability an image is a bird given user responses

First, we will calculate the Bayesian posterior probability that the true identity of an image is a bird given the user identifications of the image. Let t define the true identity of the image:

$$t = \begin{cases} 1, & \text{if image is a bird} \\ 0, & \text{if image is a plane.} \end{cases}$$

Prior distribution. Before we obtain any responses from users the prior distribution for the unknown t is expressed as $Pr(t = 1) = \rho_B$ and $Pr(t = 0) = \rho_P$ where ρ_B, and ρ_P are the bird and plane prevalences, respectively. Since, we only have bird and plane images $\rho_B + \rho_P = 1$.

Data model and likelihood. The data used to update the prior information consists of the identification responses from the users. Let $x^{(k)}$ be the response from user k where,

$$x^{(k)} = \begin{cases} 1, & \text{if user } k \text{ says image is a bird} \\ 0, & \text{if user } k \text{ says image is a plane.} \end{cases}$$

Suppose we have one image that has been classified by user one as $\{x^{(1)}\}$. We have the following likelihood function:

$$\begin{aligned} Pr(x^{(1)}|t = 1) &= \theta^{x^{(1)}} \cdot (1 - \theta)^{1-x^{(1)}} \\ Pr(x^{(1)}|t = 0) &= (1 - \theta)^{x^{(1)}} \cdot \theta^{1-x^{(1)}}. \end{aligned}$$

Thus, the probability that a user correctly classifies the image as a bird is θ and the incorrectly classifies the image as a bird (when it is a plane) is $1 - \theta$.

Directed acyclic graphs (DAGs) are commonly used in Bayesian statistics to visualise the relationships between parameters in the model. A DAG of case 1 is shown in Figure 3a.2. Square nodes are known parameters and circular nodes are unknown parameters. Directed edges indicate a conditional dependence between two nodes.

Posterior distribution. Using Bayes' Theorem we combine the information from the user response with the prior probability to calculate the posterior probability the true identity of the image is a bird given the user response:

$$\begin{aligned} Pr(t = 1|x^{(1)}) &= \frac{Pr(x^{(1)}|t = 1) \cdot Pr(t = 1)}{Pr(x^{(1)}|t = 1) \cdot Pr(t = 1) + Pr(x^{(1)}|t = 0) \cdot Pr(t = 0)} \\ &= \frac{\theta^{x^{(1)}} \cdot (1 - \theta)^{(1-x^{(1)})} \cdot \rho_B}{\theta^{x^{(1)}} \cdot (1 - \theta)^{(1-x^{(1)})} \cdot \rho_B + (1 - \theta)^{x^{(1)}} \cdot \theta^{(1-x^{(1)})} \cdot (1 - \rho_B)}. \end{aligned}$$

We extend the likelihood function and posterior distribution to include identification responses from K independent users where all users have common user accuracy θ.

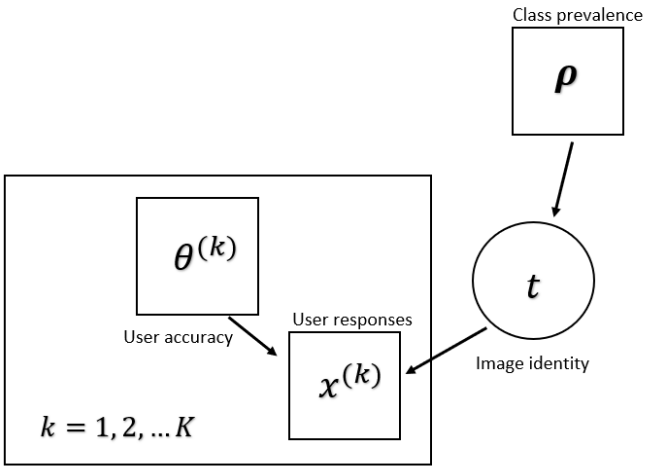


Figure 3a.2: **Model for case 1 takes known user accuracy, user responses and class prevalences to infer true image identity.** Circular nodes are unknown variables and square nodes are known parameters. Plates are over users, $k = 1, 2, ...K$. Directed edges indicate a conditional dependence between two nodes.

If the K users identified an image respectively as $\{x^{(1)}, ..., x^{(K)}\}$ using the binomial distribution we have the following likelihood functions for n of the K users saying an image is a bird:

$$\begin{aligned} Pr(\sum_{k=1}^{K} x^{(k)} = n | t = 1) &= L(n, K, \theta) \\ Pr(\sum_{k=1}^{K} x^{(k)} = n | t = 0) &= L(n, K, 1-\theta). \end{aligned}$$

Where,

$$L(n, K, \theta) = \binom{K}{n} \theta^n \cdot (1-\theta)^{K-n}.$$

Therefore our posterior probability that the image identity is a bird given n out of K users said it was a bird is:

$$Pr(t = 1 | \sum_{k} x^{(k)} = n) = \frac{\theta^n \cdot (1-\theta)^{K-n} \cdot \rho_B}{\theta^n \cdot (1-\theta)^{K-n} \cdot \rho_B + (1-\theta)^n \cdot \theta^{K-n} \cdot (1-\rho_B)}. \quad (3a.1)$$

Image classification based on user responses

We will use the posterior probability of an image being a bird to decide the collective classification of an image. We will let $\hat{t}$ define the collective classification of an image:

$$\hat{t} = \begin{cases} 1, & \text{if image is classified as a bird} \\ 0, & \text{if image is classified as a plane} \end{cases}$$

$$\hat{t} = \begin{cases} 1, & \text{if } Pr(t = 1 | \sum_k x^{(k)} = n) \geq 0.5 \\ 0, & \text{if } Pr(t = 1 | \sum_k x^{(k)} = n) < 0.5. \end{cases}$$

From Equation 3a.1 we see that we will classify an image as a bird if;

$$\begin{aligned} \theta^n \cdot (1-\theta)^{K-n} \cdot \rho_B &\geq (1-\theta)^n \cdot \theta^{K-n} \cdot (1-\rho_B) \\ n &\geq \frac{K}{2} - \frac{1}{2} \frac{\log(\frac{\rho_B}{1-\rho_B})}{\log(\frac{\theta}{1-\theta})} \qquad \text{for } \theta > 0.5, \end{aligned} \tag{3a.2}$$

where n is the number of users that said the image was a bird.

For simplicity we will refer to the RHS of equation 3a.2 as n^*, i.e. the Bayesian decision threshold. Therefore,

$$\hat{t} = \begin{cases} 1, & \text{if } n \geq n^* \\ 0, & \text{if } n < n^*. \end{cases}$$

Classification accuracy

We will let Z indicate if we correctly classify an image:

$$Z = \begin{cases} 1, & \text{if } \hat{t} = t \\ 0, & \text{Otherwise.} \end{cases}$$

Then the classification accuracy, α, of classified images is defined as the expected value of Z:

$$
\begin{aligned}
\alpha &= E(Z) \\
&= Pr(Z = 1) \\
&= Pr(\hat{t} = t) \\
&= Pr(\hat{t} = 1 \cap t = 1) + Pr(\hat{t} = 0 \cap t = 0) \\
&= Pr(\hat{t} = 1 | t = 1) \cdot Pr(t = 1) + Pr(\hat{t} = 0 | t = 0) \cdot Pr(t = 0) \\
&= Pr(Z = 1 | t = 1) \cdot Pr(t = 1) + Pr(Z = 1 | t = 0) \cdot Pr(t = 0) \\
&= E(Z | t = 1) \cdot Pr(t = 1) + E(Z | t = 0) \cdot Pr(t = 0) \\
&= \rho_B(1 - F(\lfloor n^* \rfloor, K, \theta)) + (1 - \rho_B)F(\lfloor n^* \rfloor, K, 1 - \theta).
\end{aligned}
$$

Where $F(\lfloor n^* \rfloor, K, \theta)$ is the cumulative binomial distribution function with parameters θ and K.

Figure 3a.3 shows how classification accuracy, α, varied over a range of user accuracies, θ, and number, n, of identifications per image for $\rho_B = 0.6$ and $\rho_P = 0.4$. As expected, with higher individual user accuracy, θ, fewer identifications, n, per image are needed to reach a classification accuracy, α, of 95% as shown by the 0.95 iso-accuracy curve in Figure 3a.3. When user $\theta = 0.5$, $\alpha = \rho_B$ regardless of how many users identify each image. Whereas, when $\theta = 1$ it only takes one identification to reach $\alpha = 100\%$. As expected, if $\theta > 0.5$ classification accuracy, α, will tend towards 100% as the number of identifications per image increases. This result matches the Condorcet Jury Theorem that states, provided individual accuracy is more than 50%, the probability of the group making the correct decision will tend towards 1 as the size of the group increases (Condorcet, 1785).

Given a fixed budget constraint, the results in Figure 3a.3 can be used to perform a statistical power analysis to find the number of users and user accuracy to maximise classification accuracy for a given bird prevalence. For example, given we have a budget that can afford either 1 user with user accuracy 0.9, 20 users with user accuracy 0.5, or any equivalently priced combination in between. We can use this information to draw an iso-cost line on our classification accuracy figure as shown by the dotted line in Figure 3a.3. To maximise classification accuracy along this iso-cost line we use the method of Lagrange multipliers (Lagrange, 1804) and find the point on the iso-accuracy curve furthest from the origin that is tangent to the iso-cost curve, shown in Figure 3a.3 as a solid triangle. For any budget constraint we can find an optimal combination of number of users and user accuracy.

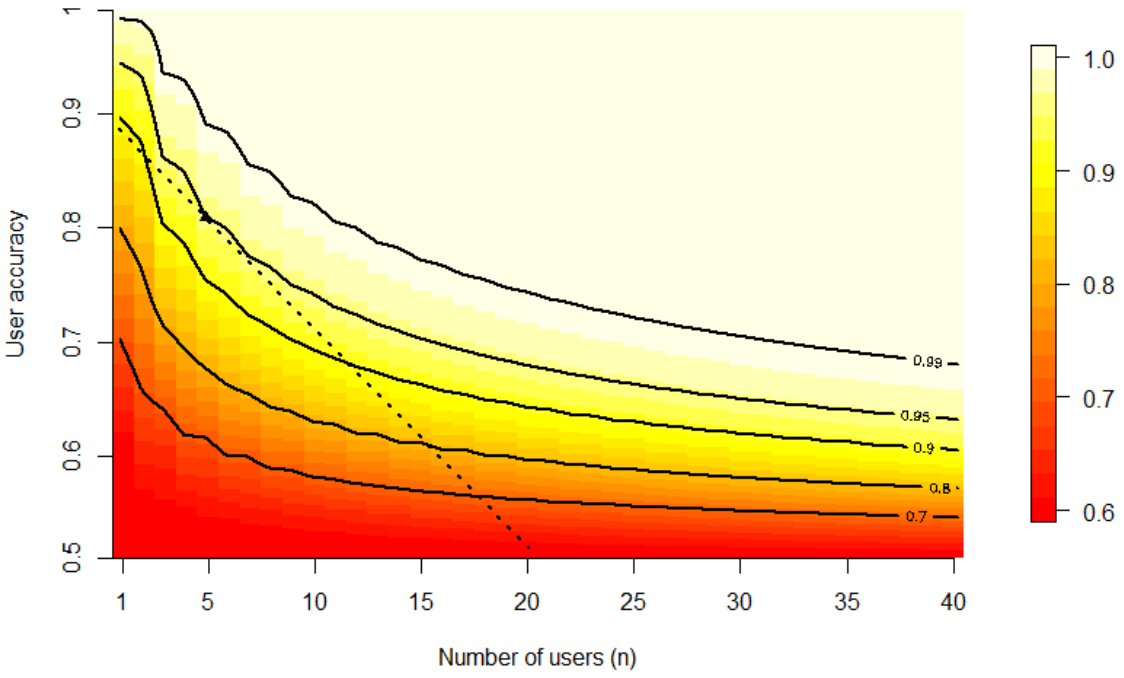


Figure 3a.3: **Higher user accuracy means fewer users (lower data redundancy) are required to identify an image to reach a fixed classification accuracy.** Background colour is classification accuracy as a function of user accuracy, θ, and data redundancy, n, for $\rho_B = 0.6$. Solid lines are iso-accuracy curves. The dotted line is an example iso-cost line. Maximum classification accuracy for the costs is achieved at 5 users with 0.82 user accuracy, shown as a solid triangle.

Bayes' formula vs. majority vote decision rule

In this subsection we examine when the Bayes' decision threshold differs from the simple majority and compare the classification accuracy of the two classification methods.

Bayes' decision threshold in Equation 3a.2 differs from majority vote by the $\frac{1}{2}\frac{\log(\frac{\rho_B}{1-\rho_B})}{\log(\frac{\theta}{1-\theta})}$ term, i.e. the majority vote threshold is $n \geq \frac{K}{2}$.

In the following cases the two decision thresholds converge, i.e. $\frac{K}{2} - \frac{1}{2}\frac{\log(\frac{\rho_B}{1-\rho_B})}{\log(\frac{\theta}{1-\theta})} \to \frac{K}{2}$:

1. We have no prior information on bird and plane image prevalence and assume $\rho_B = \rho_P = 0.5$, $\frac{1}{2}\frac{\log(\frac{\rho_B}{1-\rho_B})}{\log(\frac{\theta}{1-\theta})} = 0$.

2. As $\theta \to 1$, i.e. users have perfect accuracy, $\frac{1}{2}\frac{\log(\frac{\rho_B}{1-\rho_B})}{\log(\frac{\theta}{1-\theta})} \to 0$.

3. As the number of users, K, increases the importance of the second term, $\frac{1}{2}\frac{\log(\frac{\rho_B}{1-\rho_B})}{\log(\frac{\theta}{1-\theta})}$, of Bayes' decision threshold diminishes relative to the first term, $\frac{K}{2}$.

When Bayes' decision threshold is equal to the majority vote decision threshold both classification methods have the same classification accuracy, shown in Figure 3a.4.

There are many cases when Bayes' decision rule is different from the majority vote decision rule. For example, when birds are the dominant species, $\rho_B > 0.5$ and $0.5 < \theta < 1$ then $\frac{1}{2}\frac{\log(\frac{\rho_B}{1-\rho_B})}{\log(\frac{\theta}{1-\theta})} > 0$ and $n^* < \frac{K}{2}$, i.e., a fewer number of users is required to identify an image as a bird to make a group classification of a bird using Bayes' formula than using majority vote. When $\theta \to 0.5$, i.e. user accuracy tends towards random guessing, and birds are the dominant species (i.e. $\rho_B > 0.5$), $\frac{1}{2}\frac{\log(\frac{\rho_B}{1-\rho_B})}{\log(\frac{\theta}{1-\theta})} \to \infty$ and therefore $n^* \to -\infty$. This means that zero users need to identify an image as a bird to make a group classification of a bird. Effectively, when θ is close to random guessing ($\theta = 0.5$) user identifications are not contributing any value to the classification decision and the Bayes' classification decision is based on always voting for the dominant species. However, due to requiring the decision rules, n^* and $\frac{K}{2}$, to be rounded up to the nearest integer, the majority vote and Bayes decision thresholds are stepwise functions and are equal for more values of ρ_B and θ than just $\rho_B = 0.5$ and $\theta = 1$, shown in Figure 3a.4.

As expected, when the two classification methods have different decision thresholds using Bayes' classification rule always results in a higher classification accuracy than majority vote, Figure 3a.4. However, Bayes' formula achieves this higher classification accuracy by classifying the dominant class more accurately than the minority class. Therefore, if the objective is to maximise the minority class classification accuracy then majority vote would be used.

Varied user accuracy

In reality, it is very unlikely that all users will have the same user accuracy. In this section we now relax the common user accuracy assumption by first considering the case where there are expert and amateur users and then considering the general case where all users have a unique user accuracy.

As in the common user accuracy case we assume all users are independent, there are only two image classes, we have perfect knowledge of every user's accuracy, users have the same classification accuracy for both categories and every image is of equal difficulty to classify.

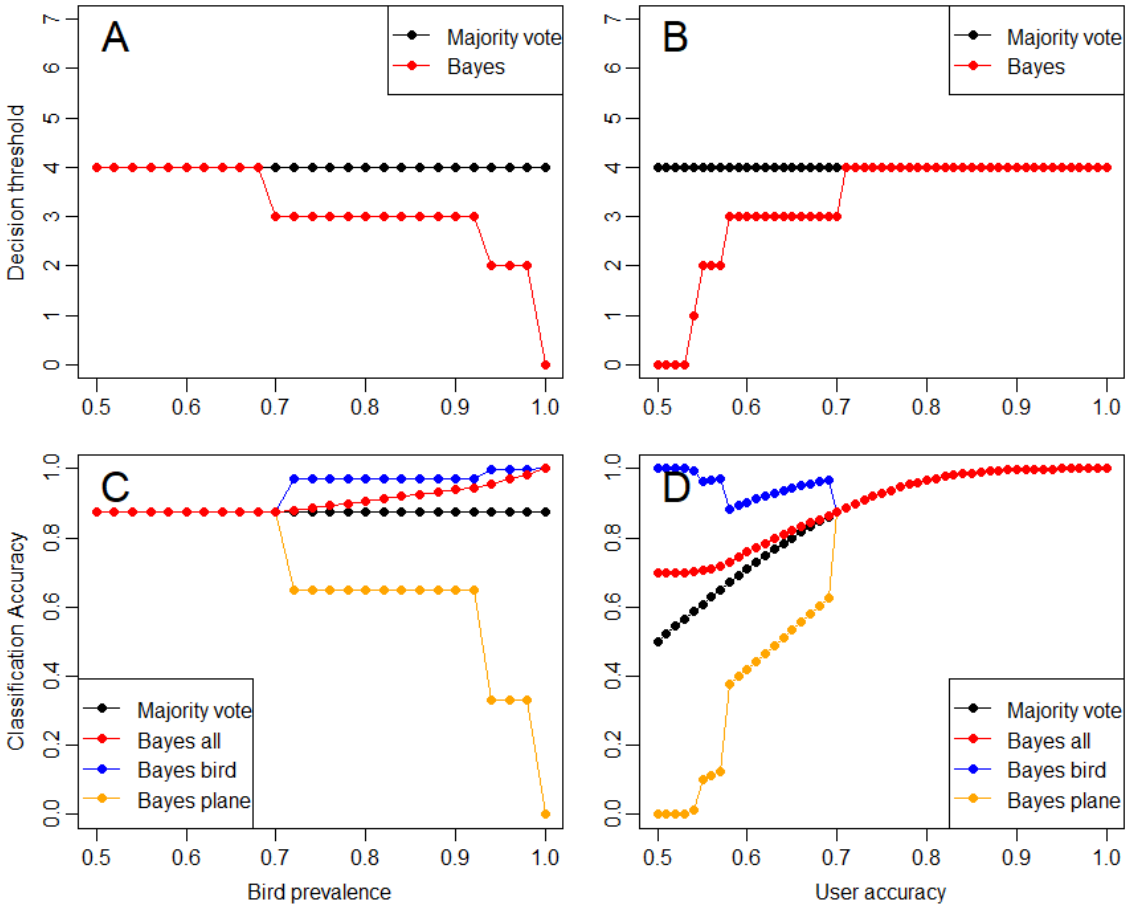


Figure 3a.4: **Using Bayes' formula to make a group classification decision always results in a higher or equivalent classification accuracy compared to using majority vote. A** Bayes' formula and majority vote decision thresholds as a function of bird prevalence for $\theta = 0.7$ and $K = 7$. **B** Bayes' formula and majority vote decision thresholds as a function of user accuracy for $\rho_B = 0.7$ and $K = 7$. **C** Bayes' formula and majority vote classification accuracy as a function of bird prevalence for $\theta = 0.7$ and $K = 7$. **D** Bayes' formula and majority vote classification accuracy as a function of user accuracy for $\rho_B = 0.7$ and $K = 7$.

Amateurs and experts

We consider the case that each user is either an amateur or an expert, i.e. some users are very accurate at identifying images (experts) and the other users (amateurs) are less accurate at identifying images. This is a relevant scenario to consider for citizen science projects like iNaturalist NZ where the users are a mix of qualified ecologists and amateur nature watchers. We assume that all amateurs have the same accuracy, θ_A, and all experts have the same accuracy, θ_E, and $\theta_A < \theta_E$. We assume we have perfect knowledge of who are the amateurs and experts and there are K_A amateurs and K_E experts.

Prior distribution. Our prior distribution for the unknown t is the same as the case where all users have the same user accuracy. That is, $Pr(t = 1) = \rho_B$ and $Pr(t = 0) = \rho_P = 1 - \rho_B$.

Data model and likelihood. The data used to update the prior information consists of the responses from the $K_A + K_E$ users. If the K_A amateur users identified an image respectively as $\{x^{(1)}, ..., x^{(K_A)}\}$ and the K_E expert users identified the image respectively as $\{x^{(1)}, ..., x^{(K_E)}\}$ we have the following likelihood functions for n_A of the K_A amateur users and n_E of the K_E expert users saying an image is a bird:

$$\begin{aligned} Pr(\sum_{k_A=1}^{K_A} x^{(k_A)} = n_A, \sum_{k_E=1}^{K_E} x^{(k_E)} = n_E | t = 1) &= L(n_A, K_A, \theta_A) \cdot L(n_E, K_E, \theta_E) \\ Pr(\sum_{k_A=1}^{K_A} x^{(k_A)} = n_A, \sum_{k_E=1}^{K_E} x^{(k_E)} = n_E | t = 0) &= L(n_A, K_A, 1 - \theta_A) \cdot L(n_E, K_E, 1 - \theta_E) \end{aligned}$$

Posterior distribution. Using Bayes' Theorem we can find the posterior probability that the image identity is a bird given n_A amateurs and n_E experts say the image is a bird from K_A amateurs and K_E experts respectively. As with the common user accuracy case we classify an image as a bird if the posterior probability an image is a bird is greater than or equal to 0.5, i.e. $Pr(t = 1 | \sum_{k_A=1}^{K_A} x^{(k_A)} = n_A, \sum_{k_E=1}^{K_E} x^{(k_E)} = n_E) \geq 0.5$.

Classification accuracy. We calculate the classification accuracy over a range of combinations of amateurs and experts for given amateur and expert accuracies by using simulations of the amateurs and experts identifying images and the resulting image classification based on Bayes' formula to find the proportion of correctly identified images.

Figure 3a.5 shows how classification accuracy varies over different combinations of amateurs and experts. As expected, we found that increasing the number of expert users increases classification accuracy more rapidly than adding additional amateur users. Similarly, to the case of a common user accuracy we can use iso-accuracy curves to show the combinations of amateurs and experts that result in the same classification accuracy. For example, in Figure 3a.5 where amateur accuracy is 0.6 and expert accuracy is 0.85 we see that one expert achieves a classification accuracy of 0.85, whereas, it takes 26 amateurs to achieve 0.85 classification accuracy. We may also perform a power analysis to find the optimal number of amateurs and experts to maximise classification accuracy given a budget constraint, as shown by the solid triangles in Figure

3a.5 where the iso-cost line is tangent to the iso-accuracy curve.

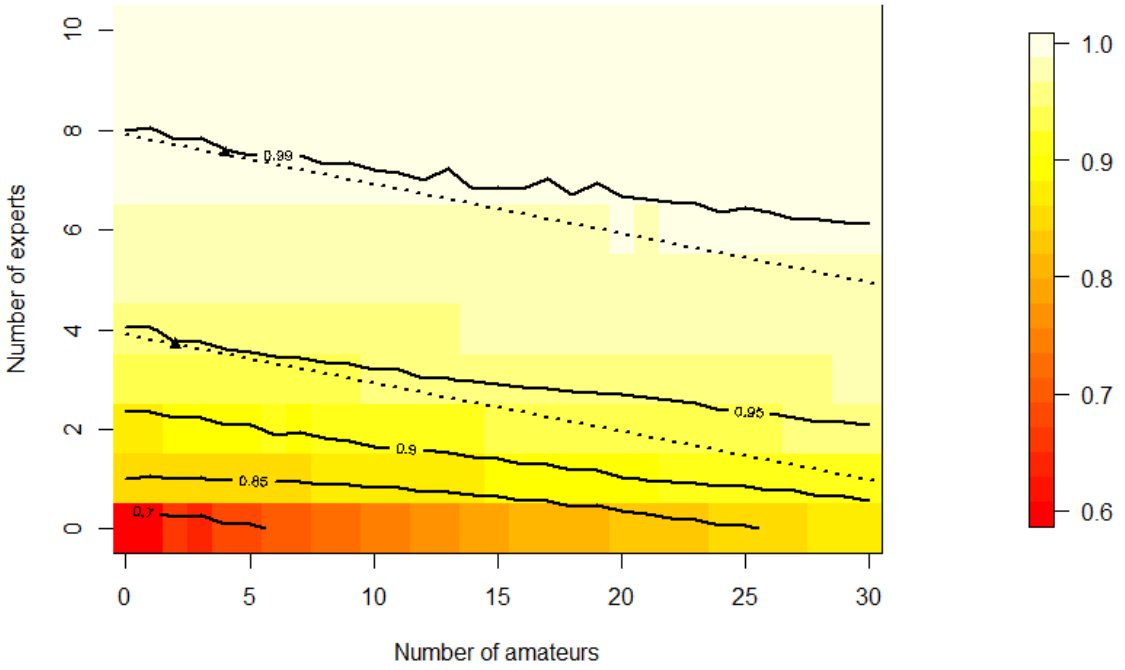


Figure 3a.5: **There is a range of combinations of expert users and amateurs users that achieve the same classification accuracy.** This plot is generated via a simulation where the background colour is classification accuracy as a function of the number of experts and number of amateurs for amateur accuracy of 0.6, expert accuracy of 0.85 and bird prevalence of 0.6. Solid lines are iso-accuracy curves. Dotted lines are Iso-cost lines for the case that 1 expert costs the same as 10 amateurs. Maximum classification accuracy given the budget constraint of 8 experts or 80 amateurs is achieved with 7 experts and 4 amateurs, shown as a solid triangle.

Generalisation to unique user accuracy for all users

We consider the most likely scenario in citizen science projects where all K users have a unique user accuracy $\{\theta^{(1)}, ..., \theta^{(K)}\}$.

Data model and likelihood. If the K users identified an image respectively as $\{x^{(1)}, ..., x^{(K)}\}$ we have the following likelihood functions for the user responses:

$$\begin{aligned} Pr(x^{(1)}, ..., x^{(K)}|t=1) &= \Pi_{k=1}^{K}(\theta^{(k)})^{x^{(k)}} \cdot (1-\theta^{(k)})^{1-x^{(k)}} \\ Pr(x^{(1)}, ..., x^{(K)}|t=0) &= \Pi_{k=1}^{K}(1-\theta^{(k)})^{x^{(k)}} \cdot (\theta^{(k)})^{1-x^{(k)}}. \end{aligned}$$

Posterior distribution. Applying Bayes' Theorem we obtain the following posterior probability that an image is a bird given the responses from the K users:

$$Pr(t=1|x^{(1)},...,x^{(K)}) = \frac{\Pi_{k=1}^{K}(\theta^{(k)})^{x^{(k)}} \cdot (1-\theta^{(k)})^{1-x^{(k)}} \cdot \rho_B}{\Pi_{k=1}^{K}(\theta^{(k)})^{x^{(k)}} \cdot (1-\theta^{(k)})^{1-x^{(k)}} \cdot \rho_B + \Pi_{k=1}^{K}(1-\theta^{(k)})^{x^{(k)}} \cdot (\theta^{(k)})^{1-x^{(k)}}(1-\rho_B)}. \quad \text{(3a.3)}$$

In the case that all users have the same user accuracy, θ, Equation 3a.3 reduces to Equation 3a.1.

Classification accuracy. As with the common user accuracy case we classify an image as a bird if the posterior probability that an image is a bird given the user responses is greater than or equal to 0.5, i.e. $Pr(t=1|x^{(1)},...,x^{(K)}) \geq 0.5$.

Similar to the amateur and expert case, classification accuracy can be calculated by simulating the users identifying a large number of images, making a group classification decision for each image, and then calculating the proportion of images they correctly identified. We can use this analysis to answer questions such as: "Given 5 users with user accuracies $\{\theta^{(1)},...,\theta^{(5)}\}$ identify an image what is the expected group classification accuracy?" or "If we increase one of the user's accuracies how much do we improve the expected group classification accuracy?"

Case 2: Estimating user accuracy when ground truth is known

In case 1 we classified images based on a Bayes' decision rule that is a function of user accuracy, user identifications and class prevalences. In this case we explore how we can use Bayesian statistics to estimate user accuracy when the ground truth of the images are known. It is common for classification based citizen science projects to initially ask users to classify images with known ground, often called a testing stage. As well as providing information on user accuracies, testing stages provide a training opportunity for users. We will refer to the classification of images with known ground truth as the testing stage and with unknown ground truth as the classification stage. In this case we introduce a two-stage classification method where we initially estimate user accuracies in a testing stage and then, using Bayes' Theorem, as in case 1, we use these user accuracy estimations to classify images in a classification stage. We explore the trade-off on classification accuracy of distributing user responses between the testing and the classification stages.

As in case 1, we assume:

- Users are independent.
- There are only two image classes, birds or planes.
- User accuracies are the same for both image classes.
- All images are of equal difficulty to identify.

However, in this case we assume we do not know the user accuracies but we do know the ground truth of the images.

We assume user k has identified $m^{(k)}$ images. The user identifications form a sequence of Bernoulli trials where $y^{(k)}$ is the number of images user k correctly identified. Each trial is independent and exchangeable as we are assuming the user does not improve or worsen their user accuracy over time. Therefore, we can model the number of successes, $y^{(k)}$, user k has from the $m^{(k)}$ trials as a binomial distribution where $\theta^{(k)}$ is the unknown user accuracy:

$$y^{(k)}|\theta^{(k)} \sim \text{Bin}(m^{(k)}, \theta^{(k)})$$

Prior distribution. We know $\theta^{(k)}$ can only take on values between 0 and 1, therefore, we will assign $\theta^{(k)}$ a Beta distribution prior as it is a continuous probability distribution defined on the interval $[0, 1]$:

$$\theta^{(k)}|a^{(k)}, b^{(k)} \sim \text{Beta}(a^{(k)}, b^{(k)}),$$

where $a^{(k)}$ and $b^{(k)}$ reflect a prior knowledge of user k's accuracy and can be interpreted as the number of correct and incorrect identifications respectively that the user makes. If we have no prior knowledge on suitable values for $a^{(k)}$ and $b^{(k)}$ we can assign a non-informative uniform prior by setting $a^{(k)} = b^{(k)} = 1$.

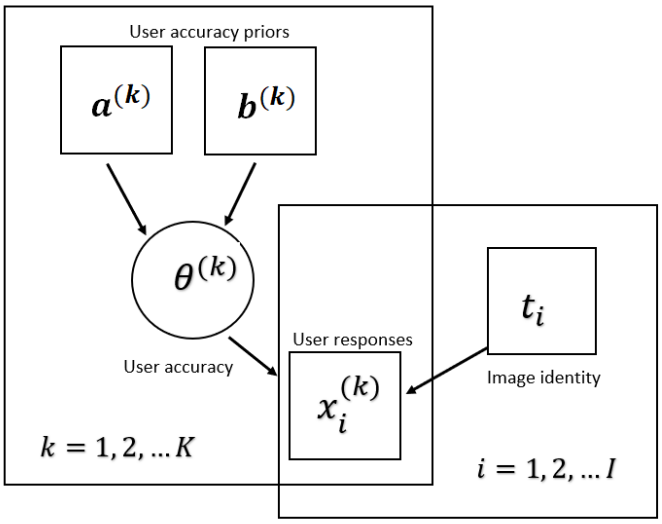


Figure 3a.6: **Model for case 2 takes known image identities, user responses and user accuracy priors to estimate posterior user accuracies.** Circular nodes are unknown variables and square nodes are known parameters. Plates are over users, $k = 1, 2, ...K$ and images, $i = 1, 2, ..., I$. Directed edges indicate a conditional dependence between two nodes.

Posterior distribution. Using Bayes' formula to combine user identifications with the prior distribution we obtain the unnormalised posterior probability distribution for $\theta^{(k)}$:

$$\theta^{(k)}|y^{(k)}, m^{(k)} \sim \text{Beta}(a^{(k)} + y^{(k)}, b^{(k)} + m^{(k)} - y^{(k)}). \tag{3a.4}$$

Every time user k makes a new identification we update the posterior distribution by adding 1 to the first parameter, $a^{(k)} + y^{(k)}$, if it was a correct identification or to the second parameter, $b^{(k)} + m^{(k)} - y^{(k)}$, if the identification was incorrect. A DAG of Case 2 in shown in Figure 3a.6. Figure 3a.7 shows how the shape of the Beta posterior distributions change as the number of user identifications increases for different prior distributions.

We use the mean of user k's Beta posterior distribution as an estimate of user k's

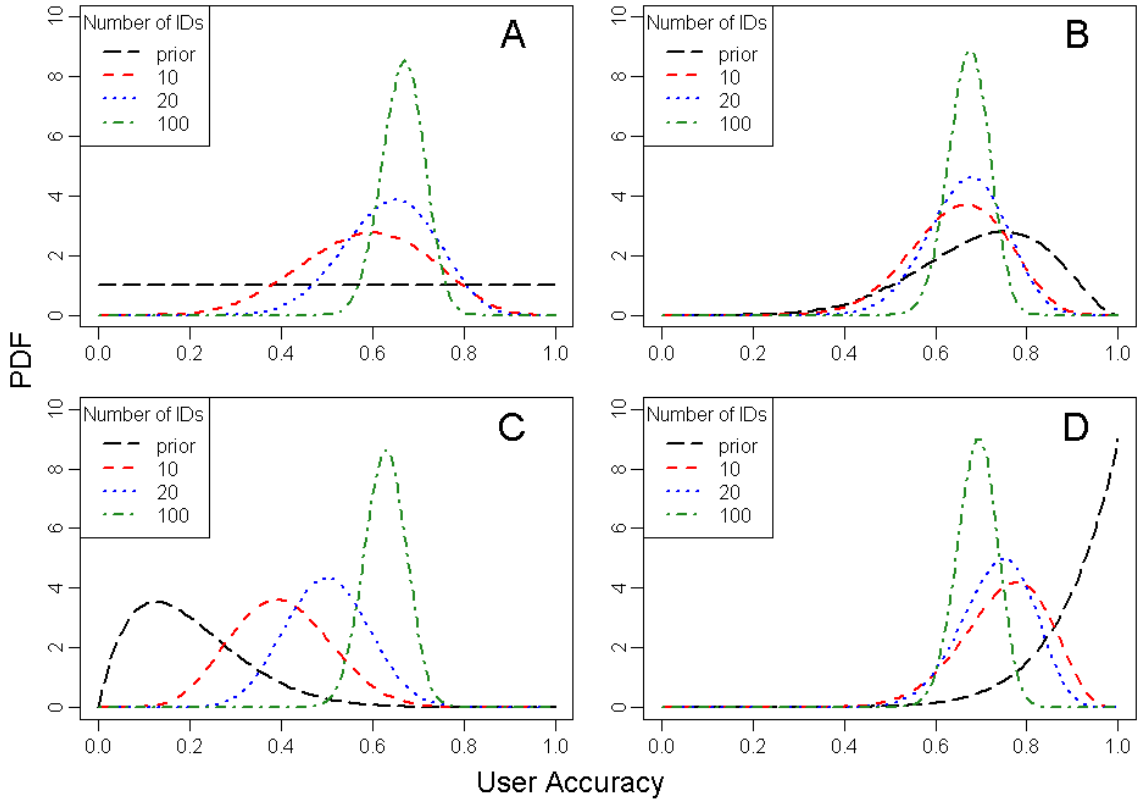


Figure 3a.7: **The prior distribution has less effect on the posterior distribution as the number of user identifications increases.** Posterior distributions of user accuracy as the number of identifications increases for a user with 0.7 user accuracy. **A** uniform prior, a=1, b=1. **B** correct informative prior, a=7, b=3. **C** incorrect informative prior, a=2, b=8. **D** incorrect informative prior, a=9, b=1.

accuracy:

$$\bar{\theta}^{(k)} = \frac{a^{(k)} + y^{(k)}}{a^{(k)} + b^{(k)} + m^{(k)}}. \tag{3a.5}$$

This is a generalisation of the standard user accuracy estimator;

$$\bar{\theta}^{(k)} = \frac{y^{(k)}}{m^{(k)}},$$

which does not include any prior information about user accuracies that we may have from similar exercises. The Bayesian estimator (Equation 3a.5), also avoids the problematic estimated values of 0, and 1, which tend to occur with small sample sizes.

Although this approach of estimating user accuracy with the Bayesian estimator

mean is computationally efficient it disregards the variance in the posterior distribution.

Two-stage classification method

In this section we outline our two-stage classification method that consists of: the testing stage, and the classification stage. In the testing stage user accuracies are estimated using Equation 3a.5. In the classification stage these user accuracy estimations are combined with user responses, and class prevalence values by Bayes' theorem (Equation 3a.3 from Case 1) to classify all images. I_T is the number of images in the testing stages, I_C is the number of images in the classification stage, and $I_T + I_C = I$. A DAG for the two-stage classification method is shown in Figure 3a.8.

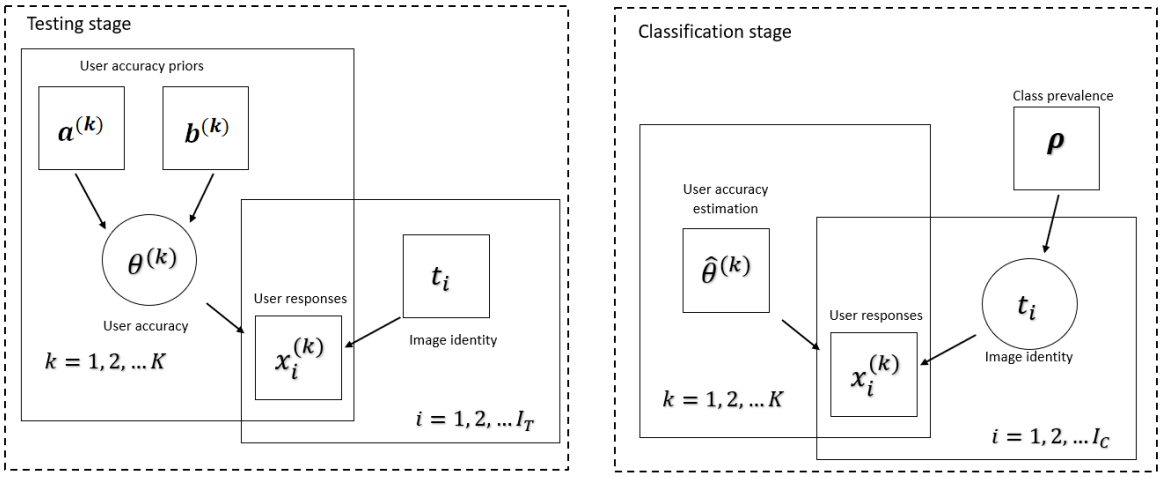


Figure 3a.8: **The two stage classification method utilises methods from case 2 to estimate user accuracy in the testing stage, and methods from case 1 to classify images in the classification stage.** Circular nodes are unknown variables and square nodes are known parameters. Plates are over users, $k = 1, 2, \ldots K$, testing stage images, $i = 1, 2, \ldots, I_T$, and classification stage images $i = 1, 2, \ldots, I_C$. Directed edges indicate a conditional dependence between two nodes.

This classification method is a relatively simple step from majority vote to a classification method that includes user accuracy estimations. However, there are two main limitations. First, there is no feedback between the two stages and therefore the model is unable to capture changes in user accuracies overtime, often referred to as learning or tiring effects. Second, there is a trade-off on classification accuracy of splitting user responses between the two stages.

Trade-off between testing stage and classification stage

In Chapter 2 we found that 75% of users that have done identifications on iNatrulaist NZ have done less than 10 identifications. It has also been reported by Sauermann and Franzoni (2015a) that this phenomenon of most users contributing a small number of responses is common to most citizen science projects. Therefore, when citizen science projects ask users to initially identify images with known ground truth (in the testing stage) there is a risk that users will do fewer identifications on images with unknown ground truth (in the classification stage). As a result we have a trade-off between requiring users to identify more testing images and therefore improving our estimate of their user accuracy, or leaving more identifications for the classification stage to have more identifications per image and therefore a higher classification accuracy.

We created a simulation to show the impact on classification accuracy of splitting user identifications between the testing and classification stages, given $\rho_b, \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}$, where bold indicates a vector. In the testing stage each user identifies m images and in the classification stage each image is identified by n users. Simulations, preformed for $m = 1, ..., 40$ and $n = 1, ..., 20$, are as follows:

1. For each user, simulate responses to m testing stage images and estimate each user's accuracy, $\hat{\theta}^{(k)}$, as in Equation 3a.5.

2. For each classification stage image, randomly select n users to identify the image and simulate their identifications, $\{x_i^{(1)}, ..., x_i^{(n)}\}$ given their true user accuracies $\{\theta^{(1)}, ..., \theta^{(n)}\}$.

3. Given the estimated user accuracies from step 1, user identifications from step 2 and ρ_B use Bayes' formula to classify each classification stage image as in case 1, Equation 3a.3.

4. Evaluate classification accuracy as the proportion of correctly classified classification stage images.

Results from the simulation outlined above are shown in Figure 3a.9 for the arbitrary example with $I = 10^4$ images, $K = 10^3$ users, $\rho_B = 0.9$, and $\theta^{(k)} = 0.7$ for $k = 1, ..., K$. The prior parameters for user accuracy were set at $a^{(k)} = 6$, and $b^{(k)} = 4$ for $k = 1, ..., K$, corresponding to low expectation of 60% user accuracy. The solid iso-accuracy curves show different combinations of m and n that result in the same classification accuracy and the dotted iso-cost lines show different combinations of m and n that result in the same classification effort. An optimal distribution of identifications between the testing and classification stage is at the tangent point of the iso-accuracy and iso-cost lines, shown with solid triangles. For example, shown in Figure 3a.9, if we are limited to 120,000 user identifications this is distributed as either

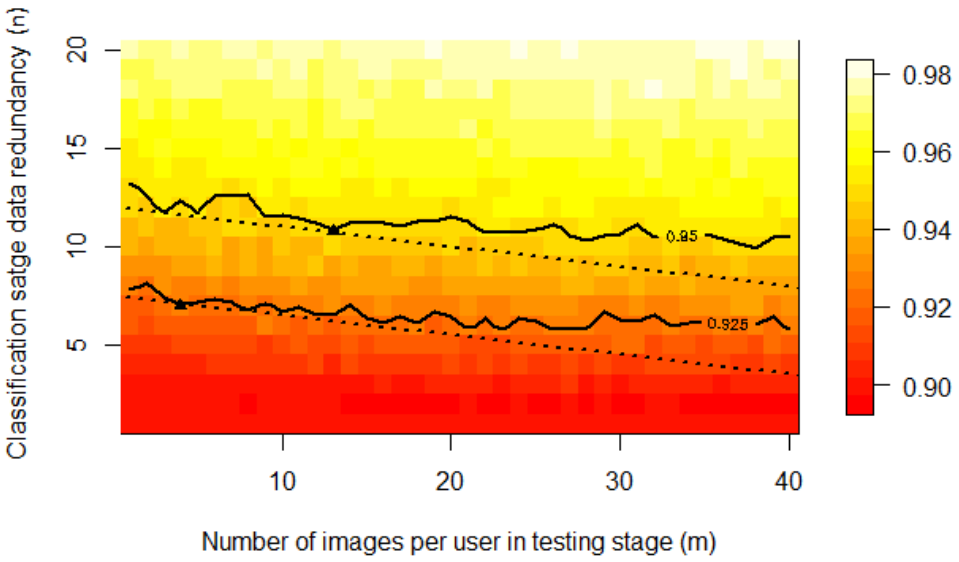


Figure 3a.9: **Given a certain number of identifications per user we can use simulations to find the optimal allocation of identifications to the testing stage and classification stage.** Colours show classification accuracy, α, for the two-stage method as a function of m and n, the number of testing images per user and the number of identifications per image in the classification stage. Solid lines show iso-accuracy curves, dotted lines show iso-cost lines, and circle markers show optimal allocation of identifications given a fixed number of identifications per user. Parameters used: $\rho_B = 0.9$, and $\theta^{(k)} = 0.7$ for $k = 1, ..., K$. The prior parameters for user accuracy were set at $a^{(k)} = 6$, and $b^{(k)} = 4$ for $k = 1, ..., K$.

120 images per user in the testing stage, 12 identifications per image in the classification stage, or any combination of m and n that satisfies $10^3 n + 10^4 m = 120,000$. The highest classification accuracy we can achieve is 95% by approximately assigning $n = 11$ (i.e. 11 identifications per image in the classification stage) and $m = 13$ (i.e. 13 images per user in the testing stage). In conclusion, given a fixed set of identifications, one may use the above simulations to find the optimal distribution of identifications between the two stages.

Case 3: Classifying images based on user accuracies when no user accuracies or ground truth are known.

In case 2 we found there is a trade-off between allocating user identifications to a testing stage to improve user accuracy estimations or to the classification stage to improve classification accuracy. To eliminate this trade-off many existing works have considered how to estimate user accuracies and infer image ground truth from user identifications alone. Dawid and Skene (1979) completed some of the early research in this field by proposing an Expectation-Maximisation (EM) algorithm for modelling individual clinician error rates and compiling patient records when the patient's true response is not available. This is analogous to citizen science projects where multiple users have identified the same image but their responses are subject to error. The EM algorithm used by Dawid and Skene (1979) has the following iterative procedure:

1. Make initial estimates of the patient response ground truths.
2. *Maximisation step.* Calculate maximum likelihood estimations for individual clinician error rates and patient response prevalences.
3. *Expectation step.* Calculate new estimates of the patient response ground truths.
4. Repeat until both the maximum likelihood estimates and ground truth estimates converge.

It has been shown that the EM algorithm provides a slow but reliable method of obtaining maximum likelihood estimates of the parameters of interest.

Inspired by the early work of Dawid and Skene (1979) many other models have been developed to infer ground truth and user accuracy from user identifications. Zheng et al. (2017) provide a detailed survey of 17 representative truth inference models, including EM, and performed a comprehensive comparison of the 17 methods using 5 real datasets. Variations between the 17 models included incorporating task difficulty and/or latent topics in the model, modelling user accuracy as a single value or a confusion matrix, including modelling of user bias and variance, modelling user confidence based on number of identifications per user, and treating users as independent or dependent.

The general approach adopted by most of the methods was similar to the EM algorithm:

1. Initialise user accuracies.
2. Infer the ground truth of each image based on user identifications and accuracies.

3. Estimate and update user accuracy based on user identifications and the inferred image classifications.

4. Repeat until both the user accuracy estimates and ground truth estimates converge.

Zheng et al. (2017) concluded the classical EM method has a relatively simple implementation and attains very good results, however, a high performing extension of the EM method is a Bayesian classifier combination model (Kim and Ghahramani, 2012). Incorporating task difficulty or latent topics did not significantly improve classification accuracy. Modelling user accuracy as confusion matrices performs significantly better than a single user accuracy value per user. Modelling other user features (e.g. user bias, variance, and confidence) did not result in significant benefits.

In this section we outline the Independent Bayesian Classier Combination (IBCC) model described by Kim and Ghahramani (2012), but, we simplify the way user accuracies are modelled. We outline a Gibbs' sampling algorithm for the IBCC model that combines our image classification methods from case 1 and user accuracy estimation methods from case 2.

Independent Bayesian classifier combination model

Kim and Ghahramani (2012) introduced the IBCC model that treats all user as independent. IBCC combines discrete, categorical responses, such as image identifications by citizen scientists to estimate image ground truth and user accuracy. IBCC extends the earlier work of Dawid and Skene (1979) to allow Bayesian inference techniques, such as Gibbs' sampling, as suggested in Kim and Ghahramani (2012).

As in case 1 and case 2 we assume:

- Users are independent.
- User accuracies are the same for all image classes.
- All images are of equal difficulty to identify.

However, in case 1 and case 2 we assumed there were only two classes of images, birds and planes. In this section we relax this assumption and generalise to J classes.

As in case 1, we have I images, for which we wish to infer a set of ground truth labels $\hat{\mathbf{t}} = \{\hat{t}_1, \hat{t}_2, ..., \hat{t}_I\}$. The ground truth, t_i, for image i takes a value $j \in \{1, ..., J\}$ where J is the number of classes. For image i, we assume the true label t_i is generated by a multinomial distribution, a generalisation of the binomial distribution to more than two outcomes, with parameters $\boldsymbol{\rho}$: $p(t_i = j|\boldsymbol{\rho}) = \rho_j$, which models the class prevalences. For image i, the identification by user k, $x_i^{(k)} \in \{1, ..., J\}$.

We simplified the IBCC model outlined by (Kim and Ghahramani, 2012) from modelling user accuracy as a confusion matrix to assuming that an identification $x_i^{(k)}$ by user k is generated from a multinomial distribution with the following probabilities:

$$p(x_i^{(k)}|t_i = j) = \begin{cases} \theta^{(k)}, & \text{if } x_i^{(k)} = t_i \\ \frac{1-\theta^{(k)}}{J-1}, & \text{if } x_i^{(k)} \neq t_i, \end{cases} \tag{3a.6}$$

where $\theta^{(k)}$ is the classification accuracy of user k. A user has the same user accuracy at classifying every class, all images have equal classification difficulty, and if a user incorrectly classifies an image they are equally likely to classify the image as one of the $J - 1$ remaining classes.

The posterior distribution over the unknown variables $\mathbf{t}$ and $\boldsymbol{\theta}$ given the user responses $\mathbf{x}$ is given by:

$$p(\mathbf{t}, \boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{t})p(\boldsymbol{\theta}|\boldsymbol{a}, \boldsymbol{b})p(\mathbf{t}|\boldsymbol{\rho}). \tag{3a.7}$$

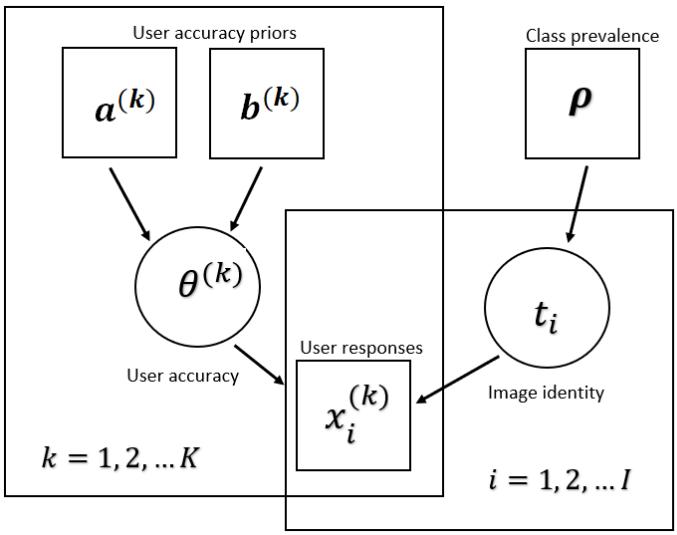


Figure 3a.10: **IBCC model takes known user accuracy priors, class prevalences, and user responses to estimate user accuracies and infer true image identities** Circular nodes are unknown parameters and square nodes are known parameters. Plates are over users, $k = 1, 2, ...K$ and images $i = 1, 2, ...I$.

Equation 3a.7 is the IBCC model and the DAG for IBCC is shown in Figure 3a.10. Noteworthy, the IBCC DAG is effectively a combination of the DAGs from case 1 and case 2.

Inference for the unknown parameters $\mathbf{t}$ and $\boldsymbol{\theta}$ can be done via Gibbs sampling.

IBCC Gibbs sampling algorithm

Gibbs sampling, introduced by Geman and Geman (1984), is a Monte Carlo Markov Chain (MCMC) algorithm that is commonly used to generate samples that approximate a posterior distribution when an analytical derivation is intractable or direct sampling is difficult. Gibbs sampling generates posterior distribution samples by sequentially sampling from the conditional probability distribution of each variable with the remaining variables fixed to their current values. Gibbs sampling is suited to models such as the IBCC model as the conditional probability distributions for all the parameters are easily computed.

Conditional probability distributions
First, we define the following indicator function;

$$I_{\{Z\}} = \begin{cases} 1, & Z \text{ is true} \\ 0, & \text{otherwise.} \end{cases}$$

From the posterior distribution function, Equation 3a.7, the conditional probability distribution function for $\mathbf{t}$ is a multinomial distribution with event probabilities given as:

$$p(t_i = j|\text{all}) \propto \rho_j \prod_{k=1}^{K} (\theta^{(k)})^{I_{\{x_i^{(k)}=j\}}} \cdot \left(\frac{1-\theta^{(k)}}{J-1}\right)^{\left(1-I_{\{x_i^{(k)}=j\}}\right)}, \tag{3a.8}$$

where Equation 3a.8 is an extension of Equation 3a.3 from Case 1 to more than two classes.

Similarly, the conditional probability distribution for $\boldsymbol{\theta}$ is given as;

$$p(\theta^{(k)}|\text{all}) \sim \text{Beta}\left(a^{(k)} + \sum_i I_{x_i^{(k)}=t_i}, b^{(k)} + \sum_i I_{x_i^{(k)}\neq t_i}\right), \tag{3a.9}$$

where Equation 3a.9 is an extension of Equation 3a.4 from Case 2 to more than two classes.

Since the conditional density of $\mathbf{t}$ is a multinomial distribution and the conditional density of $\boldsymbol{\theta}$ is a Beta distribution they can be sampled from easily in the Gibbs' sampling algorithm.

Gibbs sampling algorithm

1. Initialise the values of $\mathbf{t}$ and $\boldsymbol{\theta}$.
2. Calculate $p(t_i = j|\text{all})$ using Equation 3a.8 for $j \in 1, ..., J$ for all images.
3. Sample t_i for $i = 1, ..., I$ from the multinomial distribution given the probabilities calculated in step 2. Replace current value of t_i with new sample.
4. Sample $\theta^{(k)}$ for $k = 1, ..., K$ from the Beta distribution given in Equation 3a.9, using the current $\mathbf{t}$. Replace current value of $\theta^{(k)}$ with new sample.
5. Record current $\mathbf{t}$ and $\boldsymbol{\theta}$ sample values in a list of sample values.
6. Repeat all steps from step 2 until a sufficient approximation is obtained, or convergence is observed.

In the Gibbs sampling algorithm step 2 and step 3 use the methods from Case 1 extended to multiple classes. Step 4 uses methods from Case 2, however, rather than estimating each user's accuracy as the mean of their posterior distribution a sample is drawn from the updated posterior distribution each iteration and therefore variance in the distribution is regarded.

Convergence

For each parameter the stationary distribution of the Markov chain from the Gibbs sampling algorithm is the desired posterior distribution, where, a stationary distribution of a Markov chain is a probability distribution that remains unchanged in the Markov chain as time progresses. But, how do we know when the Gibbs sampling algorithm has converged? It may be the case that each Markov chain has converged to the stationary distribution and we may need many samples from the stationary distribution to accurately approximate the posterior distribution. This has been a well studied area of MCMC algorithms. Cowles and Carlin (1996) and Brooks and Roberts (1998) both reviewed a wide range of MCMC convergence diagnostic methods. They concluded that many theoretical convergence tests are not practical to apply, and are often unreliable. To ensure convergence it is common practice to run Gibbs sampling algorithms for a large number of iterations and visually check for convergence by inspecting trace plots of parameter values against iteration number. Due to it likely taking a while for the stationary distribution to be reached, early "burn-in" iterations are discarded. It is also good practise to run multiple chains from different starting points and check they converge to the same distribution/density plots.

We estimate t_i for $i = 1, ..., I$ and $\theta^{(k)}$ for $k = 1, ..., K$ from the sampled values by taking the mean of all the θ_i samples and the mode of all the t_i values, after removing the burn-in samples.

Modelling user accuracy with a confusion matrix

In the IBCC model outlined by (Kim and Ghahramani, 2012) user accuracies were modelled as a confusion matrix as shown in Equation 3a.10. Each row of the confusion matrix is independent and for a given ground truth value j, the distribution over responses from user k has a parameter vector from the corresponding row of the confusion matrix. Modelling user accuracy with a confusion matrix models the situation where a user has different abilities at identifying different classes and is more likely to confuse particular pairs or groups of classes. For example, on iNaturalist NZ well known species may be easily identified, whereas an uncommon species without distinctive features may be misclassified more often, and if a user incorrectly classifies a kiwi they are probably more likely to mistake it for a weka than a seal.

$$\boldsymbol{\theta}^{(k)} = \begin{bmatrix} \theta^{(k)}_{(1,1)}, & \cdots & \theta^{(k)}_{(1,J)} \\ \vdots & \ddots & \vdots \\ \theta^{(k)}_{(J,1)}, & \cdots & \theta^{(k)}_{(J,J)} \end{bmatrix} \tag{3a.10}$$

The user accuracy confusion matrix could be non symmetric in the case that there is not a 1-1 mapping between user identification classes and ground truth classes. For example, users may classify land types as either forested, or non-forested, but ground truth labels may include more detailed classes, e.g. urban, grasslands, e.t.c.

Instead of modelling user accuracy, $\theta^{(k)}$, for all classes as a Beta distribution as shown in equation 3a.9 each row of the confusion matrix would be modelled with a separate Beta or Dirichlet distribution. The Dirichlet distribution is a multivariate generalisation of the Beta distribution and a conjugate prior of the multinomial distribution. We would choose a Beta distribution in the case of only two classes, or if we wanted to model users as equally likely to misclassify an image with any of the other $J-1$ classes. Otherwise, we would model every row of the confusion matrix with a Dirichlet distribution as done by Kim and Ghahramani (2012):

$$\boldsymbol{\theta}_j^{(k)}|\boldsymbol{\alpha}_j^{(k)} \sim \text{Dir}(\boldsymbol{\alpha}_j^{(k)})$$

Where $\boldsymbol{\alpha}_j^{(k)} = \{\alpha_{(j,1)}^{(k)}, ..., \alpha_{(j,J)}^{(k)}\}$ reflects a prior knowledge of user k accuracy and $\alpha_{(j,1)}^{(k)}$ can be interpreted as the number of identifications for class 1 given the true class j.

As a result the conditional density function for $\mathbf{t}$ shown in Equation 3a.8 would become:

$$p(t_i = j|\text{all}) \propto \rho_j \prod_{k=1}^{K} \theta_{j,x_i^{(k)}}^{(k)}.$$

Similarly, the conditional density function for $\boldsymbol{\theta}$ shown in Equation 3a.9 would become:

$$p(\theta_j^{(k)}|\text{all}) \sim \text{Dir}(\boldsymbol{\alpha}_j^{(k)} + \boldsymbol{N}_j^{(k)}).$$

Where

$$\boldsymbol{N}_j^{(k)} = \big[\sum_{i|t_i=j} I_{x_i^{(k)}=1}, ..., \sum_{i|t_i=j} I_{x_i^{(k)}=J}\big].$$

Discussion

In this chapter we worked through three cases of including user accuracy in collective classification decisions. In the first case, we assumed that user accuracy was perfectly known. We used Bayes' formula to calculate the probability that the identity of the image was one of the two classes. In the second case, we used a Bayesian approach to estimate user accuracy when ground truth was known. In some citizen science projects the users identify a mix of images with known ground truth (the testing stage) and images with unknown ground truth (the classification stage). We explored the trade-off on overall classification accuracy of asking users to identify images in the testing stage versus the classification stage, In case 3, we assumed that user accuracy was unknown and there were no images with known ground truth. We outlined a Bayesian method that is able to simultaneously estimate user accuracy and the identity of the images.

Using Bayes' formula and Bayesian techniques to include user accuracy and class prevalence in collective classification decisions can improve the classification accuracy compared to simple majority vote. However, obtaining information about these extra variables, in particular user accuracy, is not without costs. Including a testing stage in a classification based citizen science project is an effective and simple method to estimate user accuracies. However, it requires allocating some of each users limited identifications to the testing stage and therefore reducing the number of identifications available in the classification stage. This method also provides no feed back between user identifications in the classification stage and user accuracy estimations and therefore cannot account for a user learning or tiring effect. Including testing images throughout a users participation on a classification based citizen science project rather than just at the beginning would allow this two-stage method to regularly update user accuracy estimations. This would enable the two-stage method to account for temporal variation in user accuracy, like the IBCC model but with significantly less computational effort. However, the trade-off between distributing user responses to test images and classification stage images must always be considered.

To address the trade-off from distributing identifications between the testing and classification stages many models and algorithms have been developed to estimate user accuracy and classify images simultaneously. In Case 3 we outlined the IBCC model and a suitable Gibbs sampling algorithm in the context of classification based citizen science. The IBCC model combines the methods from Case 1 and Case 2, and allows for information feedback between all user identifications and their user accuracy estimation. However, implementing the IBCC model with a Gibbs sampling algorithm is significantly more computationally expensive than a testing stage method or simple majority vote.

Chapter 3b

Citizen science decisions: A Bayesian approach optimises effort

All the votes were summed,
user effort was wasted.
Bayesian sorts that.

Abstract

Volunteer citizen scientists are an invaluable resource for classifying large numbers of images that are used for species monitoring. Citizen science projects often rely on the "wisdom of the crowd" through majority vote methods to produce accurate classifications and assume all volunteer citizen scientists have equal ability.

We use a Bayesian framework to estimate iNaturalist NZ user accuracies and simultaneously collectively classify the observations. We calculate the probability that the inferred observation classification from the Bayesian framework is correct for each observation given the assumed true user accuracies. We refer to this probability as the classification certainty.

Our results show that 50% of images were classified by more volunteer citizen scientists than required to reach a minimal desired collective classification certainty level and more than one third of identifications were above the number required to meet the minimal desired classification certainty.

Over 60% of observations that are yet to be considered research grade have a high classification certainty that has already surpassed the desired minimal level and could therefore be upgraded to research grade with no additional identifications.

Synthesis and applications With more sophisticated collective classification methods than a simple majority vote procedure citizen science data and volunteer citizen scientists effort could be utilised more optimally.

Introduction

Citizen science, the involvement of many individuals that are mostly not trained as scientists, in collecting, categorising, transcribing, or analysing scientific data (Bonney et al., 2014), has increased rapidly in the last decade (Kosmala et al., 2016). In particular, due to technological advances in instrumentation the amount of imagery data is accumulating faster than the processing abilities of research institutions (Porter et al., 2009). Thus, there is a growing demand for computer algorithms and human resources to assist in analysing and categorising the data (Matabos et al., 2017). For imagery data the human eye is still vastly more accurate than computer algorithms (Schoening et al., 2012; Aguzzi et al., 2009; Purser et al., 2009). Therefore, volunteer citizen scientists are invaluable to categorising the large imagery databases. For example, citizen science platform Zooniverse that largely hosts classification based citizen science projects has long relied on volunteer citizen scientists to classify enormous amounts of images, videos, and audio. Their projects have led to the classification of more than a million galaxies, the recovery of lost fragments of ancient poetry, and the classification of thousands of wildebeest (Simpson et al., 2014). However, typically a large share of the volunteer citizen scientists only participate once in the project with a low level of effort, leaving a small proportion of volunteer citizen scientists that contribute a large share of the effort (Sauermann and Franzoni, 2015b). Thus, it is important to optimally utilise volunteer citizen scientists efforts.

Classification citizen science projects vary in the level of training and guidance they provide to volunteer citizen scientists. For example, Zooniverse projects provide their users with tutorials and guides (Simpson et al., 2014). Land type classification citizen science project Geo-Wiki provides users with training material and requires users to classify some images with known ground truth to judge the quality of the contributions (See et al., 2015). On the other hand, iNaturalist is a network of citizen science projects that provides users with an online community to share observations of biota and also classify these observations. However, iNaturalist users are not given any specific classification training and the project does not include any observations with known ground truth.

Citizen science image classification relies on the "wisdom of the crowd" rather than the knowledge of an expert. Group judgements have long been noted to be able to be more accurate than individual decisions, and this concept dates back to the Condorcet Jury Theorem (Condorcet, 1785). Multiple users identify each image resulting in the citizen science project building up a database of user identifications per image. Many classification based citizen science projects then use a simple majority vote rule to combine individual user identifications into a collective classification, for example

projects from Zooniverse (e.g. Snapshot Serengeti), and the iNaturalist network (a collection of localised country websites that are fully connected to the global iNaturalist community). Majority vote is a simple method to collate user identifications but it has the potential to introduce errors due to variations in users' abilities to identify species or events (Bird et al., 2014). A simple majority vote rule weights every user's identifications equally even though there is often wide variation in user accuracies.

Many studies have considered how to enhance the quality of collective classifications by incorporating a measure or estimate of user ability. A simple approach is to acquire information on users' confidence in their vote and then use this information to weight the votes by placing more emphasis on the votes with high confidence (See et al., 2013). This method is problematic as some users overestimate their abilities while other users underestimate their abilities (Kruger and Dunning, 1999). An improvement of this basic confidence weighted method may be using the surprising popular vote approach outlined by Prelec et al. (2017) that makes a collective classification by selecting the answer that is more popular than people predict. Another common approach is to estimate user accuracy by comparing user responses to expert responses for a subset of the images. For example, Geo-wiki is experimenting with a user accuracy estimation for each user based on the identifications given for the user testing images that have known ground truth (See et al., 2015). Hsing et al. (2018) used a 'gold standard' set of classifications created by themselves to determine the accuracy of MammalWeb citizen scientists. Siddharthan et al. (2016) developed a supervised incremental Bayesian model to re-evaluate the quality of the consensus label following each species identification that also accounts for species-specific differences in the ease of identification and differential skill level among users. Other studies have considered how to estimate user accuracies and infer image ground truth from user identifications alone. Dawid and Skene (1979) completed some of the early research in this field by proposing an Expectation-Maximisation (EM) algorithm for modelling individual clinician error rates and compiling patient records when the patient's true response is not available. A high performing extension of the EM method is an independent Bayesian classifier combination (IBCC) model (Kim and Ghahramani, 2012; Zheng et al., 2017). The IBCC model assumes all user identifications are independent and combines discrete, categorical responses, such as image identifications by citizen scientists, to estimate image ground truth and user accuracy. The IBCC model extends the earlier work of Dawid and Skene (1979) to allow Bayesian inference techniques, such as Gibbs' sampling, as suggested in Kim and Ghahramani (2012).

We apply the IBCC model to the iNaturalist NZ citizen science project, the New Zealand member of the iNaturalist network, to estimate user accuracies and simultaneously collectively classify the observations in the absence of any observation with

known ground truth. Assuming the estimated user accuracies are the true user accuracies we use Bayes' formula to calculate the probability that the IBCC observation classifications are correct and refer to this probability as the classification certainty. We analyse the number of observations that meet a classification certainty threshold and could therefore be considered ready for research use with no further identifications. We also analyse the number of overclassified observations that would still meet the classification certainty threshold with fewer identifications and subsequently we assess the number of superfluous identifications.

Materials and Methods

Data

We used observation and identification data from the first five years (August 2012 - July 2017) of iNaturalist NZ. iNaturalist NZ observations are classified using a simple majority vote rule. Observations on iNaturalist NZ are assigned to one of three quality grades, casual, "needs ID", or research, depending on how many users have identified the observation and the level of consensus between the identifiers. A quality grade of casual means that the observation is considered non-verifiable, i.e. it has no photo or sound clip, or has been flagged as non-wild. Verifiable observations are labelled as "needs ID". An observation moves from "needs ID" grade to research grade when it has been identified by at least two users and the majority ($> 50\%$) of the users agree on the observation identity. iNaturalist NZ users are able to identify casual and "needs ID" observations by either agreeing with the previous user identifications or by suggesting another identification from the database of taxa. Identifications can be made at any level of taxonomic rank, however, only observations with a majority consensus at a species or subspecies level will progress to research grade. However, there is an exception that some observations with a majority consensus at a genus level may also move to research grade if someone flags the observation ID as "it's as good as it can be". Once an observation is considered research grade further identifications may still be contributed to the observation and the classification or quality grade of the observation could change. For this study we only kept identifications at a species or subspecies level as that is the desired taxonomic rank for observation final classifications. As a result 17.43% (50732) of observations were removed from the data set because they have no identifications at a species or subspecies level. Computer vision was added to the iNaturalist platform throughout the second half of 2017. The computer vision algorithm is trained on historical verifiable iNaturalist observations and provides identifiers with an automated list of taxon suggestions. Computer vision was first integrated into the iNaturalist iOS application on June 29, 2017. However, it was not fully integrated into the Android application and web observation uploader until September 2017. There were 38 observations that received an identification in July 2017 by iOS users that used a suggestion from the computer vision technology. Given the small scale of observations these identifications affected (0.1% of observations that we considered) we did not filter these out of our dataset. iNaturalist NZ observations are categorised into 13 'iconic' taxa based on the identifications. Table 3b.1 shows the number of observations, number of users that identified an observation, and number of species and subspecies taxa per iconic taxon. Some iconic taxa are nested in lower taxonomic ranks than others (e.g. Animals, Insects) and observations are assigned to the lowest matching iconic taxon.

Iconic taxa	Observations	Users	Species and subspecies	Observations with multiple IDs per user	Identification agreement: last user identification per observation
Protozoans	180	86	44	2.22%	98.89%
Amphibians	368	171	32	41.30%	98.10%
Chromista	1085	119	80	0.37%	97.70%
Ray finned fish	1633	243	445	7.59%	96.51%
Reptiles	1643	338	183	9.92%	92.94%
Animal	3697	444	496	6.84%	97.08%
Arachnids	4048	440	332	4.69%	98.72%
Mollusks	4213	306	628	14.50%	95.11%
Mammals	6717	758	236	2.80%	96.41%
Fungi	11280	599	2073	6.30%	98.46%
Insects*	29026	1540	3707	4.16%	98.72%
Birds*	50290	4005	810	6.37%	97.34%
Plants*	126134	1681	6645	5.87%	97.98%
All observations	**240314**	**5921**	**15617**	**5.93%**	**97.82%**

Table 3b.1: **iNaturalist NZ data summary by the 13 iconic taxa.** Observations: total number of observations with species and/or subspecies level identifications. Users: number of unique users that have identified an observation by iconic taxa. Species and subspecies: number of unique species or subspecies by iconic taxa. *Not included in IBCC calculations.

IBCC model

We use a simplified version of the Independent Bayesian Classification Combination (IBCC) model outlined by Kim and Ghahramani (2012) to estimate user identification accuracies and classify the iNaturalist NZ observations. Applying the original IBCC model outlined in (Kim and Ghahramani, 2012) to the first 5 years of iNaturalist NZ data is extremely computationally expensive, largely due to the large number of species/subspecies. Therefore, we apply the IBCC model separately to each iconic

taxon subset. Very few observations, 133 (0.06%), had identifications from more than one iconic taxon subset. Identifications are made by individual users and classifications are the result of collating identifications for an observation. User accuracy, denoted as θ, is the ability of a user to correctly identify an observation.

We have I observations, for which we wish to infer a set of ground truth labels $\hat{\mathbf{t}} = \{\hat{t}_1, \hat{t}_2, ..., \hat{t}_I\}$. The ground truth, t_i, for observation i takes a value $j \in \{1, ..., J\}$ where J is the number of species and subspecies. For observation i, we assume the true label t_i is generated by a multinomial distribution with parameters $\boldsymbol{\rho}$: $p(t_i = j|\boldsymbol{\rho}) = \rho_j$, which models the species and subspecies prevalences. For observation i, the identification by user k is denoted $x_i^{(k)} \in \{1, ..., J\}$.

We simplified the IBCC model outlined by Kim and Ghahramani (2012) from modelling user accuracy as a confusion matrix to assuming that an identification $x_i^{(k)}$ by user k is generated from a multinomial distribution with the following probabilities:

$$p(x_i^{(k)} = j|t_i) = \begin{cases} \theta^{(k)}, & \text{if } j = t_i \\ \frac{1-\theta^{(k)}}{J-1}, & \text{if otherwise} \end{cases} \tag{3b.1}$$

For each iconic taxon subset $\theta^{(k)}$ is the classification accuracy of user k and users have the same user accuracy at classifying every species and subspecies within the subset. All observations are assumed to have equal classification difficulty, and if a user incorrectly classifies an observation they are assumed to be equally likely to classify the observation as any one of the $J - 1$ remaining species or subspecies within the subset.

We assigned $\theta^{(k)}$ a Beta distribution prior with parameters $(a^{(k)}, b^{(k)})$ where $a^{(k)}$ and $b^{(k)}$ are chosen to reflect any existing belief or information about a user's accuracy:

$$\theta^{(k)}|a^{(k)}, b^{(k)} \sim \text{Beta}(a^{(k)}, b^{(k)}).$$

We have assumed that the underlying species prevalence, ρ is fixed and known. We have evaluated it directly from the overall iNaturalist NZ observations. In practice, the values may come from previous studies or via expert opinion elicitation. The uncertainty about the parameter could also be incorporated via an additional prior distribution.

The posterior distribution over the unknown variables $\mathbf{t}$ and $\boldsymbol{\theta}$ given the user responses $\mathbf{x}$ is given by:

$$p(\mathbf{t}, \boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{t})p(\boldsymbol{\theta}|\boldsymbol{a}, \boldsymbol{b})p(\mathbf{t}|\boldsymbol{\rho}). \tag{3b.2}$$

The directed acyclic graph for IBCC is shown in Figure 3b.1. Inference for the

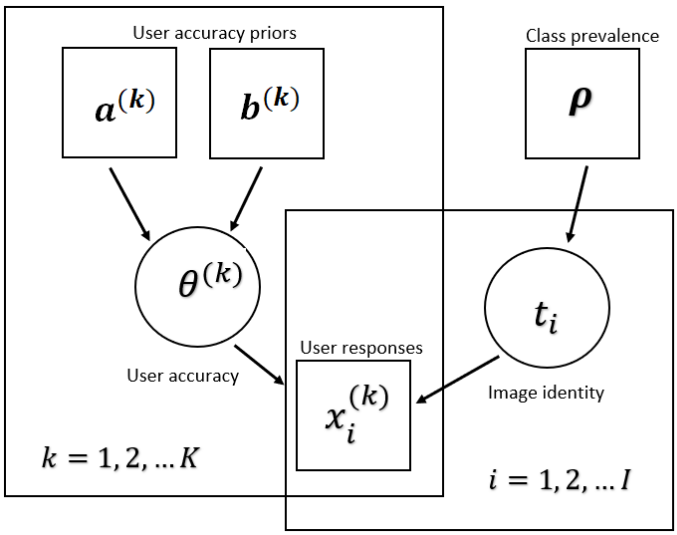


Figure 3b.1: **IBCC model takes known user accuracy priors, species and subspecies prevalences, and user identifications to estimate user accuracies and infer true observation identities** Circular nodes are unknown parameters and square nodes are known parameters or observations. Plates are over users, $k = 1, 2, \ldots K$ and observations $i = 1, 2, \ldots I$. The arrows indicate the direction of conditional dependency.

unknown parameters $\mathbf{t}$ and $\boldsymbol{\theta}$ is done via Gibbs sampling.

IBCC Gibbs sampling algorithm

First, we define the following indicator function;

$$I_{\{Z\}} = \begin{cases} 1, & Z \text{ is true} \\ 0, & \text{otherwise.} \end{cases}$$

From the posterior distribution function, Equation 3b.2, the conditional probability distribution function for $\mathbf{t}$ is a multinomial distribution with event probabilities given as:

$$p(t_i = j|\text{all}) \propto \rho_j \prod_{k=1}^{K} (\theta^{(k)})^{I_{\{x_i^{(k)}=j\}}} \cdot \left(\frac{1-\theta^{(k)}}{J-1}\right)^{\left(1-I_{\{x_i^{(k)}=j\}}\right)}, \tag{3b.3}$$

Similarly, the conditional probability distribution for $\boldsymbol{\theta}$ is given as;

$$\theta^{(k)}|\text{all} \sim \text{Beta}\left(a^{(k)} + \sum_i I_{x_i^{(k)}=t_i}, b^{(k)} + \sum_i I_{x_i^{(k)}\neq t_i}\right). \tag{3b.4}$$

Since the conditional distribution of $\mathbf{t}$ is a multinomial distribution and the conditional distribution of $\boldsymbol{\theta}$ is a Beta distribution they can be sampled from easily in a Gibbs' sampling algorithm as follows.

1. Initialise the values of $\mathbf{t}$ and $\boldsymbol{\theta}$.
2. Calculate $p(t_i = j|\text{all})$ using Equation 3b.3 for $j \in 1, ..., J$ for all observations.
3. Normalise $p(t_i = j|\text{all})$ by dividing by $\sum_{j=1}^{J} p(t_i = j|\text{all})$.
4. Sample t_i for $i = 1, ..., I$ from the multinomial distribution given the probabilities calculated in step 2. Replace current value of t_i with new sample.
5. Sample $\theta^{(k)}$ for $k = 1, ..., K$ from the Beta distribution given in Equation 3b.4, using the current $\mathbf{t}$. Replace current value of $\theta^{(k)}$ with new sample.
6. Record current $\mathbf{t}$ and $\boldsymbol{\theta}$ sample values in a list of sample values.
7. Repeat steps 2-6 until convergence of the distributions for t_i and $\theta^{(k)}$ is visually observed in the density and trace plots.

We estimate t_i for $i = 1, ..., I$ and $\theta^{(k)}$ for $k = 1, ..., K$ from the sampled values by taking the mean of all the $\theta^{(k)}$ samples and the mode of all the t_i values, after removing burn-in samples.

iNaturalist NZ data adjustment for IBCC model

iNaturalist NZ users are able to identify the same observation multiple times. This usually occurs when a user changes their mind about their previous identification or they are refining the taxonomic rank of their identification. Since the IBCC model assumes that user identifications are independent, if a user identifies an observation more than once we only keep their final identification. Overall, less than 6% of observations had multiple identifications by the same user, Table 3b.1. In general there is a high level of agreement between user identifications across all the observations, Table 3b.1, with 97.82% of observations having no identification disagreements.

We do not include the insect, bird, and plant subsets in our analysis due to the large size of these subsets and subsequent long run time of the Gibbs sampling algorithm, however, in the discussion section we discuss methods to reduce the run time on these subsets.

Prior sensitivity and convergence analysis on mollusks

The mollusk subset was used for prior sensitivity and convergence analysis as it is a relatively large subset but with a relatively short Gibbs sampling algorithm run time. We tested the sensitivity of user accuracy, $\theta^{(k)}$, and observation classifications, t_i estimations to a range of user accuracy prior distributions that reflected highly accurate,

lowly accurate, and uninformative *a priori* views of user accuracy. For example, for a highly accurate prior we assigned all users $a = 9$ and $b = 1$, which would correspond to an assumption that the user is correct in an average of 90% of assessments. For low accuracy users we assigned all users $a = 1$ and $b = 9$ and for an uninformative prior we assigned all users $a = 1$ and $b = 1$. As expected user accuracy estimations for users with a low number of identifications (< 10) are sensitive to the prior distribution parameters. However, these users contribute a small amount of the total mollusks identifications (5.31%). In contrast, user accuracy estimations for users who have individually contributed a larger number of identifications (≥ 40) are insensitive to different prior values and overall these users have contributed 85.28% of the mollusk identifications. Observation classifications are not sensitive to user accuracy prior distributions. We tested the sensitivity of parameter estimations to different species and subspecies prevalence distributions. In particular, an informative measure of species and subspecies prevalence calculated as the proportion of total identifications per species and subspecies and an uninformative distribution of an equal prevalence value for each species and subspecies. Observation classifications and user accuracy estimations are not sensitive to species and subspecies prevalence distributions. Convergence was assessed visually by inspecting trace plots and density plots. Due to the large number of estimated parameters we only inspected convergence for 100 parameters. We found that there was excellent convergence after 1500 samples and discarding the first 500 burn in samples. We fit the model on the remaining iconic taxa subsets by generating 1500 samples for each subset using our Gibbs sampling algorithm, and retain the final 1000 samples. Species and subspecies prevalence estimations are calculated as the proportion of identifications for each class, and we assign an uninformative user accuracy prior of $a = 1, b = 1$ to all users.

Classification certainty

After using the IBCC Gibbs sampling algorithm to estimate user accuracies, $\hat{\theta}^{(k)}$, and observation classifications, $\hat{t}_i$, we calculate the probability that the inferred observation classification from the Gibbs sampling algorithm is correct for each observation given the assumed true user accuracies, species and subspecies prevalence, ρ_i, and user identifications, $x_i^{(k)}$:

$$p(t_i = \hat{t}_i | \hat{\theta}^{(k)}, x_i^{(k)}, \rho_i).$$

We refer to this probability as the classification certainty for observation i.

We calculate the classification certainty of observations in casual, "needs ID", and research grade observations. We then calculate the number of observations that would

be labelled research grade if the research quality grade was based on a threshold of classification certainty rather than a majority vote. We also calculate the number of extra identifications above the number required to meet the minimum classification certainty and subsequently the number of observations that could be labelled research grade after a single user identification.

Results

Using the IBCC model with Gibbs sampling resulted in the same classification as a simple majority vote method for the vast majority (99.76%) of research grade observations. More than half of the 53 observations classified differently across the two methods are due to subtle differences between species or subspecies, e.g. *Erinaceus europaeus ssp. occidentalis* vs. *Erinaceus europaeus*. The other 24 observations had greater disagreements between species or subspecies and often more than 3 identifications.

Users that do a large number of identifications have a high estimated user accuracy within a given iconic taxon subset (Fig 3b.2). For example, users that have done less than 10 identifications within a given iconic taxon subset have an average estimated user accuracy of 0.69, whereas, users that have done 30 or more identifications within a given iconic taxon subset have an estimated expected user accuracy of 0.95.

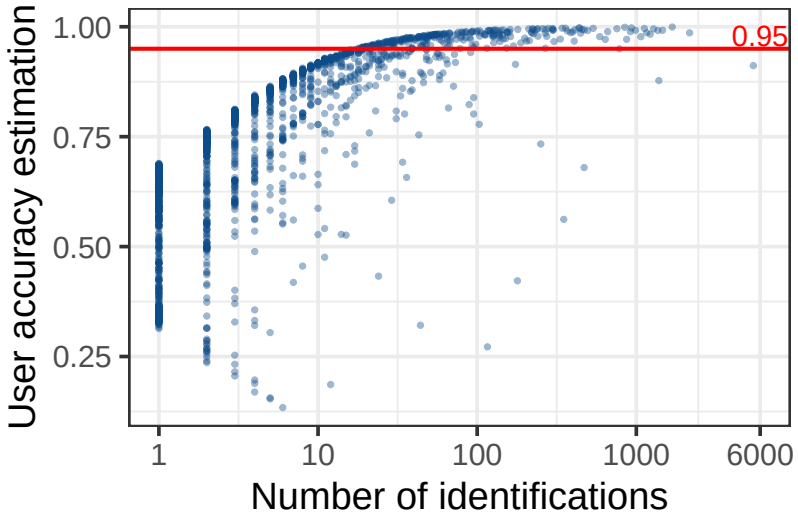


Figure 3b.2: **Users with high identification activity have the highest user accuracy estimations.** User accuracy estimations within a given iconic taxon subset
from the IBCC model Gibbs sampling output by identification activity level for all iconic taxa subsets.

Using the estimations from the IBCC model we found a lot of the casual and "needs ID" grade observations have a high classification certainty and therefore could be upgraded to research grade with no additional identifications (Fig 3b.3). For a 95% classification certainty threshold to be considered research grade, 76.25% of casual and 61.27% of "needs ID" grade observations could be upgraded to research grade, (Table

3b.2). If we required a 99% classification certainty threshold to be considered research grade these values reduce to 65.26% of casual and 41.86% of "needs ID" grade observations that could be upgraded to research grade (Table 3b.2). We have excluded observations flagged as non-wild from the casual grade observations in our results as iNaturalist have made a conscious decision to not allow non-wild observations to progress to research grade to discourage non-wild observations on the platform. As a result 27.89% of casual grade observations are excluded from the results. Interestingly, a higher proportion of casual grade observations could be upgraded to research grade than "needs ID" grade observations. This may be because many of the casual grade observations do not have any media and therefore only a single identification from the user that shared the observation due to the lack of ability for any other users to provide an identification. In contrast, observations in the "needs ID" often have conflicting identifications.

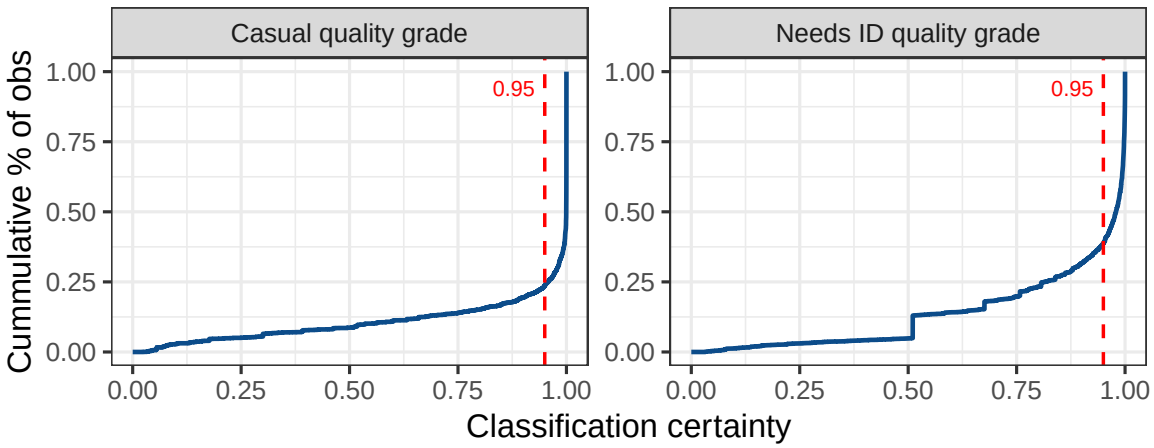


Figure 3b.3: **Given the estimated user accuracies, and observation classifications from the IBCC model a high proportion of casual and "needs ID" grade observations have classification certainties greater than 0.95.** Classification certainty is the probability that the inferred observation classification from the IBCC model is correct given the assumed true user accuracies. Classification certainty is calculated for every observation and this plot shows the cumulative distribution function for casual and need ID grade observations.

Conversely, many observations have more identifications than required to ensure their classification certainty is greater than the set threshold. For a 95% research grade classification certainty threshold, 61.80% of all observations could have reached the threshold with fewer identifications and 43.02% of identifications were not required to ensure all research grade observations met the classification certainty threshold, (Table 3b.2). Similarly, for a 99% research grade classification certainty threshold, 50.20% of all observations had unnecessary identifications and 34.85% of identifications were

	Research grade classification certainty threshold			
	95%		99%	
	proportion	(value)	proportion	(value)
Upgradeable casual grade observations	76.36%	(3319)	65.23%	(2837)
Upgradeable "needs ID" observations	61.27%	(4116)	41.86%	(2812)
Underclassified research grade observations	00.24%	(53)	00.39%	(86)
Observations with excess identifications	61.80%	(20505)	50.20%	(16656)
Superfluous identifications	43.02%	(27942)	34.85%	(22639)
Number of observations that could be research grade with 1 identification	78.80%	(26149)	57.95%	(19229)
Number of observations that could be research grade with 1 identification and user accuracy > 0.95	58.80%	(19510)	50.15%	(16641)

Table 3b.2: Total upgradeable casual and "needs ID" grade observations, under classified research grade observations, and superfluous identifications for two research grade classification certainty thresholds: 95% and 99%.

not required to ensure all research grade observations were above the classification certainty threshold (Table 3b.2).

Many observations could be considered research grade after the initial identification by using the output from the IBCC model and categorising observations as research grade when they have a classification certainty greater than a chosen threshold value. For a 95% threshold value 78.80% of observations could be considered research grade with a single identification and in the case of a 99% threshold value this proportion reduces to 57.95% of observations. The initial identifiers on these research grade observations mostly have a high user accuracy. The initial identification may be made by the user that shared the observation, or if they do not provide an identification the initial identification may be made by any other iNaturalist NZ user. For a 95% research grade classification certainty threshold, 74.61% (19510) of the observations that could be considered research grade with a single identification were initially classified by a user with an estimated accuracy greater than 95% . When the classification threshold is 99%, 86.54% (17082) of the observations that could be considered research grade with a single identification were initially classified by a user with an accuracy greater than 95%. However, some of the observations that could be considered research grade after a single identification are initially classified by users with low accuracies, sometimes less than 0.5, as the classification certainty calculation also depends on the prevalence of the identified species. For example, a user with an accuracy of 0.27 was the initial

identifier of a fungi observation that had a classification certainty of 0.98 after the first identification and they identified it as an *Amanita muscaria* which has a relatively high prevalence value (0.057) compared to the other 2072 fungi species/subspecies. This is a fairly rare situation with only 4 observations that could be considered research grade having an initial identifier with a user accuracy less than 0.5 at a 95% classification certainty threshold level and 0 observations at a 99% threshold classification certainty level. If we imposed the extra condition of requiring user accuracy to be greater than 0.95 for observations that met the classification certainty threshold value with the initial identification the proportion of observations that could be considered research grade with a single observation would reduce. At a 95% classification certainty threshold the proportion would reduce from 79% (26149) of observations to 59% (19510) of observations and at a 99% classification certainty threshold the proportion would reduce from 58% (19229) to 50% (16641) of observations (Table 3b.2).

Across iconic taxa subsets there are variations in the above key results. Results by iconic subset are in the supplementary information. For example, only 1 (6.25%) protozoan casual grade observation is upgradeable to research grade at a 95% threshold level, whereas, 90.81% (3419) of mammal casual grade observations are upgradeable to research grade (Table S1). In New Zealand, the total number of mammal species is very small and it is likely tractable for many non-expert identifiers to master identifying mammal observations, therefore, this result is unlikely to generalise to other geographic regions. The differences in upgradeable "needs ID" observations is less variable across taxa, however, there is still a large difference between upgradeable mammal observations (88.07%), and upgradeable fungi (53.00%) or protozoan (51.25%) "needs ID" grade observations at a 95% threshold level. Under classified research grade observations are very consistent across all iconic taxa subsets. Approximately 80% of chromista and mollusk observations had superfluous identifications, whereas, approximately 40% of mammal observations had superfluous identifications at a 95% threshold level. At a 95% threshold level, protozoans, amphibians, ray finned fish, fungi and reptiles all have approximately 65%-75% of observations that could be considered research grade with a single identification, whereas, more than 80% of chromista, arachnid, mollusk, and mammal observations could be considered research grade with one identification. The differences in results between a 95% and 99% threshold value are reasonably consistent across all results and iconic taxa, with the exception of upgradeable casual grade observations. There is approximately a 44% point reduction in upgradeable causal grade observations between the two threshold levels for ray finned fish, but only a 7% point reduction for mammals and no decrease for protozoans.

Discussion

Citizen science projects are often used to classify a large number of objects and rely on the valuable efforts of volunteer citizen scientists, however, our results demonstrated that these efforts could be used more optimally. In particular, for iNaturalist NZ a high proportion of observations could be considered research grade with a single identification. In the future the excess identification effort could potentially be redirected to observations that need further identifying to reach a desired classification certainty to be considered research grade. However, any redirected observations would likely need to be of a similar taxon because we do not expect that users will be equally accurate at identifying observations to a species level across different taxa. By allowing observations to move to research grade after one identification there would be less interaction between users and perhaps lost learning opportunities for more novice users. There may be an increased risk of incorrect classifications as the process is moving from the "wisdom of the crowd" to relying on the knowledge of an "expert". If this strategy of categorising observations as research grade once they reach a certain classification threshold value was adopted, extra measures could be put in place to ensure inaccurate users were not able to be the sole identifier on a research grade observation. For example, a minimum user accuracy standard could be required for a user to be a 'trusted' sole identifier of a research grade observation. This is particularly important when the initial identification is of a species with a relatively high prevalence as in this case the user may have a relatively low user accuracy and the classification certainty could still be above the research grade threshold. To mitigate the risk of "trusted" users never getting feedback on a series of incorrect identifications, we could require that some of their observations get at least a second identification as a feedback mechanism. The framework outlined in this paper could be used to select the proportion of observations that are required to have at least two identification so as to optimise the trade-off between minimising the potential for classification error and minimising the number of unnecessary identifications.

Currently, iNaturalist NZ observations are in the casual quality grade for two reasons; a non-wild observation, or no media (image or sound file) provided. We removed the non-wild observations from our results. However, if casual grade observations with no media that are contributed by "trusted" users were able to progress to research grade, this could allow more observations to be considered research grade that would otherwise be stuck in casual grade. For example, on iNaturalist NZ there are a number of casual grade observations with no media that have been contributed by established ecologists and are very likely accurately classified, however, they have no means of being verified by another user. There is also a trade-off in a user's effort between the quantity and detail of observations. Observations without media can be made faster

than those with such evidence, and so more can be made for the same amount of effort. Therefore, "trusted" users would be able to make many more research grade observations per unit effort. This would result in more knowledge and monitoring.

The IBCC model allows citizen science projects to estimate user accuracies and collectively classify observations when there is no known ground truth for any observations. The user accuracy estimations are largely based on the user accuracy prior parameters and the level of identification agreement with other users. On iNaturalist NZ there a is very high level of agreement between users at a species and subspecies level, and, especially when we only consider users' most recent identifications. Therefore, there is a risk that the IBCC model over estimated user accuracy values. Also, if all users that identify an image agree on the classification, in most cases the IBCC model considers this to be the correct answer and as a result the user accuracies increase - ignoring the possibility that they could all be incorrect. Further steps in this research would be to consider all the identifications made by users, not just the last, in particular looking at the first identification per user per observation. We could also consider using taxonomic experts to independently assess the identifications of a random subset of observations that we consider at research grade and check for incorrect identifications.

The IBCC model assumes that user identifications are independent. On iNaturalist NZ users are able to see all the previous identifications from other users when they make their identification, in fact, there is an option to simply agree with the previous identification. Given the human tendency to follow the crowd (Asch, 1956) it is very unlikely that all iNaturalist NZ identifications are independent. However, it is difficult to know which identifications would be the same if users were not able to see previous identifications. One method to improve trust in the independence of identifications would be to filter out agreements made by the initial identifier after another user identifies the observation. However, this would just address one example of non-independence between identifications. In reality most of the identifications will be non-independent and this is beneficial when it results in identifiers gradually refining the classification to the true identity.

We simplified the IBCC model outlined by (Kim and Ghahramani, 2012) by assuming a user is equally likely to correctly identify any species or subspecies within an iconic subset, and, if they incorrectly identify the observation, they are equally likely to identify it as any of the remaining species or subspecies in the subset. In contrast, Kim and Ghahramani (2012) modelled user accuracy as a confusion matrix which is a more accurate approach. For example, a user may confuse a *Porphyrio hochstetteri* (Takahē) with a *Porphyrio melanotus* (Pūkeko), two visually similar ground-dwelling

Rallidae but it is probably much less likely that they would confuse a Takahē with a *Rhipidura fuliginosa* (Pīwakawaka) a small flying forest bird. However, given the large number of species and subspecies on iNaturalist NZ and the high proportion of users that have made less that 10 identifications, there is a risk of not having enough data per user and species or subspecies to accurately model user accuracy as a confusion matrix. By calculating a separate user accuracy for each of the 13 iconic taxa subsets we reduce some of the problems that come with not modelling user accuracy as a confusion matrix and greatly reduce the computational expense.

We omitted the insect, bird, and plant subsets from our analysis due to the large size of the data sets and subsequent long running time of the Gibbs sampling algorithm. The inefficiency in the algorithm is due to step 2 of the algorithm where one must calculate $p(t_i = j|\text{all})$ using Equation 3b.3 for $j \in 1, ..., J$ for all observations. However, we could make predictions about results for these subsets based on the results from the 10 subsets we ran the IBCC model on. These predictions could be refined further by the work of Wiggins and He (2016), who found that there are notable bias in the level of identification attention by taxon on the iNaturalist platform. They also found that a heightened level of identification attention contributes to the reliability of the classifications. For example, they found bird observations received more community validation interactions than plant observations. Alternatively, we could split these subsets into smaller subsets based on the taxonomic hierarchy assuming there is a low level of identification crossover between the smaller subsets. For example we could split the plant subset into subsets based on plant phylum. Fortunately, in this study the two full conditional distributions that are sampled from in the Gibbs sampling algorithm are well known distributions, however, if this was not the case one could use suitable techniques to improve the efficiency of the Gibbs sampling algorithm. For example, the use of non-parametric proposal probability density functions for drawing from the full conditional distributions (Gilks et al., 1995), or reusing samples from a Metropolis–Hastings algorithm within a Gibbs sampler (Martino et al., 2018).

The computational efficiency of the IBCC model estimation with Gibb's sampling can be improved by using sequential updating. That means, every time a new rating is done in the iNaturalist NZ system, the model will be rerun only with this new data point rather than with the entire historical dataset. Sequential updating typically involves approximating posterior distributions by a multivariate Gaussian or, if the parameters are independent of each other, a-posteriori by suitable univariate distributions. We found low correlation between our user accuracy parameters a-posteriori and the QQ-plots of Mahalanobis distances for the posterior distribution of the logit transformation of these parameters indicated multivariate normality. It may thus be possible to devise an efficient algorithm for sequential updating, which will be the focus

of our future work.

This study was conducted on a snapshot of iNaturalist NZ data before the complete integration of computer vision to the platform. The initial computer vision model was fully integrated into the iNaturalist platform in September 2017. Subsequently, improvements have been made to the computer vision model in June 2019 and March 2020. This study provides an important baseline of iNaturalist NZ user accuracies and classification reliability prior to the addition of computer vision. There is a need for future work to validate that computer vision is improving the classification accuracy of research grade images on iNaturalist NZ. However, it is important for future data users to recognise that there have been multiple changes to the computer vision technology in this time period and therefore identifications contributed across these changes are fundamentally different due to the different data generating mechanisms. To account for the inclusion of the computer vision, one would have to specify the computer vision accuracy $\Pr(\text{computer chooses } i \mid \text{image is class } j)$ in addition to the user accuracy $\Pr(\text{user chooses } k \mid \text{image is class } j)$, and derive the conditional probability $\Pr(\text{user chooses } k \mid \text{computer chooses } i, \text{image is class } j)$. However, the exact nature of this model is beyond the scope of this paper and is subject of the future work.

This study has shown that by adopting a more sophisticated collective classification method than a simple majority vote procedure citizen science projects could optimise the citizen scientist's efforts. By making a citizen science project more effort efficient large gains could be generated in both the research value of the data and the engagement of the citizens.

Chapter 4a

A mechanistic model of a citizen scientist encountering individuals

The first building block.
A model to simulate,
photos shared per walk.

Abstract

Observation based citizen science data is often collected with minimal collection protocols, which results in significant sources of bias in the data. For example, the list of observations from a citizen science walk is likely to not be a complete list of the individuals encountered during the walk.

In this chapter we outline a mechanistic model of a citizen scientist walking through a field with a population of a given taxon that could be encountered and subsequently an observation of an individual could be shared to iNaturalist NZ. We include some of the typical biases we would expect from citizen scientists, in particular, during a walk, an observation of each encountered individual is not shared to iNaturalist NZ.

The mechanistic model has parameters for the length of the observer walk, observer's perceptive radius, home range radius of the taxon, population size in the field, speed of the taxon relative to the observer, and the probability of an observer sharing an observation to iNaturalist NZ.

We use the mechanistic model to simulate the probability distribution of the number of observations shared during an observer walk and where possible we use an analytical approach to verify the results. Based on our assumption that the parameters in each observer walk are identical, the resulting probability distribution of the number of shared observations is always binomially distributed.

This model is an important step in our wider method to use simulated iNaturalist NZ data to test the ability to make reliable ecological inferences about taxon population changes from noisy citizen science observation data.

Introduction

Over the next three chapters we switch our focus from citizen science image classification accuracy, and instead consider if these observations of taxa can be used to reliably identify biological trends in the taxa abundances. As discussed in the introduction to this thesis, ecological citizen science data may include presence-only data, or presence and absence data (Bird et al., 2014). iNaturalist is an example of a citizen science project that gathers presence-only ecological data. This type of data collection protocol improves the accessibility of the project to varying levels of participants' motivation, commitment, and skills (Bird et al., 2014). However, the lack of information on taxa absence, and no metadata about an observer's effort (e.g. the walk length or duration) increase the challenge of distinguishing between temporal changes in a taxon abundance and temporal changes in observer effort. Throughout, the next three chapters we use stochastic processes, in particular random walks and stochastic simulations, combined with statistical model fitting with the empirical iNaturalist NZ data to understand if we are able to robustly detect ecological changes in taxon abundances from noisy citizen science data.

This chapter is the first building block of our method to find the signal of abundance change in the noise of citizen science data. We outline a model of a single citizen scientist walking through a field with a population of a given taxon that they may encounter and subsequently share an observation of to iNaturalist NZ. Our model aims to include some of the typical biases we would expect from citizen scientists, particularly that observers do not share an observation of every individual they encounter. In the subsequent chapter this single citizen scientist model is scaled up to simulate multiple years of iNaturalist NZ data. We use the simulated iNaturalist NZ data to test the ability to make reliable ecological inferences about taxon abundance changes from noisy citizen science observation data. Finally, in the third chapter we use maximum likelihood model fitting techniques to gain insights into typical citizen science observation sharing behaviours.

Our model is a mechanistic model of a citizen scientist (observer) walking through a field with a population of a given taxon. An individual (i.e. a member of a given taxon) may be encountered and subsequently an observation of the individual could be shared to iNaturalist NZ. The model includes parameters for the length of the observer's walk, the perceptive radius of the observer, the number of individuals in the field, the radius of individuals' home range, and the probability of an observer sharing an observation of an individual to iNaturalist NZ. We consider three scenarios of how an individual moves within their home range. First we assume they are stationary, e.g. a tree, second, we assume they move very fast, e.g. a fast flying bird, and finally we

outline a general model where the individuals are randomly walking within their home range.

For all three cases of individual movement, we use the mechanistic model to simulate the probability distribution of the number of observations shared during an observer walk. For the first two cases an analytical approach is used to verify the simulated probability distributions. The probability distribution is a function of the observer parameters (walk length, perceptive radius, and probability of sharing an observation) and the individual parameters (number of individuals, and home range radius). For simplicity we assume that all observer walks have an identical set of parameters. This is a significant simplification. Ultimately, we are assuming that each observation is shared with the same probability and this results in the probability distribution of the number of observations shared per walk always being binomially distributed. In Chapter 4c we relax the assumption that all observations are shared with the same probability.

Models like we outline in this chapter with two groups moving in space (e.g. walker and population members) have a lot in common with foraging theory models. Such as, predator and prey models, where the predator are aiming to maximise foraging of prey and the prey are trying to avoid the predators (Pitchford et al., 2003). However, we assume the observer never deviates from their pre-defined path and that there is no spatial pattern to the individuals in the field. For example, there is no clustering or repelling of neighbouring individuals, the individual is not reacting to a citizen scientist entering their home ranges, and the citizen scientist is not adopting any optimal foraging theory in an attempt to maximise the number of individuals they encounter.

Random walk background

Random walk processes are a widely used mathematical modelling technique for animal, micro-organisms, and cell movement models (Codling et al., 2008). The insight from the rigorous mathematics that underlies random walk theory gives them the ability to distinguish underlying mechanisms from observed data. This insight has greatly improved the understanding of various movement mechanisms that occur in nature (Codling et al., 2008).

A simple random walk model is uncorrelated and unbiased, where, in this context uncorrelated means the direction of movement is completely uncorrelated with the previous directions moved, and unbiased means that there is no preferable direction. Therefore in a simple random walk model the process is Markovian with regard to the location and the direction moved at each step is completely random (Weiss and Weiss, 1994). This simple model is the basis of most of the theory of diffusive processes.

A common extension of the simple random walk model is a correlated random walk. These random walks include a correlation between successive step orientations that is referred to as 'persistence' (Patlak, 1953). This results in a local direction bias as each step tends to a point in the same direction as the previous step. However, over time the influence of the initial direction of movement diminishes. In the long term step orientations are uniformally distributed. Since animals have the tendency to move forward (i.e. persistence) correlated random walks have been frequently used to model animal paths in many contexts (Siniff and Jessen, 1969; Bovet and Benhamou, 1988; Turchin, 1998).

A biased random walk is when a global directional bias is introduced by making the probability of moving in a certain direction greater. This leads to the drift-diffusion equation. A range of factors may be causing the bias. For example, chemical gradients (Alt, 1980), mean-reversion movements within a home range (Blackwell, 1997; James et al., 2017), or external environmental factors (Hill and Häder, 1997).

Movement within a confined area can be modelled by introducing a repelling or reflecting boundary condition. An absorbing boundary condition can be used to model walkers leaving the system upon reaching a given point. These models can also used to model development or growth where different life stages are reached, (Pitchford and Brindley, 2001; Pitchford et al., 2005).

In our model we assume that the individuals are performing unbiased and uncorrelated random walk with a reflecting boundary condition on the home range perimeter.

Mechanistic model

Figure 4a.1 shows the layout of the mechanistic model where there are n individuals randomly placed in the field of area Ω. Each individual is confined to a home range with radius R_H. The observer walks a length L from point A to point B and has a perceptive radius of R_P.

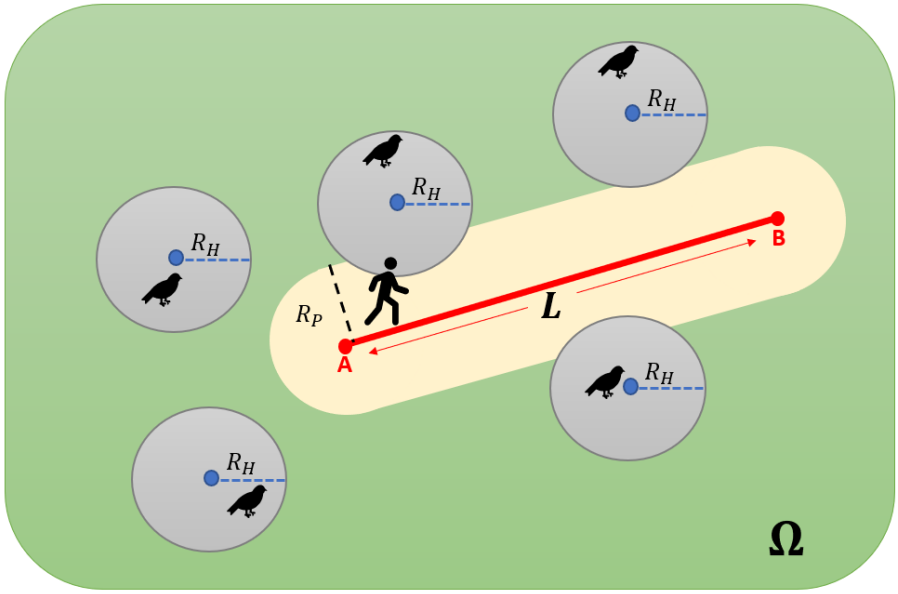


Figure 4a.1: **Model diagram of an observer walking in a field containing individuals.** The observer walks a distance L from point A to point B and has a perceptive radius of R_P and therefore has a perceptive area as shown in pale yellow. The individuals are randomly placed in the field and are confined to a circular home range with radius R_H, shown in grey. The field has an area Ω.

Case 1: Individuals do not move

First we consider the case where the individuals are stationary while the observer walks a path from point A to point B. For example, plant species would be modelled by this case. In this case the individual's home range is not relevant as the individual is unable to move. Figure 4a.2 illustrates this simplified layout of the mechanistic model.

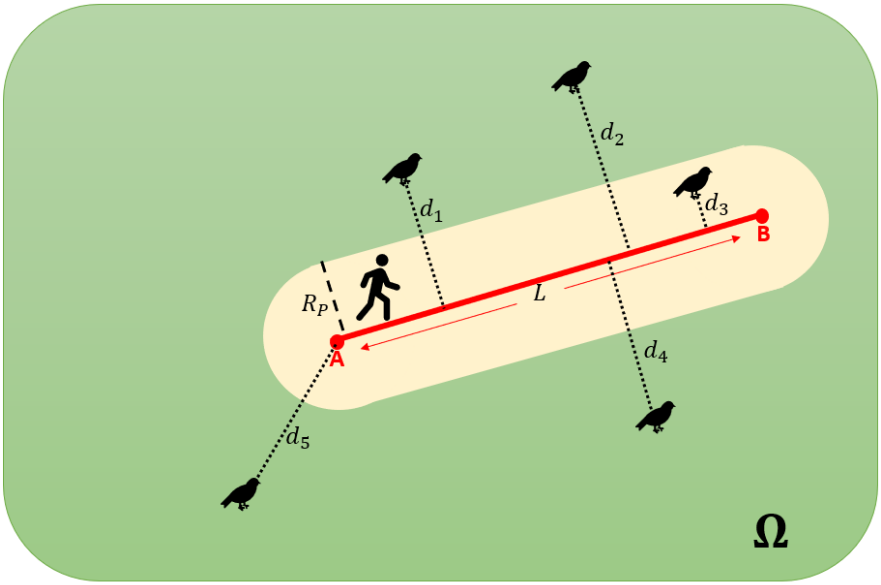


Figure 4a.2: **The home range is not applicable for stationary individuals.** The individuals do not move in the time it talks the observer to walk from point A to point B and therefore the home ranges of the individuals are not required.

Analytical solution of probability of encountering a single individual

Assuming individuals are independent, in this case the distribution of individual observations per observer walk of length L follows a binomial distribution with parameters n and p. Where n is the number of individuals and p is the probability that an individual will be encountered and shared to iNaturalist NZ by the observer. The probability, p, a single individual is encountered and is shared on iNaturalist NZ is given by the probability a single individual is encountered, p_{enc}, multiplied by the probability an observer photographs and shares the photograph to iNaturalist NZ, p_{shared}. The probability a single individual is encountered, p_{enc} is analytically calculated as the proportion of the field that is covered by the perceptive area of the observer (pale yellow in Figure 4a.2):

$$p_{\text{enc}} = \frac{2R_P L + \pi R_P^2}{\Omega}. \tag{4a.1}$$

$$p = p_{\text{enc}} \cdot p_{\text{shared}}$$

Clearly, if the observer's perceptive radius, R_P, increases or the length of the walk, L, increases, the observer's perceptive area will increase and therefore the probability of an observation, p, will increase, assuming the field area, Ω, is unchanged. In practise we assume that the observer's perceptive radius is relatively small and the length of the walk is relatively long, therefore the $\frac{\pi R_P^2}{\Omega}$ term in Equation 4a.1 is negligible and Equation 4a.2 can be used as an approximation for the perceptive area when the path walked is not perfectly straight.

$$p_{\text{enc}} \approx \frac{2R_P L}{\Omega}. \tag{4a.2}$$

The expected number of individual observations per citizen science walk of length L is given as:

$$E(\text{individual observations}) = np$$

Numerical verification of analytical solution

The analytical results for this simple case can be easily verified by a simple simulation as follows:

1. Randomly place n individuals on a grid with an area Ω.
2. Define an observer walk of length L.
3. Calculate the minimum distance, d_n, from each individual to the path of the observer's walk. Where the minimum distance is the length of a perpendicular line from the individual's location to the citizen scientist's walking path.
4. Sum the number of points that are within the observer's perceptive area, i.e. $\sum_n (d_n \leq R_P)$. Each observation is shared with probability p_{shared}, independent of other observations.
5. Repeat the simulation many times and record the number of encounters per simulation (i.e. citizen scientist walk).

Figure 4a.3 shows the analytically calculated and simulated distribution of individual encounters when, $R_p = 15$, $L = 250$, $\Omega = 500^2$, and $n = 50$, with $p_{\text{enc}} = 0.03$.

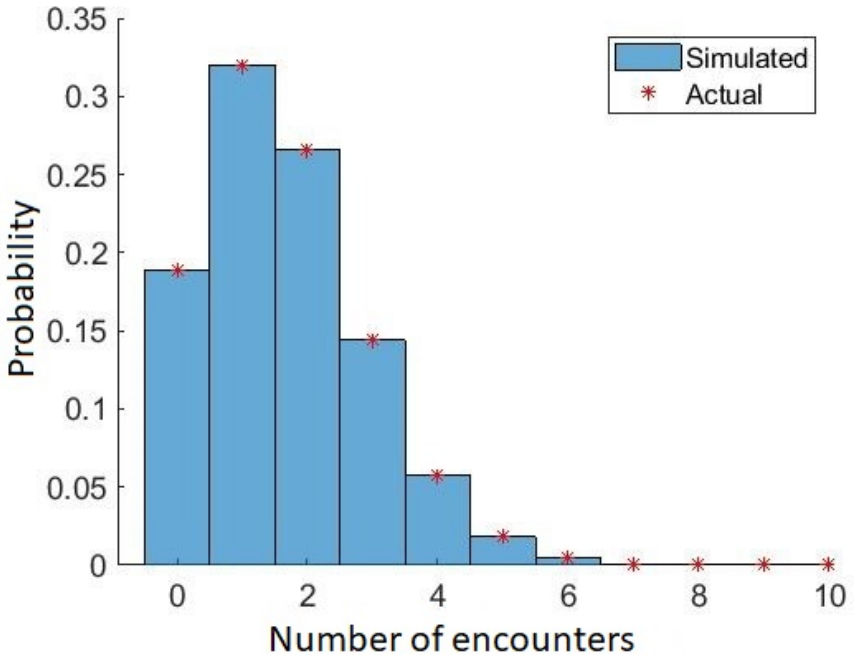


Figure 4a.3: **Individual encounters are binomially distributed when individuals do not move and all observers have the same observation behaviour.** The analytically calculated and simulated distribution of individual encounters during one citizen science walk when, $R_p = 15$, $L = 250$, $\Omega = 500^2$, and $n = 50$, with $p_{\text{enc}} = 0.03$.

Case 2: Individuals move very fast

At the other end of the scale is the possibility that the individuals move so fast they cover their entire home range every time the observer takes a step. For example, a *Rhipidura* (Pīwakawaka) or a *Petroica australis* (South Island Robin) move so fast that you are almost certain to observe them as they fly around their home range so quickly and frequently. We model this case in a similar way to the first case where individuals did not move. However, now we want to know if any section of the individual's home range overlaps with the observer's path. Therefore, rather than calculating if an individual is within the perceptive radius, R_P, of the observer's path we want to calculate if the origin of an individuals home range is within the combined radius of the perceptive radius and home range radius, $R_P + R_H$, of the path. Figure 4a.4, illustrates the layout of this model. In particular, Figure 4a.4 shows three possible scenarios for the origin of the individual's home range:

1. The distance, d, from the origin of the individual's home range to the path is more than the combined radius of the perceptive radius and home range radius, i.e. $d > R_P + R_H$. In this case there is no overlap of the individual's home range and the perceptive area and therefore the individual is not encountered.

2. $d = R_P + R_H$, at the point where the home range and perceptive area are tangent the individual is encountered.

3. $d < R_P + R_H$, there is an overlap between the individual's home range and the perceptive area and therefore the individual is encountered.

Analytical solution of probability of encountering a single individual

Similar to the stationary individual case, the distribution of individual observations per observer walk of length L follows a binomial distribution with parameters n and p, where n is the number of individuals and p is the probability that an individual will be encountered and shared to iNaturalist NZ by the observer. The probability an individual is encountered, p_{enc}, is analytically calculated as the proportion of the field that is covered by sweeping out an area surrounding the observer's path of width $R_P + R_H$ (pale yellow and blue in Figure 4a.4):

$$p_{\text{enc}} = \frac{2(R_P + R_H)L + \pi(R_P + R_H)^2}{\Omega}. \tag{4a.3}$$

Again, as in the case with stationary individuals this encounter probability, p, will increase if the observer's perceptive radius (R_P) increases, the length of the walk (L) increases, the probability of sharing an observation (p_{shared}) increases, or in this case if

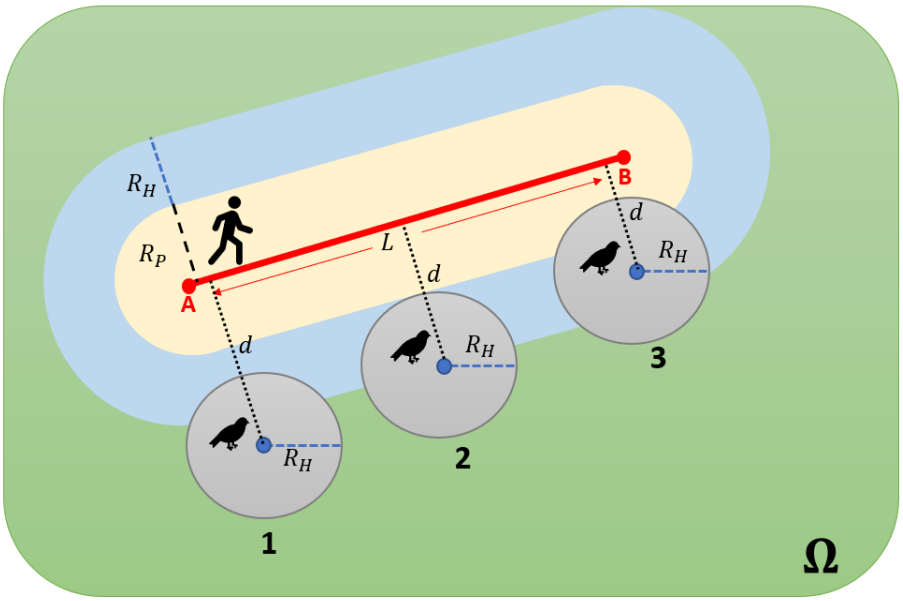


Figure 4a.4: **Model diagram of an observer walking in a field containing fast moving individuals.** Individuals more very fast and cover their entire home range every time the observer takes a step. If the distance, d, from the origin of the individual's home range to the path of the observer is less than or equal to the combined radius of the perceptive radius and home range radius the individual will be encountered. Individual 1 is not encountered, individual 2 and 3 are encountered.

the home range radius (R_H) of the individual increases, assuming the field area (Ω) is unchanged.

Numerical verification of analytical solution

Similar to the case of stationary individuals the analytical results for this case can be easily verified by a simple simulation as follows:

1. Randomly place n individuals on a grid with an area Ω.
2. Define an observer walk of length L.
3. Calculate the minimum distance, d_n, from each individual to the path of the observer's walk.
4. Sum the number of points that are within the combined radius of the perceptive radius and home range radius, i.e. $\sum_n (d_n \leq R_P + R_H)$. Each observation is shared with probability p_{shared}, independent of other observations.
5. Repeat the simulation many times and record the number of encounters per simulation. 4a.3.

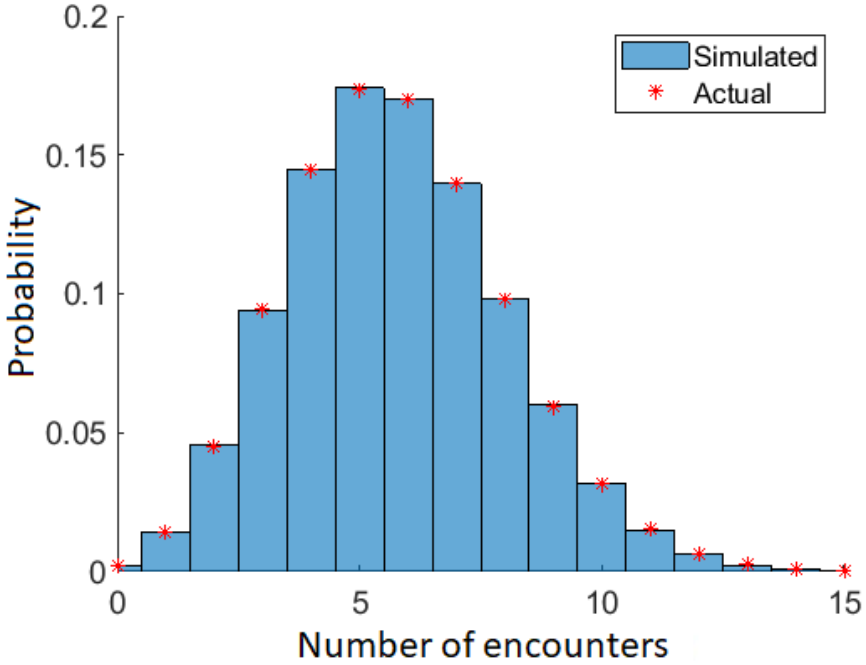


Figure 4a.5: **Individual observations are binomially distributed when individuals move very fast and all observers have the same observation behaviour.** The analytically calculated and simulated distribution of individual observations during one citizen science when, $R_p = 15$, $R_H = 30$, $L = 250$, $\Omega = 500^2$, and $n = 50$, with $p_{\text{enc}} = 0.115$.

Figure 4a.5 shows the analytically calculated and simulated distribution of individual encounters when, $R_p = 15$, $R_H = 30$, $L = 250$, $\Omega = 500^2$, and $n = 50$, with $p_{\text{enc}} = 0.115$. Note that all the applicable parameters are the same as the example for the stationary individual case but as expected the probability, p_{enc}, of an individual encounter is higher. Figure 4a.6 shows how the probability of an individual observation increases as either the perceptive radius or the home range radius increases when $L = 250$, and $\Omega = 500^2$, and $p_{\text{shared}} = 1$. The walker will encounter the individual if the walker is within the individual's home range. Therefore, even when the walker's perceptive radius is zero the probability of an encounter increases as the size of the individual's home range increase.

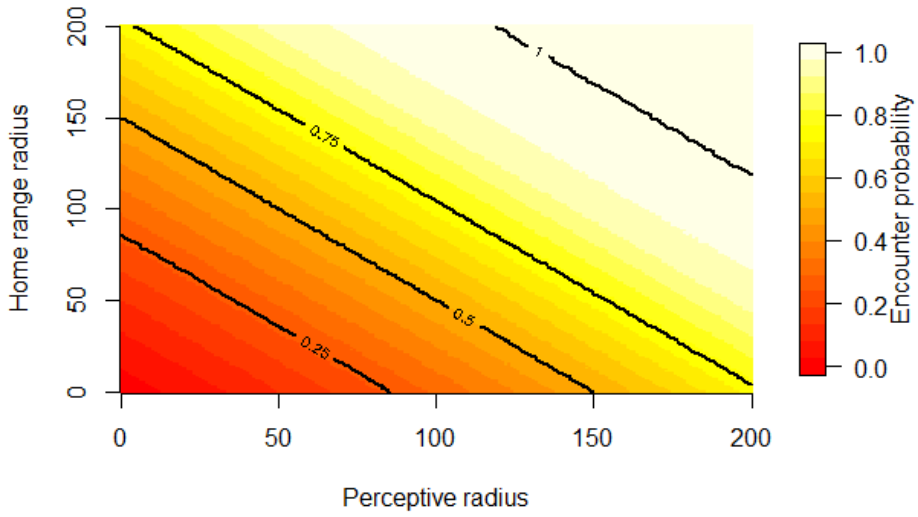


Figure 4a.6: **For a very fast moving individual the probability of an encounter increases as either the observer's perceptive radius increases or the individual's home range radius increases.** A unit increase in an observer's perceptive radius, or in the individual's home range radius equally increase the individuals encounter probability.

Case 3: Individuals do a random walk

A more general scenario is that the individuals are doing a random walk inside their home range as the observer walks through the field. To build this into the mechanistic model we assume the individuals are performing a simple unbiased and uncorrelated random walk with a reflecting boundary condition on the home range perimeter. A diagram of this scenario is shown in Figure 4a.7. The individual is confined to a home range with radius R_H. Each step the individual moves a distance δx and τ is the time between steps. The diffusion coefficient is given as $D = \frac{\delta x^2}{2\tau}$, (Codling et al., 2008). The diffusion coefficient is held constant as we alter the individual's step size, δx, and therefore τ to ensure the results from the random model are convergent. The citizen scientist travels at a constant speed V_W along a path of length L from point A to point B and has a perceptive range of R_P.

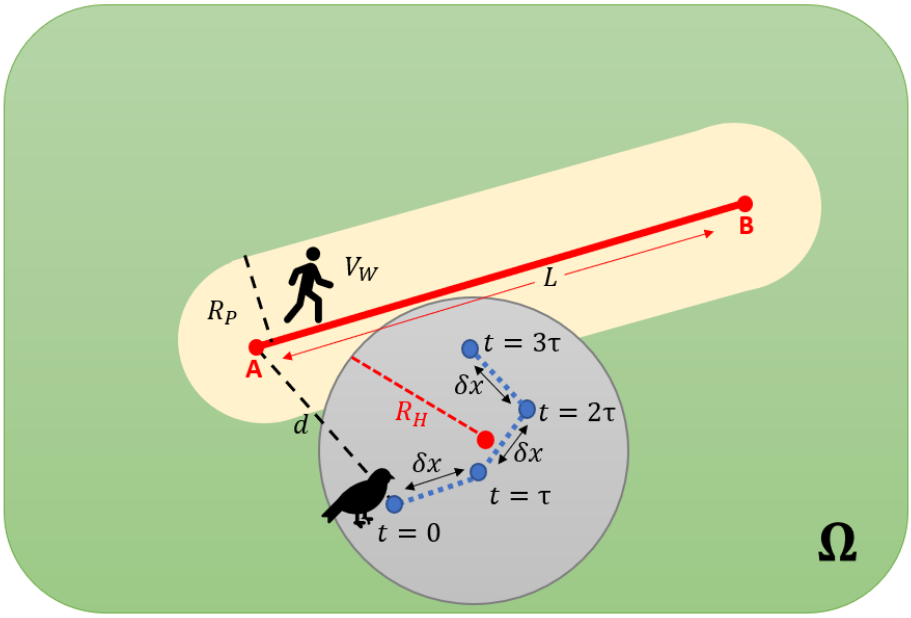


Figure 4a.7: **The generalised case where the individual is doing a random walk within it's home range.** The individual is doing a simple uncorrelated and unbiased random walk with a reflective boundary condition. The individual random walk step size is δx and the time between steps is τ.

Simulated probability of encountering a single individual

In the previous two cases we calculated p_{enc} analytically and numerically verified the calculation. In this case we use simulations to find the probability the observer will encounter the individual as below:

1. Uniformly select a random point in the field as the origin of the individual's home range.

2. Uniformly select a random starting point for the individual within the home range.

3. Calculate the distance, d, from the individual's starting point to the start point A of the citizen scientist's walk.

4. If d is less than the observer's perceptive radius, $d < R_P$, record that the individual has been encountered.

5. Perform a step in the individual's random walk:

   - Randomly select an angle, $\theta \in [0, 2\pi]$, for the direction of the individual's step and calculate the new location of the individual after taking a step of distance δx.
   - If the new individual location falls outside of the individual's home range change the step direction to be in the opposite direction to impose a reflective boundary condition on the perimeter of the home range.

6. Calculate the location of the observer given t and V_W.

7. Calculate the distance, d, from the individual's new location to the updated location of the observer.

8. If d is less than the observer's perceptive radius, $d < R_P$, record that the individual has been encountered. Note, we are assuming the observer recognises if they have already encountered the individual and therefore the individual is either encountered or not encountered during the walk.

9. Repeat steps 5 - 8, $n_{\text{steps}} = \frac{V_W L}{\tau}$ times until the observer has finished their walk.

10. Repeat the simulation many times and each simulation record if the individual was encountered.

11. $Pr(\text{individual encounter}) = \frac{\text{number of simulations with an individual encounter}}{\text{number of simulations}}$.

The initial location of the individual within their home range is uniformally distributed. Therefore, in the short and long term the distribution of locations the individual has visited within their home range is uniform. In our simulations the individual typically does 1500 steps and only covers a small portion of their home range. This is why we chose this random walk specification as it models the walker encountering the individual at a random time and location. We could have chosen an alternative

specification, for example a mean-reversion random walk, where the individual is likely to move back towards the centre of the home range as they stray toward the edge of the home range (James et al., 2017).

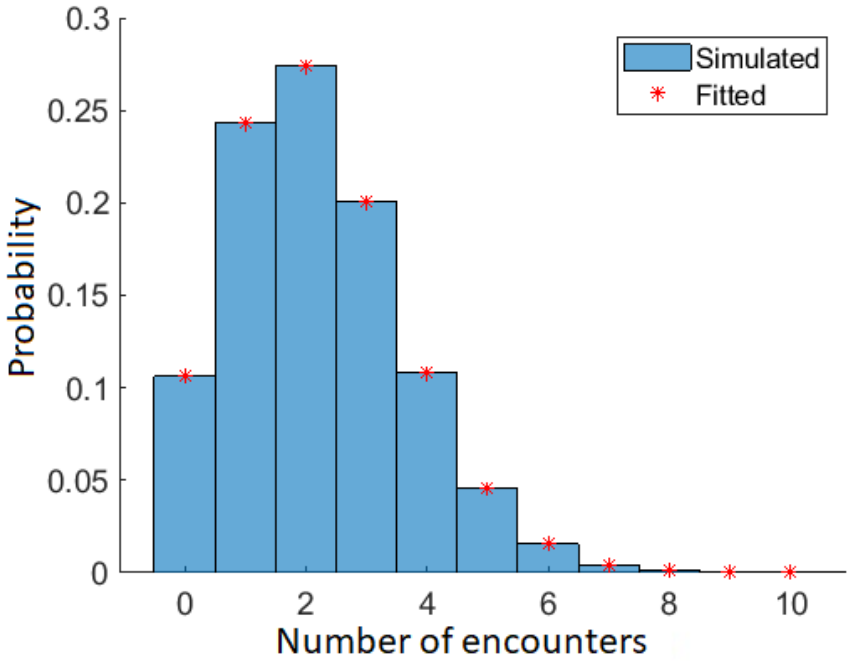


Figure 4a.8: **Individual encounters are binomially distributed when individuals randomly walk within their home range and all observers have the same observation behaviour.** The simulated distribution and a fitted binomial distribution of individual encounters when, $n = 50$, $R_p = 15$, $R_H = 30$, $L = 250$, $\Omega = 500^2$, $V_W = 2$, $\delta x = 0.2$, $D = 0.0625$, and no. steps $= 1562$ with $p_{\text{enc}} = 0.036$.

Similar to the stationary and fast moving individual cases the distribution of individual encounters per observer walk of length L follows a binomial distribution with parameters n and p_{enc}. The probability, p_{enc}, of an individual being encountered by the observer is calculated using the above simulation and is dependent on the length of the observer's walk, L, the individuals home range radius, R_H, and the observer's perceptive radius, R_P.

Figure 4a.8 shows the simulated distribution and a fitted binomial distribution of individual encounters where the simulation parameters are: $n = 50$, $R_p = 15$, $R_H = 30$, $L = 250$, $\Omega = 500^2$, $V_W = 2$, $\delta x = 0.2$, $D = 0.0625$, and no. steps $= 1562$ with $p_{\text{enc}} = 0.036$. As expected, the probability of encountering an individual converges to approximately 0.036 as the size of δx is reduced but $D = 0.0625$ is held constant.

In this example the parameters that are also applicable to the stationary and fast

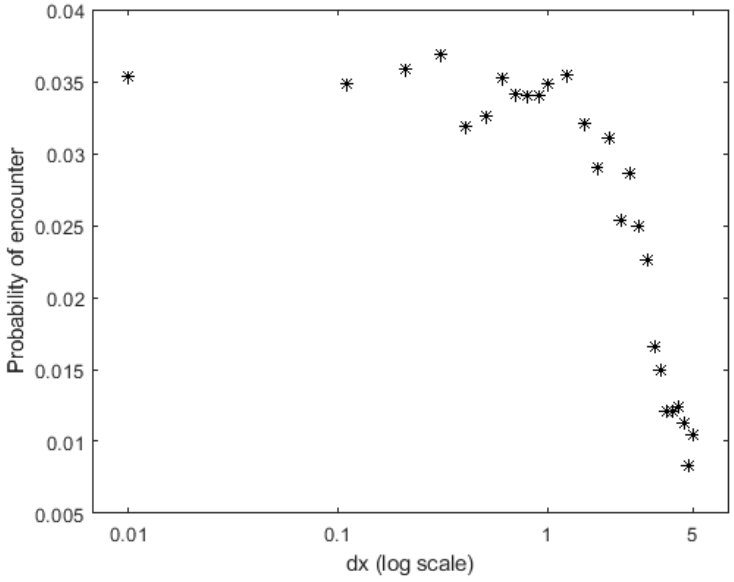


Figure 4a.9: **The probability of encountering an individual converges as we reduce the individual step size in the random walk model.** As expected, the probability of encountering an individual converges to approximately 0.036 if we fix $D = 0.0625$ and reduce the random walk step size δx.

moving individual cases are the same but as expected the probability, p_{enc}, of encountering an individual falls between the p_{enc} from these two extreme cases. Figure 4a.10, shows how the probability of encountering a random walking individual converges to the slow and fast individual cases as we vary the relative speed of the individual compared to the citizen scientist walking speed. For example, when the individual's speed is 1/10 the speed of the citizen scientist the probability of an individual encounter is close to the case of the stationary individuals where the relative speed is zero. As the individual's speed becomes much faster than the citizen scientist's speed the probability of an individual encounter increases. The upper-bound on the encounter probability is given by the case where the individual is moving very fast. In Figure 4a.10 we also plotted the results for an example with a larger perceptive radius. In this example, the lower-bound and upper-bound are higher than the previous example, and as expected the probability of encounter from the random walk model are also higher.

Figure 4a.11 shows how the probability, p_{enc}, of a individual encounter increases as either the observer's perceptive radius increases or the individual's home range radius increases, with the exception of when the walker's perceptive radius is zero. When the individual is doing a random walk, if the walker's perceptive radius is zero the only time the walker will see the individual is if they are both on exactly the same x, y coordinate. The probability of this occurring is effectively zero and therefore regardless

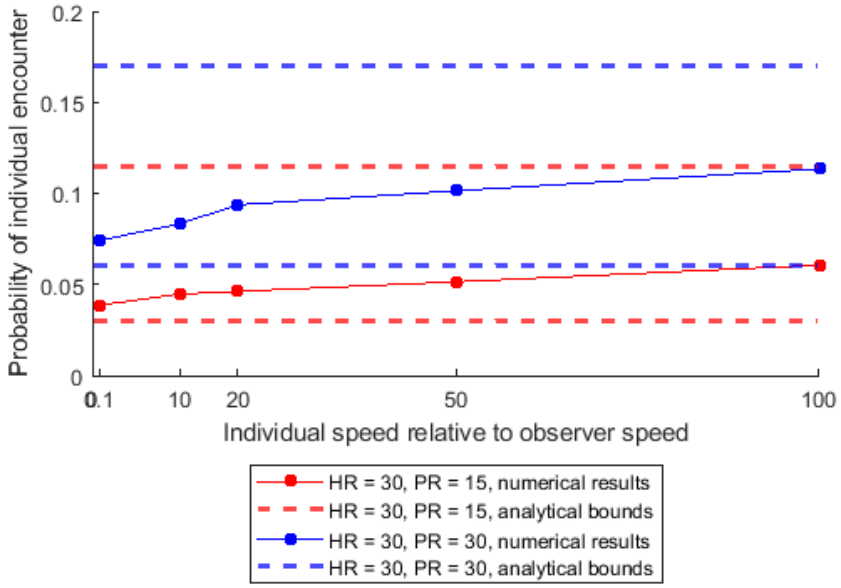


Figure 4a.10: **The probability of encountering an individual converges to the stationary and very fast moving cases as we vary the relative speed of the individual compared to the observer.** The dashed lines are the analytical lower and upper-bounds that are given by the stationary and fast moving individual cases. When the individual's speed is 1/10 the speed of the observer the probability of an individual encounter is close to the case of stationary individuals where the relative speed is zero. As the individual's speed becomes much faster than the observer the probability of an individual encounter increases. Increasing the walker's perceptive radius shift all the probabilities up.

of the size of the individual's home range, if the walkers perceptive radius is zero, there is zero chance of an encounter. This is in contrast to the fast moving individual case where if the walker is within the individual's home range the walker will encounter the individual, regardless of the size of the walker's perceptive radius.

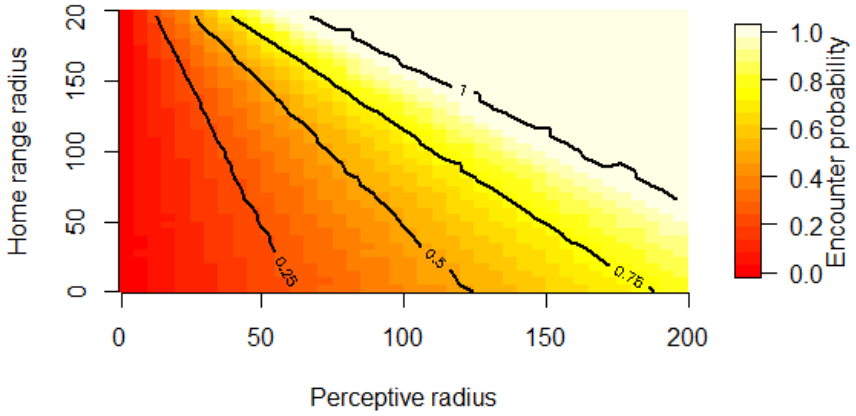


Figure 4a.11: **For an individual randomly walking within their home range, the probability of encountering an individual increases as either the observer's perceptive radius increases or the individual's home range radius increases.** With the exception of when the walker's perceptive radius is zero, where there is a zero probability of the walker encountering the individual regardless of the size of the individual's home range.

Discussion

In this chapter we built a simple mechanistic model to describe the citizen science scenario of an observer walking in a field and encountering individuals that they may subsequently share an observation of to iNaturalist NZ. We made a bold assumption that all observers walks have the same set of parameters. This is a particularly simplifying assumption with regard to the length of the observer walk, number of individuals in the field, and the probability that an observer will share an observation of each individual taxon they encounter. We also assumed that an observer never shares more than one observation of the same individual, but is equally likely to share subsequent observations of different individuals of the same taxon. We assumed there is no spatial pattern to the individuals within the field. As a result the probability distribution of the number of observations shared per observer walk is always binomially distributed. The probability parameter in the binomial distribution encompasses both the probability that an encounter occurs and that an observation of the encounter is subsequently shared to iNaturalist NZ.

The probability an individual is encountered depends on the home range radius of the individual, the perceptive radius of the observer, the length of the observer's walk, and the relative speed of the individual compared to the observer. We considered the two extreme relative speed cases where individuals are either stationary or moving very fast compared to the walker. The stationary case models plant and fungi species where the observer will not encounter the individual unless it is within their perceptive radius. Whereas, the very fast moving individual case models species like the fantail or South Island robin where if any part of the individual's home range is within the observer's perceptive radius they will encounter the individual as the individual is constantly covering their entire home range.

In this chapter we assumed that an observer has the same probability of sharing an observation of each individual they encountered. In reality this is unlikely to be the case for the majority of iNaturalist NZ observers. It is more likely observers have a range of different behaviours. For example, some observers may regularly share many observations, other observers may only share one observation per walk, or some observers may have a decaying probability of sharing an observation as the walk and number of encounters progress. In Chapter 4c we revisit this assumption by fitting multiple candidate models of the probability of sharing an observation to empirical data from iNaturalist NZ.

The mechanistic model in this chapter was built to simulate iNaturalist NZ observers sharing observations of individuals they have encountered in the wild. This is

the first building block in our larger method to understand if we are able to robustly detect temporal ecological changes in species abundance from noisy citizen science data. In Chapter 4b we use the result from this chapter - that the number of observations shared per walk is binomially distributed - to simulate multiple years of iNaturalist NZ data from multiple observers. We use the simulated iNaturalist NZ data to test the ability to make reliable ecological inferences about taxon abundance changes from noisy citizen science observation data.

Chapter 4b

Extracting the change signal from noisy ecological citizen science data

Annual trends are found.
Are they species driven or,
observer changes?

Abstract

Observed counts in ecological monitoring programmes, particularly when citizen scientists are involved, are the result of two linked stochastic processes. The first stochastic process is that of the true biological state, and the second is the observation process that consists of the variation in the observer's behaviour.

Inference is desired on the first stochastic process about the true biological state. However, when there are limited measurable covariates about observer behaviour, it is extremely challenging to specify a statistical model and make inferences about the biological state without falsely drawing conclusion based on changes in observer behaviour.

In this chapter we generate computer simulated citizen science data based on the iNaturalist NZ dataset with sources of variation in observer effort. We use the simulated data to test the ability of statistical models to accurately infer annual trends in species abundance.

We find that with the current number of years of observation data on iNaturalist NZ, and the minimal amount of information about variation in observer behaviour, it is difficult to use statistical methods to reliably estimate annual trends in species' abundances.

We apply our method in more detail to the kiwi taxa and show that data from iNaturalist NZ is unlikely to be able to detect a 2% annual increase in kiwi over a 10 year period with the current level of kiwi observation activity on iNaturalist NZ. However, over a 20 year period iNaturalist NZ data would be able to reliably detect a 2% annual increase in the combined kiwi taxa.

The work in the chapter highlights the importance of a citizen science project having measurable covariates about observer behaviour if the aim is to use the collected data to make ecological inferences about temporal changes in species abundances.

Introduction

Monitoring changes in species' populations is crucial to identifying extinction risk (Mace and Lande, 1991), evaluating the performance of conservation efforts (McKinley et al., 2017), monitoring ecological responses to climate change, and reporting against national and international targets (Butchart et al., 2010). Species monitoring by trained experts and in standardised forms is resource intensive and costly. These practices struggle to provide the information required for extensive ecological monitoring as they are often biased towards developed countries in temperate regions (Hudson et al., 2014; McGeoch et al., 2010; Martin et al., 2012), and monitoring schemes typically do not deliver long-term data or at regional scales, making it difficult to monitor biodiversity change across space and time (Hudson et al., 2014; Schmeller et al., 2009). Citizen science, on the other hand, provides access to an abundance of labour, skills and computational power (Cohn, 2008b; Silvertown, 2009). Often citizen science is the only practical way to answer ecological questions at the scales relevant to species range changes, migration patterns, disease spread and impacts of climate change (Tulloch et al., 2013; Miller-Rushing et al., 2012). Furthermore, citizen science has exploded in popularity in recent years. For example, the birdwatching citizen science project, eBird, has formed one of the largest biodiversity datasets in the world (Wiggins, 2011). However, there is ongoing debate and doubt on the precision and usefulness of citizen science data in scientific research due to the uncertainty of the variability in participants abilities and effort, and the opportunity for persistent bias in the data (Crall et al., 2011; Dickinson et al., 2010).

Many ecological citizen science projects began with the goal of engaging participants in the natural world, rather than collecting data for scientific research (Tonachella et al., 2012). As a result, many citizen science projects have collection protocols that limit the ability to apply post-hoc statistical manipulation to the data. This may influence the types of research questions particular citizen science data are able to answer and/or make it difficult to draw ecological inferences (Bird et al., 2014; Crall et al., 2011; Wiggins et al., 2011; Tonachella et al., 2012). For example, the variation in participants' sampling efforts spatially, temporally, and across taxa are often unknown and difficult to infer, making it difficult to isolate true ecological changes from changes in participants' efforts and abilities. However, collection protocols vary widely across citizen science projects and some citizen science projects collect an extensive amount of metadata on the participants' behaviours. For example eBird requires observers to select one of four different protocols to follow while counting birds. Three of these are effort-based sampling protocols: travelling count, stationary count, and area count. These require participants to supply associated information, for example the amount of time spent observing and the distance travelled. The fourth protocol is a less rig-

orous option, called "casual observation" and only requires date, location, and species observed to describe the sampling event. Once an observer has selected a protocol and entered location information, a checklist is displayed of the species most likely to be observed at the reporting location on the selected date. The participant then provides the number of individuals seen of each species and submits the completed checklist to the eBird database for further verification (Sullivan et al., 2009). On the other hand, the iNaturalist citizen science project is completely curiosity driven with no standard sampling structure. iNaturalist observers share observations when they desire, where they desire, and of what they desire and there is no need or option to record species absences, or participant recording effort. In this chapter we show that this lack of information about recorder effort severely limits the ability to use the iNaturalist NZ dataset to draw ecological inferences about species abundances.

The debate on the ability of citizen science to produce data as useful as that from standardised collection methods has been ongoing in the literature (Cohn, 2008b; Galloway et al., 2006). In some cases there is substantial variation in the quality and usability of data collected by volunteers versus professionals, and at other times there are negligible differences. The variation largely depends on the scope of the project and collection protocols, the volunteer skill level, and the professional data collection methods (Galloway et al., 2006). Some research has shown the combination of the large size of the datasets and the use of appropriate statistical and analytical tools means that any bias and noise in the data can be minimised, and citizen science data can provide similar information to professionally collected and designed monitoring programs (Szabo et al., 2012; Hochachka et al., 2007; Szabo et al., 2010). For example, Szabo et al. (2012) compared two datasets of bird observations from Mount Lofty Ranges, South Australia. The first dataset was from a weakly structured national bird atlas collected by volunteer surveyors who were free to choose where and when to visit. The second dataset was from monitoring surveys by experienced and paid surveyors that are carried out twice a year, using a stratified sampling design to determine the location of the survey. Szabo et al. (2012) found that the two independent datasets were highly correlated and minimum population estimates from the two datasets agreed very well. On the other hand the ability of citizen scientists (even those who have received training) to produce data that has a comparable quality to data collected by professionals has been questioned in the scientific literature (Anderson, 2001; Fitzpatrick et al., 2009). For example, Fitzpatrick et al. (2009) conducted an experiment to compare the ability of volunteer and experienced observers to detect low-density populations of the hemlock woolly adelgids (*Adelges tsugae*), an insect pest of eastern hemlock (*Tsuga Canadensis*) trees. They found that volunteers who received 15 minutes of training had a lower probability than experienced observers at detecting the low densities of the actively spreading invasive species in hermlock trees.

Appropriate statistical and analytical tools are often required to address variation/bias in the data due to recorder behaviour or data quality issues in citizen science data. However, the types of tools that are useful and reliable are very dependent on the nature and the specifications of the citizen science project. Isaac et al. (2014) constructed computer simulated citizen science data and compared type I error rates (the false rejection of the null hypothesis) of 11 statistical methods that aim to estimate the temporal trend of a focal species in noisy ecological data. The sources of bias they simulated were: an increase in site visits over time, an observer bias towards sites with a high proportion of the focal species, a decline in sampling effort per site visit, species becoming more detectable over time, and a decline in non-focal species over time. The methods of trend estimation they considered included a simple approach of a Poisson generalised linear model (GLM) that included no mechanism to control for variation in recorder activity. They also modelled the response variable as a proportion of visits in a given year that produced a record of the focal species by using a binomial GLM with year as a covariate. Modelling the focal species as a proportion is expected to make the trend estimate robust to variations in recording effort. They added extra components to the simple binomial GLM to account for other sources of variation in the data. For example, including a covariate of the number of unique species per site visit to control for uneven sampling effort. Their simulations and subsequent calculation of type I errors for all 11 methods found that no method was wholly robust under all variation scenarios, but the more detailed models - for example the models that included covariates for species list lengths - are more robust to additional sources of variation. A successful example of using a relatively sophisticated method to model citizen science count is by Tonachella et al. (2012). They used a mixed effect generalised linear model (GLMM) to analyse data from the Great Whale Count, an annual citizen science event where humpback whale (*Megaptera novaeangliae*) sightings in Maui County are counted by volunteers for one day during the breeding season. Observer, year, and site effects were included in the GLMM to account for the bias and variation these effects contribute to the count data. Their random effects model estimated an increase of humpback whale sightings in the Maui coastal waters of 5.2% per year which is similar to other Hawaiian humpback whale counts (Mobley et al., 2001).

It is not always a given that suitable covariates/explanatory variables about observer behaviour will be recorded throughout the citizen science data collection process. This is especially the case for ecological citizen science projects that did not initially have an aim of using the collected data to make inferences about changes in species abundances. In this chapter, we investigate the ability to use statistical models to detect ecological changes in species abundances when there are minimal measurable covariates that describe observer behaviour. We follow a similar method to Isaac et al.

(2014) of using computer simulated citizen science data to understand the statistical properties of candidate models that aim to find the signal of annual species abundance change in the noise of citizen science data. The chapter has the following layout.

Chapter outline

iNaturalist NZ data. We summarise and visualise the relevant data from iNaturalist NZ that will be used to both parametrise the simulated data, and in the model fitting on iNaturalist NZ data section.

Statistical model background. We provide background information on generalised linear models (GLM), and in particular Poisson GLMs as this is the family of model we use throughout this chapter.

Model fitting on simulated data. We extend the single citizen scientist simulations from Chapter 4a to construct simulated citizen science data from multiple observers over multiple walks and years. Where possible our simulated data is parametrised using iNaturalist NZ data. We add variation into the simulated data by annually increasing the number of observers, allowing variation in the annual number of walks per observer, and allowing variation in the number of individuals per observer walk. For each additional source of variation in the simulated data we test if we are able to specify a suitable Possion GLM by measuring the type 1 error rate.

iNaturalist NZ trends. We fit the most comprehensive candidate model given the explicitly known meta-data about citizen science behaviour to all the iNaturalist NZ species that have at least 7 years of data.

Case study: Can citizen science data be used to reliably detect a 2% annual increase in a kiwi species abundance? The 2018 - 2028 Kiwi Recovery Plan in New Zealand outlines the goal to reach 100,000 kiwi by 2030 by growing all kiwi species by at least 2% per year (Germano et al., 2018). We use simulated data to examine the power of our most comprehensive model to detect a genuine 2% increase in a species. In this section the simulated data is parametrised to reflect the iNaturalist NZ kiwi data. We explore the relationship between the model power and years of citizen science data.

Discussion. We discuss the findings of this chapter and potential paths forward to be able to reliably use iNaturalist NZ data to make ecological inferences about temporal changes in species abundances.

iNaturalist NZ data

In this section we outline and visualise the key data from iNaturalist NZ that will be used to parametrise the simulated data. We consider research grade observations for the 7 years from January 2013 to December 2019. Over the 7 years there have been 449,512 research grade observations by 10,499 observers. For the remainder of this chapter we refer to research grade observations as, simply, observations. Both the numbers of observations and observers have increased steadily over the last 7 years, as shown in Table 4b.1. However, the average number of observations per observer has been relatively stable over the 7 years.

Year	Number of active observers	Number of observations	Average observations per observer
2013	514	17033	33
2014	699	25786	37
2015	1356	34418	25
2016	1571	48314	31
2017	2549	68267	27
2018	3472	101907	29
2019	4974	153787	31
Total	**10499**	**449512**	**43**

Table 4b.1: **The number of active observers and observations per year has increased steadily over the last 7 years.** This table shows the annual number of iNaturalist NZ observers, research grade observations and the average number of observations per observer from 2013 to 2019.

Within the list of observations there are 11407 unique taxa and Figure 4b.1 shows the distribution of total observations per taxa. A large number of taxa have had very few observations over the 7 years, with 25% of the taxa having had only 1 observation, and 68% of taxa have had on average 2 or less observations per year. At the other end of the scale, Kererū is the taxon with the largest number of observations, and there have been 8404 Kererū observations over the 7 years.

Figure 4b.2 shows the distribution of taxon observations in the first year the taxon was observed by an iNaturalist NZ observer. Most taxa had very few observations in the first year they were observed by an iNaturalist NZ observer with 62% of taxa only having one observation in the first year and 96% of taxa having 10 or fewer observa-

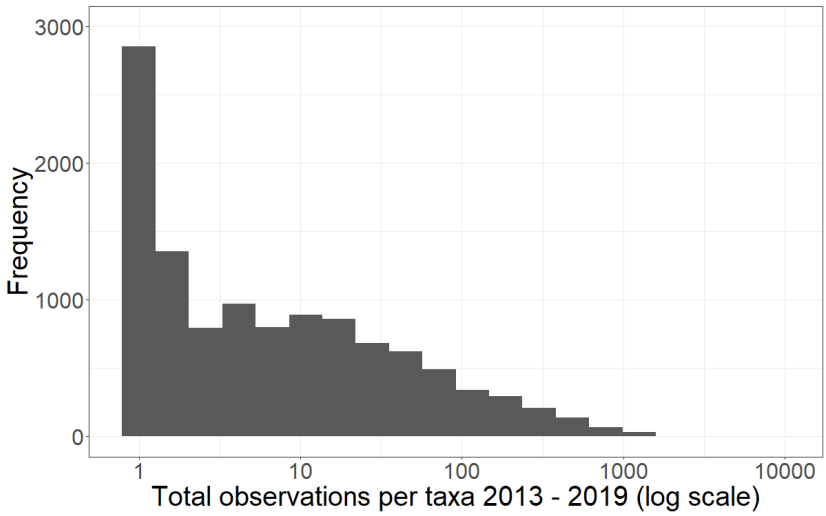


Figure 4b.1: **The majority of taxa have had fewer than 100 observations on iNaturalist NZ.** This figure shows the distribution of the total number of observations per taxon on iNaturalist NZ from 2013 to 2019.

tions in the first year. Not all taxa were first observed in the first year of our data, in fact only 30% of the taxa were first observed in year 1, Figure 4b.3. In the 7th year of data, 2019, 12% (1359) of the taxa were first observed. For the purposes of this chapter we only consider taxa that were first observed on iNaturalist NZ in year 1 (2013). We impute a zero count if there is a year they did not have any observations. After removing these taxa and grouping any subspecies observations with their parent species, 3218 taxa remain.

An important metric for parametrising the simulations in this chapter is the number of walks an observer does per year. We use the iNaturalist NZ data to find the number of unique days per year an iNaturalist NZ observer shares an observation. We use the number of days an observer shares an observation per year as a proxy for the number of observer walks per year. From the iNaturalist NZ data we have an empirical distribution of the number of observer walks per year, shown in Figure 4b.4. In this empirical distribution the probability of an observer doing one walk per year is 0.46, and the probability of the observer doing less than 15 walks per year is 0.90.

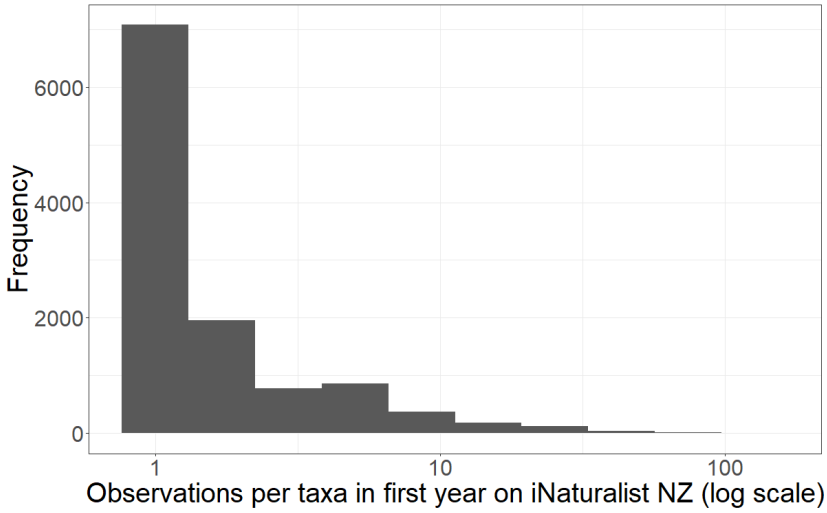


Figure 4b.2: **More than half of the taxa only had one observation in the first year they were observed on iNaturalist NZ.** This figure shows the distribution of the number of observations per taxa in the first year they were observed on iNaturalist NZ.

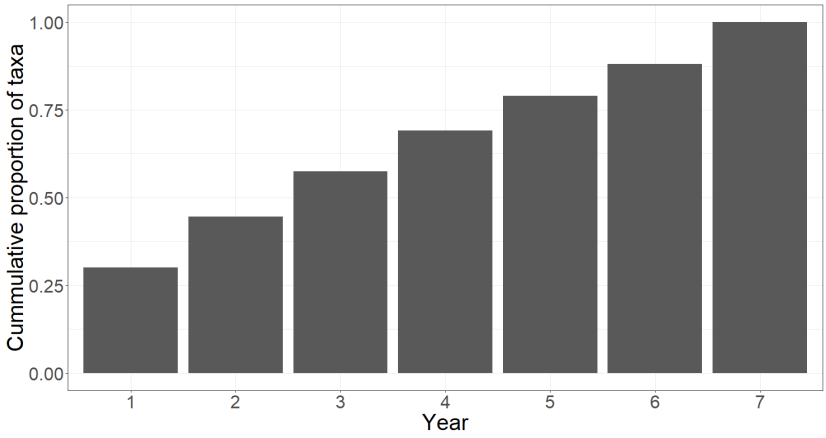


Figure 4b.3: **Just over a quarter of taxa were first observed on iNaturalist NZ in the first year of the data (2013).** This figure shows a cumulative plot of the proportion of taxa that were observed in the first 7 years of iNaturalist NZ data.

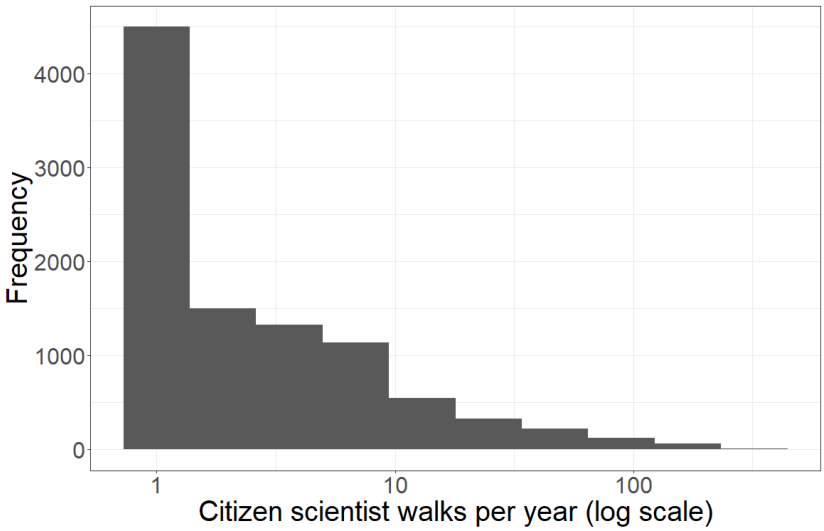


Figure 4b.4: **Most observers only share observations to iNaturalist NZ from one walk per year.** This figure shows the empirical distribution of the number of unique days an observer shares an observation to iNaturalist NZ per year. We consider this a proxy for the number of walks an observer does per year that results in an observation being shared to iNaturalist NZ. The probability of an observer doing one walk per year is 0.46, and the probability of the observer doing less than 15 walks per year is 0.90.

Statistical model background

In this section we introduce the statistical models we use to model the observation count data in the simulated iNaturalist NZ data. First, we introduce generalised liner models (GLMs) and in particular a Poisson GLM that is useful for modelling count data. In general, a generalised linear model (GLM) consists of three steps, (Zuur et al., 2009):

1. An assumption about the distribution of the response variable, x_i.
2. The specification of the systematic component in terms of explanatory variables.
3. The relationship/link between the mean value of the response variable and the systematic component.

For example, for a Poisson GLM:

1. We assume x_i is Poisson distributed with mean and variance μ_i
2. The systematic component is given as $\eta(X_{i1}, ..., X_{iq}) = \alpha + \beta_1 X_{i1} + ... + \beta_q X_{iq}$. Where, $X_{i1}, ..., X_{iq}$ are the explanatory variables and $\alpha, \beta_1, ..., \beta_q$ are the unknown parameters.
3. There is a logarithmic link between the mean of x_i and the systematic component $\eta(X_{i1}, ..., X_{iq})$. A logarithmic link ensures that the fitted values are always non-negative.

Therefore, we have:

$$\begin{aligned} & x_i \sim \text{Pois}(\mu_i) \\ & E(x_i) = \mu_i \\ & \log(\mu_i) = \eta(X_{i1}, ..., X_{iq}) \quad \text{or} \quad \mu_i = e^{\eta(X_{i1}, ..., X_{iq})} \end{aligned} \tag{4b.1}$$

Where, for a Poisson distribution the variance is equal to the mean, μ_i.

A simple, but naive, Poisson GLM model for iNaturalist NZ observations would be:

$$\begin{aligned} & x_t \sim \text{Pois}(\mu_t) \\ & E(x_t) = \mu_t \quad \text{and} \quad \text{var}(x_t) = \mu_t \\ & \log(\mu_t) = \alpha + \beta t \quad \text{or} \quad \mu_t = e^{\alpha + \beta t} \end{aligned} \tag{4b.2}$$

Where, x_t is the number of observations in year t, α and β are the unknown intercept and slope parameters, and $t \in \{1, ..., 7\}$. This is implicitly assuming the mean number of observations per year is either exponentially growing or decaying, or constant if $\beta = 0$ with time.

Offset number of observers per year

Annually the number of iNaturalist NZ observers, K, has been increasing, Table 4b.1. Therefore, we need to account for the number of observers in the Poisson GLM as there may be a larger number of observations in one year simply because the number of observers was larger. We could work with densities, e.g. $\frac{\text{number of observations}}{\text{number of observers}}$. However, by using densities we lose information on the magnitude of the number of observations and the number of observers. Another option is to use the number of observers as an explanatory variable, however, this would mean we are modelling a functional relationship between the number of observers and the number of observations. A neater method is to use the number of observers as an offset. We now assume that the number of observations, x_t, in year t is Poisson distributed with mean $\mu_t \times K_t$. K_t is also called the exposure or intensity parameter of the Poisson process, and μ_t, is the expected number of observations per observer in year t. In our case K_t is the number of observers in year t. This leads to the following GLM with the number of observers as an offset:

$$\begin{aligned} x_t &\sim \text{Pois}(\mu_t \times K_t) \\ E(x_t) &= \mu_t \times K_t \\ \log(\mu_t) &= \alpha + \beta t \quad \text{or} \quad \mu_t = e^{\alpha+\beta t} \\ E(x_t) &= K_t e^{\alpha+\beta t} \end{aligned} \tag{4b.3}$$

Overdispersion

In a Poisson distribution the mean is equal to the variance. Overdispersion is present when the variance is larger than the mean. Overdispersion may cause standard errors of the estimates to be underestimated, i.e. a variable may appear to be a significant predictor when it is in fact not significant. Hilbe (2011) distinguishes between apparent and real overdispersion. Apparent overdispersion is mainly due to model misspecification, for example, missing covariates or interactions, non-linear effects of covariates entered as linear terms in the systematic part of the model, or the wrong choice of link function. Whereas, real overdispersion is when no model misspecification can be identified because the variation in the data really is larger than the mean. A model may be overdispersed if the value of the Pearson χ^2 statistic divided by the degrees of freedom is greater than 1.0. The quotient of either is called the dispersion parameter. Small amounts of overdispersion are of little concern and the tolerance of what is considered small is dependent on the number of responses in the model.

If it is evident that the overdispersion in the model is not apparent overdispersion, a quasi-Poisson GLM can be tried to deal with the overdispersion before considering more

complicated methods like a negative binomial GLM. A quasi-Poisson GLM consists of the following steps:

1. $E(x_t) = \mu_t$ and $\text{var}(x_t) = \phi \times \mu_t$
2. $\log(\mu_t) = \eta(X_{i1}, ..., X_{iq})$ or $\mu_t = e^{\eta(X_{i1},...,X_{iq})}$

The key difference between a Poisson GLM and quasi-Poisson GLM is that in a quasi-Poisson GLM a Poisson distribution is not explicitly specified, instead only a relationship between the mean and variance of x_t is specified. The parameter estimates are not changed by using a quasi-Poisson GLM, however, the standard errors of the parameters are multiplied with the square root of the dispersion parameter, ϕ. If the dispersion parameter $\phi = 1$ in the quasi-Poisson GLM the results for the estimated parameters and standard errors will be the same as the Poisson GLM. As a rule of thumb, if ϕ is larger than 15, then other methods should to be considered to deal with overdispersion, e.g. the negative binomial GLM or zero-inflated models, (Zuur et al., 2009).

Model fitting on simulated data

In this section we generate simulated citizen science observation data. Subsequently we specify a GLM model that models the annual number of taxon observations as a function of observer behaviour covariates and the temporal changes in the taxon abundance. In the simulated data we have control over the sources of variation due to citizen scientist behaviour that are present in the data. In Chapter 4a we introduced a simple mechanistic model of an observer walking through a field with individual population members that could be encountered. These encounters could result in observations being contributed to iNaturalist NZ. In this section we scale that simple single observer model up to simulated iNaturalist NZ data by considering the following elements:

- 7 years of iNaturalist NZ observation data.
- Annually increasing number of iNaturalist NZ observers.
- Variation in the number of walks per observer per year.
- Variation in the expected number of individual population members per observer walk.

The null hypothesis is that there is no annual change in a given taxon abundance. We simulate data iNaturalist NZ observation data given the null hypothesis. We begin with a base scenario where there are 7 years of data but none of the above sources of variation are present. We then build up to a scenario where all of the above sources of variation are present. For each case of simulated data we specify a Poisson GLM. We test the validity of the model by measuring the type 1 error rate, where a type 1 error is the false rejection of the null hypothesis. If the model is adequately modelling the variation in citizen science behaviour the type 1 error rate should by definition be α, where α is the statistical significance level. If the type 1 error rate is larger than α the model is falsely finding significant trends in the annual species abundance at a greater rate than expected. We use the following process for a given expected number of observations in the first year:

1. Simulate 7 years of observation data.
2. Fit a Poisson GLM model to the annual number of observations.
3. Record the annual trend parameter, and corresponding p-value.
4. Repeat steps 1-3 multiple times.
5. Calculate the type 1 error rate, i.e. the number of simulations with a significant p-value for the annual trend estimate at a 0.05 significance level.

6. Repeat the above steps for a range of expected number of observations in the first year.

Table 4b.2 outlines the assumptions we enforce and relax per case in this section.

The probability of an individual observation is calculated using the mechanistic random walk model from Chapter 4a. Recall that this probability is a combination of the probability an individual is encountered and the probability an observation is then shared to iNaturalist NZ. The probability that an individual in encountered is dependent on the individual's home range radius, perceptive radius of the observer, the length of the path the observer walks, and the relative speed of the individual to the observer. We are assuming these variables remain constant over time and therefore the probability of an individual observation being shared to iNaturalist NZ is constant over the years of the scaled up simulation.

Assumption	**Base case**	**Variation in # individuals**	**Variation in # observers**	**Variation in expected # walks per observer**	**Combined case**
7 years of data	✓	✓	✓	✓	✓
Fixed number of observers per year	✓	✓	✗	✓	✗
Equal number of walks per observer per year	✓	✓	✓	✗	✗
Equal number of individuals in the field per walk	✓	✗	✓	✓	✗
Equal probability of individual observation per observer	✓	✓	✓	✓	✓

Table 4b.2: **Summary of assumptions per simulation case.**

Base case: no variation

In this case we assume the following:

- 7 years of data.
- Fixed number of observers per year.
- Equal number of walks per observer per year.
- Equal number of individuals in the field per walk.
- Equal probability of individual observation per observer.

Simulation set-up

We ran a very simple simulation as follows to calculate the type 1 error rate given the true null hypothesis of no annual change in the species abundance:

1. Set number of observers K, probability of an observation p, number of individuals n, number of walks per observer, per year W. In total there are W_t walks per year, where $W_t = KW$.
2. Every year the number of individual observations, x_t, is calculated by:
   (a) Randomly sample the number of observations per observer walk, x_t^w, from a Poisson distribution Pois(np), where np is the expected number of individual observations per walk, and $w \in 1, ..., W_t$. Note we are using a Poisson approximation to a binomial random variable, this is a good approximation as n is large and p is small.
   (b) Sum the number of observations per observer walk to get the total number of observations in year t: $x_t = \sum_{w=1}^{W_t} x_t^w$.
3. Fit a Poisson GLM as follows:

$$\begin{aligned} & x_t \sim \text{Pois}(\mu_t) \\ & E(x_t) = \mu_t \quad \text{and} \quad \text{var}(x_t) = \mu_t \\ & \log(\mu_t) = \alpha + \beta t \quad \text{or} \quad \mu_t = e^{\alpha+\beta t} \end{aligned} \tag{4b.4}$$

4. Record if the p-value for the trend estimate parameter, β, is significant at a 0.05 significance level.
5. Repeat steps 2 - 4, 10^4 times and calculate the type 1 error rate, i.e. the proportion of simulations where the true null hypothesis of no trend in species abundance is falsely rejected.

6. Repeat the above steps for a range of expected number of observations in year 1 by adjusting n, p, K, or W, where $E(\text{Observations in year 1}) = npKW$.

In this case the number of observations x_t in year t is Poisson distributed (an approximation of the binomial distribution due to the small value of p), and the mean number of observations per year is constant. Therefore, the model in Equation 4b.4 is perfectly specified and $\beta = 0$.

Simulation results

As an example, we set $p = 0.0002$, 20 observers per year, and assumed every observer just did one walk per year. We varied n in steps of 250 between 250 and 7500 to simulate datasets with the expected number of observations in year 1 ranging from 1 to 30. As expected the type 1 error rate is approximately 5% at a 0.05 significance level. Figure 4b.5 shows the distribution of type 1 error rates for multiple repeated runs of this simulation and this is typical for a range of parameter values.

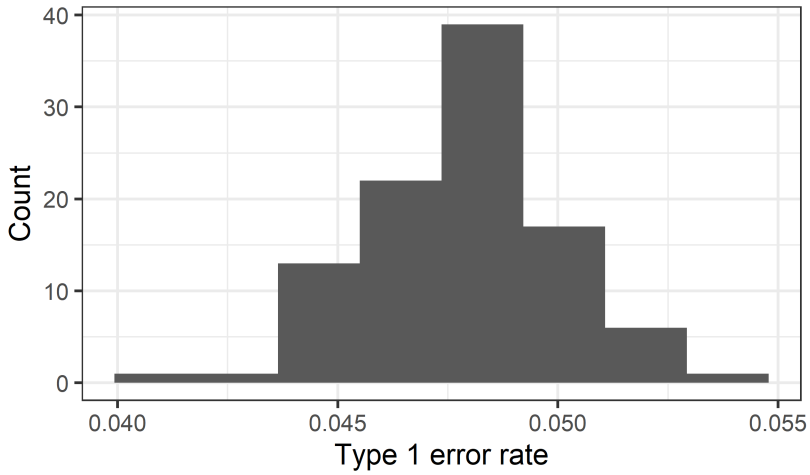


Figure 4b.5: **At a 0.05 significance level the type 1 error rate for the base case is near 5%.** The distribution of type 1 error rates is generated from multiple runs of the base case simulation.

Variation in number of individuals per walk

In the base case we assumed that on every observer walk there is the same number of individuals in the field to be observed. In reality, this is effectively assuming all observer walks are in the same location and of the same length. To simulate some spatial variation and variation in the length of the observer walks we allow the actual number of individuals in the field to vary between walks. We do not relax any of the other assumptions as shown in Table 4b.2.

Simulation set-up

We change the method to calculate the number of observations per walk in step 2 of the base case simulation to the following:

1. For each observer walk, w, randomly sample from a Poisson distribution the number of individuals, $n^{(w)}$, in the field. Where,

$$n^{(w)} \sim \text{Pois}(n).$$

   Note, we are assuming that $n^{(w)}$ are independent random variables and are independent of the observer, i.e. there is no tendency for some observers to go on longer walks or to visit densely populated areas.

2. Randomly sample the number of observations per observer walk, x_t^w, from the Poisson distribution $\text{Pois}(n^{(w)}p)$, where $n^{(w)}p$ is the expected number of individual observations per walk.

3. Sum the number of observations per observer walk to get the total number of observations in year t: $x_t = \sum_{w=1}^{W_t} x_t^w$.

Therefore, the simulation to calculate the type 1 error rate is the same as the base case with the exception that step 2 is replaced with the above steps. Figure 4b.6 shows the distribution of observations in the first year for this case and also the base case. In the base case we know that the number of observations per year is Poisson distributed (an approximation of the binomial distribution due to the small value of p). In this case we found numerically that the number of observations per year also follows a Poisson distribution. Therefore this simulation is not different from the base case. However, it could be different if we selected a broader distribution for the number of individuals in the field per year (e.g. a geometric or negative-binomial distribution).

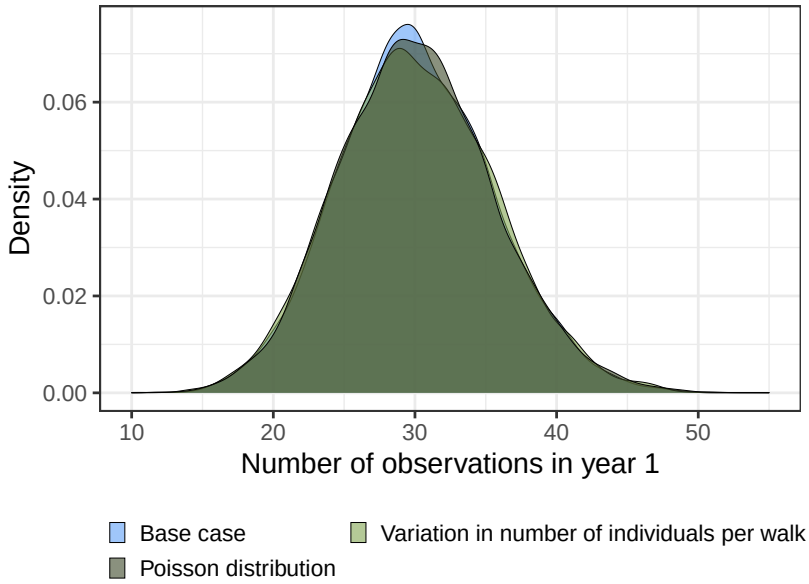


Figure 4b.6: **The number of observations in the first year of the simulated data follows a Poisson distribution for the base case, and the case with variation in the number of individuals per walk.** In the base case the number of observations per year is the sum of samples from a Poisson distribution with a fixed mean, np. Whereas, in the case where the number of individuals varies per walk the number of observations per year is the sum of samples from a Poisson distribution that has a Poisson distributed mean.

Simulation results

Given that randomly sampling the number of individuals in the field per citizen science walk from a Poisson distribution contributes little variance to the simulation, the results from this simulation are the same as the base case. On average the type 1 error rate is 5% at a 0.05 significance level across the range of expected number of observations in year 1.

Variation in annual number of observers

Annually the number of iNaturalist NZ observers has increased, shown in Table 4b.1. In this case we include this variation in our simulation and assume all other sources of variation are absent, as outlined in Table 4b.2. Our scaled up simulation is simulating the annual number of observations of a single species. It is very unlikely that every iNaturalist NZ observer is searching for each species therefore we scale the number of observers per species down to 5% of the total iNaturalist NZ observers.

As with the base case we use a simulation to calculate the type 1 error rate when the null hypothesis is there is no temporal trend in the species abundance. However, we now have a different number of observers every year and we offset by a proxy for this number in the Poisson GLM. We use a proxy because in the iNaturalist NZ data we do not have any knowledge of the walks that observers go on that result in zero observations. Therefore, we offset by the number of observers that shared at least one observation each year, because this is something we can extract from the iNaturalist NZ data. The simulation is as follows:

1. Set the number of observers per year K_t as 5% of the total annual iNaturalist NZ observers (shown in Table 4b.1), the probability of an individual observation p, number of individuals n, and number of walks per observer per year W. In total there are W_t walks per year, where $W_t = K_t W$

2. Every year the number of individual observations, x_t, is calculated by:

   (a) Randomly sample the number of observations, x_t^w, per observer walk from the Poisson distribution Pois(np), where np is the expected number of individual encounters per walk.

   (b) Sum the number of observations per observer walk to get the total number of observations in year t: $x_t = \sum_{w=1}^{W_t} x_t^w$.

3. Fit a Poisson GLM with an offset for the annual number of observers that shared at least one observation, K_t^*, as follows:

$$\begin{aligned} & x_t \sim \text{Pois}(\mu_t) \\ & E(x_t) = \mu_t \times K_t^* \quad \text{and} \quad \text{var}(x_t) = \mu_t \times K_t^* \\ & \log(\mu_t) = \alpha + \beta t + \log(K_t^*) \quad \text{or} \quad \mu_t = K_t^* e^{\alpha+\beta t} \end{aligned} \tag{4b.5}$$

4. Record if the p-value for the trend estimate parameter, β, is significant at a 0.05 significance level.

5. Repeat steps 2 - 4 10^4 times and calculate the type 1 error rate, i.e. the proportion of simulations where the true null hypothesis of no temporal trend in species abundance is falsely rejected. Repeat the above steps for a range of expected number of observations in year 1 by adjusting n, p, or W,

Simulation results

As with the base case the type 1 error rate is on average 5% at a 0.05 significance level as the added variation of increasing the number of observers per year is captured in the Poisson GLM by the observer offset.

Variation in expected annual number of walks per observer

In the base case we assumed that all observers did the same number of walks per year. However, in reality some observers go walking in search of individuals to observe often, whereas, other observers only go as little as once per year. In the data section of this chapter we outlined a proxy for the distribution of observer walks per year and here we include that information in the simulation below. Here we vary the number of annual observer walks. No other sources of variation are present (Table 4b.2).

1. For every observer, k, randomly sample an annual expected number of walks, $E(W^{(k)})$ from the empirical distribution of observer walks per year, Figure 4b.4.

2. Every year of the simulation we randomly sample the actual number of walks, $W_t^{(k)}$, for the observer from the Poisson distribution:

$$W_t^{(k)} \sim \text{Pois}(E(W^{(k)})).$$

Therefore the total number of walks for year t from all observers is:

$$W_t = \sum_{k=1}^{K} W_t^{(k)}.$$

Note, we are assuming the expected number of walks per year for each observer does not change over time, i.e., the observer does not become more or less active over time. But we are assuming that the number of walks per year for any observer is Poisson distributed.

As we did with the base case, we use a simulation to calculate the type 1 error rate given the true null hypothesis of no temporal trend in the species abundance. However, we now have a different number of walks per observer with some variation in the total number of walks per year due to the stochastic nature of the number of annual walks an observer does. In reality we are not going to know exactly how many walks an observer did per year as we have presence-only data on iNaturalist NZ, i.e. we are not aware of walks observers do that result in zero observations. Therefore, we cannot use the information about the number of walks per year to fit the Poisson GLM model. We fitted a Poisson GLM model as we did in the base case, as the only deviation in this simulation from the base case is the difference in the number of walks each observer does. However, overdispersion was detected. This overdispersion is most likely apparent overdispersion due to a misspecification of the model as we have no covariate to account for the variation in the number of walks a person does per year. However, we corrected the standard errors using a quasi-Poisson GLM model where the variance is given by $\phi \times \mu$, where μ is the mean and ϕ is the dispersion parameter.

Therefore we have the following simulation to calculate the type 1 error rate when there is variation in the number of walks per observer:

1. Set number of observers K, probability of an individual observation p, number of individuals n.

2. Randomly sample the annual expected number of walks per observer, $E(W^{(k)})$, from the empirical distribution of observer walks per year, shown in Figure 4b.4.

3. For each year, $t = 1, ..., 7$:

   (a) Randomly sample from a Poisson distribution the actual number of walks per user, $W_t^{(k)}$. Where,

   $$W_t^{(k)} \sim \text{Pois}(E(W^{(k)})).$$

   (b) Calculate the total number of annual observer walks as,

   $$W_t = \sum_{k=1}^{K} W_t^{(k)}.$$

   (c) Randomly sample the number of observations, x_t^w, per observer walk from the Poisson distribution Pois(np), where np is the expected number of individual encounters per walk.

   (d) Calculate the total number of annual observations as,

   $$x_t = \sum_{w=1}^{W_t} x_t^w.$$

4. Fit a quasi-Poisson GLM as follows:

   $$\begin{aligned} E(x_t) = \mu_t \quad &\text{and} \quad \text{var}(x_t) = \phi \times \mu_t \\ \log(\mu_t) = \alpha + \beta t \quad &\text{or} \quad \mu_t = e^{\alpha+\beta t} \end{aligned} \tag{4b.6}$$

5. Record if the p-value for the trend estimate parameter, β, is significant at a 0.05 significance level.

6. Repeat steps 2 - 4 10^4 times and calculate the type 1 error rate, i.e. the proportion of simulations where the true null hypothesis of no temporal trend in species abundance is falsely rejected. Repeat the above steps for a range of expected number of observations in year 1 by adjusting n, p, or K.

Simulation results

In this simulation scenario the type 1 error rate was on average 5% at a 0.05 significance level across a range of expected number of observations in the first year. Figure 4b.7 shows the distribution of type 1 error rates for multiple repeated runs of this simulation. This is a similar result to the base case, however, there is slightly more variance in the distribution of type 1 error rates,

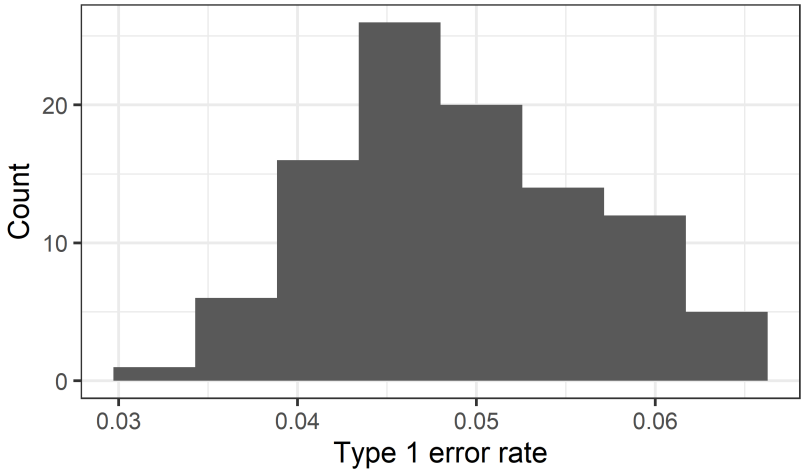


Figure 4b.7: **At a 0.05 significance level the type 1 error rate for the case of annual variation in the expected annual number of observer walks is near 5%.** The distribution of type 1 error rates is generated from multiple runs of the simulation with annual variation in the expected annual number of observer walks.

Combined case: all sources of variation

Finally, we combine all the above sources of variation to create a simulation that includes some of the sources of variation that are likely present in the iNaturalist NZ data. However, we still assume the probability of sharing an observation of an individual is fixed over time and equal across observers. Therefore, we have the following sources of variation in this combined simulation (Table 4b.2):

- Increasing number of observers per year.
- Variation in number of walks per observer per year.
- Variation in number of individuals in the field per walk.

Simulation set-up

To include all the three forms of variation into the combined case we have the following simulation to calculate the type 1 error rate:

1. Set the following parameters:
    - Probability of observing an individual, p.
    - Number of observers per year K_t as 5% of the total annual iNaturalist NZ observers shown in Table 4b.1.
    - Number of individuals, n
2. Randomly sample the annual expected number of walks per observer, $E(W^{(k)})$, from the empirical distribution of observer walks per year shown in Figure 4b.4.
3. For each year, $t = 1, ..., 7$:
    (a) Randomly sample from a Poisson distribution the actual number of walks per user, $W_t^{(k)}$. Where,
    $$W_t^{(k)} \sim \text{Pois}(E(W^{(k)})).$$
    (b) Calculate the total number of annual observer walks as,
    $$W_t = \sum_{k=1}^{K} W_t^{(k)}.$$
    (c) For every walk, w:

i. Randomly sample from a Poisson distribution the number of individuals, $n^{(w)}$, in the field. Where,

$$n^{(w)} \sim \text{Pois}(n).$$

Note, we are assuming that $n^{(w)}$ are independent random variables and are independent of the observer, i.e. there is no tendency for some observers to go on longer walks or to visit densely populated areas.

ii. Randomly sample the number of observations per observer walk, x_t^w, from the Poisson distribution $\text{Pois}(n^{(w)}p)$, where $n^{(w)}p$ is the expected number of individual observations per walk.

(d) Calculate the total number of annual observations as,

$$x_t = \sum_{w=1}^{W_t} x_t^w.$$

4. Fit a quasi-Poisson GLM with an offset for the number of observers that shared at least one observation per year as follows:

$$\begin{aligned} &E(x_t) = \mu_t \times K_t^* \quad \text{and} \quad \text{var}(x_t) = \phi \times (\mu_t \times K_t^*) \\ &\log(\mu_t) = \alpha + \beta t + \log(K_t^*) \quad \text{or} \quad \mu_t = K_t^* e^{\alpha+\beta t} \end{aligned} \tag{4b.7}$$

5. Record if the p-value for the trend estimate parameter, β, is significant at a 0.05 significance level.

6. Repeat steps 2 - 4 10^4 times and calculate the type 1 error rate, i.e. the proportion of simulations where the true null hypothesis of no temporal trend in species abundance is falsely rejected.

Simulation results

Figure 4b.8 shows three examples of situations that may arise when fitting a Poisson GLM to simulated annual observation data where there is no annual increase in the given taxon abundance. Case A is the most frequent occurrence where no significant annual trend is detected. The number of observations per year is still increasing because the number of observers per year is increasing. However, the number of observations per observer is relatively flat over the 7 years. In case B, a significant positive trend is found, despite there being no annual increase in the taxon abundance. In this case the number of observations per observer has an increasing trend over the 7 years. This is due to the stochastic nature of the simulation. In case C, a significant negative trend is found. In this case the total number of observations per year is increasing, but the number of observations per observer in decreasing.

On average, across the range of expected number of observations in year 1, the type 1 error rate for the annual trend estimate parameter is above 5% at a 0.05 significance level, shown in Figure 4b.9. This indicates that the Poisson GLM is not capturing all the variation in the simulated data that is due to citizen scientist variation, and therefore some of this variation is being mistaken for a change in the underlying species abundance. The type 1 error rate converges to approximately 12%. For low expected number of observations in year 1, the type 1 error rate is less than 12%. This is because, for these lower values, the Poisson distribution that the number of observations per walk is being sampled from has a small mean, and therefore there is not a possibility of sampling a value less than that observed in year 1. This is eliminating the random possibility of a downward trend in the annual number of observations per observer. Therefore, for these lower expected number of observations in year 1, the type 1 errors are only occurring when there is randomly an annual increase in the number of observations per observer.

The inflated type 1 error rate for this case shows that a Poisson GLM is a poor choice to model citizen science data when the only known meta-data about observer effort is the number of active observers per year. On the contrary, if we off-set by the number of walks per year, including walks that resulted in zero observations, the type 1 error rate returns to the expected 5% at a 0.05 significance level. This is not a source of information that is currently collected by iNaturalist NZ. However, other observation sharing citizen science project, e.g. eBird, do collect this information and therefore a Poisson GLM may be a suitable choice to model annual eBird observation data.

Given the inability to specify a suitable Poisson GLM to model simulated iNatural-

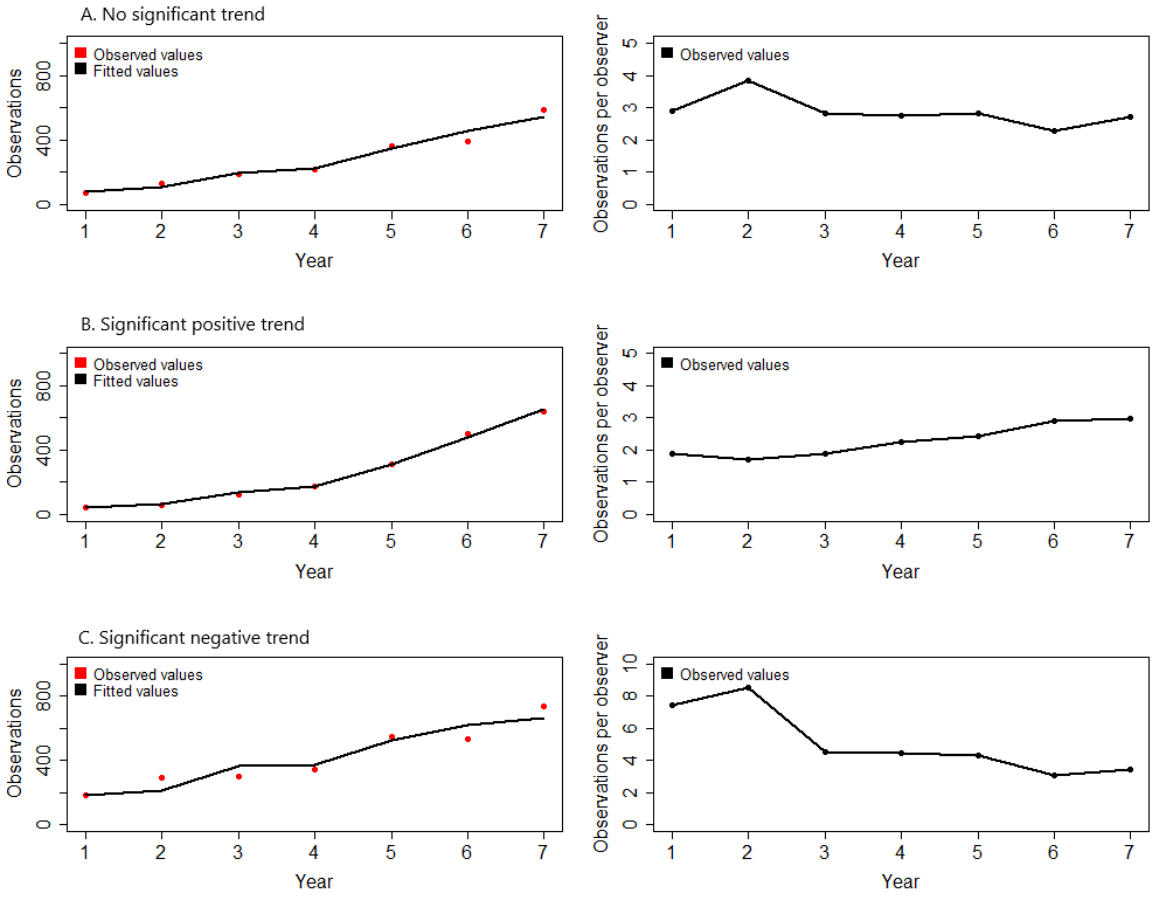


Figure 4b.8: **Given a true null hypothesis of no annual change in the taxon abundance, the stochastic nature of citizen science may still result in significant positive and negative trends being found in a fitted GLM.** All these plots are examples from the combined case of all sources of variation, and the true null hypothesis of no annual change in the taxon abundance. The number of observers is increasing annually. The figures on the right show an annual increase in the number of observations, despite case A having no annual change in the number of observations per observer, and case C having an annual decline in the number of observations per observer.

ist NZ data, due to a lack of meta-data about changes in observer behaviour, we do not consider any other type of statistical model. All models will suffer the same problem of poor performance due to the model specification not covering the applicable range of explanatory variables. Before additional candidate models are pursued, we would need to gather additional citizen science behaviour data either explicitly from the citizen science project, or infer the information from the already collected data.

If we naively apply this model to the 3218 taxa in the iNaturalist NZ dataset with 7 years of observation data to test the null hypothesis that the annual taxa abundance

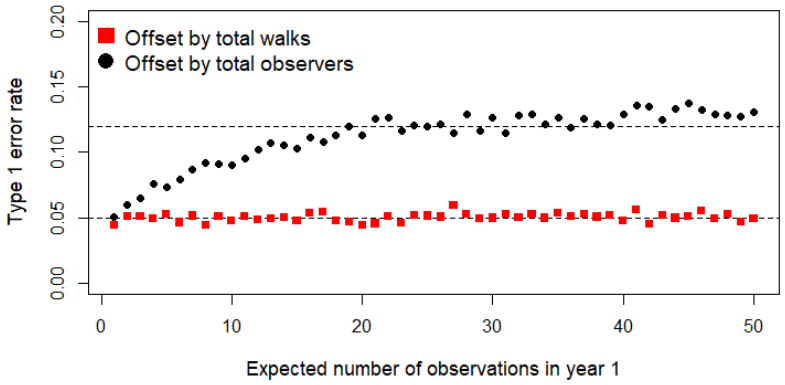


Figure 4b.9: **For the combined case with three sources of variation, when we include an observer offset in the Poisson GLM, the type 1 error rate for the trend estimate parameter is above 5% at a 0.05 significance level.** In these simulations the null hypothesis of no annual change in a given taxon abundance is true. Therefore a type 1 error occurs when the GLM model detects a significant annual trend. If we offset by the total number of walks the type 1 error rate is on average 5% at a 0.05 significance level. However, the number of total walks (including walks with zero observations) is not known on iNaturalist NZ.

is not changing, we will expect to find more than 5% of taxa with false significant trends (type 1 errors) at a 5% significance level. It will be difficult to know which taxa have shown a significant trend due to a type 1 error, versus which taxa have had an ecological annual change in their abundance. Even with more information about annual changes in citizen scientist behaviour, it is still very unlikely that a statistical model will be able to be specified that captures the wide range of variation in citizen science data. Therefore, any significant results from a statistical model applied to annual citizen science data should be analysed with a degree of caution as it is very likely that there will be an inflated level of type 1 errors.

Later in this chapter we measure the power of the model to detect genuine annual abundance trends. Given only 7 years of observation data there is also a high risk of type 2 errors, where we fail to reject the null hypothesis of no annual trend, when there is actually an annual trend in the species abundance, i.e. we do not find a significant trend when one exists.

iNaturalist NZ trends

In this section we naively fit the quasi-Poisson GLM model with an observer offset, outlined in Equation 4b.7, individually to each of the 3218 species in the cleaned iNaturalist NZ dataset that have 7 years of observation data. This is naive for many reasons, the most obvious being that the model already had an inflated level of type 1 errors on the simulated data with limited and known sources of citizen science variation. It is also naive because we are assuming that all species have the same mechanisms that affect their annual number of observations. For example in the quasi-Poisson GLM model we offset by the total annual number of iNaturalist NZ observers. However, it is naive to assume all species have had the same annual growth in observer effort. For example, there could be a species that has become boring to observers over the years and as a result every year fewer observers share an observation of that species.

Results

We found 17% of the species that have 7 years of iNaturalist NZ data had a significant annual abundance trend at a 0.05 significance level, and 13% had both an acceptable dispersion parameter ($0.1 < \phi < 15$) and significant trend (Table 4b.3). Figure 4b.10 shows the proportion of significant trends by the expected number of observations in year 1, versus the percentage of type 1 errors in the simulated data. On average, in the real iNaturalist NZ data we found more significant trends than the expected type 1 error rate, given the results from the simulated data. However, this does not provide any conclusive evidence that some of these significant trends are not type 1 errors, as it is very likely that the in real iNaturalist NZ data there is more variation due to citizen scientist behaviour than in the simulated data. Therefore, we would expect a higher rate of type 1 errors when fitting a quasi-Poisson GLM with an observer offset to the real iNaturalist NZ data.

Figure 4b.11, shows the distribution of the significant trend values given the number of observations in the first year. Each vertical slice of the plot is a histogram of the annual trend estimates, given the number of observations in year 1. We have grouped together all the significant trends for taxa that had more the 11 observations in the first year, as there were few taxa in this category, while keeping in mind that the majority of these significant trends may be type 1 errors. The majority of the significant trends are negative. This is likely because the specified GLM is offsetting by an annually increasing number of observers. Therefore, if the observation counts are static over the 7 years, the model will conclude a negative trend. In the case that annual observations were static, the Poisson GLM model would find an annual trend value of -0.375 (shown as a dashed line on Figure 4b.11). The lower the number of observations in the first

	Number of species	Percentage of species
Total species	3218	-
Acceptable dispersion	3047	95%
Significant trend	557	17%
Acceptable dispersion and significant trend	423	13%

Table 4b.3: **13% of the 3218 iNaturalist NZ species with 7 years of observation data were found to have a significant trend with an acceptable dispersion parameter from the fitted GLM model.** A quasi-Poisson GLM with an observer offset was individually fitted to each of the iNaturalist NZ taxa with 7 years of observation data. An acceptable dispersion parameter is assumed to be $0.1 < \phi < 15$. Significant trends are at a 0.05 significance level.

year the wider the spread of estimated significant trend values. This is likely because for small observation counts the type one errors are very noisy. This plot highlights the fact that with only 7 years of observation data it is difficult to find a significant trend of a small magnitude.

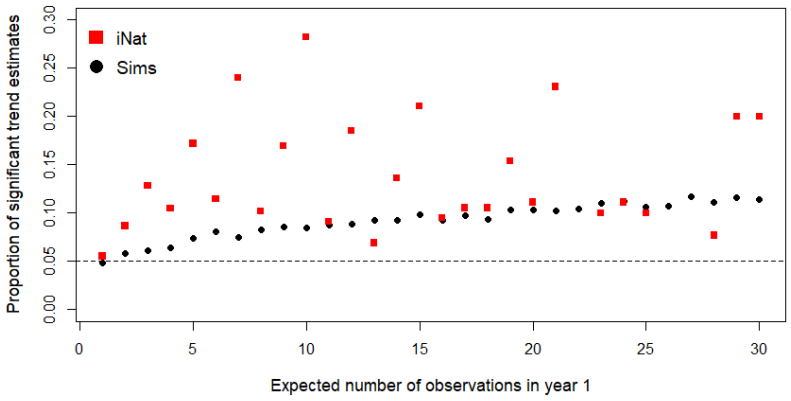


Figure 4b.10: **There are a higher proportion of significant trends in the real iNaturalist NZ data than in the simulated data that had a true null hypothesis of no annual change in species abundances.** The red points are the proportion of significant trends found in the iNaturalist NZ data. The black points are the type 1 error rates in the simulated citizen science data with a true null hypothesis of no annual change in a given taxon abundance.

Figure 4b.12 shows the annual observations and annual observations per observer for three iNaturalist NZ taxon with significant annual trend values. In all three examples the annual number of observations is increasing, despite two of the taxon having a

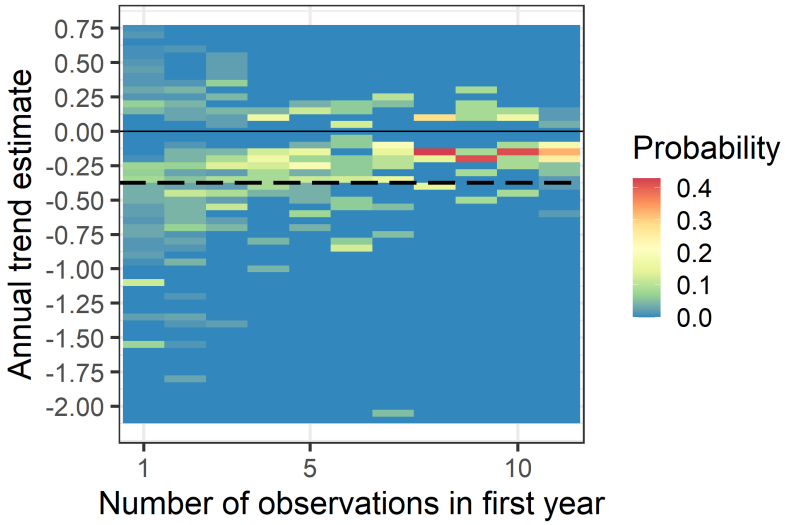


Figure 4b.11: **Most significant trends found in the iNaturalist NZ data are negative.** Significant annual trends are from the Poisson GLM fitted to iNaturalist NZ data at a 0.05 significance level. Each vertical slice of this plot is a histogram of the significant trend estimate given the number of observations in the first year. Significant trends for taxa with at least 11 observations in the first year have been grouped together. The dashed line is the expected annual trend estimate if there is no annual variation in the number of observations.

significant negative annual trend value. This is because the annual number of observers is also increasing and at a faster rate than the observations per observer are decreasing. In all three examples a dashed line is plotted that represents the expected number of observations if the null hypothesis of no annual change in the taxa abundance was true.

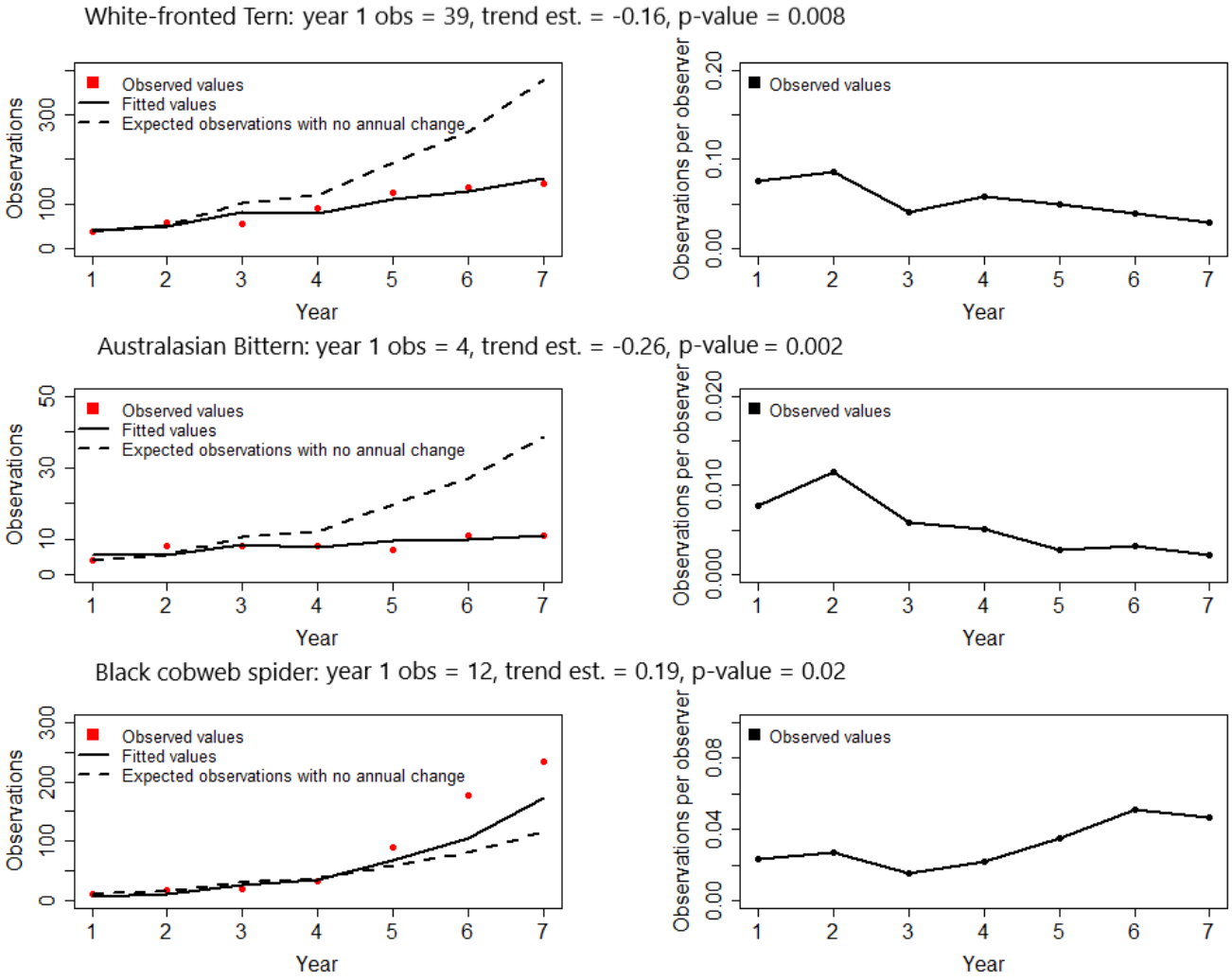


Figure 4b.12: **Example plots of iNaturalist NZ taxon with significant annual trends.** In all examples the annual number of observations is increasing, despite the first two examples having a significant negative annual trend value. This is because the number of observers in increasing each year. On the left hand plots, the dashed lines show the expected number of annual observations if there was no annual change in the number of observations per observer. The solid line on the left hand plots show the fitted values from the Poisson GLM with an observer offset. In the first two examples the annual trend estimate is negative. In the third example the annual trend estimate is positive. The left hand plots show the expected number of observations per observer.

Case study: Can citizen science data be used to reliably detect a 2% annual increase in a kiwi species abundance?

The 2018 - 2028 Kiwi Recovery Plan by the New Zealand Department of Conservation outlines the goal to reach 100,000 kiwi by 2030 by growing all kiwi species by at least 2% per year (Germano et al., 2018). Furthermore, in this plan it is identified as an issue (Issue 4.4) that the full potential of synergies between national survey programmes (e.g. DOC national biodiversity monitoring, iNaturalist NZ, eBird) and kiwi monitoring are not being realised. In this case study we take a closer look at the iNaturalist NZ kiwi data and use simulations to explore if a 2% increase is possible to detect in noisy citizen science data for taxa, like kiwi, that have relatively low densities and detection rates on iNaturalist NZ. However, recall that the specification of the GLM in the previous section was problematic and therefore this result does not guarantee a true biological trend. Also, increases in the number of observations per year (as seen in Brown Kiwi and Great Spotted Kiwi) may be explained by the annually increasing number of observers (Table 4b.1).

Kiwi data

In 2018 there were approximately 70,000 kiwi in New Zealand, split among 5 species and 14 taxa, (Germano et al., 2018). The 5 species are: Little spotted kiwi, Great spotted kiwi, Brown kiwi, Rowi, and Tokoeka. The Rakiura kiwi is a subspecies of the Tokoeka kiwi and is a relatively prevalent kiwi taxon. All 5 kiwi species and the Rakiura kiwi have research grade observations on iNaturalist NZ between January 2013 and December 2019. Figure 4b.13 shows the annual iNaturalist NZ observations per kiwi taxa between 2013 and 2019. In the previous section of this chapter we found that the Little Spotted kiwi was the only kiwi taxa with a significant trend value of -0.42 in the results from the GLM model. Recall that the Poisson GLM model we fitted has an offset for the annual number of observers. Therefore, a significant trend is indicating that there is an annual change in the number of observations per observer. Figure 4b.14, shows the annual number of kiwi observations per iNaturalist NZ observer for the 6 kiwi taxa. As expected, the number of observations per observer for the Little Spotted kiwi declined over the 7 years.

Kiwi are nocturnal, which makes it challenging in a mostly photo based citizen science project for observers to share a kiwi observation. However, photos do not need to be of the bird. Instead photo evidence of a kiwi, e.g. a footprint, feather, or burrow, may be adequate. On iNaturalist NZ there is also the ability for an observation to

reach research grade with a sound file.

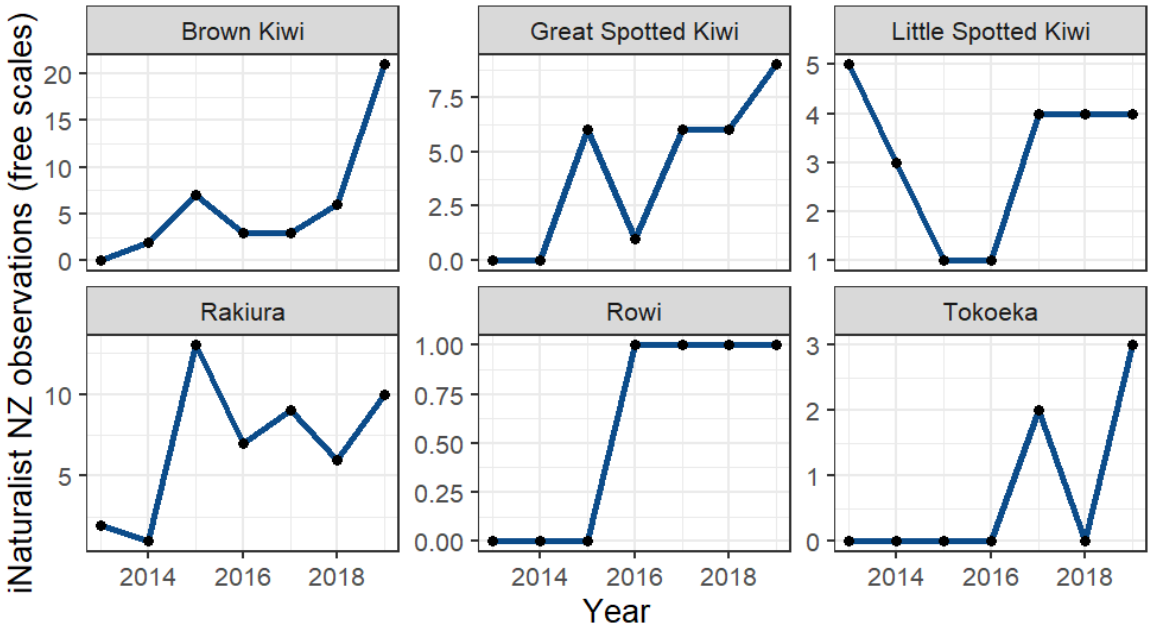


Figure 4b.13: **There have been iNaturalist NZ kiwi observations for 6 kiwi taxa between 2013 and 2019**. These graphs show the annual number of research grade kiwi observations by kiwi taxa between 2013 and 2019. Note the y-axis scales are different for each taxa.

Table 4b.4 shows the 2018 abundance estimates for each kiwi taxa (Germano et al., 2018) alongside the corresponding number of 2018 iNaturalist NZ research grade observations. With knowledge of the species abundances, and assuming the number of kiwi observations per observer walk is binomially distributed, we can estimate the probability of sharing a kiwi observation (this is a combination of the probability of encountering and sharing an observation of a kiwi) for each kiwi taxa as:

$$\Pr(\text{Kiwi observation shared to iNat NZ per walk}) = \frac{\text{2018 observations}}{\text{2018 abundance} \times \text{number of walks}}$$

For example, if 3000 walks took place in 2018 with the possibility of sharing a kiwi observation then the encounter probabilities for each kiwi taxa are very small as shown in Table 4b.4. 3000 walks is an arbitrary number but it is assuming that every year 20% of the total iNaturalist NZ observers would potentially share a kiwi observation. This is a useful calculation for parametrising the following simulations. However, we are assuming that all 70,000 kiwi are available to observe. In reality, many of these kiwi probably spend most of their time in an area that is not accessible to humans. Further, these kiwi taxa are limited to particular locations within New Zealand. For

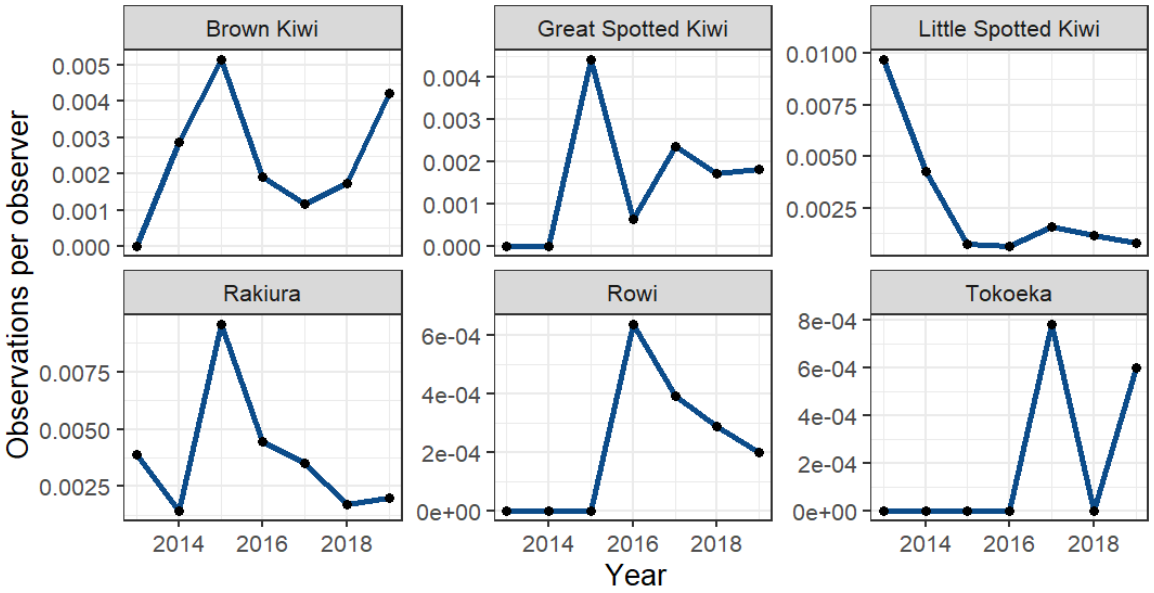


Figure 4b.14: **The number of observations per observer for the Little Spotted kiwi declined between 2019 and 2019.** The Poisson GLM model we fitted to iNaturalist NZ data in this chapter had an annual offset for the number of observers. Therefore, a significant trend is indicating that there is an annual change in the number of observations per observer. The only kiwi taxa with a significant trend was the Little Spotted kiwi with a negative significant trend. Note the y-axis scales are different for each taxa.

example the Rakiura kiwi is only found on Rakiura Island (Stewart Island), and the Rowi kiwi is only found in Ōkarito.

Methods

To test if iNaturalist NZ data would be able to be used to identify a 2% increase in a kiwi taxa abundance we use the simulation framework outlined earlier in this chapter, but parametrise it to reflect the kiwi taxa. We set an annual increase of 2% in the initial abundance, and consider a range of encounter probabilities that span the above probability estimations of encountering and sharing an observation for each kiwi taxa.

We run the simulations over 10 years to reflect the time frame of the Kiwi Recovery Plan, and over 20 years to understand the impact that more years of data has on the power to identify an annual 2% increase in kiwi abundance. We extrapolate the number of iNaturalist NZ observers to 2038 (20 years after the start to the Kiwi Recovery Plan) by fitting a linear model to the known 7 years of iNaturalist NZ observers. We assumed that every year 20% of the total iNaturalist NZ observers would potentially share a kiwi observation if they encountered a kiwi. This is a fairly modest assumption given

Taxon	iNaturalist NZ observations 2018	iNaturalist NZ walks estimate	Actual abundance estimate 2018	Observation probability
Little spotted kiwi	4	3000	1900	7.0×10^{-7}
Great spotted kiwi	6	3000	14000	1.4×10^{-7}
Brown kiwi	6	3000	25100	8.0×10^{-8}
Rowi	1	3000	600	5.6×10^{-7}
Tokoeka*	2	3000	12550	5.3×10^{-8}
Rakiura	6	3000	12300	1.6×10^{-7}
Total	25	3000	66450	1.3×10^{-7}

Table 4b.4: **All kiwi taxa have a low probability of being encountered and shared as an observation to iNaturalist NZ.** The 2018 iNaturalist NZ research grade kiwi observations and 2018 kiwi abundance estimations are used to estimate the probability of an iNaturalist NZ observer encountering and sharing a kiwi observation by taxa for an assumed number of annual walks. In 2018 there were no Tokoeka kiwi observations so the 2017 count of 2 Tokoeka kiwi observations was used in the probability calculation as an example of a non-zero probability.

the popularity of kiwi taxa in New Zealand, and gives an average of 3000 walks in 2018 (the first year of the simulation).

Model power

Let β be the type 2 error rate at a 0.05 significance level (i.e. the rate at which a false null hypothesis is not rejected). The power of a statistical test is the probability of rejecting the null hypothesis when the alternative hypothesis is true:

$$\text{power} = Pr(\text{reject } H_0 | H_1 \text{ is true}).$$

Therefore,

$$\text{power} = 1 - \beta.$$

Isaac et al. (2014) suggested an alternative power calculation to allow power comparisons across model with varying levels of inflated type 1 error rates:

$$\text{power} = 1 - \beta - \alpha,$$

where α is the type 1 error rate. Therefore, when a model has an inflated type 1 error this bias towards finding significant trends when they do not exist is not contributing to the power calculation. For this case study, we checked the type 1 error rates across the parameter space of interest and found that is was consistently near 5% therefore

we use the traditional calculation of power $= 1 - \beta$.

Results

Over a 10 year period, and with the current level of kiwi observation activity on iNaturalist NZ, it is unlikely the data could be used to detect a 2% increase in a kiwi taxa abundance. The power of the model was less than 10% within the parameter space that the 6 kiwi taxa fall within (the parameters are the p and n in Table 4b.4), Figure 4b.15. That is, less than 10% of simulations were able to find a significant trend value when there was a genuine 2% annual increase in the kiwi abundances. For all kiwi taxa data pooled together the power of the model was approximately 20%.

Figure 4b.15 may be used to indicate the observation probability required to be able to spot a significant trend when there is an annual 2% increase in the underlying abundance. For example, for each individual kiwi taxa the observation probability would have to be much larger to be able to spot a significant trend 75% of the time when there is an annual 2% increase in the underlying abundance. Whereas, for all the kiwi taxa combined if the observation probability was 8 times larger there would be a 75% chance of spotting a significant trend when there is a 2% increase in the kiwi abundance and given 10 years of kiwi observation data.

After 20 years of iNaturalist NZ kiwi data the power to detect a 2% annual increase in the kiwi abundance is much larger than after 10 years of data, Figure 4b.16. For all the kiwi taxa combined the power of the model to detect a 2% annual increase is 95%. For all the individual kiwi taxa the power is approximately 50%. However, we have assumed that the number of observers increased linearly over this period. There is no certainty that this observer effort increase will be the case for all iNaturalist NZ observers, and further that it will apply to kiwi observations.

Figure 4b.17 shows the distribution of the estimated significant trend values when the simulations are parametrised for the combined kiwi taxa point with 10 years of observation data. When the null hypothesis of no annual trend is true, the estimated significant trend values (i.e. type 1 errors) are symmetrically and bi-modally distributed about zero. The bi-modal shape is due to there being insufficient data to detect very small trends. However, when there is a genuine 2% annual increase in the kiwi abundance, there are firstly a higher proportion of significant trends, 19% versus 5.5% when the null hypothesis was true. Furthermore, the majority of the estimated significant trends are positive, however the mean is approximately 4%. Figure 4b.18, shows the same plot but for 20 years of simulated data. In the scenario where the null hypothesis of no annual trend is true there are just over 6% of simulations where the fitted GLM

resulted in a type 1 error. Again, these estimated significant trends are symmetrically and bi-modally distributed about zero. When there was a genuine 2% annual increase in the kiwi abundance, 97% of the simulations resulted in the fitted GLM returning a significant trend estimate and these are symmetrically distributed around 2%.

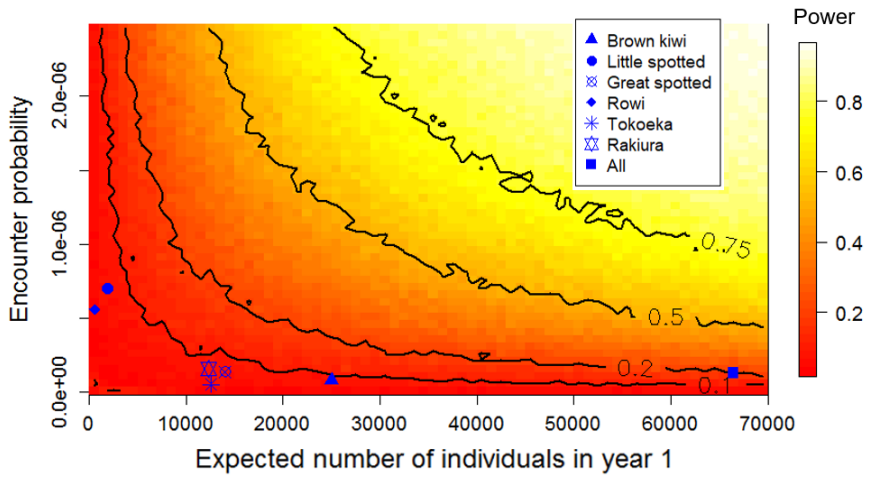


Figure 4b.15: **With the current level of kiwi observation activity on iNaturalist NZ it is unlikely a significant trend could be found in the data.** The heatmap shows the percentage of simulations with a significant estimated trend when the actual kiwi abundance trend was increasing by 2% annually and the simulations were run over 10 years. Points for the 6 kiwi taxa that have iNaturalist NZ observations are plotted, as well as a point for the combination of all 6 taxa.

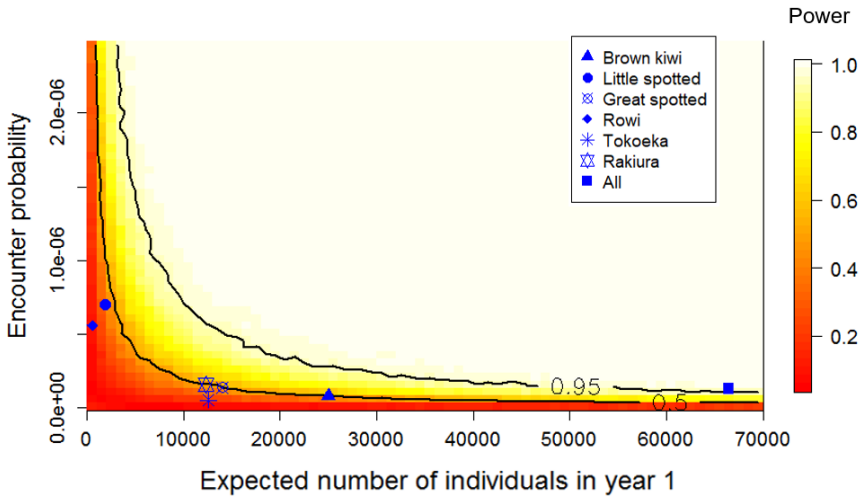


Figure 4b.16: **If there were 20 years of kiwi data on iNaturalist NZ there is a high probability of identifying a significant trend in the data.** The heatmap shows the percentage of simulations with a significant estimated trend when the actual kiwi abundance trend was increasing by 2% annually and the simulations were run over 20 years. Points for the 6 kiwi taxa that have iNaturalist NZ observations are plotted, as well as a point for the combination of all 6 taxa.

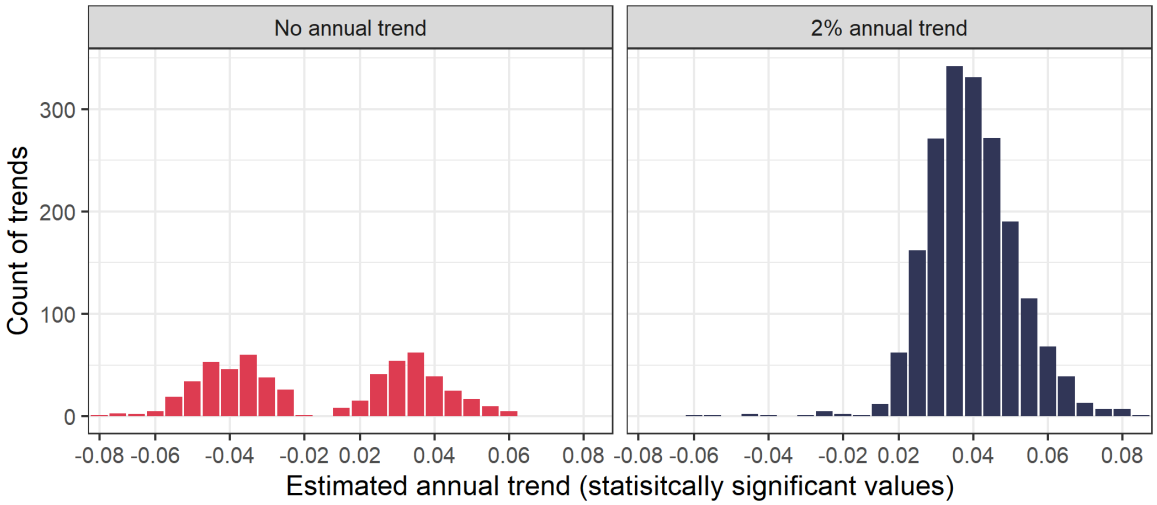


Figure 4b.17: **Significant trend estimations from 10 years of simulated data parametrised for all kiwi** Left: null hypothesis is true. Right: null hypothesis is false and there is a 2% annual increase in the kiwi abundance.

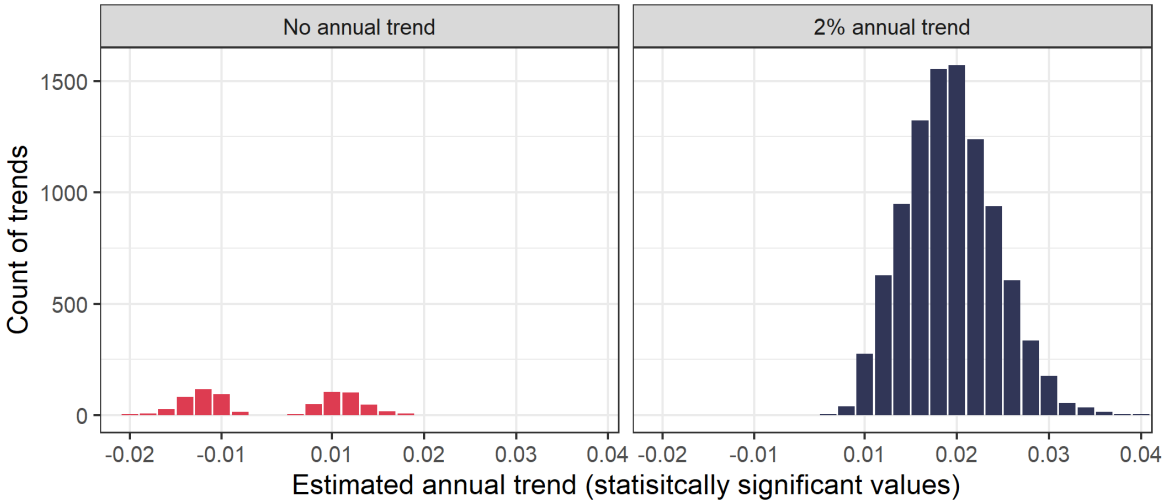


Figure 4b.18: **Significant trend estimations from 20 years of simulated data** Left: null hypothesis is true. Right: null hypothesis is false and there is a 2% annual increase in the kiwi abundance.

Discussion

This chapter has three areas of focus to explore the feasibility of using statistical models to estimate annual trends in species abundances from citizen science data with both minimal data collection protocols, and limited metadata on observer effort. As we have done throughout this thesis, we used the iNaturalist NZ dataset as a case study. The iNaturalist NZ data meets both the criteria of having a data collection protocol with minimal guidelines for the observers, and a minimal amount of metadata data on observer effort. First, we generated computer simulated citizen science data with various sources of variation in citizen scientist observer effort. We used the simulated data to test the performance of statistical models to detect annual trends in a species abundance. Second, we fitted the same statistical model to every species in the iNaturalist NZ dataset that had 7 years of observation data to estimate the annual change in the species abundance. Third, we explored if citizen science data from iNaturalist NZ could be used to detect a 2% annual increase in a kiwi species abundance.

The often relaxed data collection protocols in citizen science, coupled with no or limited metadata on citizen science recording effort, make it particularly difficult to use statistical methods to isolate ecological changes from changes in citizen science behaviour. Even in standardised ecological monitoring programmes, observed counts are the result of two linked stochastic processes (Kery et al., 2009). The first stochastic process is that of the true biological state, and the second is the observation process that consists of the variation in the observer's behaviour. Inference is desired on the first stochastic process about the true biological state. However, when there are no or limited measurable covariates about observer behaviour, it is extremely challenging to specify a statistical model and make inferences about the biological state without falsely drawing conclusion based on changes in observer behaviour. In citizen science projects like iNaturalist NZ with minimal data collection protocols, and minimal explicit metadata about citizen science effort, this problem of disentangling ecological change from citizen scientist change is difficult and largely unexplored in the literature.

Simulated data

The ability of a statistical model to reliably detect the ecological change in a species abundance from the noise of citizen science can be tested with simulated citizen science data. In this chapter, we followed a similar approach to Isaac et al. (2014). We used computer simulated citizen science data with varying sources of variation from the citizen scientists to test the performance of statistical models to detect trends in species abundances. We simulated multiple years of observations for a particular species by

scaling up the single citizen scientist simulation from Chapter 4a. Where possible we parametrised the simulated data to resembled iNaturalist NZ data. For example, based on the iNaturalist NZ data the simulated data had: 7 years of data, an annually increasing number of observers, and variation in the number of walks per year for each observer. In the simulated data the annual species abundance was set to no annual change. We fitted a Poisson GLM to the simulated datasets and measured the type 1 error rate, given the true null hypothesis of no annual change in the species abundance.

We found that for measurable levels of variation in citizen science behaviour, it is possible to specify a model that reasonably captures the variation present in the data. For example, if the only source of variation is an annual increase in the number of active citizen scientists, then this can be included as an off-set in the model specification and, as expected, the resulting type 1 error rate of 5% and at 0.05 significance level. However, if there are other sources of variation present in the data that are not currently recorded in the iNaturalist NZ metadata, for example the number of walks a citizen scientist does per year, then this cannot be specified in the model and the type 1 error rates are greater than 5% at a 0.05 significance level. This indicates that there is more variation in the data than the model is specifying, or the data is overdispersed. As we outlined earlier in this chapter there is apparent overdispersion and real overdispersion. Apparent overdispersion may be due to a model misspecification, and real overdispersion is when the variation in the data really is larger than the mean (in the case of a Poisson GLM). In the case of annual observation data from citizen scientists it is very likely that a large component of overdispersion in the model is apparent overdispersion because there is often a lack of information about annual changes in citizen science behaviour and therefore no ability to capture this variation with a covariate in the model.

iNaturalist NZ data

Testing the validity of a statistical model on simulated iNaturalist NZ citizen science data provides insight into how reliable the statistical model will be when applied to the real iNaturalist NZ dataset. Despite an inflated type 1 error rate when using a quasi-Poisson GLM with an observer offset to model annual citizen science observation data, we fitted the model to 3218 species with seven years of iNaturalist NZ observation data. For each species, the fitted GLM provided an estimate of the annual change in the species abundance. However, we found for a large proportion of the species the GLM either had an unacceptable dispersion parameter or did not detect a significant trend.

We fitted the same Poisson GLM to each species - a mega-fitting approach - even

with a well specified model, there are problems with this approach. A mega-fitting approach neglects the reality that there are differences in the factors influencing the annual observation data of each species. For example, we applied the same increasing observer offset to the Poisson GLM we fitted to each species. However, we do not know that this increase in observer effort has been distributed evenly across all species. Also, fitting the same model multiple times inevitably results in type 1 errors where the null hypothesis is falsely rejected, i.e. we falsely find a significant trend estimate when there was no annual change in the species abundance. For a well specified model we would expect a type 1 error to occur in 5% of the species or realisations of the simulation. For the simulated data, when we had control of all the sources of additional variation but omitted our knowledge of the total number of walks per year (as this is not information that is available in the iNaturalist NZ dataset), we found that more then 5% of the GLM fits resulted in a type 1 error. In the actual iNaturalist NZ data the type 1 error rate is likely to be larger as there are likely additional sources of variation beyond the annual change in the species abundance and those that we have accounted for in the GLM model. Therefore, it is possible that the 13% of species that had a significant trend estimate and acceptable dispersion parameter are all type 1 errors.

Detecting a 2% annual increase in a kiwi abundance

As well as type 1 errors, there is also the risk of type 2 errors. A type 2 error occurs when we fail to reject the false null hypothesis of no annual change in the species abundance, i.e. we fail to identify a significant annual trend that is actually present in the species abundance. As a case study we considered the ability to detect a 2% annual increase in the kiwi abundance given 10 and 20 years of iNaturalist NZ data. We used simulated data to show that with only 10 years of data where there is a true 2% annual increase in the taxa abundance we could not reliably detect an annual increase for any of the kiwi taxa, including all the taxa combined together. However, with 20 years of kiwi citizen science data there was a relatively high probability of detecting a 2% annual increase in the underlying kiwi abundance for each of the kiwi taxa, and especially for all the taxa combined together. This result is based on the assumption the number of observers continued to increase annually over the 20 years. Observations are consistently being shared to iNaturalist NZ, and we have shown as the number of years that data is collected for increases, the ability of this data to make temporal ecological inferences about a species abundance will improve.

The kiwi case study also raises the issue of the appropriateness of applying the same model specification to each species. For each species we offset by the total number of observers in New Zealand. However, some species are confined to very small geograph-

ical locations within New Zealand. For example the Rakiura kiwi is only found on Rakiura Island (Stewart Island), and the Rowi kiwi is only found in Ōkarito. Therefore in these cases it would be more sensible to offset by the total number of observers that have shared an observations from Rakiura Island, and Ōkarito.

Gathering more metadata about observer effort

More metadata about citizen scientists' observation behaviour is required to form well specified models. We have shown in this chapter that with the current limited metadata about annual changes in citizen science observer effort, i.e. just the annual change in number of active observers, it is not possible to specify a statistical model for changes in annual observation data. There are two options to rectify this problem: (1) explicitly collect additional metadata from the iNaturalist NZ observers, e.g. the number of walks they do per year, length of walks, time spent observing, etc; (2) infer this metadata from the already collected information.

The first option of explicitly collecting this additional metadata is done on the eBird citizen science project. Observers have the option of selecting a collection protocol where they share information about the time spent observing, distance travelled, and species they did not observe. eBird, also has the option of an observer sharing a "casual observation" which is similar to the protocol iNaturalist NZ currently uses. The ability for the observer to select different observing protocols means that the citizen science project and observers get the best of both worlds. If the citizen scientist wants to share casual observations they are able to, and these observations could be excluded from any analysis of annual changes in species abundances. On the other hand, if an observer is willing to share a more complete set of data, they are able to select a suitable protocol. The additional information that is collected in these effort-based sampling protocols can then be included in the specification of a suitable statistical model. iNaturalist NZ may be in a good position to adopt these different collection protocols. For example, iNaturalist NZ has already built up a solid base of observers that may be willing to contribute the additional information. Furthermore, iNaturalist NZ has an expansive observation dataset that could be used to create lists of expected species at different locations to aid with the collection of absence data.

The second option of inferring metrics of observer effort from the already collected dataset reduces the burden on the observer, but may be subject to error. For example, Isaac et al. (2014) used records of other species to control for variation in observer effort, by either assuming that the record of one species indicates the absence of others, or as a means for estimating observer sampling effort. However, this largely relies on the assumption that species are recorded in assemblages and therefore the failure to

record a species is interpreted as a non-detection, instead of a "not searched for". In the case of iNaturalist NZ, this is likely a false assumption. As we showed in Chapter 2, most iNaturalist NZ observers share very few observations to the project, and it is very unlikely that these observers did not encounter more taxa during the days they shared an observation.

Areas for future work

Throughout this chapter we made the simplifying assumption to not include variation in observer probability of sharing an observation. This is almost certainly not the case as we are assuming every observer has the same probability of sharing an observation, and that throughout a walk an observers' probability of sharing an observation does not change. However, this assumption greatly simplified our simulation process as for ever observer walk we were able to draw from a binomial distribution the number of individuals that were observed given the probability of sharing an observation and the number of individuals present during the walk. Recall from Chapter 4a that the probability of sharing an observation is a combination of the encounter probability and the probability of sharing an observation to iNaturalist, where the encounter probability is a function of the length of the observer walk, the observer's perceptive radius, the home range of the individual, and the relative speed of the individual to the observer. In the next chapter, we revisit this assumption that there is no variation in the probability of sharing an observation.

Inferences about species abundances from citizen science data may be used as an early warning system to identify species that need further monitoring. In practice, the estimations of species trends we found from the iNaturalist NZ data could be used as an early warning sign of species that may require more robust monitoring. For example, our model estimated that the little spotted kiwi abundance is declining. In the kiwi case study we found that, based on less than 10 years of data, this trend estimate may not accurately reflect the biological changes in the kiwi abundance. However, it may warrant further investigation with more standardised practices and also close monitoring as more years of kiwi data build up on iNaturalist NZ. However, it is important to be aware of the type 1 and type 2 error rates when using this method in practice. A type 1 error could result in spending unnecessary resources monitoring a species that was not experiencing any annual change in abundance. However, a type 2 error of failing to identify a species that either has an annually increasing or decreasing abundance could be much more problematic. For example, it could lead to the unnoticed decline of a native species abundance or an undetected explosion of an unwanted invasive species abundance. The results from the simulated datasets provide insights

into the likelihood of a species being subject to a type 1 or type 2 error and this could be used to assign a level of confidence about trend estimates from the iNaturalist NZ dataset

Conclusion

Observational citizen science data without standardised data collection protocols is subject to noise in the data from many sources of variation in the observer behaviours. Without measurable covariates about the observation process it is difficult to use a statistical model to estimate annual changes in a species abundance without making spurious conclusions based on the noise of the citizen scientist behaviour. However, these problems could be alleviated by introducing the explicit collection of data on observer effort, or inferring metrics on observer effort from the already recorded observation data. In the complete absence of being able to use observational ecological citizen science data to reliably monitor changes in a species population, there are many other purposes and benefits of the citizen science projects, for example encouraging participants to explore and engage with nature. This is a very important quality as research has shown that if people are actively engaged in nature they are more eager to support policies and efforts to enhance and restore nature (Toomey et al., 2020; McKinley et al., 2017).

Chapter 4c

Modelling variation in citizen scientists' contribution behaviours

Data and models,
combined to find insights, show
observers get bored.

Abstract

In the previous two chapters we built up a method to understand when noisy ecological citizen science, e.g. iNaturalist, may be used to detect actual temporal abundance changes for a taxon. However, until now we have assumed the probability of an observer sharing an observation of an individual member of a taxon is homogeneous across observers, walks, and throughout a walk.

In this chapter we outline three candidate models that include varying levels of observer variation to generate the probability distribution of the number of observations shared per observer walk. We include a model that allows an observer's probability of sharing an observation of an individual to decay as the walk and number of individual encounters progresses.

We use maximum likelihood model fitting techniques to fit the three models to empirical probability distributions of the number of observations shared on walks in the iNatualist NZ dataset. We consider the empirical probability distributions of the number of observations an observer shares of a particular taxon during a walk, any taxa within a particular iconic taxa subset, and just any taxa during a walk.

For all subsets of the iNaturalist NZ data the model that allows the probability of sharing an observation to decay as a walk progresses almost always has support from the empirical data.

The binomial model that assumes observer behaviour is homogeneous across observers, walks, and throughout a walk has the least support of the three models we consider.

The results from this chapter give insights into when our assumption of homogeneous observation behaviour was valid in the previous two chapters. Furthermore, this work provides viable alternatives to our previous assumption of homogeneous observation behaviour. Future work would involve incorporating the results from this chapter into the methodology outlined in the previous two chapters.

Introduction

In previous two chapters (Chapter 4a and Chapter 4b) when we modelled iNaturalist NZ observers sharing observations of individuals, we assumed all observers had the same probability of encountering and sharing observations. We assumed this probability was homogeneous across different observers, different walks, and throughout a given walk. However, in reality it is unlikely all observers will be homogeneous with respect to observation behaviour. For example, Sauermann and Franzoni (2015b) examined daily contribution patterns of citizen scientists in seven different citizen science projects hosted on the platform Zooniverse. They found that a small number of participants contributed a large share of the contributions. In particular, a large share of participants only contributed once to a citizen science project. They also found that participation frequency declined over time for even the highly active users. We found in Chapter 2 that iNaturalist NZ does not deviate from this typical effort profile. Over the 7 years of iNaturalist NZ data from 2013 - 2019, 10% of observers have shared 90% of the observations. Furthermore, 35% of observers only shared one observation to iNaturalist NZ. Therefore, it is clear that assuming all observers have the same observation behaviour is a great simplification of reality. However, it was also a practical simplification to build up the methods in the previous two chapters.

In this chapter we aim to link the distributions of daily observations per observer to an underlying mechanism that is driving the observer variation. We model both inter-observer variation (variation between different observers/walks) and intra-observer variation (variation within one observer/walk). Recall, we have assumed that observers do at most one walk per day. Therefore, the number of unique days per year an observer shares an observation is also the number of walks the observer does per year. Our approach is to outline candidate models that describe a physically plausible mechanism within our context, rather than blindly fitting statistical models. For simplicity we only consider three models. The first model is the binomial base case model we explored in Chapter 4c, where all observers have the same observation behaviour during a walk and across different walks, i.e. no inter or intra-observer variation. The second model, a beta-binomial model, allows inter-observer/walk variation in the probability of sharing an observation of an individual. The third model, a decay model, allows both inter and intra-observer variation in the probability of sharing an observation of an individual. The decay model allows the probability of an observer sharing an observation to decay as the walk progresses and they encounter more individuals. We are only considering three models, therefore in an absolute sense none of these models may be the best fit or describe the complexity underlying the sharing of observations to iNaturalist NZ. However, we will at least be able test if the binomial model in Chapter 4c is supported by the empirical data from iNaturalist NZ.

Although not in the context of citizen science data, Lloyd-Smith et al. (2005) carried out a similar analysis in their work on fitting three candidate models to empirical probability distributions of the number of secondary cases caused by each infectious individual from eight directly transmitted diseases. They considered three plausible candidate models: (1) a Poisson model that neglects individual variation in the reproductive number; (2) a geometric model that models constant per capita rates of leaving the infectious state; (3) a negative binomial model as a general model to incorporate variation in individual infectious histories. Lloyd-Smith et al. (2005) estimated the parameters of each model using maximum-likelihood methods. They compared candidate models using Akaike's information criterion (AIC_c) modified for a small sample size.

In this chapter we follow a similar format to the work by Lloyd-Smith et al. (2005). We consider the empirical probability distributions of the number of observations per observer walk for three cases: (1) the number of observations an observer shares of members of a particular taxon per walk; (2) the number of observations an observer shares of any taxa within a particular iconic taxa subset per walk; (3) the number of observations an observer shares of any taxa per walk. For all three cases of empirical probability distribution, we use maximum-likelihood methods to estimate the parameter values for our three candidate models. We compare the goodness of fit across the three candidate models using AIC. As a result, we gain insights into the types of behaviours iNaturalist NZ observers may have and how these behaviours vary across different taxon and observers. This work shows us that our assumption of homogeneous inter-observer and intra-observer behaviour in Chapter 4a and Chapter 4b was not always supported by the empirical data.

Materials and Methods

Models

In this section we outline the three candidate models that we fit to the empirical probability distributions of the number of observations shared per observer walk. There is only data for walks where at least one observation was recorded, i.e. all the walks with zero observations are absent from the iNaturalist NZ dataset. Therefore, in all the subsequent model fitting we truncate the models to remove any zero values.

Model 1: Homogeneous observers

This is the naive model we assumed in Chapter 4a and Chapter 4b. Figure 4c.1 shows an outline of this model. All observers have the same probability of sharing an observation of each individual they encounter during a walk. This probability does not vary during a walk, between walks, or between observers. In other words, inter-observer and intra-observer behaviour is homogeneous. The number of individuals that may be observed per walk is fixed at n. The number of observations $y^{(i,j)}$ by observer i on walk j is modelled by a binomial distribution:

$$y^{(i,j)}|p \sim \text{Bin}(n, p).$$

Where p is the probability of sharing an observation of one of the n individuals. Because we do not have any data for walks with zero observations we truncate the binomial distribution to make it exclusive of zeros:

$$Pr(y^{(i,j)} = k|p, n) = \frac{\binom{n}{k} p^k (1-p)^{n-k}}{1-(1-p)^n}, \; 1 \leq k \leq n$$

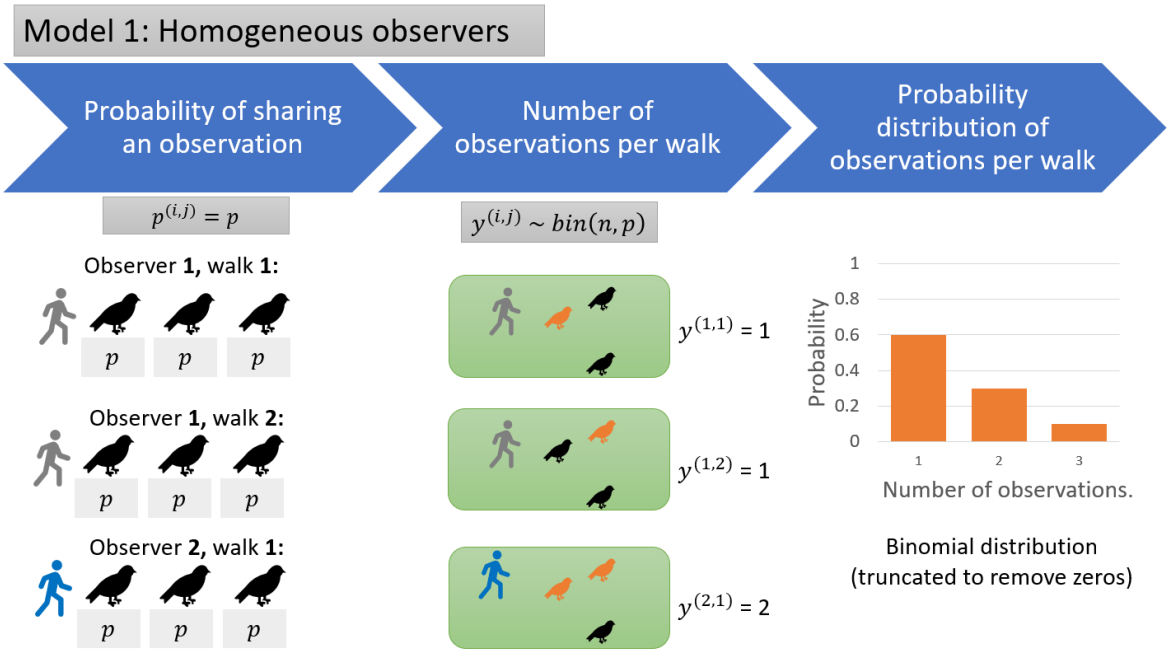


Figure 4c.1: **Model 1: Homogeneous observers.** All observers have the same probability of sharing an observation of each individual they encounter during a walk. This probability does not vary during a walk, between walks, or between observers. The resulting probability distribution of the number of observations per walk is binomially distributed.

Model 2: Inter-observer/walk variation

In this model we relax the assumption that the probability of sharing an observation of an individual is homogeneous across walks and observers. However we still assume the probability of sharing an observation does not vary during a walk. In other words, we have have introduced inter-observer/walk variation in the probability of sharing an observation. Figure 4c.2 shows a diagram of this model. The probability $p^{(i,j)}$ of observer i on walk j sharing an observation of each individual they encounter is sampled from a Beta distribution. A Beta distribution is a continuous probability distribution defined on the interval $[0, 1]$:

$$p^{(i,j)}|a, b \sim \text{Beta}(a, b).$$

Where a and b are the shape parameters of the beta distribution. Then,

$$\pi(p|a, b) = \frac{p^{a-1}(1-p)^{b-1}}{B(a, b)},\ 0 \leq p \leq 1.$$

Where $B(a, b)$ is the beta function.

The number of observations, $y^{(i,j)}$, by observer i on walk j follows a binomial distribution:

$$y^{(i,j)}|p^{(i,j)}, n \sim \text{Bin}(n, p^{(i,j)}).$$

And

$$Pr(y^{(i,j)} = k|p, n) = L(p|k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

The beta distribution is a conjugate distribution of the binomial distribution. This means that there is an analytically tractable compound distribution and one can think of the parameter p in the binomial distribution as being randomly sampled from a beta distribution. The compound beta-binomial distribution is given by

$$\begin{aligned} f(k|n, a, b) &= \int_0^1 L(p|k)\pi(p|a, b)dp \\ &= \binom{n}{k} \frac{1}{B(a, b)} \int_0^1 p^{k+a-1}(1-p)^{n-k+b-1}dp \\ &= \binom{n}{k} \frac{B(k+a, n-k+b)}{B(a, b)} \end{aligned} \tag{4c.1}$$

by using $B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$.

We truncate the beta-binomial probability distribution to be exclusive of zeros. Clearly, if $p^{(i,j)} = p$ for all i and j then the beta-binomial distribution is equivalent to the binomial distribution and therefore model 2 is equivalent to model 1.

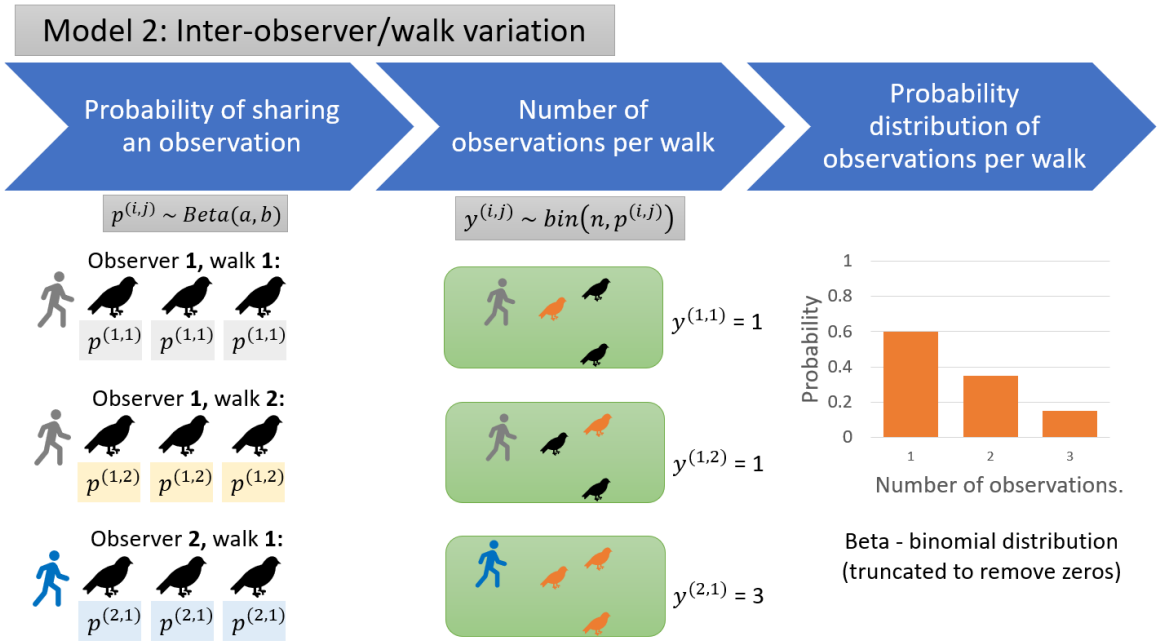


Figure 4c.2: **Model 2: Inter-observer/walk variation.** The probability that an observer shares an observation of each individual they encounter during a walk is sampled from the beta distribution. Therefore, this probability may vary across walks and observers and we have introduced inter-observer/walk variation in the probability of sharing an observation. The resulting probability distribution of the number of observations per walk follows a beta-binomial distribution.

Model 3: Inter-observer/walk and intra-walk variation

In this model we introduce observation fatigue or boredom per walk. As an observer progresses through their walk the probability of sharing an observation of each newly encountered individual decays. Figure 4c.3 shows a diagram of this model. Each observer begins a walk with a probability $p_1^{(i,j)}$ of sharing an observation of the first individual they encounter. There are multiple options for selecting this first probability. For simplicity we assume all observers always share an observation of the first individual they encounter. We could assume all observers have a probability p of sharing an observation of the first individual. Or we could also assume all observers have a probability p drawn from a beta distribution of sharing an observation of the first individual. However, because we would truncate any resulting distribution to remove the walks with zero observations all these cases are equivalent to assuming the observer always share an observation of the first individual.

As each subsequent individual is encountered the probability of sharing an observation decays according to a decay parameter $\lambda^{(i,j)}$:

$$\begin{aligned} p_2^{(i,j)} &= p_1^{(i,j)} e^{-\lambda^{(i,j)}}, \\ p_3^{(i,j)} &= p_1^{(i,j)} e^{-2\lambda^{(i,j)}}, \\ &\vdots \\ p_n^{(i,j)} &= p_1^{(i,j)} e^{-(n-1)\lambda^{(i,j)}}. \end{aligned}$$

The decay parameter $\lambda^{(i,j)}$ for observer i on walk j is drawn from a gamma distribution:

$$\lambda^{(i,j)} \sim \text{Gamma}(k, \theta).$$

Where, k is the shape parameter and θ is the scale parameter of the gamma distribution.

Therefore, the number of observations $y^{(i,j)}$ shared by observer i on j is modelled as,

$$y^{(i,j)} \sim \sum_n \text{Bernoulli}(p_n^{(i,j)}).$$

We do not need to truncate this distribution as we have assumed that every walk has at least one observation. This model would be equivalent to model 1 if $\lambda^{(i,j)} = 0$ for all i and j and $p_1^{(i,j)} = p$ for all i and j. Similarly, this model would be equivalent to model 2 if $\lambda^{(i,j)} = 0$ for all i and j and $p_1^{(i,j)} \sim \text{Beta}(a, b)$ for all i and j.

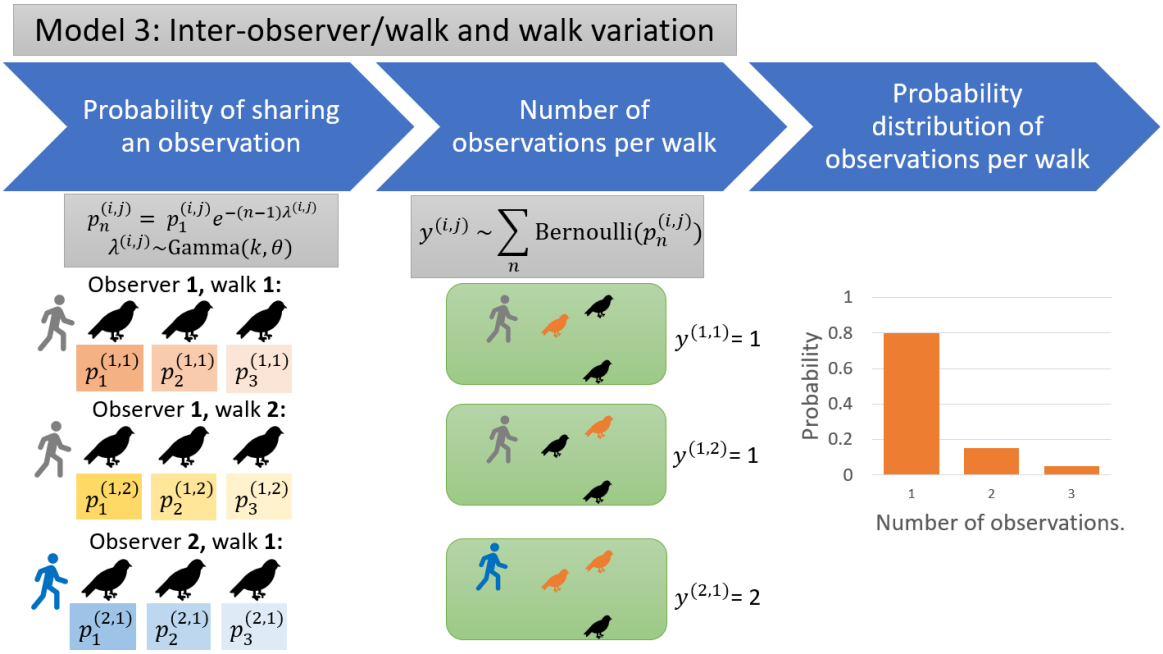


Figure 4c.3: **Model 3: Inter-observer/walk and intra-walk variation**. The probability that an observer shares an observation of each individual they encounter during a walk decays according to a decay parameter that is drawn from a gamma distribution. This decay parameter may vary across walks and observers. Therefore, we have inter-observer/walk variation and also intra-walk variation in the probability of sharing an observation. The resulting probability distribution of the number of observations per walk does not follow a standard distribution.

Data

As we did in Chapter 4b, we consider iNaturalist NZ observations for the 7 years from January 2013 to December 2017. However, in this chapter we include verifiable observations, i.e. we also include observations that have attached media but have not progressed from the "needs id" quality grade to "research grade". In total over the 7 years there are 632,916 observations, of 18,417 unique taxa that have been shared by 12,555 observers.

In this chapter we are modelling the number of observations an observer shares per 'walk'. Recall, from Chapter 4b that we assume observers do at most one walk per day. Therefore, we consider the distribution of observations per day as a proxy for the distribution of observations per walk.

We fit the three observer behaviour candidate models to the following three variations of the empirical probability distributions of observations per walk from the iNaturalist NZ data. Table 4c.1 summarises the three variations of empirical distributions.

1. The probability distribution of the number of daily observations shared by each observer for a unique taxon, e.g. house sparrow (*Passer domesticus*). We refer to these probability distributions as taxon distributions. Figure 4c.4 shows two examples of these distributions.

2. The probability distribution of the number of daily observations shared by each observer for each iconic taxon. Recall that iNaturalist NZ has 13 iconic taxa and some iconic taxa are nested in lower taxonomic ranks than others (e.g. Animals, Insects). Observations are assigned to the lowest matching iconic taxon. We refer to these probability distributions as as the iconic taxon distributions. Figure 4c.5 shows two examples of these distributions.

3. The probability distribution of the number of daily observations shared by each observer for any taxa. We refer to these probability distributions as the observer distributions. Figure 4c.6 shows two examples of these distributions.

To ensure we have sufficient data for modelling fitting we only consider empirical probability distributions that have at least 30 data points. For the taxon distributions this means that for a taxon to be included in the analysis there needs to be at least 30 daily counts of observations shared by any observer for that taxon. For the observer distribution, an observer needs to have at least 30 unique days of observation data. There are only 13 iconic taxa (amphibians, animals, arachnid, birds, chromista, fungi, insects, mammals, mollusks, plant, protozoan, ray-finned fish, reptile), and all iconic

	Taxon distributions	Iconic taxon distributions	Observer distributions
Data pooled by	Each unique taxon	Each unique iconic taxon	Each unique observer
Data point	Daily observation count by an observer for a given taxon	Daily observation count by an observer for a given iconic taxon	Daily observation count for a given observer
Total distributions	18,417	13	12,555
Distributions with more than 30 data points	3,314 (18%)	13 (100%)	623 (5%)

Table 4c.1: **Summary of the three empirical probability distributions used in candidate observer behaviour model fitting.** We fit the candidate observer behaviour models to 3,314 taxon distributions, 13 iconic taxon distributions and 623 observer distributions.

taxa have at least 30 daily observation counts by any observer. 30 data points is an arbitrary value, however, we test that the results still hold with 100 data points. Table 4c.1 outlines the number of distributions we would have for each of the three cases if we did not have a lower limit on the number of data points, and the number that remain after applying the limit. Only 18% of taxon distribution and 5% of observer distributions remain. However, the 18% of taxon distributions still include data from 95% of the observers and 85% of the observations. The 5% of observer distributions include data from 86% of the taxa and 54% of the observations.

We remove all daily counts of observations that are greater than 30 observations. This is an attempt to remove the instances where an observer bulk uploads multiple observations on the same day, but the observations were either taken across multiple days or by multiple people. We fit the three candidate models to the taxon and iconic taxon empirical probability distributions with and without the daily counts from observers that have only shared one observation to iNaturalist NZ. This exclusion of some observers has no impact on the total number of taxon and iconic taxon distributions we consider for model fitting.

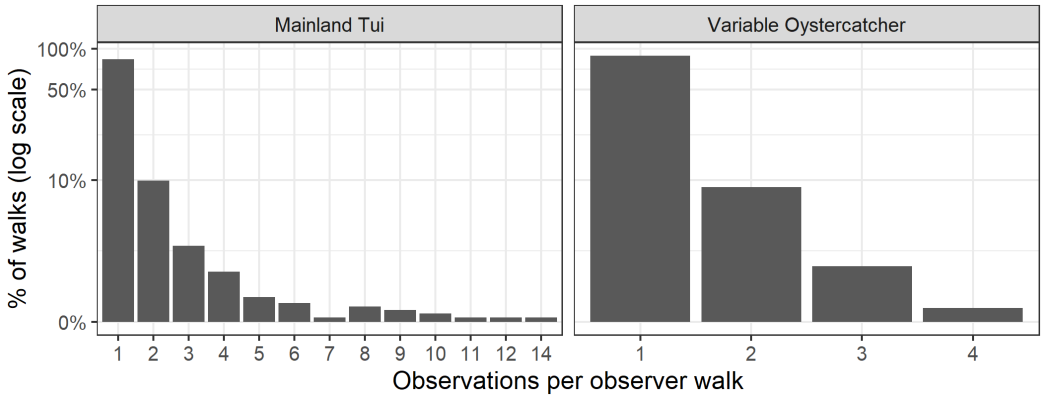


Figure 4c.4: **The majority of observers share one observation of a particular taxa per walk.** These are example distributions for the daily number of observations by each observer for a unique taxon. On the majority of walks where an observer shared an observation of a Mainland Tui *Prosthemadera novaeseelandiae*, only one Mainland Tui observation was shared, however, on some of those walks upto 14 Mainland Tui observations were shared. For the Variable Oystercatcher *Haematopus unicolor* no observer has shared more than 4 observations on a single walk.

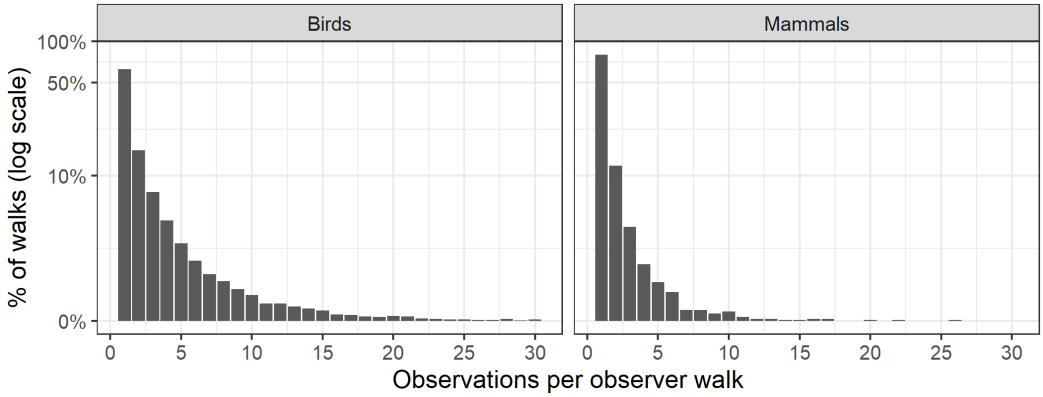


Figure 4c.5: **The majority of observers share fewer than 5 bird or 5 mammal observations per walk.** Example distributions for the daily number of observations by each observer for each iconic taxon. On the majority of walks where an observer shared an observation of any bird taxon the majority of observers shared fewer than 5 observations of any bird taxon on the same walk. This is also the case for mammal observers. However, some observers have shared up to our truncation point of 30 bird observations per walk.

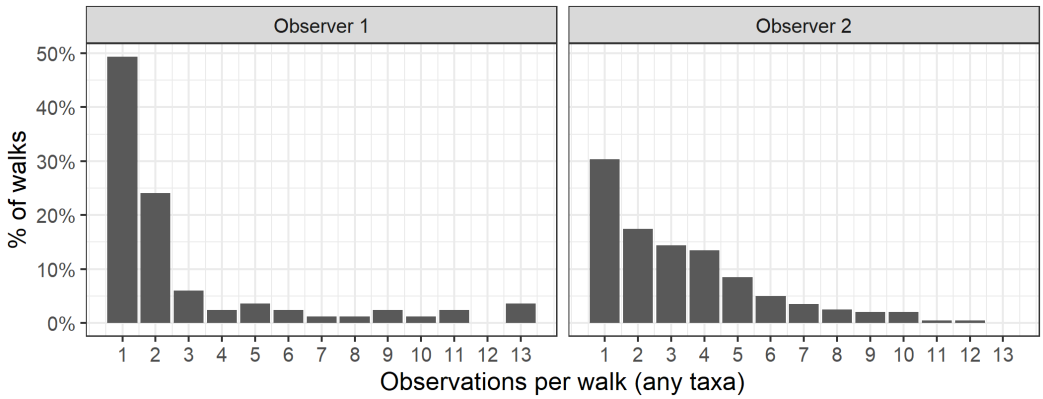


Figure 4c.6: **The distribution of observations per walk by an observer varies largely between observers.** Example distributions for the daily number of observations by each observer of any taxa. Observer 1 most often shared 1 or 2 observations of any taxa per walk, however, sometimes they shared up to 13 observations in a single walk. Observer 2 often shared up to 5 observations of any taxa per walk, and occasionally they shared 12 observations on a single walk.

Model fitting

We estimate the parameters for each model by using maximum likelihood estimation. For model 1, we estimate the probability p of sharing an observation of each encountered individual. For model 2, we estimate the shape parameters a and b of the beta distribution that the probability parameters $p^{(i,j)}$ are sampled from. For model 3, we estimate the shape parameters, k and θ, of the gamma distribution that the decay parameters $\lambda^{(i,j)}$ are sampled from. For all three models we assume a fixed value for the number of individuals n.

Model comparison

We use the Akaike information criterion (AIC) to compare the three candidate models. Akaike (1974) defined an information criterion (AIC) that is a rigorous way to estimate the Kullback-Leibler (K-L) distance based on the empirical log-likelihood function at its maximum point as:

$$\text{AIC} = -2\log(\mathcal{L}(\hat{\theta}|y)) + 2K.$$

Where $\log(\mathcal{L}(\hat{\theta}|y))$ is the numerical value of the log-likelihood at its maximum point and K is the number of estimated parameters. The preferred model is the candidate model with the minimum AIC value. AIC rewards goodness of fit (as assessed by the likelihood function), while penalising based on the number of estimated parameters. The parameter penalty discourages overfitting from increasing the number of parameters in the model, which almost always improves the goodness of the fit.

Δ_i	Level of empirical support for model i
0 - 2	Substantial
4 - 7	Considerably less
> 10	Essentially none

Table 4c.2: **Rules of thumb for interpreting Δ_i.** Table from Burnham and Anderson (2002).

AIC estimates the quality of each model relative to each other model in the set and therefore AIC does not inform anything about the absolute quality of the model. Furthermore, an individual AIC value, by itself, is not interpretable. Thus, the AIC differences Δ_i for model i are very important and useful:

$$\Delta_i = \text{AIC}_i - \text{AIC}_{min}.$$

The model estimated to be best has $\Delta_i \equiv min \equiv 0$. The larger Δ_i is, the less plausible it is that model i is also a good fit given the empirical data. Some rough rules of thumb given by (Burnham and Anderson, 2002) are outlined in Table 4c.2.

Results

By taxon

In this case we are modelling the number of observations of a particular taxon that an observer shares per walk. For example, the number of house sparrow *Passer domesticus* observations an observer shares per walk. We considered each walk (by any observer) that involved an observation of taxon x. From these walks we found the empirical probability distribution of the number of observations of taxon x shared per walk. We fitted the three candidate models to the 3,314 taxon distributions with more than 30 data points.

Model fit comparison

Table 4c.3 shows the decay model has empirical support (i.e. AIC difference, $\Delta_i < 7$) for all 3,314 of the taxon distributions we considered in the model fitting. The beta-binomial model has empirical support for 77% of the taxon distributions and the binomial model has empirical support for 66% of the taxon distributions. The decay model is the best fit (i.e. AIC difference, $\Delta_i = 0$) for 51% of the taxa, the binomial model is the best fit for 47% of the taxa, and the beta-binomial model is the best fit for 11% of the taxa. Note that for some taxa both the decay model and beta-binomial model had $\Delta_i = 0$. The AIC calculation has the $2K$ term that penalises the AIC score based on the number of parameters, K, in the model. For most taxa the maximum log-likelihood value of the beta-binomial model and binomial model differ by less than 2. Therefore, the AIC score is lower for the binomial model as it has one less parameter than the beta-binomial model. These results are consistent with and without observations from observers that have only done one iNaturalist NZ observations. These results are also consistent if we only consider taxon distributions with more than 100 data points.

We sub-setted these results by the 13 iNaturalist iconic taxa (plants, insects, birds, fungi, arachnids, animals, mollusks, mammals, ray-finned fish, reptiles, chromista, amphibians, and protozan). The following iconic taxa subsets contained less than 100 taxa and were left out of the following analysis; mammals, ray-finned fish, reptiles, chromista, amphibians, and protozan. Table 4c.3 shows that there were only minor differences in the model fit results across the iconic taxa groups.

	Total	Plant	Insect	Fungi	Bird	Arachnid	Animal	Mollusk
Decay model is best fit ($\Delta_{\text{Decay}} = 0$)	51% (1,703)	51%	50%	45%	67%	58%	47%	49%
Decay model has substantial support ($0 \leq \Delta_{\text{Decay}} \leq 2$)	98% (3,254)	98%	97%	99%	99%	97%	99%	99%
Decay model has support ($\Delta_{\text{Decay}} < 7$)	100% (3,314)	100%	100%	100%	100%	100%	100%	100%
Beta-binomial model is best fit ($\Delta_{\text{Beta binomial}} = 0$)	11% (350)	12%	13%	10%	4%	3%	4%	13%
Beta-binomial model has substantial support ($0 \leq \Delta_{\text{Beta binomial}} \leq 2$)	61% (2,009)	64%	63%	65%	38%	53%	57%	48%
Beta-binomial model has support ($\Delta_{\text{Beta binomial}} < 7$)	77% (2,550)	76%	83%	86%	60%	74%	74%	73%
Binomial model is best fit ($\Delta_{\text{Binomial}} = 0$)	47% (1,558)	47%	49%	54%	31%	42%	51%	45%
Binomial model has substantial support ($0 \leq \Delta_{\text{Binomial}} \leq 2$)	55% (1,835)	54%	58%	62%	40%	51%	63%	53%
Binomial model has support ($\Delta_{\text{Binomial}} < 7$)	66% (2,195)	64%	70%	74%	55%	71%	70%	66%
Total number of taxa	3,314	1660	635	313	182	146	136	116

Table 4c.3: **The decay model has support from the empirical data for the large majority of the taxa distributions.** Summary of Akaike information criterion difference (Δ_i) for the three fitted models fitted to the taxon distributions. The taxon distributions are the probability distribution of the number of observations an observer shares per walk of a particular taxon.

Decay parameter analysis

Since the decay model (model 3) with inter-observer/walk variation and intra-walk variation has empirical support for all the taxon distributions we did further analysis on the decay rates. For each taxon x, we have the maximum likelihood estimation of k and θ, the two parameters for the gamma distribution that describes the decay rate, $\lambda^{(x)}$. In this analysis we considered the expected decay rate for taxon x,

$$E(\lambda^{(x)}) = k^{(x)}\theta^{(x)}.$$

Figure 4c.7 shows the distribution of the expected decay rates of each taxon per walk. Most taxa have an expected decay rate close to 5, which translates to approximately a probability of 0.007 of sharing a second observation given a second individual has been encountered. We found that there is no evidence of a correlation between the number of observations for a taxon and the expected decay rate for the taxon.

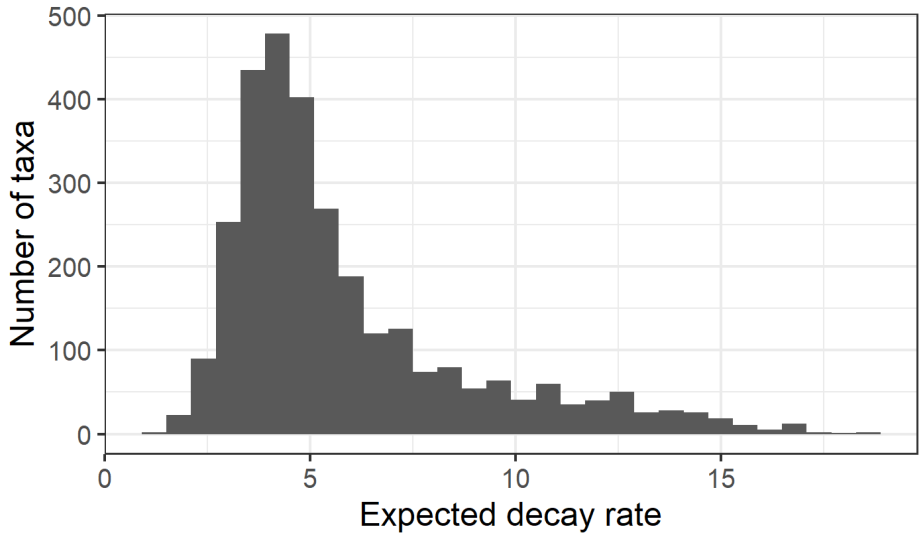


Figure 4c.7: **The majority of taxa have a decay rate close to 5.** A decay rate of 5 results in a probability of 0.007 of sharing a second observation given a second individual has been encountered.

We grouped the expected decay rates by the 13 iconic taxa subsets. We compared the distribution of decay rates across the iconic taxa subsets that have more than 100 individual taxa (7 of the 13 iconic taxa). Figure 4c.8 A shows box plots of the distributions of expected decay rates by each of these iconic taxa subsets. A box plot shows the median, 25th and 75th percentile, the maximum, and minimum points of the data. The distribution of expected decay rates for the 7 iconic taxa subsets appear to be similar. However, taxon within the mollusk iconic taxa subset appear to have slightly lower decay rates and taxon within the arachnid iconic taxa subset appear to

have slightly larger decay rates.

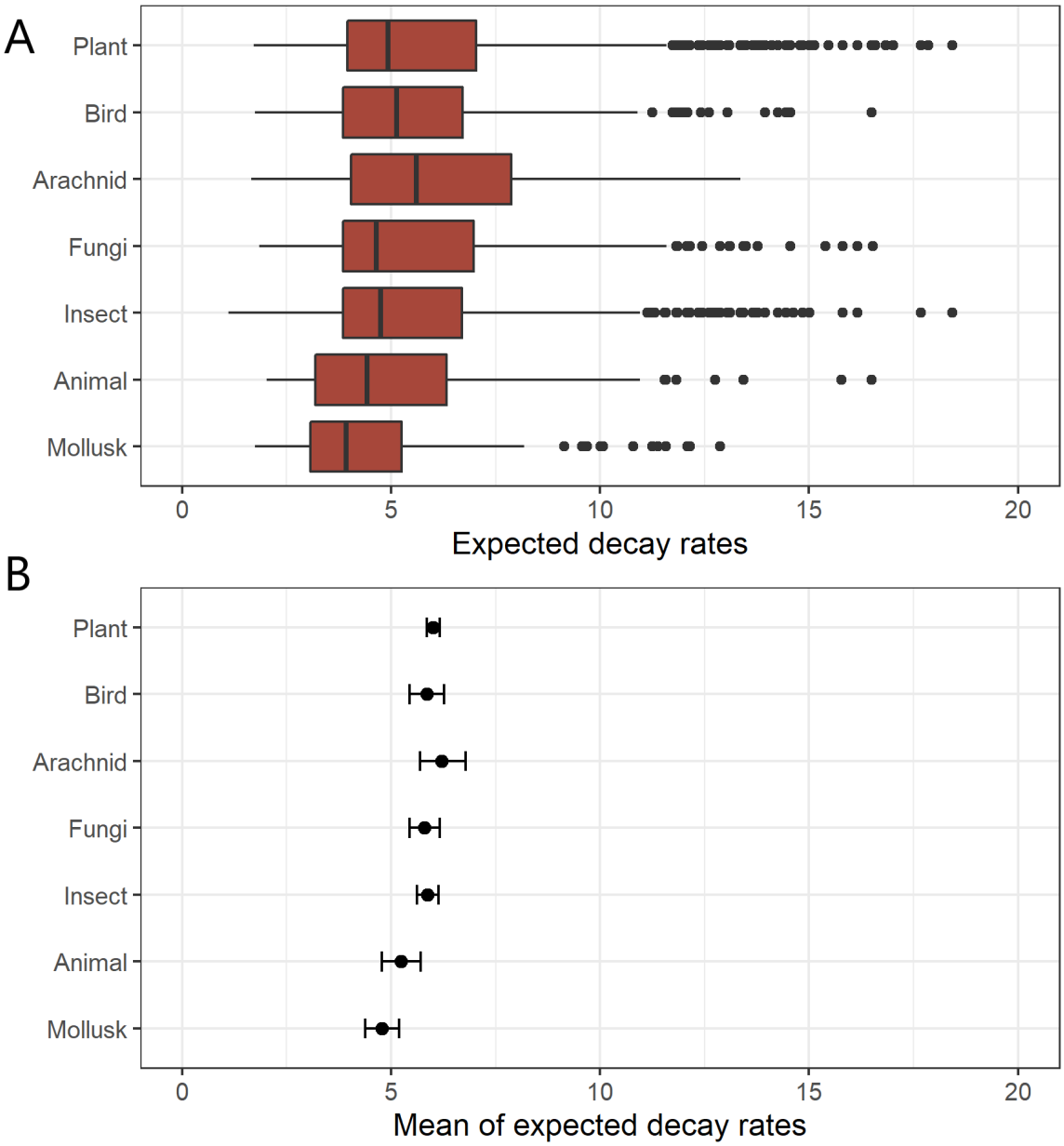


Figure 4c.8: **The distribution of expected decay rates for the 7 iconic taxa subsets with more than 100 individual taxa appear to be similar.** A, Box plots (median, 25th and 75th percentile, the maximum, and minimum) of the expected decay rates for each of the iconic taxa with model fit results for more than 100 individual taxa. B, the mean of the expected decay rates. Error bars show 95% confidence intervals around the mean and were calculated by using bootstrap samples of the datasets for each iconic taxa subgroup. We have excluded iconic taxa subsets that had fewer than 100 taxa.

Figure 4c.8 B shows a 95% confidence interval on the mean of expected decay rates

within each iconic taxa subset. The 95% confidence intervals around the mean were calculated by using bootstrap samples of the datasets for each iconic taxa subgroup. The mean of the expected decay rates for taxa within the mollusk iconic taxa subsets are smaller than the taxa within the other iconic taxa subsets.

We split the taxa into subsets by the establishment means; endemic, native, and introduced. For 660 (20%) of the taxa we do not have data on their establishment mean and they are left out of this analysis. For the remaining taxa 40% are endemic, 30% are native, and 30% are introduced. We further filtered these results to just plant and just bird taxa because for these two iconic taxa subsets it is likely that most iNaturalist NZ observers will have some knowledge of the taxa that are native and/or endemic, verses introduced to New Zealand. For bird taxa, 43% are endemic, 36% are native, and 21% are introduced. For plant taxa, 42% are endemic, 25% are native, and 34% are introduced. Figure 4c.9 shows the mean of the expected decay rates for taxa within these three subsets. Across all the data, introduced taxa, on average have a higher expected decay rate than taxa that are either endemic or native. This result is also the case when we just consider the 1,330 plant taxa with establishment means data. For the 144 bird taxa with establishment means data the result is less definitive but still follows the same trend.

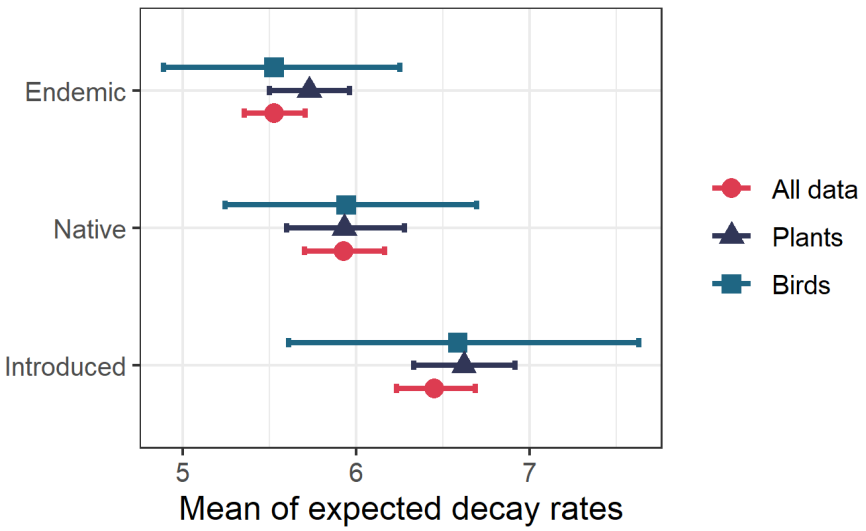


Figure 4c.9: **On average endemic and native taxa have a lower expected decay rate than introduced taxa.** The error bars are 95% confidence intervals on the mean and were calculated by using bootstrap samples of the datasets for each subset of data. We repeated this analysis for all the data, just the bird taxa, and just the plant taxa.

By iconic taxa

In this case we are modelling the number of observations of any taxa within an iconic taxon subset an observer shares per walk. For example, the number of bird observations an observer shares per walk. We considered each walk that involved an observation of any taxon within iconic taxon subset x_I. From these walks we found the empirical probability distribution of the number of observations of any taxa within the iconic taxon subset x_I shared per walk. We fitted the three candidate models to each of the 13 iconic taxa distributions.

Model fit comparison

For all 13 iconic taxa the decay model is the best fit and the binomial or beta-binomial model have no support. That is, $\Delta_{\text{Decay}} = 0$, $\Delta_{\text{Binomial}} > 7$, and $\Delta_{\text{Beta-binomial}} > 7$ for all 13 iconic taxa. This result is the same with and without the daily observation counts from observers that have only shared one observation to iNaturalist NZ.

Decay parameter analysis

As we did with the taxon distribution analysis, we analyse how the decay parameters vary across the 13 iconic taxa distributions. For each iconic taxon we have a maximum likelihood estimation of k and θ, the two parameters for the gamma distribution that the decay rate parameter, λ, is sampled from for each iconic taxon. Figure 4c.10 shows the relationship between the number of taxa per iconic taxon subset and the expected decay rate of any taxon within a particular iconic taxon subset (adjusted r-squared = 0.24). However, the fitted linear model does not have a significant slope. Note that the number of taxa does not necessarily correlate with the number of individuals. For example, an iconic taxa subset could have a lot of different rare taxa, whereas, another iconic taxa subset could have few unique taxa but they might be very prevalent. Therefore, a more insightful investigation would be the correlation between expected decay rates and the population size of the iconic taxa subset.

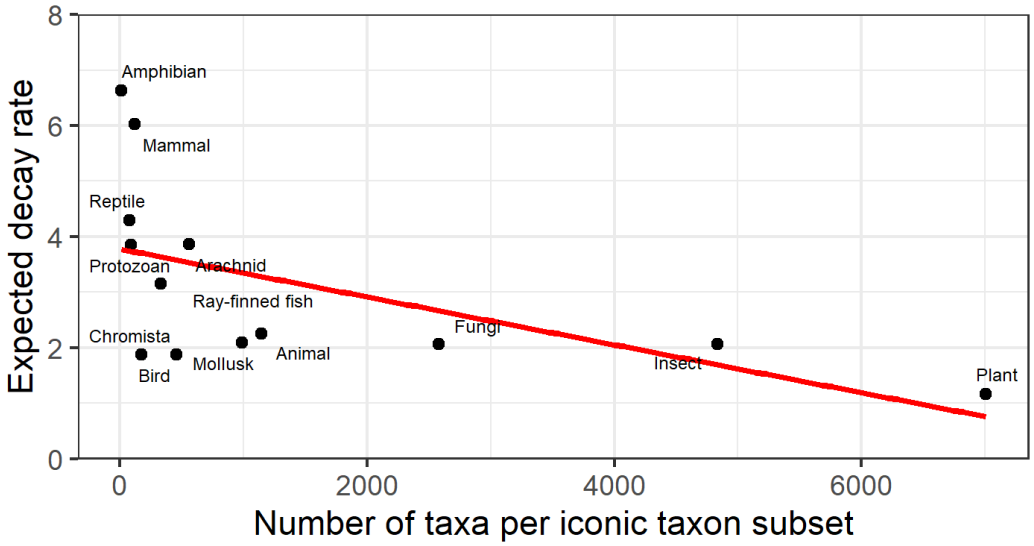


Figure 4c.10: **The relationship between the number of taxa per iconic taxon subset and the expected decay rate of any taxon within a particular iconic taxon subset.** We fitted a liner model between the number of taxa in an iconic taxon subset and the expected decay rate of a taxa within that iconic taxon subset set. The slope parameter is not statistically significant and the adjusted r-squared is 0.25.

By observer

In this case we are modelling the number of observations a particular observer shares of any taxa per walk. For each observer we found the empirical probability distribution of the number of observations of any taxa they shared on a walk. We fitted the three candidate models to each of these observer distributions with more than 30 data points (i.e. observers that have done more than 30 walks).

Model fit comparison

Table 4c.4 shows the decay model has empirical support (i.e. AIC difference, $\Delta_i < 7$) for 93% (577) of the observer distributions we considered in the model fitting. The beta-binomial model has empirical support for 43% of the observer distributions and the binomial model has empirical support for 13% of the observer distributions. The decay model is the best fit (i.e. AIC difference, $\Delta_i = 0$) for 75% of the observers, the beta-binomial model is the best fit for 19% of the observer, and the binomial model the best fit for 6% of the observers. These results are also consistent if we only consider observer distributions with more than 100 data points.

	Proportion	Value
Decay model is best fit ($\Delta_{\text{Decay}} = 0$)	75%	468
Decay model has substantial support ($0 \leq \Delta_{\text{Decay}} \leq 2$)	83%	520
Decay model has support ($\Delta_{\text{Decay}} < 7$)	93%	577
Beta-binomial model is best fit ($\Delta_{\text{Beta-binomial}} = 0$)	19%	119
Beta-binomial model has substantial support ($0 \leq \Delta_{\text{Beta-binomial}} \leq 2$)	29%	183
Beta-binomial model has support ($\Delta_{\text{Beta-binomial}} < 7$)	43%	269
Binomial model is best fit ($\Delta_{\text{Binomial}} = 0$)	6%	36
Binomial model has substantial support ($0 \leq \Delta_{\text{Binomial}} \leq 2$)	8%	51
Binomial model has support ($\Delta_{\text{Binomial}} < 7$)	13%	78

Table 4c.4: **The decay model has support from the empirical data for the large majority of the observers distributions.** Summary of Akaike information criterion difference (Δ_i) for the three fitted models fitted to the observer distributions. The observer distributions are the probability distribution of the number of observations an observer shares per walk of any taxa.

Decay parameter analysis

Figure 4c.11 shows the distribution of decay rates for each observer distribution where the decay model has empirical support (i.e. AIC difference, $\Delta_i < 7$). The expected decay rate is centered around 2 which translates to approximately a probability of 0.14 of sharing a second observation, given a second individual in encountered. These decay rates are lower than when we considered taxa individually.

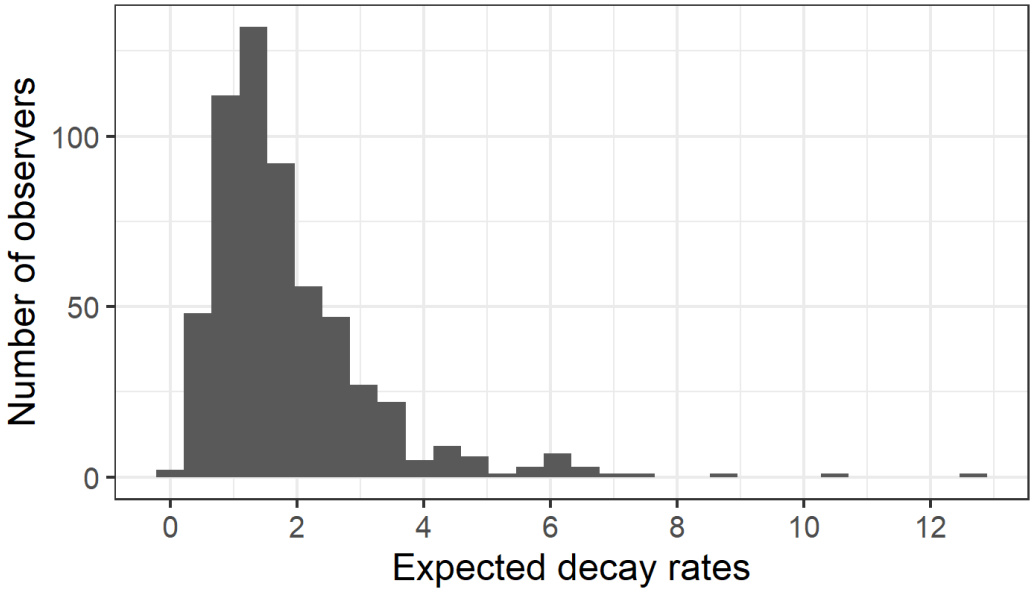


Figure 4c.11: **The majority of observers have an expected decay rate less than 4 when we consider sharing observations of any taxa to iNaturalist NZ on a walk.** This histogram shows the distribution of expected estimated expected decay rates across the 623 observers that had at least 30 data points and were therefore included in the model fitting.

We found that there is a negative correlation between the number of observations an observer has shared to iNaturalist NZ and their estimated expected decay rate on a particular walk, Figure 4c.12. A linear model between the log of number of observations an observer has shared and the expected decay rate has a statistically significant slope (p=2e-9) and an adjusted R-squared value of 0.1. For this reason we did not subset the decay results any further, for example into subgroups based on the number of observations an observer has shared.

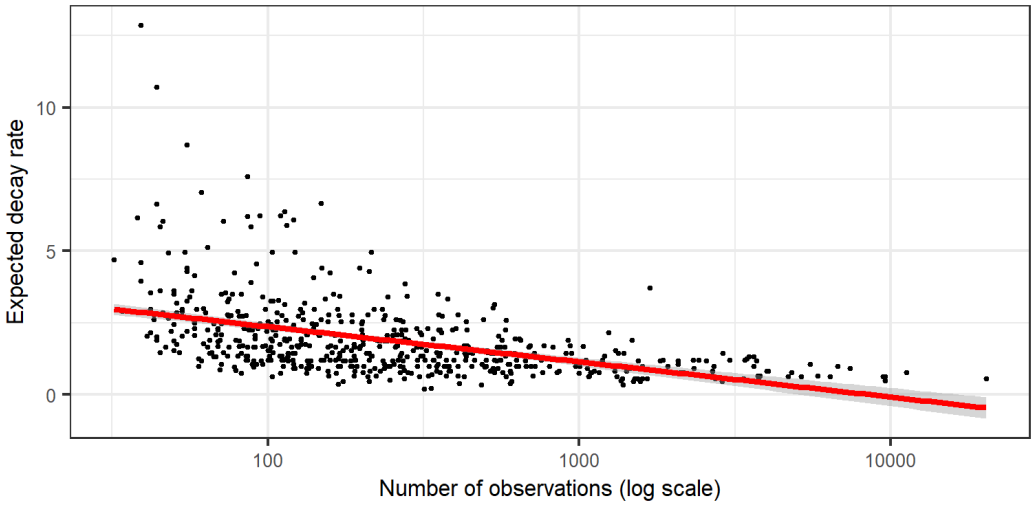


Figure 4c.12: **There is a negative correlation between the number of observations an observer has shared to iNaturalist NZ and the expected decay rate on a walk.** We fitted a liner model between the number of observations an observer has shared to iNaturalist NZ and the expected decay rate. The slope parameter is statistically significant (p=2e-9 and the adjusted r-squared is 0.1).

Discussion

In this chapter we revisited our assumption in Chapter 4a and Chapter 4b that an observer's probability of sharing an observation of an individual they encounter is homogeneous across walks, between different observers, and also for the duration of the walk. We outlined three candidate models that included physically plausible sources for the variation in this probability of sharing an observation. The first model was the model with no variation that we assumed in Chapter 4a and Chapter 4b. The second model introduced variation between walks and observers in the probability of sharing an observation. This was modelled by for each walk, sampling the probability of sharing an observation from a beta distribution. The third model introduced variation during the walk in the probability of sharing an observation of an encountered individual. This was modelled by including a decay parameter that was sampled from a gamma distribution. After every encounter of an individual the probability of sharing an observation was reduced according to the decay parameter.

We used maximum-likelihood model fitting methods to fit these three models to the empirical probability distributions of the number of observations shared per walk from iNaturalist NZ. We considered three slices of the empirical data of the number of observations shared per walk. First, we considered the number of observations observers share of each taxon per walk. For example, the number of observations of a house sparrow per walk. Second, we considered the number of observations observers share of any taxa within an iconic taxon subset per walk. For example, the number of observations of any bird taxa per walk. Third, we considered the number of observations observers share of any taxa per walk.

We compared the goodness of fit of the three models to the empirical data by comparing the AIC values. AIC is useful in selecting the best model in the set of candidate models. However, if all the models are very poor, AIC will still select the model estimated to be best. But even that relatively best model might be poor in an absolute sense. Therefore, there is no guarantee that there is not a model outside of the three models we considered that is a better fit than all three models.

The model fitting results were similar for all three slices of empirical probability distributions from the iNaturalist NZ data. The model that allowed an observer's probability of sharing an observation to decay throughout the walk was most often the best fitting model and was almost always supported by the empirical data. The binomial model that we have assumed to be the case in 4a and Chapter 4b had support from the empirical data for 2/3 of the taxa. However, when we considered the number of observations shared per walk of any taxa within an iconic taxa subset, or any taxa, the

binomial model does not have very much support from the empirical data. In Chapter 4a we were considering the observations shared per walk of a particular taxon. Therefore, on average, our assumption that the number of observations shared per walk by a given observer is binomially distributed is supported by the empirical data for 66% of the taxa.

In this chapter we have assumed that the variation in the number of observations shared per observer walk is due to variation in an observer's probability of sharing an observation of each individual they encountered. However, there are many other factors that could be contributing to the variation in the number of observations shared per walk. For example, the distance of the walk, the duration of the walk, or the number of individuals present on a walk. We have also only modelled two forms of variation beyond our base case of no variation in the probability of an observer sharing an observation. Despite not exhaustively modelling all the potential sources of variation during an observer walk we have still learnt the fundamental point, that regardless of the reason, it is not always supported by the empirical data that the probability distributions of the number of observations shared per walk follow a binomial distribution. Further, we have learnt that observers are more likely to just share one observation per walk than would be expected from a binomial model. A model that allows a decay in observation sharing enthusiasm throughout a walk is well supported by the empirical data.

There are multiple factors that could be contributing to the tendency of observers to share just one observation per walk. For example, many observers may take multiple photographs throughout a walk, but they are often not sharing observations to iNaturalist NZ until after the walk is complete, and therefore it is quite likely they would just select the best photo to share. We also did not consider the probability of encountering multiple individuals and therefore the ability to share more than one observation. Future work would involve including the decay model in the simulated data in Chapter 4a. This would enable us to quantify the impact this decaying behaviour and therefore tendency to just share one observation per walk has on the ability to robustly detect temporal changes in the underlying taxa abundances. This analysis would provide an insight into how important it is to encourage citizen scientists to share multiple observations and particularly multiple observations per walk.

Chapter 5

Final Discussion

Options were explored.
Suggestions are given to,
improve certainty.

Citizen science has grown rapidly in recent years and collectively citizen scientists have formed some of the largest biodiversity datasets in the world (Bonney et al., 2014, 2009; Baker, 2016; Pocock et al., 2017; Wiggins, 2011). However, there are ongoing concerns on the ability to use the data in a scientific framework (Bird et al., 2014; Dickinson et al., 2010). This thesis has explored a range of mathematical and statistical techniques that may be used at different stages and with different types of citizen science projects to improve the scientific usefulness of citizen science outputs.

In Chapter 2 we described the iNaturalist platform, and presented a range of data from the New Zealand chapter of the platform. iNaturalist NZ has minimal protocols for participants to follow. This presents challenges when attempting to make use of the data for scientific purposes. The iNaturalist NZ citizen science project was used as a case study to develop and test our methods and these methods may be generalised to other similarly structured citizen science projects.

In Chapter 3a and 3b we focused on the image classification element of the iNaturalist NZ citizen science project. Currently, iNaturalist uses a simple majority vote method to combine individual citizen scientists' classifications for an image into a final classification. We outlined a Bayesian approach that estimates and utilises a measure of each participant's ability to classify an image. This approach optimises the citizen scientists' classification efforts while also ensuring a desired level of certainty in final classification.

In Chapter 4a and 4b we focused on the observation sharing stage of iNaturalist NZ. We used random walk theory to explore the ability to make reliable ecological inferences about temporal changes in species abundances from noisy citizen science observation data. We found that without sufficient meta-data about observer behaviour it is difficult to specify an appropriate statistical model. Therefore, it is challenging to differentiate between changes in species abundances due to variation in observer behaviour, versus ecological changes in species abundances.

In Chapter 4c, we used maximum likelihood model fitting techniques to gain insights into typical citizen science observation sharing behaviours. We found that the probability of an observer sharing an observation decays throughout a citizen scientist walk, and that this probability was well supported by the iNaturalist NZ data. In fact, many citizen scientist walks on iNaturalist NZ only result in the sharing of one observation.

The design of a citizen science project greatly determines its usefulness. The design can impact on the ability to apply post-hoc statistical and mathematical manipulations

to the data, and the conclusions that can be reached. Some designs allow user identification accuracy to be efficiently and precisely estimated. In Chapter 3a, we showed that combining individual identifications with Bayes' formula, when the accuracy of the users is known, always results in a higher or equivalent classification accuracy compared to using a majority vote. Designs allowing metadata about observer recording behaviour are also useful. In Chapter 4b we showed that the ability to use the breadth of citizen science observation data to robustly monitor ecological changes in species abundances is dependent on knowledge of variation in citizen scientists' recording behaviours. Thus, the design of a citizen science project or platform can greatly impact future reliability and usefulness of the data for scientific purposes.

Modern technology is changing the landscape of citizen science (Newman et al., 2012). Historically, iNaturalist relied solely on humans to classify the breadth of observations shared to the platform. A small portion of iNaturalist NZ participants reported the vast majority of total identifications. However, in 2017 a computer vision algorithm that was trained on historical iNaturalist research grade observations was integrated into the platform. With an increasing rate of new observations, all of which require identification, computer vision has the potential to take much of this work burden. An important piece of future work will be to evaluate the impact that the addition of computer vision has had on the image classification process on iNaturalist NZ. Our work in Chapter 3b estimating user accuracies and classification accuracies on the first five years of iNaturalist NZ data, before the addition of computer vision, will provide a useful baseline for future evaluations.

Users of research grade iNaturalist observations (e.g. computer vision algorithms using the data for training) rely on these observations to be accurate. Our Bayesian approach, applied to the first five years of iNaturalist NZ data, was able to move an observation to research grade once the certainty of the collective classification was above a set threshold, rather than requiring a majority vote. In theory, this improves and quantifies the accuracy of research grade observations. However, within the iNaturalist NZ data there is a high level of agreement between identifiers on the vast majority of observations. As a result, the estimated user accuracies are high for most users. This may be truthful, but it may also be inflated due to the high level of agreement among identifiers. Therefore, important future work will be to validate the results from this Bayesian approach by asking expert panels to classify a selection of images.

Observation-based citizen science projects are often the only practical way to gather data at the scales required to answer many ecological and conservation questions (Tulloch et al., 2013; Miller-Rushing et al., 2012). We explored using iNaturalist NZ data for temporal monitoring of species abundance. This work could be adapted to monitor-

ing changes in species distributions. However, if we did so, it would still be necessary to differentiate between ecological changes in a species' distribution while accounting for variation in recorder behaviour. For example, a species distribution could appear to be expanding, but that observation may simply be due to an increasing number of participants making observations across a larger area of New Zealand, and recording them in iNaturalist NZ. Currently in the scientific literature, iNaturalist data has largely been used for discovering new species (Lebel et al., 2020) and monitoring the arrival and spread of invasive species (New Zealand bio-recording network trust, 2020). Thus, there is a lot of unexplored potential of the use of iNaturalist NZ observation data, and this thesis makes significant headway in tapping into that potential.

Recruiting and sustaining citizen science participation is challenging (Tinati et al., 2017). In Chapter 2 we showed that a small portion of participants make the vast majority of both observation and identification contributions to iNaturalist NZ. For many taxa, a citizen scientist will share at most one observation per walk. Further, in Chapter 4c we showed that the probability of an observer sharing a second observation of the same species in a walk is very low. This adds to the challenge of being able to use observation data to detect temporal changes in the abundance of a species population. If the majority of observers only share at most one observation of one species per walk, it will be very difficult to detect changes in abundance, as this may have no influence on the number of recorded observations.

In Chapter 3b we showed that adopting a Bayesian approach for the classification of research grade observations optimised the allocation of identifications. As a result, with the same number of identifications, more observations could reach research grade. If this method was implemented on iNaturalist NZ, it could aid our work in Chapter 4b, as there would be more research grade observations per species and therefore an increased ability to detect a temporal ecological trend in the species population.

Recommendations for iNaturalist NZ

To improve the ability to use iNaturalist NZ data for scientific research questions, we make the following recommendations for the platform.

Image classification

Research grade classification accuracy could be improved by accounting for differences in user accuracies. We have outlined a Bayesian approach that simultaneously estimates user accuracies and image classifications. This approach is suitable for iNaturalist NZ as there is no requirement that users have identified any images with a known identity. However, the Bayesian approach is computationally expensive and therefore may be

problematic to use in operation across the iNaturalist platform. An estimation of user accuracy may be efficiently obtained by requiring users to identify some images with a known identity. This could be a one-off when the user first joins iNaturalist or periodically. Given there are currently no images that have been identified with known ground truth on the iNaturalist NZ platform, one method going forward would be to add some images with a known identity. Alternatively, images that have already been classified to research grade could be validated by an expert panel.

Observation sharing

To improve the ability to use research grade observation data for species population monitoring, we also recommend that iNaturalist introduces a method to collect information on observer behaviour. This could be done by adopting a similar approach to eBird, where there are multiple recording protocols for the observer to select from and some require the observer to also log information about their recording effort. Alternatively further work could be done on inferring recording effort from data that are already collected, or on investigating the minimal amount of additional information required from observers to obtain a reliable measure of variation in recorder effort.

Sustaining participation

iNaturalist NZ, like many citizen science projects, faces the problem that there are a few regular participants make the vast majority of the contributions. This limits the usefulness of the data collected from the majority of observers and identifiers due to the inability to make an inference about observer behaviour, and insufficient information to estimate identifier accuracy. We recommend that iNaturalist investigates and implements mechanisms to increase participant retention and activity. For example, the critical challenge of recruiting and sustaining participation (Tinati et al., 2017) has led to some pursuing gamification of citizen science (Tinati et al., 2017; Iacovides et al., 2013; Eveleigh et al., 2013).

Conclusion

Ecological citizen science has seen rapid growth in recent years. For many projects the focus has been more on fostering an interest in nature and building connected communities of nature watchers, rather than ensuring the data will be able to be robustly used for scientific purposes. With the solid base of participants that ecological citizen science projects have built up, and the expanding literature on using citizen science for scientific purposes, now is a great time for ecological citizen science projects to make changes to their platforms that fill the gaps of any information they are failing to collect that would allow them to answer desired scientific questions.

References

Aguzzi, J., C. Costa, Y. Fujiwara, R. Iwase, E. Ramirez-Llorda, and P. Menesatti (2009). A novel morphometry-based protocol of automated video-image analysis for species recognition and activity rhythms monitoring in deep-sea fauna. *Sensors 9*(11), 8438–8455.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control 19*(6), 716–723.

Alt, W. (1980). Biased random walk models for chemotaxis and related diffusion approximations. *Journal of mathematical biology 9*(2), 147–177.

Anderson, D. R. (2001). The need to get the basics right in wildlife field studies. *Wildlife Society Bulletin (1973-2006) 29*(4), 1294–1297.

Asch, S. E. (1956). Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological monographs: General and applied 70*(9), 1.

Baker, B. (2016). Frontiers of citizen science: Explosive growth in low-cost technologies engage the public in research. *Bioscience 66*(11), 921–927.

Bautista-Puig, N., D. De Filippo, E. Mauleón, and E. Sanz-Casado (2019). Scientific landscape of citizen science publications: Dynamics, content and presence in social media. *Publications 7*(1), 12.

Bird, T. J., A. E. Bates, J. S. Lefcheck, N. A. Hill, R. J. Thomson, G. J. Edgar, R. D. Stuart-Smith, S. Wotherspoon, M. Krkosek, J. F. Stuart-Smith, G. T. Pecl, N. Barrett, and S. Frusher (2014, 5). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation 173*, 144–154.

Blackwell, P. (1997). Random diffusion models for animal movement. *Ecological Modelling 100*(1-3), 87–102.

Bonney, R., C. B. Cooper, J. Dickinson, S. Kelling, T. Phillips, K. V. Rosenberg, and J. Shirk (2009, 12). Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience 59*(11), 977–984.

Bonney, R., J. L. Shirk, T. B. Phillips, A. Wiggins, H. L. Ballard, A. J. Miller-Rushing, and J. K. Parrish (2014, 3). Citizen science. Next steps for citizen science. *Science (New York, N.Y.) 343*(6178), 1436–7.

Bovet, P. and S. Benhamou (1988). Spatial analysis of animals' movements using a correlated random walk model. *Journal of theoretical biology 131*(4), 419–433.

Brooks, S. P. and G. O. Roberts (1998). Convergence assessment techniques for markov chain monte carlo. *Statistics and Computing 8*(4), 319–335.

Burnham, K. P. and D. R. Anderson (2002). Model selection and multimodel inference.

Butchart, S. H., M. Walpole, B. Collen, A. Van Strien, J. P. Scharlemann, R. E. Almond, J. E. Baillie, B. Bomhard, C. Brown, J. Bruno, et al. (2010). Global biodiversity: indicators of recent declines. *Science 328*(5982), 1164–1168.

Codling, E. A., M. J. Plank, and S. Benhamou (2008). Random walk models in biology. *Journal of the Royal Society Interface 5*(25), 813–834.

Cohn, J. P. (2008a, 3). Citizen Science: Can Volunteers Do Real Research? *BioScience 58*(3), 192–197.

Cohn, J. P. (2008b). Citizen science: Can volunteers do real research? *BioScience 58*(3), 192–197.

Condorcet, N. D. (1785). Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. *Paris*.

Cowles, M. K. and B. P. Carlin (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association 91*(434), 883–904.

Crall, A. W., G. J. Newman, T. J. Stohlgren, K. A. Holfelder, J. Graham, and D. M. Waller (2011). Assessing citizen science data quality: an invasive species case study. *Conservation Letters 4*(6), 433–442.

Dawid, A. P. and A. M. Skene (1979). Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied Statistics 28*(1), 20.

Department of Conservation and Ministry for the Environment (2000). The new zealand biodiversity strategy.

Dhondt, A. A., D. L. Tessaglia, and R. L. Slothower (1998, 4). Epidemic mycoplasmal conjunctivitis in house finches from eastern North America. *Journal of Wildlife Diseases 34*(2), 265–280.

Diamond, J. M. (1990). New zealand as an archipelago: an international perspective. *Ecological restoration of New Zealand islands 2*, 3–8.

Dickinson, J. L., J. Shirk, D. Bonter, R. Bonney, R. L. Crain, J. Martin, T. Phillips, and K. Purcell (2012). The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment 10*(6), 291–297.

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

Droege, S. (2007). Just because you paid them doesn't mean their data are better. In *Citizen Science Toolkit Conference. Cornell Laboratory of Ornithology*, pp. 13–26.

Eveleigh, A., C. Jennett, A. Blandford, P. Brohan, and A. L. Cox (2014). Designing for dabblers and deterring drop-outs in citizen science. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2985–2994.

Eveleigh, A., C. Jennett, S. Lynn, and A. L. Cox (2013). "i want to be a captain! i want to be a captain!" gamification in the old weather citizen science project. In *Proceedings of the first international conference on gameful design, research, and applications*, pp. 79–82.

Fitzpatrick, M. C., E. L. Preisser, A. M. Ellison, and J. S. Elkinton (2009). Observer bias and the detection of low-density populations. *Ecological applications 19*(7), 1673–1679.

Follett, R. and V. Strezov (2015). An analysis of citizen science based research: usage and publication patterns. *PloS one 10*(11), e0143687.

Foody, G., L. See, S. Fritz, I. Moorthy, C. Perger, C. Schill, and D. Boyd (2018). Increasing the Accuracy of Crowdsourced Information on Land Cover via a Voting Procedure Weighted by Information Inferred from the Contributed Data. *ISPRS International Journal of Geo-Information 7*(3).

Foody, G. M., L. See, S. Fritz, M. Van der Velde, C. Perger, C. Schill, and D. S. Boyd (2013, 12). Assessing the Accuracy of Volunteered Geographic Information arising from Multiple Contributors to an Internet Based Collaborative Project. *Transactions in GIS 17*(6), 847–860.

Franzoni, C. and H. Sauermann (2014, 2). Crowd science: The organization of scientific research in open collaborative projects. *Research Policy 43*(1), 1–20.

Galloway, A. W., M. T. Tudor, and W. M. V. HAEGEN (2006). The reliability of citizen science: a case study of oregon white oak stand surveys. *Wildlife Society Bulletin 34*(5), 1425–1429.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*(6), 564–584.

Gengler, S. and P. Bogaert (2016). Integrating crowdsourced data with a land cover product: A Bayesian data fusion approach. *Remote Sensing*.

Germano, J., S. Barlow, I. Castro, R. Colbourne, M. Cox, C. Gillies, K. Hackwell, J. Harawira, M. Impey, A. Reuben, et al. (2018). Kiwi recovery plan 2018–2028 mahere whakaora kiwi 2018–2028.

Gilks, W. R., N. G. Best, and K. Tan (1995). Adaptive rejection metropolis sampling within gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 44*(4), 455–472.

Gura, T. (2013). Citizen science: Amateur experts. *Nature 496*(7444), 259–261.

Hickling, R., D. B. Roy, J. K. HILL, R. FOX, and C. D. THOMAS (2006, 3). The distributions of a wide range of taxonomic groups are expanding polewards. *Global Change Biology 12*(3), 450–455.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hill, N. and D.-P. Häder (1997). A biased random walk model for the trajectories of swimming micro-organisms. *Journal of theoretical biology 186*(4), 503–526.

Hochachka, W. M., R. Caruana, D. Fink, A. Munson, M. Riedewald, D. Sorokina, and S. Kelling (2007). Data-mining discovery of pattern and process in ecological systems. *The Journal of Wildlife Management 71*(7), 2427–2437.

Hsing, P.-Y., S. Bradley, V. T. Kent, R. A. Hill, G. C. Smith, M. J. Whittingham, J. Cokill, D. Crawley, M. volunteers, and P. A. Stephens (2018). Economical crowdsourcing for camera trap image classification. *Remote Sensing in Ecology and Conservation 4*(4), 361–374.

Hudson, L. N., T. Newbold, S. Contu, S. L. Hill, I. Lysenko, A. De Palma, H. R. Phillips, R. A. Senior, D. J. Bennett, H. Booth, et al. (2014). The predicts database: a global database of how local terrestrial biodiversity responds to human impacts. *Ecology and evolution 4*(24), 4701–4735.

Iacovides, I., C. Jennett, C. Cornish-Trestrail, and A. L. Cox (2013). Do games attract or sustain engagement in citizen science? a study of volunteer motivations. In *CHI'13 extended abstracts on human factors in computing systems*, pp. 1101–1106.

Isaac, N. J., A. J. van Strien, T. A. August, M. P. de Zeeuw, and D. B. Roy (2014). Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution 5*(10), 1052–1060.

James, A., J. C. McLeod, C. Rouco, K. S. Richardson, and D. M. Tompkins (2017). Spatial utilization predicts animal social contact networks are not scale-free. *Royal Society open science 4*(12), 171209.

Katsagounos, I., D. D. Thomakos, K. Litsiou, and K. Nikolopoulos (2021). Superforecasting reality check: Evidence from a small pool of experts and expedited identification. *European journal of operational research 289*(1), 107–117.

Kery, M., R. M. Dorazio, L. Soldaat, A. Van Strien, A. Zuiderwijk, and J. A. Royle (2009). Trend estimation in populations with imperfect detection. *Journal of Applied Ecology 46*(6), 1163–1172.

Khatib, F., F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, et al. (2011). Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology 18*(10), 1175–1177.

Kim, H.-C. and Z. Ghahramani (2012). Bayesian Classifier Combination. *Int. Conf. on Artificial Intelligence and Statistics*, 619–627.

Kosmala, M., A. Wiggins, A. Swanson, and B. Simmons (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment 14*(10), 551–560.

Kruger, J. and D. Dunning (1999). Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of Personality and Social Psychology 77*(6), 121–1134.

Kurvers, R. H., S. M. Herzog, R. Hertwig, J. Krause, M. Moussaid, G. Argenziano, I. Zalaudek, P. A. Carney, and M. Wolf (2019). How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science advances 5*(11), eaaw9011.

Lagrange, J. L. (1804). *Leçons sur le calcul des fonctions*, Volume 5. Imperiale.

Lebel, T., K. Syme, M. Barrett, and J. A. Cooper (2020). Two new species of asproinocybe (tricholomataceae) from australasia. *Muelleria 38*, 77–85.

Lloyd-Smith, J. O., S. J. Schreiber, P. E. Kopp, and W. M. Getz (2005). Superspreading and the effect of individual variation on disease emergence. *Nature 438*(7066), 355–359.

Mace, G. M. and R. Lande (1991). Assessing extinction threats: toward a reevaluation of iucn threatened species categories. *Conservation biology 5*(2), 148–157.

Martin, L. J., B. Blossey, and E. Ellis (2012). Mapping where ecologists work: biases in the global distribution of terrestrial ecological observations. *Frontiers in Ecology and the Environment 10*(4), 195–201.

Martino, L., V. Elvira, and G. Camps-Valls (2018). The recycling gibbs sampler for efficient learning. *Digital Signal Processing 74*, 1–13.

Matabos, M., M. Hoeberechts, C. Doya, J. Aguzzi, J. Nephin, T. E. Reimchen, S. Leaver, R. M. Marx, A. Branzan Albu, R. Fier, et al. (2017). Expert, crowd, students or algorithm: who holds the key to deep-sea imagery 'big data'processing? *Methods in Ecology and Evolution 8*(8), 996–1004.

McGeoch, M. A., S. H. Butchart, D. Spear, E. Marais, E. J. Kleynhans, A. Symes, J. Chanson, and M. Hoffmann (2010). Global indicators of biological invasion: species numbers, biodiversity impact and policy responses. *Diversity and Distributions 16*(1), 95–108.

McKinley, D. C., A. J. Miller-Rushing, H. L. Ballard, R. Bonney, H. Brown, S. C. Cook-Patton, D. M. Evans, R. A. French, J. K. Parrish, T. B. Phillips, et al. (2017). Citizen science can improve conservation science, natural resource management, and environmental protection. *Biological Conservation 208*, 15–28.

Miller-Rushing, A., R. Primack, and R. Bonney (2012). The history of public participation in ecological research. *Frontiers in Ecology and the Environment 10*(6), 285–290.

Miller-Rushing, A. J. (2017). Reflections on the strong growth of citizen science: An interview with abe miller-rushing. *Maine Policy Review 26*(2), 92–93.

Mobley, J., S. Spitz, R. Grotefendt, P. Forestell, A. Frankel, and G. Bauer (2001). Abundance of humpback whales in hawaiian waters: Results of 1993-2000 aerial surveys. *Report to the Hawaiian Islands Humpback Whale National Marine Sanctuary 9*.

Mugford, J., E. Moltchanova, M. Plank, J. Sullivan, A. Byrom, and A. James (2021). Citizen science decisions: A bayesian approach optimises effort. *Ecological Informatics*, 101313.

New Zealand bio-recording network trust (2020). inaturalist nz–mātaki taiao annual report.

Newman, G., A. Wiggins, A. Crall, E. Graham, S. Newman, and K. Crowston (2012). The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment 10*(6), 298–304.

Owen, G., B. Grofman, and S. L. Feld (1989, 2). Proving a distribution-free generalization of the Condorcet Jury Theorem. *Mathematical Social Sciences 17*(1), 1–16.

Patlak, C. S. (1953). Random walk with persistence and external bias. *The bulletin of mathematical biophysics 15*(3), 311–338.

Pearce, J. L. and M. S. Boyce (2006). Modelling distribution and abundance with presence-only data. *Journal of applied ecology 43*(3), 405–412.

Pitchford, J. W. and J. Brindley (2001). Prey patchiness, predator survival and fish recruitment. *Bulletin of Mathematical Biology 63*(3), 527–546.

Pitchford, J. W., A. James, and J. Brindley (2003). Optimal foraging in patchy turbulent environments. *Marine Ecology Progress Series 256*, 99–110.

Pitchford, J. W., A. James, and J. Brindley (2005). Quantifying the effects of individual and environmental variability in fish recruitment. *Fisheries Oceanography 14*(2), 156–160.

Pocock, M. J. O., J. C. Tweddle, J. Savage, L. D. Robinson, and H. E. Roy (2017, 4). The diversity and evolution of ecological and environmental citizen science. *PLOS ONE 12*(4), e0172579.

Porter, J. H., E. Nagy, T. K. Kratz, P. Hanson, S. L. Collins, and P. Arzberger (2009). New eyes on the world: advanced sensors for ecology. *BioScience 59*(5), 385–397.

Prelec, D., H. S. Seung, and J. McCoy (2017, 1). A solution to the single-question crowd wisdom problem. *Nature 541*(7638), 532–535.

Purser, A., M. Bergmann, T. Lundälv, J. Ontrup, and T. W. Nattkemper (2009). Use of machine-learning algorithms for the automated detection of cold-water coral habitats: a pilot study. *Marine Ecology Progress Series 397*, 241–251.

Raddick, M. J., G. Bracey, P. L. Gay, C. J. Lintott, P. Murray, K. Schawinski, A. S. Szalay, and J. Vandenberg (2009). Galaxy zoo: Exploring the motivations of citizen science volunteers. *arXiv preprint arXiv:0909.2925*.

Reed, J., M. J. Raddick, A. Lardner, and K. Carney (2013). An exploratory factor analysis of motivations for participating in zooniverse, a collection of virtual citizen science projects. In *2013 46th Hawaii International Conference on System Sciences*, pp. 610–619. IEEE.

Rotman, D., J. Preece, J. Hammock, K. Procita, D. Hansen, C. Parr, D. Lewis, and D. Jacobs (2012). Dynamic changes in motivation in collaborative citizen-science projects. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pp. 217–226.

Russell, J. C., J. G. Innes, P. H. Brown, and A. E. Byrom (2015). Predator-free new zealand: conservation country. *BioScience 65*(5), 520–525.

Sauermann, H. and C. Franzoni (2015a, 1). Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences of the United States of America 112*(3), 679–84.

Sauermann, H. and C. Franzoni (2015b). Crowd science user contribution patterns and their implications. *Proceedings of the national academy of sciences 112*(3), 679–684.

Schmeller, D. S., P.-Y. HENRY, R. Julliard, B. Gruber, J. Clobert, F. Dziock, S. Lengyel, P. Nowicki, E. Deri, E. Budrys, et al. (2009). Advantages of volunteer-based biodiversity monitoring in europe. *Conservation biology 23*(2), 307–316.

Schoening, T., M. Bergmann, J. Ontrup, J. Taylor, J. Dannheim, J. Gutt, A. Purser, and T. W. Nattkemper (2012). Semi-automated image analysis for the assessment of megafaunal densities at the arctic deep-sea observatory hausgarten. *PloS one 7*(6), e38179.

See, L., A. Comber, C. Salk, S. Fritz, M. van der Velde, C. Perger, C. Schill, I. McCallum, F. Kraxner, and M. Obersteiner (2013, 7). Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts. *PLoS ONE 8*(7), e69958.

See, L., S. Fritz, C. Perger, C. Schill, I. McCallum, D. Schepaschenko, M. Duerauer, T. Sturn, M. Karner, F. Kraxner, et al. (2015). Harnessing the power of volunteers, the internet and google earth to collect and validate global spatial information using geo-wiki. *Technological Forecasting and Social Change 98*, 324–335.

Siddharthan, A., C. Lambin, A.-M. Robinson, N. Sharma, R. Comont, E. O'mahony, C. Mellish, and R. V. D. Wal (2016). Crowdsourcing without a crowd: Reliable online species identification using bayesian models to minimize crowd size. *ACM Transactions on Intelligent Systems and Technology (TIST) 7*(4), 1–20.

Silvertown, J. (2009, 9). A new dawn for citizen science. *Trends in Ecology & Evolution 24*(9), 467–471.

Simpson, R., K. R. Page, and D. De Roure (2014). Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd international conference on world wide web*, pp. 1049–1054. ACM.

Siniff, D. B. and C. Jessen (1969). A simulation model of animal movement patterns. In *Advances in ecological research*, Volume 6, pp. 185–219. Elsevier.

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling (2009). ebird: A citizen-based bird observation network in the biological sciences. *Biological conservation 142*(10), 2282–2292.

Szabo, J. K., R. A. Fuller, and H. P. Possingham (2012). A comparison of estimates of relative abundance from a weakly structured mass-participation bird atlas survey and a robustly designed monitoring scheme. *Ibis 154*(3), 468–479.

Szabo, J. K., P. A. Vesk, P. W. Baxter, and H. P. Possingham (2010). Regional avian species declines estimated from volunteer-collected long-term data using list length analysis. *Ecological Applications 20*(8), 2157–2169.

Tinati, R., M. Luczak-Roesch, E. Simperl, and W. Hall (2017). An investigation of player motivations in eyewire, a gamified citizen science project. *Computers in Human Behavior 73*, 527–540.

Tonachella, N., A. Nastasi, G. Kaufman, D. Maldini, and R. W. Rankin (2012). Predicting trends in humpback whale (megaptera novaeangliae) abundance using citizen science. *Pacific Conservation Biology 18*(4), 297–309.

Toomey, A., L. Strehlau-Howay, B. Manzolillo, and C. Thomas (2020). The place-making potential of citizen science: Creating social-ecological connections in an urbanized world. *Landscape and Urban Planning 200*, 103824.

Treynor, J. L. (1987). Market Efficiency and the Bean Jar Experiment. *Financial Analysts Journal 43*(3), 50–53.

Tulloch, A. I., H. P. Possingham, L. N. Joseph, J. Szabo, and T. G. Martin (2013, 9). Realising the full potential of citizen science monitoring programs. *Biological Conservation 165*, 128–138.

Turchin, P. (1998). *Quantitative analysis of movement: measuring and modeling population redistribution in animals and plants*. Sinauer Associates.

Weiss, G. H. and G. H. Weiss (1994). *Aspects and applications of the random walk*. Elsevier Science & Technology.

Wiggins, A. (2011). ebirding: technology adoption and the transformation of leisure into science. In *Proceedings of the 2011 iConference*, pp. 798–799.

Wiggins, A. and K. Crowston (2011). From conservation to crowdsourcing: A typology of citizen science. In *2011 44th Hawaii international conference on system sciences*, pp. 1–10. IEEE.

Wiggins, A. and Y. He (2016). Community-based data validation practices in citizen science. In *Proceedings of the 19th ACM Conference on computer-supported cooperative work & social computing*, pp. 1548–1559.

Wiggins, A., G. Newman, R. D. Stevenson, and K. Crowston (2011, 12). Mechanisms for Data Quality and Validation in Citizen Science. In *2011 IEEE Seventh International Conference on e-Science Workshops*, pp. 14–19. IEEE.

Zheng, Y., G. Li, Y. Li, C. Shan, and R. Cheng (2017, 1). Truth inference in crowdsourcing. *Proceedings of the VLDB Endowment 10*(5), 541–552.

Zuckerberg, B. and W. F. Porter (2010). Thresholds in the long-term responses of breeding birds to forest cover and fragmentation. *Biological Conservation*.

Zuur, A., E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith (2009). *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media.