# Examining spatial biases in the community science platform, iNaturalist, using British Columbia, Canada, as a case study

by

Ellyne M. Geurts

B.Sc. (Hons.), University of Manitoba, 2019

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree of

MASTER OF SCIENCE

in the School of Environmental Studies

© Ellyne M. Geurts, 2023

University of Victoria

All rights reserved. This thesis may not be reproduced, in whole or in part, by photocopy or other means, without the permission of the author.

*We acknowledge and respect the lək̓ʷəŋən peoples on whose traditional territory the university stands and the Songhees, Esquimalt and W̱SÁNEĆ peoples whose historical relationships with the land continue to this day.*

**Supervisory Committee**

Examining spatial biases in the community science platform, iNaturalist, using British Columbia, Canada, as a case study

By

Ellyne M. Geurts

B.Sc. (Hons.), University of Manitoba, 2019

**Supervisory Committee**

Dr. Brian Starzomski, Supervisor

School of Environmental Studies

Dr. John Reynolds, Committee Member

School of Environmental Studies

**Abstract**

The ever-growing interest in community science platforms like iNaturalist and eBird is ushering in a new era of biodiversity and ecology research where researchers are overflowing with data across large geographical and temporal scales. However, these big and often unstructured data come with a cost, biases. These biases include temporal, spatial, and taxonomic biases in opportunistically collected community science datasets like the popular biodiversity platform, iNaturalist. There is a need to improve our knowledge of the biases on these platforms, so that the data can be used effectively. My thesis tackles this gap by examining spatial biases on the iNaturalist platform. My first study uses Maxent to model broad-scale spatial bias in iNaturalist observations in British Columbia, Canada. I ask: *Where are iNaturalist users primarily observing?* and *What landscape features best explain the spatial bias?* I find that distance to roads is the most important landscape variable explaining spatial bias. In my second chapter, I experimentally tested whether fine-scale spatial biases of trails affected taxonomic richness estimates on iNaturalist using paired timed transects with a team of iNaturalist observers. I found greater taxonomic richness on trails compared to away from trails and no difference in rare species observations between on and off trails, suggesting there is no loss of information by primarily surveying along trails. Overall, this research shows important variables to include to control for spatial bias when using iNaturalist data and provides reassuring evidence that fine-scale bias does not impede biodiversity surveying from community scientists.

## Table of contents

## List of tables

List of figures

**Acknowledgements**

I am grateful for this wonderful opportunity of being able to learn and study at the School of Environment Studies and experience the beautiful and diverse biodiversity and ecosystems of British Columbia. As someone who grew up in the prairies, I cherish this chance to conduct research in tidal pools, coastal rainforests, northern alpine, and sagebrush habitats, and many more.

Huge thank you to the BC Parks iNaturalist team: Lena Dietz Chiasson, Jason Headley, Kate McKeown, Tori Miller, and Erin Springinotic for their tireless work in helping with my experiments and for sharing their knowledge of lichens, birds, herps, and critters.

Thank you to the amazingly welcoming iNaturalist community for kindly correcting my identifications, sharing their tips and tricks for observations, and volunteering their time identifying the project's observations. I began using and interacting with the iNaturalist community as part of this research, and now it has become an integral part of who I am, and what I do on my free time. My time spent learning and being part of the iNaturalist community has been a blast and I am incredibly grateful this experience.

Thank you John and Brian for taking me under your wings and helping to guide and wrangle the many ideas I had for this project. Thank you for lending me cameras and insect collecting gear and letting me go feral learning about the diversity of British Columbia. I am a bit sad that this chapter of my life is ending. Thank you for spreading your iNaturalist passion bug to me!

In addition, thank you to the community at the School of Environmental Studies for expanding my worldviews and perspectives and being supportive of my research.

This work would not have been possible without the support from the Ministry of Water, Land and Resource Stewardship, BC Parks, the Sitka Foundation, The Pacific Wildlife Foundation. NSERC, and School of Environmental Studies. Also, thank you UVic Faculty of Graduate Studies and CUPE 4163 for financial assistance and enabling me to attend conferences and form bonds with folks that help guided my research and future path.

I am deeply indebted to my labmates, peers, friends, and family for all of their support. Huge thank you to the folks from the Starzomski and Reynolds labs – Gabe, Geneviève, Julie, Nicole, Kyle, Kalina, Nathan, Dan, Allison, Deb, Celeste, Nico, Sean, and Jane, for their insight, moral support, and stats advice! Thank you to my roommates – Cole, Aaron, Laura, Macgregor, and Savita for their encouragement and brainstorming sessions. Lastly, thank you to my friends, peers, and family who supported me and graciously listened to me talking about iNaturalist and waited for me on trails while I snapped that picture.

# Chapter 1: General introduction

## *1.1 Brief history and evolution of biodiversity community science*

### 1.1.1 What is community science?

Community science – also known as citizen science – is the partnership between professional scientists and the public to engage in scientific research of the natural world (Miller-Rushing et al. 2012). Community science is used in many different fields such as astronomy (Ponciano et al. 2014; Marshall et al. 2015), environment and health impacts (Walker et al. 2021; Mahajan et al. 2022), agriculture (Ryan et al. 2018), and ecology (Miller-Rushing et al. 2012; Kobori et al. 2016). Biological fields like ecology and conservation may benefit greatly from information from community science (Kullenberg & Kasperowski 2016), and will be the context I use when discussing community science hereafter.

Community science takes on many different forms. It includes opportunistically collected data via crowdsourcing as in iNaturalist, semi-structured projects with sampling effort and presence and absence data recorded as in eBird, and strictly structured projects with trained volunteers and the timing and location of surveyed controlled as performed in the North American Breeding Bird Survey (Dickinson et al. 2010). Community science projects are categorized by their goals such as targeted monitoring (*i.e.,* hypothesis driven) versus surveillance monitoring (*i.e.*, no specific hypothesis) (Nichols & Williams 2006; Dickinson et al. 2010). Community science projects are also sorted by the level of involvement the public has during the scientific research process; with up to five different levels proposed (Bonney et al. 2009; Shirk et al. 2012). The levels include *Contractual*, *Contributory*, *Collaborative*, *Co-Created*,

and *Collegial* projects, with *Contractual* projects having the least amount of community engagement and *Collegial* projects have the highest amount, that is they are completely community-run (Shirk et al. 2012; Ries & Oberhauser 2015). Out of these five project types, *Contributory* projects – community members primarily contributing data to designed scientific projects – are the most common form and the one most often assumed when discussing community science (Shirk et al. 2012; Ries & Oberhauser 2015; Walker et al. 2021).

### 1.1.2 Brief history of community science

Despite the growing use of community science in research this past decade, it is important to note that community science is not a recent phenomenon: it has been occurring for centuries around the world from passionate, curiosity-driven people interested in documenting the natural world around them (Greenwood 2007; Miller-Rushing et al. 2012). Community science conducted by amateurs (*i.e.,* non-professional scientists) was the mainstay for scientific research until the professionalization of science in the late 19$^{th}$ century became the norm (Vetter 2011; Miller-Rushing et al. 2012). In North America, some of the earliest records of community science are in ornithology. For example, the North American Bird Phenology Program began in 1881 (and later dissolved in 1970) and the Christmas Bird Count began in 1900 (Fig. 1.1; Dickinson et al. 2010). In a global context, the first known large scale collaborative bird community science project was in Finland (Fig. 1.1; Greenwood 2007). As technology advanced and made cameras, smartphones, internet, data storage capacity, and big data analysis software widely available, the number of community science platforms and projects grew rapidly beyond birds to other taxa such as butterflies, amphibians, fish, and plants (Dickinson et al. 2010; Prudic et al. 2018). The combination of technology advancement

and the professionalization of science has led to a shift in the dominant type of community science, from more *Co-Created* projects to more *Contributory* projects in recent years (Miller-Rushing et al. 2012).

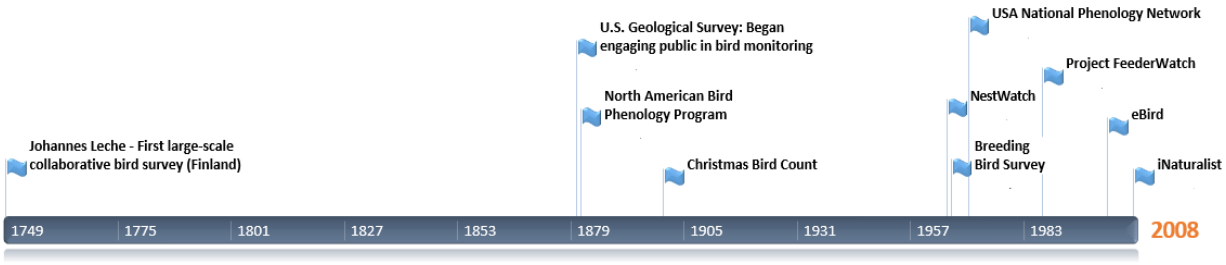



Figure 1.1 Timeline of when notable ornithological community science projects were formed with a primary focus on North American initiatives.

*1.2 Benefits of community science*

Community science projects are producing enormous amounts of biodiversity and ecological data on spatial and temporal scales that are not feasible with traditional scientific methods (Theobald et al. 2015; Pocock et al. 2018; Zhang 2020; Loarie 2022a). These data contribute to answering questions of phenology (Barve et al. 2020a; Nowak et al. 2020), species distribution changes (Johnston et al. 2021), species response to climate change (Cooper et al. 2014; Crimmins & Crimmins 2022), phenotypic variation (Drury et al. 2019; Lehtinen et al. 2020), species interactions (Saldivar et al. 2022), population trends (Neate-Clegg et al. 2020), species richness (Roberts et al. 2022), species health (Hamilton et al. 2021), functional traits (Wolf et al. 2022) and more. These data are also aiding in the discovery of new species (Jain et al. 2022) and monitoring and managing exotic and disease vector species (Werenkraut et al. 2020; Cull 2021; Hausdorf et al. 2021). These projects also play a vital role in conservation by engaging the public and helping shape local governmental environmental policies (Loss et al. 2015; MacPhail

et al. 2020). Community science projects contribute to democratizing science (Dickinson et al. 2012; Kullenberg & Kasperowski 2016; de Sherbinin et al. 2021), increasing ecological literacy among the public, and inspiring local environmental stewardship (Reynolds & Lowman 2013; Loss et al. 2015; MacPhail et al. 2020).

Community science projects are diverse in their strengths and benefits for scientific research and conservation management. Data from structured sampling design community science projects like the North American Breeding Bird Survey can produce accurate population trend data (Smith et al. 2020). Data derived from semi-structured projects like eBird, where users contribute complete checklists (*i.e.,* all species and their abundances recorded) have also been useful for estimating status and trends such as species range estimations, abundances, and phenology (*e.g.,* spring arrivals of migratory birds). For data from opportunistic collection projects, their strength lies in the massive amount of species occurrence and ecological data they produce from volunteered contributed verifiable records. Complete checklists provide invaluable abundance and presence/absence data but often lack associated verifiable records such as collected specimens, photos, or audio for further study and confirmation. On the other hand, opportunistic collections often do not provide abundance metrics or true species absence data. They may instead provide verifiable records that allow for correction of species identification, and secondary data for further ecological study across large spatial and temporal scales. Examples of secondary data from geotagged and timestamps photos include studying flowering phenology of *Yucca* sp. (Barve et al. 2020a), moult timing of mountain goats (*Oreamnos americanus*) (Nowak et al. 2020), geographical colour morphism in odonates (Drury et al. 2019), and documenting food plants that painted lady butterflies visit (Nymphalidae)

(Saldivar et al. 2022). Of the many opportunistic community science platforms available for biodiversity study, iNaturalist is one of the most popular and has been showing great potential in the scientific world (Mesaglio & Callaghan 2021).

*1.3 iNaturalist*

iNaturalist is a global social network and community science platform that collects crowdsourced species occurrence and biodiversity data of all taxa. iNaturalist was created in 2008 and is maintained as a joint initiative between the California Academy of Science and National Geographic Society. There are currently more than 125 million verifiable observations of 415,000 species from the efforts of 2.5 million observers around the world, and the platform continues to grow rapidly (Loarie 2022a).

The core of iNaturalist is to connect people with nature and help them learn about their local environment (Loarie 2022b). Collecting biodiversity and ecological data is a secondary goal of the platform. As a result, iNaturalist is open to anyone to participate, from the general public to hobbyist naturalists to professional scientists and biologists.

iNaturalist operates by someone uploading a photo(s) or audio recording of a wild organism they encounter via the desktop website or mobile app with the location, date, and time observed. The observer can attach a tentative identification to the observation with the help of suggestions from computer vision software that iNaturalist features. After uploading, the iNaturalist community can provide crowdsourced species identification for the observation, with the option to add annotations such as life stage and sex to increase information. Once the observation reaches majority agreement on a species-level identification by the community,

the observation is upgraded from “Needs-ID” to “Research-Grade” by iNaturalist. If the “Research-Grade” observation has a “CC-BY-NC”, “CC-BY”, or “CC0” license, it is then exported to the Global Biodiversity Information Facility (GBIF.org) for archiving.

The data collection protocol on iNaturalist is simple, in order to maximize engagement with the public. iNaturalist is not a hypothesis-driven community science platform and has no structured sampling guidelines for their users. There are no field for reporting sampling effort (*e.g.,* duration spent searching, distance traveled, and size of party surveying), and it does not allow for self-reporting of identification skill level. As a result, iNaturalist data are considered opportunistic (*i.e.,* presence-only data) and may contain sampling biases that interfere with effective data usage (Brown & Williams 2019; Di Cecco et al. 2021).

*1.4 Shortcomings of iNaturalist – sampling biases*

Sampling biases are an issue in all biological datasets from crowdsourced community science platforms to museum and herbaria collections (Meyer et al. 2016; Troudet et al. 2017; Cornwell et al. 2019; Shirey et al. 2021). Professional datasets are not exempt from biases, though they typically exhibit biases to a lesser degree (Kosmala et al. 2016). Sampling biases are expressed in many ways such as spatial, taxonomic, and temporal biases. Community science data are often spatially biased towards points of access like roads and footpaths, high human population areas, and nature preserves (Rocchini et al. 2011; Mair & Ruete 2016; El-Gabbas & Dormann 2018; Shirey et al. 2021). Community science records, in particular observations from opportunistic and semi-structured projects, are temporally biased towards weekends (Courter et al. 2013; Di Cecco et al. 2021), national holidays (Surmacki 2005), and organized bioblitzes

such as the “City Nature Challenge” hosted by iNaturalist (Surmacki 2005; Di Cecco et al. 2021). In terms of taxonomic bias, the highest recorded taxa tend to be eye-catching (*e.g.,* bright colours), relatively abundant, easy to photograph, and particularly interesting (*e.g.,* rare or charismatic species) (Troudet et al. 2017; Callaghan et al. 2021; Di Cecco et al. 2021; Stoudt et al. 2022). Taxonomic biases may result from over-reporting of rare large species and under-reporting of common species (Dickinson et al. 2010). It is also important to note that the strength of bias can vary among observers within the same community science project. For instance, more casual participants tend to exhibit stronger spatial bias towards developed regions, *i.e.,* where the observers live and work, than more active participants (Geldmann et al. 2016; Di Cecco et al. 2021).

Barring severe biases, biases are not a barrier to using these data, but they do mean additional attention and standardizing of the data is required before use. There are a few methods to deal with biases such as filtering data to reduce temporal and spatial autocorrelation in the data (Meyer et al. 2016; Van Eupen et al. 2022), and explicitly accounting for known biases within models (Courter et al. 2013; Fithian et al. 2015; Johnston et al. 2018). However, to use these methods for effective use of community science data, we need to know the extent and strength of the biases, and what drives them.

Some community science projects, like eBird and the North American Breeding Bird Survey, have a good understanding of the limitations and biases of their datasets (Betts et al. 2007; Harris & Haskell 2007; van Wilgenburg et al. 2015; Zhang 2020; Tang et al. 2021; Scher & Clark 2023), but other platforms like iNaturalist have limited analyses of their biases (Arazy &

Malkinson 2021; Di Cecco et al. 2021). This gap in our knowledge reduces the full and effective use of these datasets.

*1.5 Thesis Objectives*

In this thesis, I endeavour to improve our understanding of the biases present in the popular community science platform iNaturalist. I use a combination of field experiments and big data modeling techniques to investigate broad- and fine-scale spatial bias using British Columbia, Canada, as a case study. I examine how different environmental features influence the distribution of iNaturalist observations across the landscape. I then narrow my focus to trail bias and how this fine-scale spatial bias could impact community science biodiversity observations.

In Chapter Three, I investigate broad-scale spatial bias on iNaturalist in British Columbia using Maxent, a species distribution modeling technique for presence-only data. I focus on two objectives: 1) model the spatial biases of iNaturalist observations in British Columbia and predict where observations are likely to occur; and 2) determine the nature of the relationships between environmental variables and probability of observations. I predict that distance to roads and human population density will be the most important environmental variables biasing the distribution of observations, with land cover type such as urban and agricultural regions and tourist locations (*e.g.*, parks) having a lesser effect. I also predict an exponential negative relationship between distance to roads and probability of observation, and a positive linear relationship between human population density and probability of observation. This work will help determine which environmental variables are crucial to account for in broad-

scale spatial bias when modeling with iNaturalist data, and which regions are likely to be under sampled by community scientists.

In Chapter Four, I test experimentally if trail spatial bias affects the number of taxa observed on and off trails by community scientists by using timed field experiments and trained iNaturalist observers. I compare taxonomic richness estimates between on and off trail observations in provincial parks in British Columbia, using generalized linear mixed models. I also investigate whether there is a difference in exotic species richness between on and off trail. I explore differences in number of rare and vulnerable species detected on and off trail. I predict overall higher taxonomic richness estimates on-trail compared to off-trail. In addition, I predict there are more exotic species observed on-trails. This study explores whether trail bias is a significant concern for biodiversity data quality on iNaturalist.

This thesis answers important questions regarding spatial bias on the iNaturalist community science platform. These results may help scientists who use iNaturalist data to account for spatial bias in species distribution and biodiversity models. In addition, my work provides the first study to experimentally examine spatial bias on iNaturalist, and provides valuable information regarding the quality of biodiversity data from iNaturalist in parks for conservation management and research. The study of bias on community science platforms allows for more effective use of these massive datasets to answer novel ecological and conservation questions.

## Chapter 2: Study region - British Columbia

### *2.1 Biodiversity and geography*

British Columbia (BC) is the westernmost province of Canada. It borders the Pacific Ocean and contains several mountain ranges due to its complex geological history. There are over 50,000 described species, 16 biogeoclimatic zones – a broad ecosystem classification system – , and approximately 611 distinct ecological communities within BC (Meidinger and Pojar, 1991; Austin et al., 2008). Examples of different ecosystems can be seen in Figure 2.1. British Columbia also hosts some of the greatest diversity of breeding birds, mammals, butterflies, mosses, and vascular plants in Canada, with a high number of rare and threatened species (Austin et al. 2008; Canadian Endangered Species Conservation Council 2022). See Figure 2.2 for examples of threatened and vulnerable species in British Columbia.

British Columbia has high variability of human population densities across the province. These range from <10 people/$km^2$ to >900 people/$km^2$ (Environmental Reporting BC 2018a). The total area of British Columbia is 947,800 $km^2$, with total land and freshwater areas 929,730 $km^2$ and 18,070 $km^2$ respectively (Government of Canada 2017). There are approximately 719,000 kilometers of roads, both paved and unpaved (Fig. 2.3 b-d), with 34% of the land area accessible by roads (Environmental Reporting BC 2018b). Accessible is defined here as the area within 500 m of roads. Percent area accessible by roads (*i.e.,* road network density) varies across ecoregions from 0.2% to 87% (Environmental Reporting BC 2018b). In addition, British Columbia has an extensive recreation trail network across the province, with over 30,000 kilometers of formally recognized and managed trails and hundreds of thousands of kilometers of unmaintained trails (The Ministry of Forests, Lands and Natural Resource Operations 2013).

British Columbia benefits greatly from nature preservation and tourism, and currently has a total of 1,037 provincial parks, protected areas, ecological reserves, and conservancies, and seven national parks (BC Parks 2020). British Columbia also has a resource-intensive economy, supporting forestry, mining, agriculture, and fisheries industries. Human impacts from population growth, road infrastructure, and industry disproportionately affects the different biogeoclimatic zones in British Columbia (Shackelford et al. 2018).

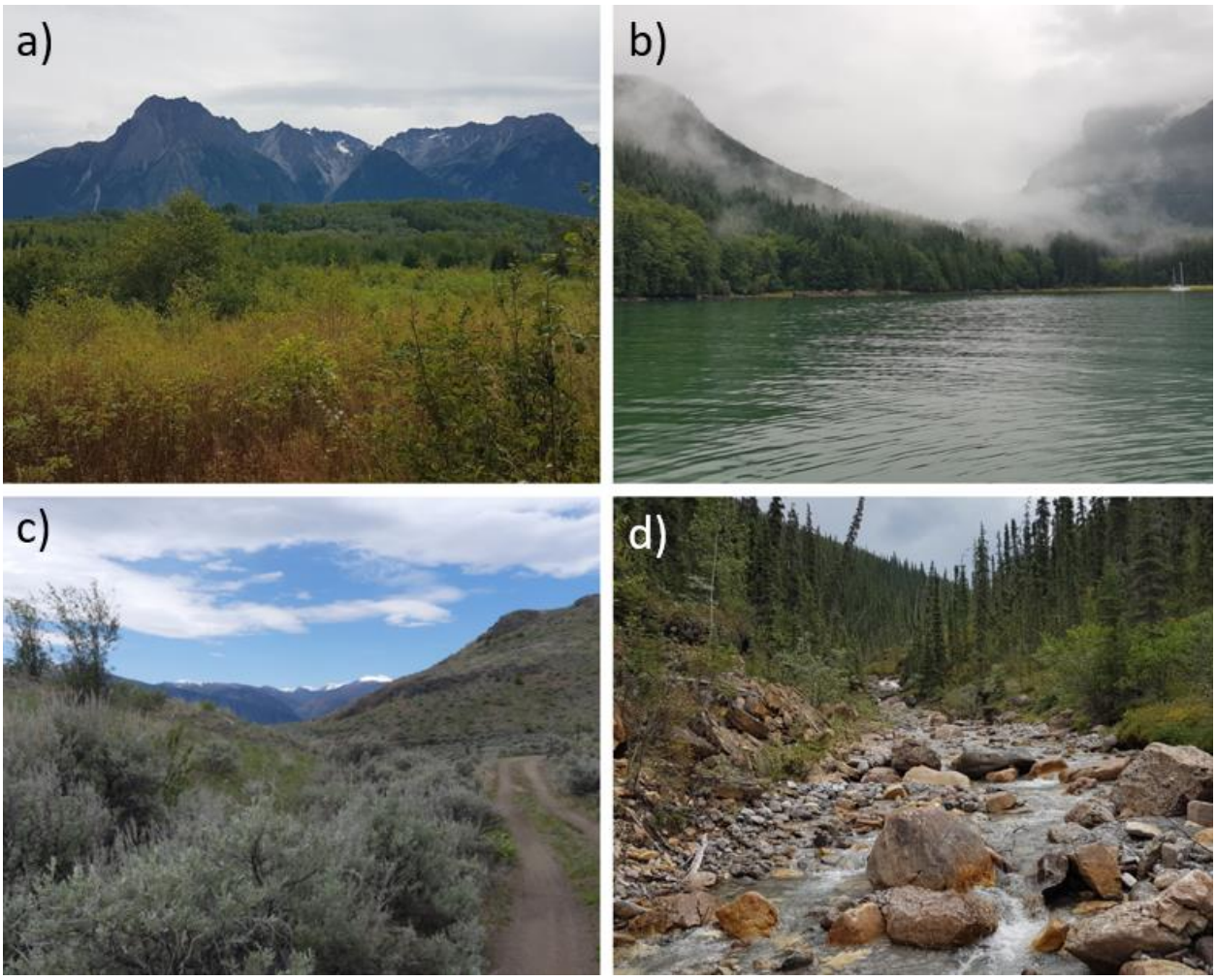



Figure 2.1 Examples of the diverse ecosystems present in British Columbia, Canada. a) Anderson Flats Provincial Park. b) Shearwater Hot Springs Conservancy. c) South Okanagan Grasslands Protected Area. d) Muncho Lake Provincial Park. Photos taken by Ellyne Geurts.

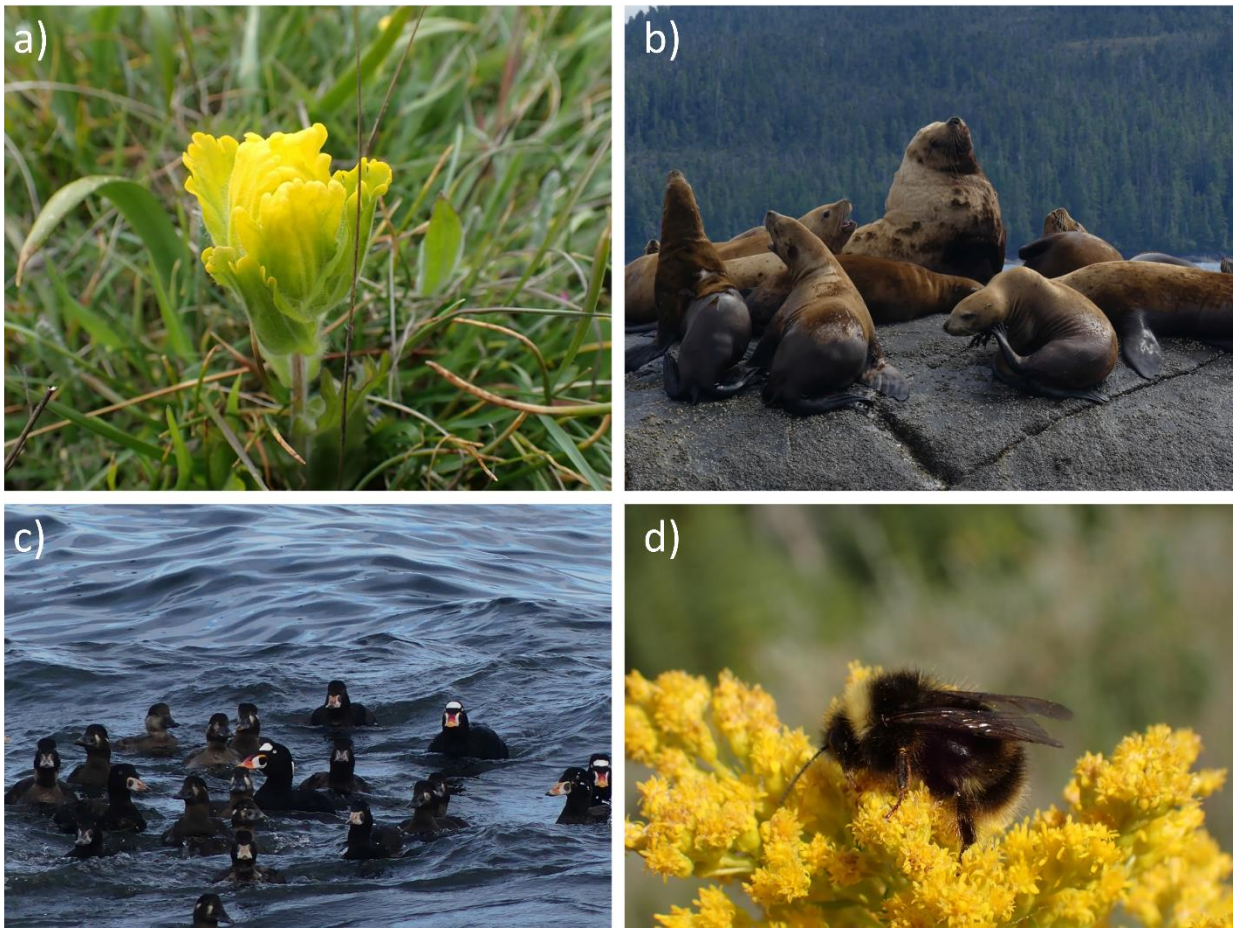



**Figure 2.2** British Columbia, Canada, hosts many rare, vulnerable, and threatened species. Provincial Conservation Status ranks are from NatureServe Explorer (https://explorer.natureserve.org/). **a)** Golden Indian Paintbrush (*Castilleja levisecta)* - Critically Imperiled (S1). **b)** Steller Sea Lion (*Eumetopias jubatus*) - Vulnerable (S3). **c)** Surf Scoter (*Melanitta perspicillata*) - Vulnerable (S3). **d)** Western Bumble Bee (*Bombus occidentalis*) - Vulnerable (S3). Photos taken by Ellyne Geurts.

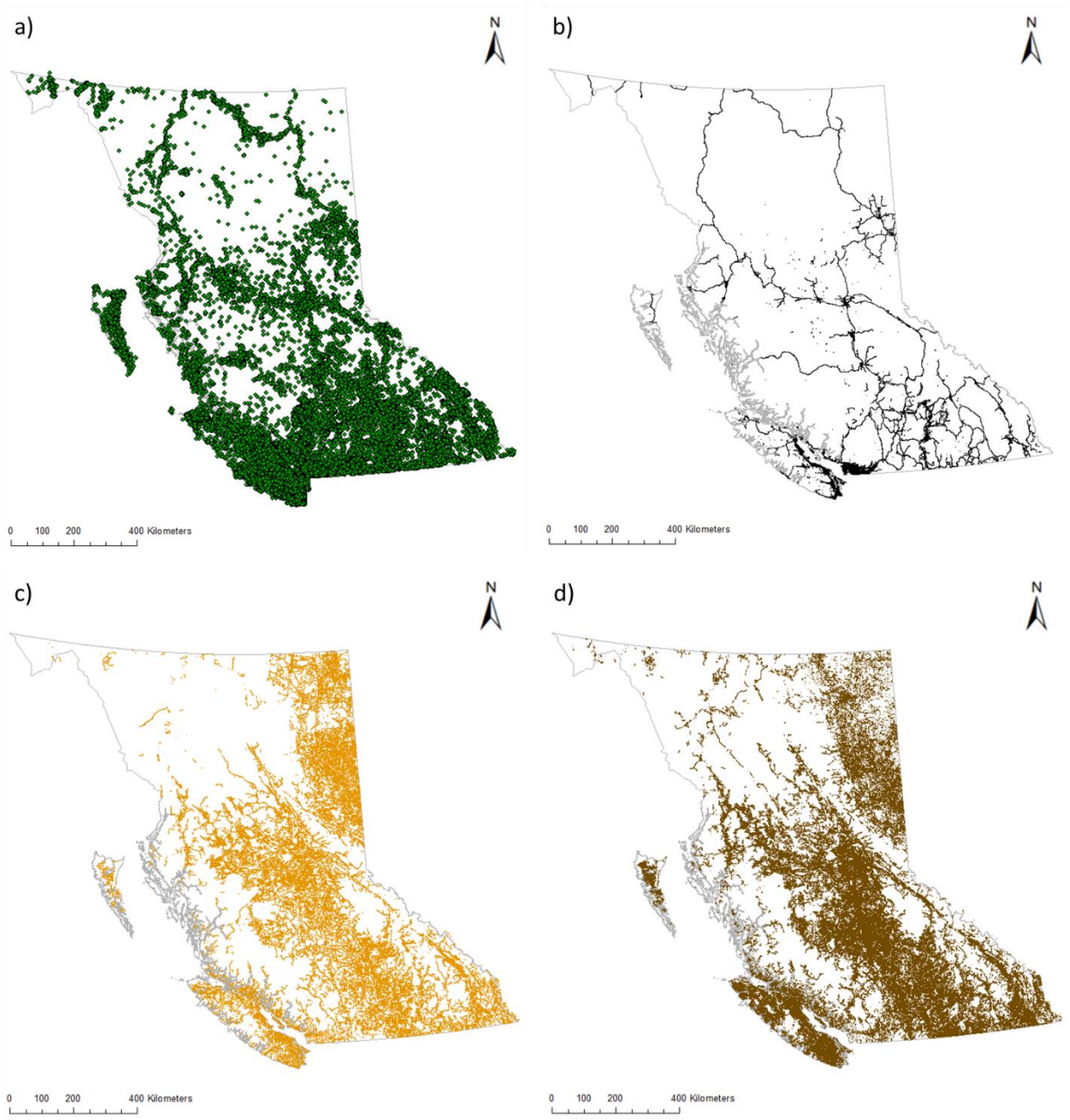



Figure 2.3 iNaturalist occurrence points and roads in British Columbia, Canada. All road data are from November 2020. a) All iNaturalist observations from 2008 to 2020. N = 1,005,653 observations. b) Map of paved roads. c) Map of maintained gravel roads. d) Map of unmaintained gravel and dirt roads. Data sources: iNaturalist and FLNRORD – GeoBC.

*2.2 Community science presence*

British Columbia has a large number of people participating in community science. Many passionate volunteers spend numerous hours contributing biodiversity observations to

community science online platforms such as eBird, Bumblebee Watch, Beetle Watch, BC Cetacean Sightings Network, eButterfly, Frog Watch, and iNaturalist. Many also participate in structured monitoring programs like the Christmas Bird Count, Hawk Watch, North American Breeding Bird Survey, and British Columbia Waterbird Survey. A rapidly growing biodiversity community science platform in British Columbia is iNaturalist (Fig. 2.4). The iNaturalist database in BC currently contains over 2.5 million georeferenced observations across the taxonomic spectrum from protozoans to mammals. These data came from over 43,000 observers from 2008 to 2023, with the number of observations growing exponentially (Fig. 2.4; iNaturalist 2022). iNaturalist is attracting a lot of attention in British Columbia from initiatives by government agencies and non-profit societies encouraging the use of the platform in parks (BC iNaturalist Program 2021; Parks Canada 2022; Strathcona Wilderness Institute 2022).

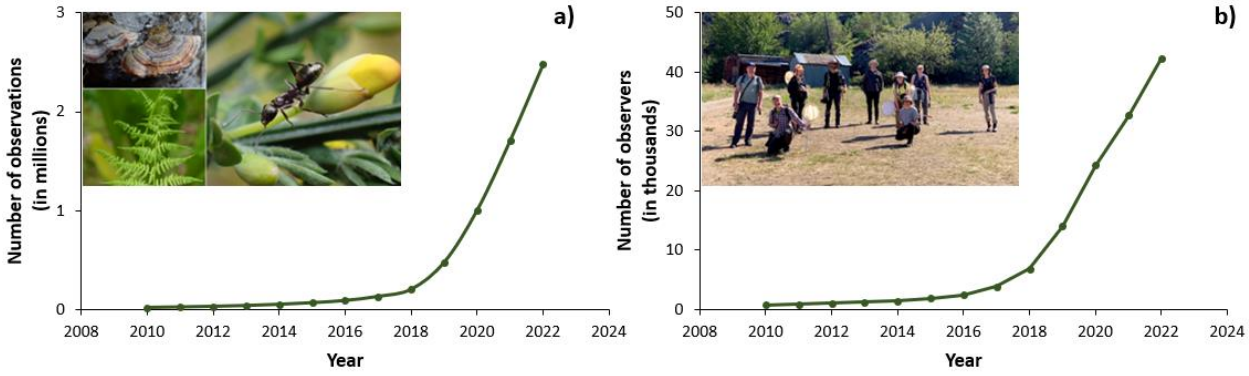



Figure 2.4 Trends of iNaturalist activity since 2010 in British Columbia, Canada. a) The number of observations in millions over time. b) The number of iNaturalist observers in thousands over time. Pictures taken by Ellyne Geurts (a) and by a generous dogwalker at Skaha Bluffs Provincial Park (b).

## Chapter 3: Turning observations into biodiversity data: broad-scale spatial biases in community science

### *3.1 Abstract*

Biodiversity community science projects are growing rapidly in popularity. The enormous amounts of data generated by these programs are transforming how we conduct ecological research and conservation management. However, as with other biodiversity surveys, community science datasets suffer from biases in time and locations of observations. To better use these data, I modeled the spatial biases present in the popular community science platform, iNaturalist. iNaturalist uses crowdsourcing to collect georeferenced and time-stamped observations of all taxa worldwide. With its wealth of biodiversity data, iNaturalist is now being used to answer a broad range of questions in ecology and conservation, but little is known about the platform's spatial biases. I focus on the more than 1.75 million iNaturalist observations available (to December 2021) from British Columbia, Canada, a region with a strong community science presence and diversity of ecosystems. Using machine learning and species distribution modeling, I examined which landscape factors (*e.g.*, protected areas, roads, human population density, habitat zones, elevation) were most important in determining where observations are taken, and I created a predicted probability map revealing how likely different regions are to be sampled by community scientists. I found strong road biases for observations in iNaturalist, with over 94% of observations within 1 km of roads. In addition, human population density and broad habitat ecosystem zones played a large role in predicting where iNaturalist observation occur across the landscape. These methods demonstrate tools for modeling the effects of spatial biases in large opportunistic datasets that can then be used

to produce more accurate species distribution and biodiversity models from community science data.

*3.2 Introduction*

The use of community science to collect data on biodiversity is growing rapidly with advances in technology and online platforms like eBird and iNaturalist (Sullivan et al. 2009; Miller-Rushing et al. 2012; Pocock et al. 2018; Loarie 2020). These community science (also known as citizen science) platforms range from user-driven opportunistic data collections (*e.g.*, iNaturalist) to standardized surveys with volunteers helping collect specific data (*e.g.*, Breeding Bird Surveys). Community science platforms are producing massive amounts of biodiversity and ecological data across large geographical and temporal scales (Pocock et al. 2018; Zhang 2020; Loarie 2022a). These data are being used to answer questions of phenology (Barve et al. 2020b; Nowak et al. 2020), species distributions (Johnston et al. 2021), population trends (Neate-Clegg et al. 2020), phenotypic variation (Drury et al. 2019; Lehtinen et al. 2020), health (Hamilton et al. 2021), and species interactions (Saldivar et al. 2022). In addition, these community science platforms are contributing to the discovery of new species and monitoring and management of exotic and rare species (Werenkraut et al. 2020; Hausdorf et al. 2021; Jain et al. 2022; Roberts et al. 2022). The contribution of these datasets to our biodiversity knowledge hinges on whether we understand the associated observer and spatial biases and account for them explicitly in analyses (Dickinson et al. 2010; Isaac et al. 2014; Johnston et al. 2018; Brown & Williams 2019).

Spatial biases, taxonomic biases, and variability in observer sampling effort are common limitations in community science biodiversity datasets. Observations are often concentrated in

regions with high human population densities (Ballesteros-Mejia et al. 2013; Ruete 2015; Speed et al. 2018) and in areas that are easily accessible such as roads and tourist locations (Kadmon et al. 2004; Oliveira et al. 2016). The data often also exhibit taxonomic biases, favouring large charismatic taxa such as birds and mammals over cryptic taxa like spiders (Isaac & Pocock 2015; Troudet et al. 2017). Furthermore, variability in sampling effort is common in crowdsourced community science, where there are little to no sampling guidelines, resulting in large variability in distances surveyed, duration of surveys, and intensity of observations, which may (*e.g.*, eBird) or may not (*e.g.*, iNaturalist) be accounted for (Isaac et al. 2014; Ruete 2015). Note that biases exist in all datasets, even professional surveys (Kosmala et al. 2016), and there are methods to explicitly account for spatial biases in species distributions (Fithian et al. 2015; Zizka et al. 2020), observer variability in species occupancy-detection and biodiversity modeling (Isaac et al. 2014; Kelling et al. 2015; Meyer et al. 2016; Johnston et al. 2018), and temporal biases in phenology studies (Courter et al. 2013). However, these methods require knowledge of the extent and strength of the biases and the factors affecting them before use.

Despite the importance of understanding the extent and strength of biases in community science diversity datasets, many projects have limited information on biases and errors, with the exception of bird-focussed programs, such as eBird and breeding bird surveys (van Wilgenburg et al. 2015; La Sorte & Somveille 2020; Zhang 2020). The popular platform iNaturalist has the largest number of participants with over 2.4 million users and the broadest taxonomic coverage in the world with more than 120 million observations (Callaghan et al. 2020; iNaturalist 2022). iNaturalist also displays strong evidence of spatial biases and observer variability due to its opportunistic data collection (Loarie 2020; Di Cecco et al. 2021; iNaturalist

2022). Current research has only begun to scratch the surface of the biases within iNaturalist data (Callaghan et al. 2020; Di Cecco et al. 2021; Mesaglio & Callaghan 2021). Studies so far have focused on describing the broad taxonomic, temporal, and spatial biases within iNaturalist such as taxonomic specialization of iNaturalist observers, weekend temporal biases, and spatial bias towards developed land (Di Cecco et al. 2021; Mesaglio & Callaghan 2021). However, there is currently no example framework available for investigators interested in visualizing and modeling the spatial sampling biases on iNaturalist to show where predicted community science activity will be high versus low. In addition, we do not know which landscape features are the most influential for where iNaturalist users make observations and how these are related to the probability of an observation being made at a particular location.

My study addresses this knowledge gap on biases in community science diversity datasets by modelling spatial biases in the iNaturalist database using the large and geographically diverse province of British Columbia (BC), Canada as a case study. I use Maxent, a popular machine learning software that produces species distribution models with presence-only data (Phillips, Dudík, and Schapire 2020). The software can also be used to model sampling effort (*i.e.,* spatial sampling biases) and test which variables such as habitats and distance to roads influence observer behaviour (Merow et al. 2013; Barber et al. 2022). I developed a workflow that could be applied anywhere to: 1) model spatial biases of iNaturalist observations across BC and predict where observations are likely to occur across the landscape; and 2) determine the strength and direction of relationships between environmental variables and probability of iNaturalist observations. Given the opportunistic nature of iNaturalist observations and previous studies of community science projects, I predicted that distance to

roads and human population density will be the most important environmental variables biasing the distribution of observations, with land cover type such as urban and agricultural regions and tourist locations (*e.g.*, parks) having a lesser effect. I predicted an exponential negative relationship between distance to roads and probability of observation, and a positive linear relationship between human population density and probability of observation.

### *3.3 Methods*

#### 3.3.1 Background – iNaturalist and study area

iNaturalist is a social network platform where users upload their own georeferenced and time-stamped photos and audio recordings of organisms for community identification (iNaturalist 2022). The platform is designed for users of all skill levels with the primary goals of education and connecting people with nature. iNaturalist has no sampling guidelines; people select what, when, and where they want to observe nature. This leads to large variability in sampling effort and spatial, temporal, and taxonomic coverage of observations (Di Cecco et al. 2021). As a result, observations on iNaturalist are considered presence-only data, which require specialized statistics for analysis (Dickinson et al. 2010).

British Columbia (BC) is an excellent study region with its wide range of habitats and population densities (Meidinger & Pojar 1991; Environmental Reporting BC 2018a), strong iNaturalist community (iNaturalist 2022), and publicly available fine resolution spatial data (BC Data Catalogue: https://catalogue.data.gov.bc.ca/). The BC iNaturalist database currently contains over 2 million georeferenced observations across the taxonomic spectrum from 40,000 observers from 1937 to 2022, with the number of observations growing exponentially (iNaturalist 2022).

### 3.3.2 Study datasets

I downloaded iNaturalist observations (n = 1,769,501) directly from the iNaturalist website (https://www.inaturalist.org/) on December 2, 2021. I selected distance to roads as a landscape feature that I expected to influence the spatial coverage of iNaturalist observations (Reddy & Dávalos 2003; Kadmon et al. 2004; Stolar & Nielsen 2015; Tye et al. 2017). I included provincial and national parks because many parks are popular tourist spots in BC (BC Parks 2018) and tourist spots can bias number of community science records in a region (Boakes et al. 2010). In addition, there are increasing initiatives by government agencies and non-profit societies encouraging the use of the iNaturalist platform in parks in BC (BC iNaturalist Program 2021; Parks Canada 2022; Strathcona Wilderness Institute 2022). I expected human population density to be related to the spatial distribution of iNaturalist observations (Ballesteros-Mejia et al. 2013; Ruete 2015; Speed et al. 2018). I selected the fine-scale land cover type (*e.g.*, cropland, urban, and mixed forest) from MODIS (IGBP global vegetation classification scheme; Appendix S1 Table S1) because this can cause spatial bias in volunteer community science (Geldmann et al. 2016; Di Cecco et al. 2021; Petersen et al. 2021). I also included data from the broader scale Biogeoclimatic Ecosystem Classification system (BEC), used in BC to classify landscape-level ecosystems (Meidinger & Pojar 1991; Tulloch & Szabo 2012; Geldmann et al. 2016). I also analyzed elevation as mountainous regions are less accessible than lower elevation sites, which likely causes spatial bias in observations (Fernández & Nakamura 2015; Mair & Ruete 2016). See Table S2 in Appendix S1 for further information on the spatial datasets.

### 3.3.3 Spatial data preparation

I removed marine observations (n = 141,147), resulting in 1,628,354 terrestrial observations for analysis. I further refined the data for Maxent analysis by spatially filtering the observations following Kass et al. (2022) where I retained only one observation per grid cell (277 m resolution) to reduce spatial autocorrelation. The final cleaned occurrence dataset contained 152,785 terrestrial observations. I cleaned and processed environmental spatial layers to ensure identical spatial projection, cell size, extent, and origin. I clipped the roads dataset to the BC terrestrial boundary to remove boat routes and Yukon highways. I created a Euclidean distance to road raster layer with cell size of 25 m. I rasterized the national and provincial park polygon layers, converted them to binary surfaces, and combined them to create a raster layer of park land versus non-park land. I took the log of human population density following Mair and Ruete (2016) and Barber et al. (2022) to better examine the effect of population density changes at very low densities. I projected all environmental layers in BC Albers projection, cropped and masked using the BC terrestrial polygon, then resampled to cell size of 277 m to match the coarsest raster layer (MODIS land cover). This resolution is comparable or even finer than other studies of spatial sampling bias in community science (Stolar & Nielsen 2015; Mair & Ruete 2016; El-Gabbas & Dormann 2018). I used bilinear resampling for continuous raster layers (distance to road, human population size, and elevation). I conducted spatial data preparations in R (version 4.1.3), RStudio (R Core Team 2021; RStudio Team 2021), and ArcMap 10.6.1. (Redlands 2017). I used the following R packages: *bcmaps* (Teucher et al. 2021), *MODIStsp* (Busetto & Ranghetti 2016), *sf* (Pebesma 2018), *dplyr* (Wickham et al. 2021), *raster* (Hijmans 2021a), *fasterize* (Ross 2020), and *Terra* (Hijmans 2021b).

### 3.3.4 Statistical analyses

I quantified the number of terrestrial iNaturalist observations near roads using an empirical cumulative distribution function, which I compared to a random null model. I selected one million random points across BC, extracted Euclidean distance to road for each random spatial point, then took the mean of the million random points. I bootstrapped those million distances 10,000 times. Mean and standard error were measured for each bootstrap sample. I then selected one million data points randomly from the observed terrestrial iNaturalist observations and took bootstrapped samples in the same manner. I chose a sample size that was similar to the observed number of observations in British Columbia. I compared the observed mean distances between the two distributions using a Welch Two Sample t-test. I conducted these analyses in RStudio using the *stats*, *sf*, *raster*, and *dplyr* packages (Wickham 2016; Pebesma 2018; Hijmans 2021a; R Core Team 2021; RStudio Team 2021; Wickham et al. 2021).

I used the species distribution modeling software, Maxent, via the *ENMeval* and *dismo* packages to investigate which environmental variables are strong predictors of where iNaturalist observations are made across the province (Kass et al. 2021; Hijmans et al. 2021; Phillips, Dudík, and Schapire 2020). I selected Maxent because it handles presence-only data (Elith et al. 2011), it is widely used (Fourcade et al. 2014), and is ranked as one of the top species distribution models for presence-only data for predictiveness (Valavi et al. 2022). In addition to modeling species distributions, Maxent can also be used to create a biased prior of sampling effort that can then be fed into a species distribution model to correct for sampling and geographical bias (Phillips, Dudík, and Schapire 2020; Elith et al. 2011; Barber et al. 2022).

Creating a biased prior is usually based on target group sampling, where occurrence data of species within the same taxonomic category of the focal species with similar sampling biases are pooled together and modeled with different covariates related to observer behaviour *e.g.*, distance to roads and urban centers (Phillips et al. 2009; Merow, Smith, and Silander Jr 2013). I adapted Maxent to model the distribution of iNaturalist observers across BC. People who make an observation are thus my "species". I pooled all terrestrial observations to model probability of occurrence of an observation (*i.e.*, sampling spatial bias) across the province and to determine which environmental variables best explain where observations are made (Elith et al. 2011; Phillips 2017). I selected the R package *ENMeval* to run Maxent as it allows multiple tuning parameters to be tested at once, produces reproducible code, provides metrics such as AICc to allow comparison of different maxent models, and has the function to test Maxent models against a null model (Kass et al. 2021). The null model analysis for species distribution modeling is based on Bohl et al. (2019).

I conducted a Pearson correlation matrix analysis for continuous raster layers using the *ENMTools* package (Warren & Dinnage 2022) and Cramer's V for similarity association measurements of categorical layers using the *rcompanion* R package (Mangiafico 2022) to ensure the environmental variables were not highly correlated (correlation metric < 0.50) (Merow et al. 2013; Fourcade et al. 2014). I produced similarity matrices using Schoener's D (*i.e.*, niche overlap) of the predicted Maxent values among the different Maxent models in geographic space (Schoener 1968; Kass et al. 2022).

### 3.3.5 Maxent settings

I ran Maxent models with the following tuning arguments and settings. I used the Randomkfold partition method with k = 5 and the 'maxent.jar' algorithm following the default Maxent GUI settings (Kass et al. 2021). I used 300,000 background points and made the model randomly select the background points from across BC (Phillips 2017; Kass et al. 2021; Valavi et al. 2022). I did not restrict background area to buffer zones around presence points because I am interested in making inferences for the entire province (Fourcade et al. 2014; Kass et al. 2021). Maxent removed 4,664 occurrence points with NA predictor variable values. I tested regularization multiplier values from 0.5 to 2 by increments of 0.5. The feature classes I included were linear (L), quadratic (Q), hinge (H), and product (P), with the following combinations tested: L, Q, H, P, LQ, LQH, LQP, and LQHP. I tested 32 different Maxent models. A prediction map for the top model was produced using the cloglog output format. The top Maxent model was selected using AICc and AUC validation (Kass et al. 2021). It was then compared to a null model using the 'ENMnulls' function from *ENMeval* package with 100 iterations implemented (Kass et al. 2021). I used the tuning arguments from the top model as inputs for the Maxent GUI to produce the response curves and variable importance graphs (Phillips 2017).

I examined variable importance of the top maxent model using percent contribution, permutation importance, and a jackknife test of regularized training gain (Phillips 2017). I analysed the relationships of the environmental variables with the Maxent predicted probability of iNaturalist observations using marginal and isolated variable response curves. Marginal response curves show the relationship between predicted probability of observation

with an environmental variable across its range while all other variables are held at their average sample value. Isolated variable response curves consider only one variable at a time.

### *3.4 Results*

#### 3.4.1 Where are iNaturalist observations likely to occur?

The highest probabilities of observations are in southern British Columbia and along highways (Fig. 3.1). The lowest predicted probabilities are in the north, particularly in the northwest (Fig. 3.1). These probabilities are based on the top Maxent model in terms of both AICc and AUC validation values ($AUC_{val}$ = 0.906, delta AICc = 0, number of coefficients = 176; Table 3.1). This was also the most complex model with all four feature classes included (= linear, quadratic, hinge, and product) and a regularization multiplier value of 0.5 (*i.e.*, low penalty towards complexity). There was high "niche overlap" among the 32 Maxent model predictions, suggesting the different Maxent settings produced similar maps. The similarity matrices using Schoener's D of the predicted Maxent model values in geographic space ranged from $D$ = 0.818 to $D$ = 0.998. Lastly, the top Maxent model was significantly different from the null model using the AUC validation metric (Z = 15.25, $p < 0.00001$; Appendix S1 Table S3 and Fig. S1).

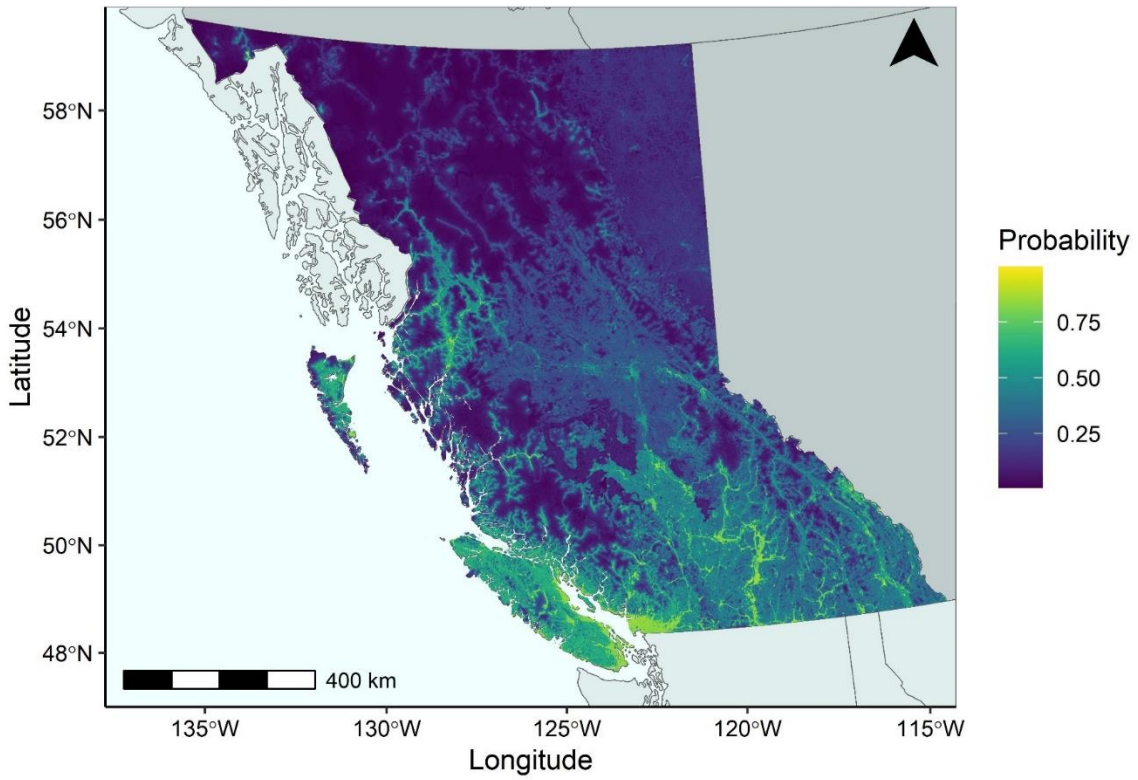



Figure 3.1 Maxent predicted probability of presence for iNaturalist observers making observations in British Columbia, Canada. Predictions are based on the cloglog output format. Cell resolution = 277 m.

Table 3.1 Maxent outputs of the different models tested using *ENMeval* R package for the distribution of iNaturalist observations in British Columbia, Canada. Feature classes: L = Linear, Q = quadratic, H = hinge, and P = product.

| Feature class tested | Regularization multiplier | Average AUC validation value | Number of coefficients | AICc | Delta AICc | AIC weight |
|---|---|---|---|---|---|---|
| **LQHP** | 0.5 | 0.906 | 176 | 4486437 | 0 | 1 |
| **H** | 0.5 | 0.905 | 143 | 4488373 | 1937 | 0 |
| **LQH** | 0.5 | 0.905 | 183 | 4488660 | 2223 | 0 |
| **LQHP** | 1 | 0.905 | 118 | 4488985 | 2548 | 0 |
| **LQH** | 1 | 0.905 | 134 | 4491003 | 4567 | 0 |
| **H** | 1 | 0.905 | 127 | 4491004 | 4567 | 0 |
| **LQHP** | 1.5 | 0.905 | 89 | 4491452 | 5016 | 0 |
| **LQH** | 1.5 | 0.904 | 105 | 4493235 | 6798 | 0 |
| **H** | 1.5 | 0.904 | 116 | 4493359 | 6923 | 0 |
| **LQHP** | 2 | 0.905 | 76 | 4493955 | 7519 | 0 |
| **LQH** | 2 | 0.904 | 94 | 4495533 | 9096 | 0 |
| **H** | 2 | 0.904 | 129 | 4495710 | 9274 | 0 |
| **LQP** | 0.5 | 0.898 | 33 | 4510179 | 23742 | 0 |
| **LQ** | 0.5 | 0.897 | 29 | 4513439 | 27003 | 0 |
| **LQP** | 1 | 0.897 | 29 | 4515680 | 29244 | 0 |
| **LQ** | 1 | 0.897 | 26 | 4518143 | 31706 | 0 |
| **LQP** | 1.5 | 0.896 | 27 | 4519937 | 33500 | 0 |
| **L** | 0.5 | 0.895 | 29 | 4521003 | 34566 | 0 |
| **LQ** | 1.5 | 0.896 | 24 | 4522466 | 36029 | 0 |
| **LQP** | 2 | 0.896 | 25 | 4524454 | 38018 | 0 |
| **L** | 1 | 0.894 | 26 | 4524770 | 38334 | 0 |
| **LQ** | 2 | 0.895 | 25 | 4526896 | 40460 | 0 |
| **L** | 1.5 | 0.894 | 24 | 4528472 | 42036 | 0 |
| **L** | 2 | 0.893 | 23 | 4531824 | 45388 | 0 |
| **P** | 0.5 | 0.886 | 32 | 4547360 | 60924 | 0 |
| **P** | 1 | 0.884 | 30 | 4554220 | 67783 | 0 |
| **P** | 1.5 | 0.882 | 28 | 4560714 | 74277 | 0 |
| **Q** | 0.5 | 0.875 | 33 | 4563962 | 77525 | 0 |
| **P** | 2 | 0.881 | 25 | 4566936 | 80500 | 0 |
| **Q** | 1 | 0.871 | 31 | 4571971 | 85534 | 0 |
| **Q** | 1.5 | 0.869 | 29 | 4578378 | 91942 | 0 |
| **Q** | 2 | 0.867 | 27 | 4584099 | 97662 | 0 |

### 3.4.2 Which environmental features best predict where iNaturalist observations occur?

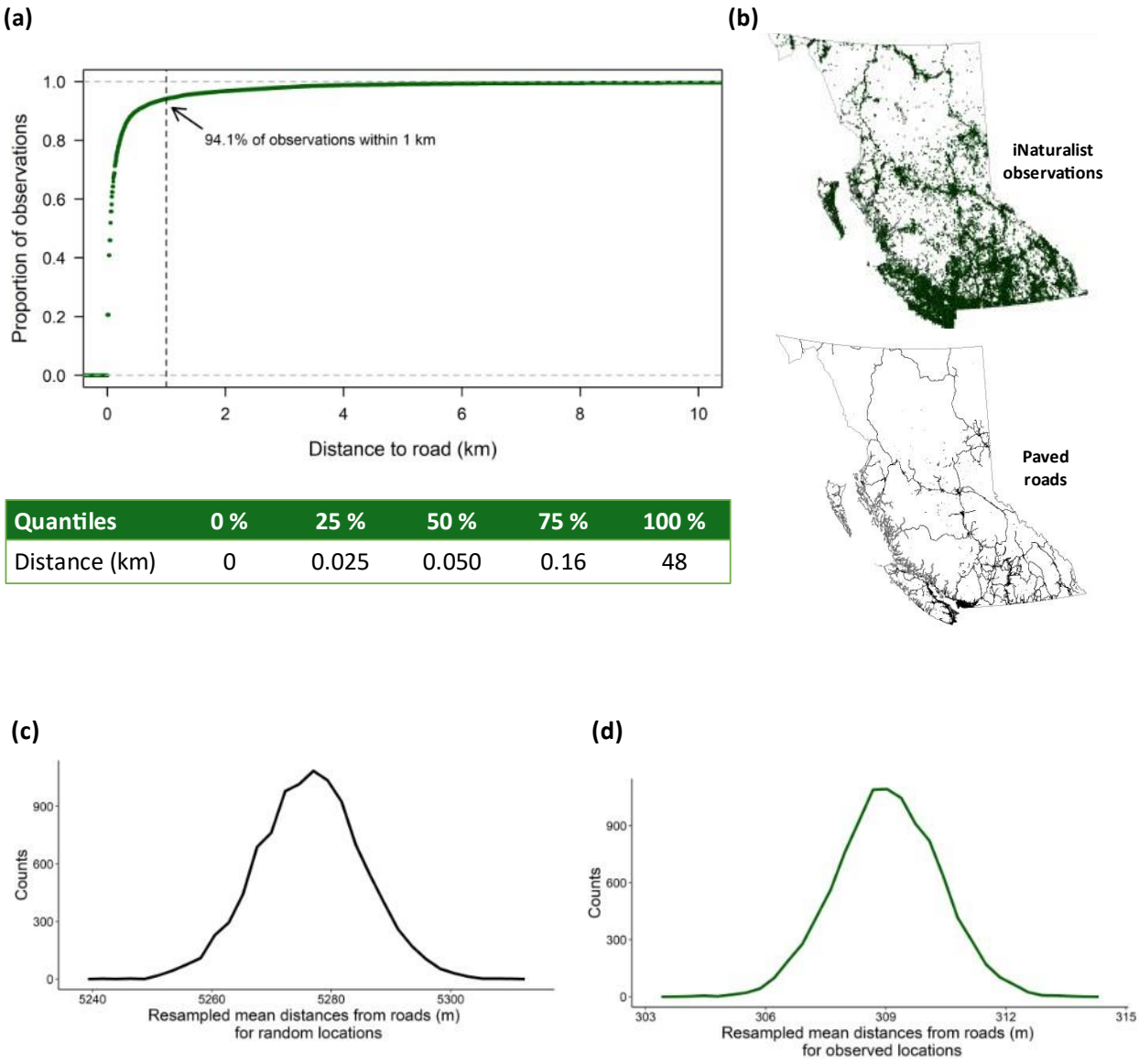
For the top Maxent model, distance to road was the most influential variable with a permutation importance value of 51.6%, while Biogeoclimatic Ecosystem Classification (BEC)

was the second most important variable at 28.7% (Table 3.2). Human population density was much less important (permutation importance value 9.2%), though its percent contribution was similar to distance to roads (34%; Table 3.2). Since percent contribution values are pathway (*i.e.*, algorithm) dependent and permutation importance values are derived from the final Maxent model (Phillips 2017), this drop in relative importance suggests that population density was important for the Maxent algorithm to create the model, but ultimately distance to roads and BEC play a larger role for predicting where iNaturalist observations occur (Table 3.2). Land cover type was the least important variable with permutation importance of 1.2% and percent contribution of 0.6% (Table 3.2). See Appendix S1 for jackknife test results of variable importance (Fig. S2).

Table 3.2 Two measures of variable importance for the top Maxent model (Feature classes = LQHP, regularization multiplier = 0.5) selected by the *ENMeval* R package for the distribution of iNaturalist observations in British Columbia, Canada. Permutation importance values are derived from the final Maxent model. Percent contribution values are algorithm (*i.e.*, Maxent) dependent.

| **Variable** | **Permutation importance (%)** | **Percent contribution (%)** |
|---|---|---|
| Distance to roads (m) | 51.6 | 34.2 |
| Biogeoclimatic zones | 28.7 | 19.3 |
| Human population density | 9.2 | 34.1 |
| Park vs. non-park land | 7.0 | 7.2 |
| Elevation (m) | 2.3 | 4.6 |
| Land cover type | 1.2 | 0.6 |

The strong bias of observations towards roads can be seen in Figure 3.2. A total of 75% of observations were within 160 m and 94% of observations within 1 km of roads (Fig. 3.2). The mean distance from roads was 309 ± 0.01 m (SE) from roads, whereas random points were 5,277 ± 0.09 m from roads (Welch two sample t-test: $t = 55695$, $df = 10422$, $p\text{-value} < 0.001$; Fig. 3.2c-d).




| Quantiles | 0 % | 25 % | 50 % | 75 % | 100 % |
|---|---|---|---|---|---|
| Distance (km) | 0 | 0.025 | 0.050 | 0.16 | 48 |

Figure 3.2 (a) Empirical cumulative distribution function and quantiles of observed distance to roads for terrestrial iNaturalist observations in British Columbia, Canada. Analysis includes all road types: paved, maintained, and unmaintained. (b) Maps of iNaturalist observations and paved roads. (c) Frequency polygon plot of the mean distances from roads for random locations (n = 10, 000 bootstrapped samples). (d) Frequency polygon plot of the mean distances from roads for observed locations (n = 10,000 bootstrapped samples). Mean distance was calculated for each bootstrapped sample. Each sample contained one million resampled data points.

### 3.4.3 How do the environment variables affect predicted probability of iNaturalist observation?

Marginal response curves provide another way of assessing and visualizing the roles of individual predictor variables in biasing the locations of observations by holding all other predictors at their average sample value. As expected, there was higher predicted probability of an iNaturalist observation for locations close to roads (Fig. 3.3). When controlling for other variables, for example human population density, the biogeoclimatic zones with the highest predicted probability of observation were the Coastal Mountain-heather Alpine and Mountain Hemlock zones, and the lowest probabilities were the Boreal White and Black Spruce and Sub-Boreal Pine – Spruce zones (Fig. 3.3). When the other variables are not controlled for, the highest predicted BEC zones are Coastal Douglas Fir and Ponderosa Pine (Appendix S1 Fig. S3). As predicted, there was a positive relationship between predicted probability of observation and human population density, and a higher predicted probability of iNaturalist observations within parks than outside parks (Fig. 3.3). See Appendix S1 for marginal and isolated variable response curves for all six environmental variables (Fig. S3).

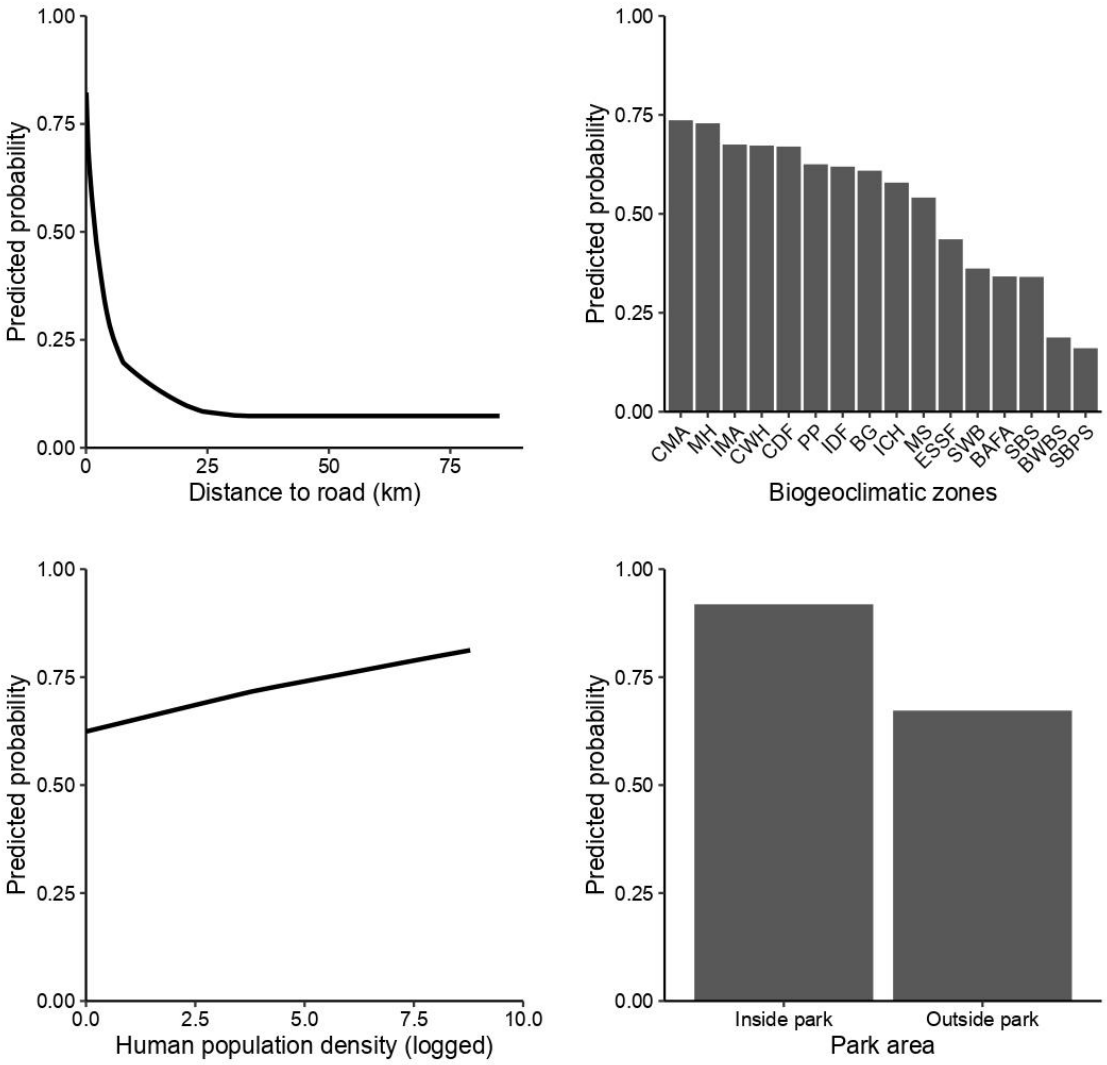



Figure 3.3 Marginal response curves for the top four ranked environmental variables in the Maxent model. These show the relationship between predicted probability of observation with an environmental variable while all other variables are held at their average sample value. Biogeoclimatic zones: BAFA = Boreal Altai Fescue Alpine, BG = Bunchgrass, BWBS = Boreal White and Black Spruce, CDF = Coastal Douglas-fir, CMA = Coastal Mountain-heather Alpine, CWH = Coastal Western Hemlock, ESSF = Engelmann Spruce – Subalpine Fir, ICH = Interior Cedar – Hemlock, IDF = Interior Douglas-fir, IMA = Interior Mountain-heather Alpine, MH = Mountain Hemlock, MS = Montane Spruce, PP = Ponderosa Pine, SBPS = Sub-Boreal Pine – Spruce, SBS = Sub-Boreal Spruce, and SWB = Spruce – Willow – Birch.

*3.5 Discussion*

My results provide a workflow to visualize the spatial sampling biases present within iNaturalist, the world's largest community science biodiversity platform. Using this method, I was able to determine the important environmental drivers behind those spatial biases. This workflow demonstrates a method to identify spatial sampling bias in presence-only data (*e.g.*, iNaturalist) that could then be subsequently incorporated into species distributions models using these data, as well as other community science summary analyses. As predicted, distance to roads played the largest part in influencing where people make iNaturalist observations (Table 3.2). Unexpectedly, the broad habitat variable (*i.e.*, Biogeoclimatic Ecosystem Classification - BEC) was more important than human population density for predicting where iNaturalist observations occur (Table 3.2).

These results align with other studies examining variable importance in presence-only datasets (Geldmann et al. 2016; Mair & Ruete 2016; El-Gabbas & Dormann 2018). Geldmann et al. (2016) also found distance to roads, population density, and land cover type (*e.g.*, urban) are important factors influencing spatial biases in four different community science projects in Denmark, however, they did not look at any additional variables. Mair and Ruete (2016) found road access and population density were consistently the most important variables in the Swedish LifeWatch platform (Mair and Ruete, 2016). Lastly, El-Gabbas and Dormann (2018) found accessibility covariates (*e.g.*, distances to roads, cities, and protected areas) better accounted for spatial biases than other environmental and effort variables for opportunistically collected data on bats in Egypt.

Although elevation did not appear to have a large influence on where iNaturalist observations occur (Table 3.2), this may be because distance to road was the better measure of accessibility. Mair and Ruete (2016) also found that roads were more important than elevation for accessibility. There are other potential metrics of accessibility such as steepness (Mair & Ruete 2016), terrain ruggedness (Stolar & Nielsen 2015), and travel time to major cities (Barber et al. 2022). Lastly, protected areas (*i.e.*, national and provincial protected parks) and land cover types explained relatively little of the spatial biases (Table 3.2). This was unexpected, considering they have been important in other opportunistically collected datasets (Rocchini et al. 2011; Stolar & Nielsen 2015; Geldmann et al. 2016; El-Gabbas & Dormann 2018; Petersen et al. 2021). In particular, it was interesting to see the land cover type variable ranking so low when Di Cecco et al. (2021) found evidence of iNaturalist observations being biased across different land cover categories in the United States. However, they did not look at multiple spatial variables simultaneously. Thus, iNaturalist observations are likely biased by land cover type, but accessibility (*i.e.*, distance to roads) is a more important factor for predicting where observations will occur. The small effect of parks (7%; Table 3.2), may be due to many of the parks, in particular large ones, being in remote northern regions with no year-round road access. Thus, I recommend including distance to roads, human population density, and broad habitat classification when accounting for spatial biases when using iNaturalist data and consider including protected areas and land cover land type if available.

The negative exponential relationship in predicted probabilities with distance from roads (Fig. 3.3) mirrors the sharp cumulative curve with 94% of iNaturalist observations within 1 km of roads in British Columbia (Fig. 3.2). This relationship was similarly found in Kadmon et al.

(2004) with 61% of plant observations in Israel within 500 m of roads, and 97% of observations within 4 km. Predicted probability increasing with human population density (Fig. 3.3) supports other studies of opportunistic collected datasets (Mair & Ruete 2016). Although the effect of parks on observation distribution was smaller than the other variables I tested, there was higher predicted probability of observations in parks than outside parks (Fig. 3.3). Lastly, for the marginal response curve for the Biogeoclimatic Ecosystem Classification variable, it is likely that the highest predicted zones *e.g.*, Coastal Mountain-heather Alpine and Mountain Hemlock (Fig. 3.3) were due to popular provincial parks occurring within these zones that have intense community science activity (Egan 1997; BC iNaturalist Program 2021; Strathcona Wilderness Institute 2022).

### *3.6 Conclusion*

With its exponential growth, community science continues to be of increasing importance in supplying data for analyses of biodiversity patterns and processes. This work shows how researchers can identify and account for spatial biases in such data. There are additional biases that need attention, including taxonomy of species and inter-observer variability in where people go and what they record. I feel it is important to remember that no dataset is without bias, from community science to professionally collected data, and that it would be a disservice to our pursuit of ecological knowledge to view community science as unusable due to strong biases. I hope that further studies building on my findings can improve the scientific value of community science platforms, including testing the limits of inference that are possible.

## Chapter 4: Not all who wander are lost: trail bias in community science

### *4.1 Abstract*

The exponential growth and interest in community science programs is producing staggering amounts of biodiversity data across broad temporal and spatial scales. Large community science datasets such as iNaturalist and eBird are allowing ecologists and conservation biologists to answer novel questions that were not possible before. However, the opportunistic nature of many of these enormous datasets leads to biases. Spatial bias is a common problem, where observations are biased towards points of access like roads and trails. iNaturalist – a popular biodiversity community science platform – exhibits strong spatial biases, but it is unclear how these biases affect the quality of biodiversity data collected. Thus, I tested whether fine-scale spatial bias due to sampling from trails affects taxonomic richness estimates. I compared timed transects with experienced iNaturalist observers on and off trails in British Columbia, Canada. Using generalized linear mixed models, I found higher overall taxonomic richness on-trails than off-trails. In addition, I found more exotic as well as native taxa on-trails than off-trails. There was no difference between on and off trail observations for species that are rarely observed. Thus, fine-scale spatial bias from trails does not reduce the quality of biodiversity measurements, a promising result for those interested in using iNaturalist data for research and conservation management.

### *4.2 Introduction*

Community science, also called citizen science, involves data collected in collaboration between community members and scientists. It is growing rapidly in popularity in biodiversity research and management (Miller-Rushing et al. 2012; Pocock et al. 2018; Loarie 2022a). The

advancements of technology and accessible platforms like eBird and iNaturalist have increased participation and are producing massive amounts of biodiversity and ecological data around the world (Sullivan et al. 2009; Loarie 2020). These data have been used to answer questions of phenology (Barve et al. 2020a; Nowak et al. 2020), species range changes (Johnston et al. 2021), species migrations (Walker & Taylor 2017), phenotypic variation (Drury et al. 2019; Lehtinen et al. 2020), species interactions (Saldivar et al. 2022), and body condition and disease (Hamilton et al. 2021). The data from these platforms are also helping researchers discover new species and track invasive species (Bois et al. 2011; Werenkraut et al. 2020), and improve species distribution models (Feldman et al. 2021; Matutini et al. 2021). In addition to contributing large amounts of ecological data, community science platforms are educational resources that connect the public with their environments and with other like-minded individuals, cultivating a community of naturalists and environmental stewards (Dickinson et al. 2012; Kobori et al. 2016; Pocock et al. 2018). As a result, researchers and environmental managers are keen to promote these platforms to collect more and better data.

iNaturalist engages the largest number of people globally of any biodiversity community science platform, with more than 125 million observations of taxa around the world, serving as a huge voucher photo and audio data repository (Loarie 2022a). The photo vouchers provide primary data on species occurrences, but also valuable secondary data such as life stage and sex, species interactions, and body condition. Users of iNaturalist are self-directed (*i.e.,* no sampling guidelines) and can make a species observation anywhere in the world. This open-geographic feature allows users to contribute data across broad geographic scales that could never be obtained from traditional scientific surveys. Nevertheless, reservations remain

regarding the quality of the data from iNaturalist due to its lack of sampling guidelines, which may lead to spatial, temporal, and taxonomic biases (Di Cecco et al., 2021, Geurts et al., in review).

Virtually all biodiversity datasets suffer from biases on spatial, temporal, taxonomic, and observer levels (Isaac & Pocock 2015; Ruete 2015; Oliveira et al. 2016; Troudet et al. 2017; Speed et al. 2018). Community science data are frequently biased spatially by level of accessibility such as proximity to roads and trails (Stolar & Nielsen 2015; Geldmann et al. 2016; El-Gabbas & Dormann 2018). Observations are also biased towards tourist locations such as parks and protected areas (Boakes et al. 2010; Rocchini et al. 2011) and by human population density (Ballesteros-Mejia et al. 2013; Ruete 2015; Speed et al. 2018). These biases need to be understood and properly accounted for (Courter et al. 2013; Fithian et al. 2015; Johnston et al. 2018). Some observational studies of the spatial, taxonomic, observer, and temporal biases in iNaturalist data have been conducted (Di Cecco et al. 2021), but no experimental studies of iNaturalist biases and their impacts on biodiversity records have been conducted.

To address this gap, I tested fine-scale spatial impacts on taxonomic richness estimates. I compared observations made along trails and away from trails, because in many terrestrial habitats observations are biased toward trails due to accessibility (Jackson et al. 2015; Geldmann et al. 2016; Callaghan et al. 2020). This bias may be reinforced by rules in parks, where people are not allowed to leave trails (Filazzola et al. 2022). I asked the question “Does trail spatial bias affect the number of taxa observed in parks by community scientists?”. I conducted timed field experiments using experienced iNaturalist observers to compare taxonomic richness estimates between on-trail versus off-trail observations in provincial parks

in British Columbia, Canada. The experiments emulated how an iNaturalist user would observe their environment with no distance or taxonomic restriction for each transect. I implemented a duration limit to aid comparability between the two treatments. I predicted higher taxonomic richness estimates on-trail compared to off-trail (Root-Bernstein & Svenning 2018; Wedegärtner et al. 2022). In addition, I predicted the species compositions to vary between on and off trail, with more exotic species observed on-trails (Liedtke et al. 2020). Lastly, I examined differences in rare and vulnerable species observations between on and off trail.

## *4.3 Methods*

### 4.3.1 Study area and data

Sampling took place in 22 provincial parks and protected areas in British Columbia, Canada (Appendix S2 Fig. S1). I conducted transects in three habitat types: grasslands, open-canopy forests, and closed-canopy forests. Figure 4.1 shows example habitats and Appendix S2 Table S1 shows dominant vegetation observed. Grasslands lack trees, open-canopy forests are dominated by pine trees (*Pinus* sp.), and closed-canopy forests are dominated by non-pine trees. The full list of parks and habitats surveyed can be found in Appendix S2 Table S1. I surveyed a variety of trail substrate types: bare ground (n = 63), gravel (n = 33), and asphalt (n = 1). Trail widths ranged from 0.5 m to 10 m with a mean of 3.1 m. Note that the 10 m trail was a groomed cross-country trail.

I defined taxonomic richness as the number of unique taxa per transect. This allowed for higher classification level observations to be included, as not all observations (*e.g.,* spiders and lichens) could be identified to species level by photographs alone (McMullin & Allen 2022; Mesaglio et al. 2023). I analyzed the number of exotic, vulnerable, and rare species detected on

and off trails. Exotic species are species that are established outside of their native range via anthropogenic means. The list of exotic species (*i.e.,* not-native to British Columbia) was downloaded from the BC Species & Ecosystems Explorer from the Conservation Data Centre (B.C. Conservation Data Centre 2022). Vulnerable species have a status in British Columbia that ranges from "special concern" (S3) to "historical species or possibly extirpated communities" (SH). The full list can be found on the "BC rarities project" on iNaturalist.ca (https://inaturalist.ca/projects/bc-rarities?tab=about). I define rare species as species that have only one record on iNaturalist in the park where the transects were done. I downloaded the transect data from iNaturalist using its CSV Export Function on November 18, 2021 (n = 9,931 observations), and downloaded all observations recorded in the 22 parks on January 30, 2023 (n = 94,777 observations).


**Figure 4.1** Examples of habitats surveyed in British Columbia, Canada. **a)** Closed-canopy forest (Beatton Provincial Park). **b)** Closed-canopy forest (Nairn Falls Provincial Park). **c)** Grassland (Steelhead Provincial Park). **d)** Open-canopy forest (Ellison Provincial Park). Photos taken by Kate McKeown (a-c) and Ellyne Geurts (d).

### 4.3.2 iNaturalist

iNaturalist is a crowd-source type community science platform that was founded in 2008 and has since grown to over 2.5 million users globally (iNaturalist 2022). It is hosted by the California Academy of Sciences and National Geographic Society. Users can upload georeferenced photo and audio recordings of any wild organism (*i.e.,* a verifiable observation). Photo identification can be aided by iNaturalist's computer vision program as well as input from the iNaturalist community. If the verifiable observation reaches majority agreement in species identification from the community, the quality grade of the observation is upgraded from "Needs ID" to "Research Grade", and depending on license restrictions by the observer, the

data are incorporated in the Global Biodiversity Information Facility (https://www.inaturalist.org/pages/help). iNaturalist produces opportunistic presence-only data.

### 4.3.3 Field experiments

I conducted field experiments of trail spatial biases with a team of six iNaturalist observers from the BC Parks iNaturalist Program (BC iNaturalist Program, 2021; https://inaturalist.ca/projects/bc-parks and https://inaturalist.ca/projects/bc-parks-inat-team-big-summer-2021). The observers were active users with at least 3,500+ observations on iNaturalist and had prior experience surveying biodiversity in British Columbia. Observers did surveys together for the first two weeks with professional naturalists before beginning the field experiments to help standardize identification and search patterns. Data collection took place between May and August 2021 (Appendix S2 Table S1). I conducted 96 paired transects. I photographed the habitat at the beginning and end of each transect to help refine habitat categories. I conducted the paired transects in teams of three, with two observers and one facilitator for efficient data recording. One observer walked on the trail while the other observer walked parallel to the trail approximately 20 m away within the same habitat. I selected 20 m because trail effects from disturbance generally fade after 5 m away from the trail (Avon et al. 2010; Swart et al. 2019). The facilitator followed behind the observer pair to keep time, gather all metadata for each transect, and ensure the off-trail observer stayed at least 20 m from the trail. The observer pairs photographed all plants, animals, and fungi that they encountered for 30 minutes in each transect. They were instructed simply to try to maximize the number of species that they photographed. They recorded their straight-line

distances using Garmin GPS units (*etrex 20*) as a proxy for distance traveled. I did not measure true distance traveled due to GPS signal interferences in forests and ravines resulting in unreliable GPS track distances. Transects did not have set distances to better emulate a typical iNaturalist user but I considered the effects of distances traveled in the statistics. The on-trail transects included a 2 m buffer on either side of the trail to mimic typical iNaturalist user behaviour in photographing species from trails. I selected trail segments that did not contain switchbacks to allow parallel surveying without backtracking. The side of the off-trail survey was selected randomly except when there were safety concerns.

During the on and off trail transects, the facilitator recorded trail width and type, and the dominant habitat type and mesohabitats that the observers encountered (Newmaster et al. 2005; McMullin & Wiersma 2017). Mesohabitats included: cliff faces (large vertical rock), seeps (damp depressions), streams (narrow flowing water), and water pools (small areas with standing water). The observers employed the floristic habitat sampling method also referred to as the "intelligent meander" (Selva 2003), where observers seek different mesohabitats and microhabitats (*e.g.,* logs and rocks) during transects. This sampling method observes more species, especially rarer species, than the traditional plot sampling method and is thought to more closely resemble the behaviour of iNaturalist users (Newmaster et al. 2005). To ensure a balanced sampling design, each observer surveyed on and off trail an equal number of times throughout the summer. In addition, observer pairs switched regularly. Each observer surveyed with each partner the same number of times. To avoid temporal clustering of sampling for both trail position and observer partner, the rotations were spaced out evenly throughout the

summer. Surveys were done opportunistically throughout the summer when parks with adequate trails were available and weather permitted.

### 4.3.4 Statistical analyses

I compared on versus off trail taxonomic richness using generalized linear mixed models with a Poisson distribution using the "glmer" function from the *lme4* R package (Bates et al. 2015). Taxonomic richness is defined as the number of unique taxa recorded per transect. I included observer identity as a random effect and the trail variable of on- and off- position as a fixed effect. I also included location of transect as a random effect with trail name nested within park name. In addition, I considered habitat type, number of observations, and distance traveled as additional explanatory variables. I standardized the distance traveled variable to have a mean of zero and standard deviation of one. I excluded the variable of mesohabitats encountered as most of the transects encountered zero mesohabitats (n = 168 out of 192 individual transects), meaning there was only one dominant habitat type throughout the transect. I used random intercepts for the variance structure. I did not test for random slopes for different observers due to high ratio of parameters to sample size. I compared four models with different fixed-effect structures using AICc: (1) a full model with trail position (on vs. off), habitat type, and distance traveled, (2) a full model with an interaction term for distance traveled and trail position, as distance traveled could be affected by whether the observer is on or off trail, (3) a model with only the trail position variable, and (4) a null model. I then repeated these steps with exotic species observations removed from the analysis (n = 391 observations removed) to compare native taxonomic richness.

I compared the number of rare species observed on or off trails using a Wilcoxon sign rank test. I began by filtering all observations in each park to only contain species-level observations (n = 57,843), then I removed all "casual" quality grade observations from the dataset (n = 57,547). For each park I created individualized rare species lists by selecting species that only had one record in the park. This allowed us to count the number of rare species observations recorded by us on and off trail by park. I treated each park as a pair (n = 22). I also counted the number of vulnerable species recorded on and off trails across all parks. Similarly, I compared the number of exotic species observed between on and off trail using a Wilcoxon sign rank test. I grouped transects by park and summed the number of exotic species observations recorded on and off trail. I also calculated the proportion of observations in each transect that contained exotic species.

I conducted the statistical analyses in R (version 4.1.2.) and RStudio (R Core Team 2021; RStudio Team 2021). The packages I used were: *lme4* (Bates et al. 2015), *lubridate* (Grolemund & Wickham 2011), *dplyr* (Wickham et al. 2021), *bbmle* (Bolker & R Development Core Team 2022), *sp* (Pebesma & Bivand n.d.), *rgdal* (Bivand et al. 2021), *adehabitatLT* (Calenge 2006), and *ggplot2* (Wickham 2016).

## *4.4 Results*

### 4.4.1 Trail bias – taxonomic richness estimates

The observers recorded a total of 9,931 observations (on-trail = 5,107, off-trail = 4,824) of 1,323 taxa across the 96 paired transects. As predicted, there was higher taxonomic richness on-trails than off-trails. The mean number of taxa on-trail was 40.8, while off-trail was 37.6.

The top model was the full model with all variables included and no interaction between trail position and distance traveled (Table 4.1). Both the trail position and habitat variables were significant in that model (Fig. 4.2). The grassland habitat had a lower number of observed taxa than the closed-canopy forest habitat (Fig. 4.2, 4.3). The number of taxa observed was highly correlated with number of observations (Pearson's product-moment correlation: $r = 0.90$, $t = 29.27$, $df = 190$, $p < 0.001$). Distance traveled did not affect the number of taxa observed (Fig. 4.2). Mean and standard error of distance traveled by observers on-trail was 114 ± 9 m and 89 ± 6 m for off-trail.

When I restricted the analysis to native taxa, the same top model explained taxonomic richness (Table 4.1). The pattern of variable importance is also similar, except 'open-canopy forest' now has an effect (Fig. 4.2).

Table 4.1 Models tested to explain taxonomic richness observed along transects. Total taxonomic richness is the total number of unique taxa observed per transect. Native taxonomic richness is total taxonomic richness with exotic species removed (n = 391 removed). Trail position indicates whether the observer was on or off trail. Distance = straight-line distance traveled during a transect. Habitat is classified into three broad categories reflecting general vegetation structure. All models had observer and trail name nested within park name as random effects.

| Model name | Variables tested | ΔAICc | DF | Weight |
|---|---|---|---|---|
| ***Total taxonomic richness*** | | | | |
| **Full model** | Trail position + distance + habitat | 0 | 8 | 0.671 |
| **Full model with distance interaction** | Trail position + distance + habitat + trail position:distance | 2.1 | 9 | 0.231 |
| **Trail position only** | Trail position | 3.9 | 5 | 0.097 |
| **Null model** | 1 | 15.1 | 4 | < 0.001 |
| ***Native taxonomic richness*** | | | | |
| **Full model** | Trail position + distance + habitat | 0 | 8 | 0.729 |
| **Full model with distance interaction** | Trail position + distance + habitat + trail position:distance | 2.1 | 9 | 0.256 |
| **Trail position only** | Trail position | 8.3 | 5 | 0.012 |
| **Null model** | 1 | 10.4 | 4 | 0.004 |

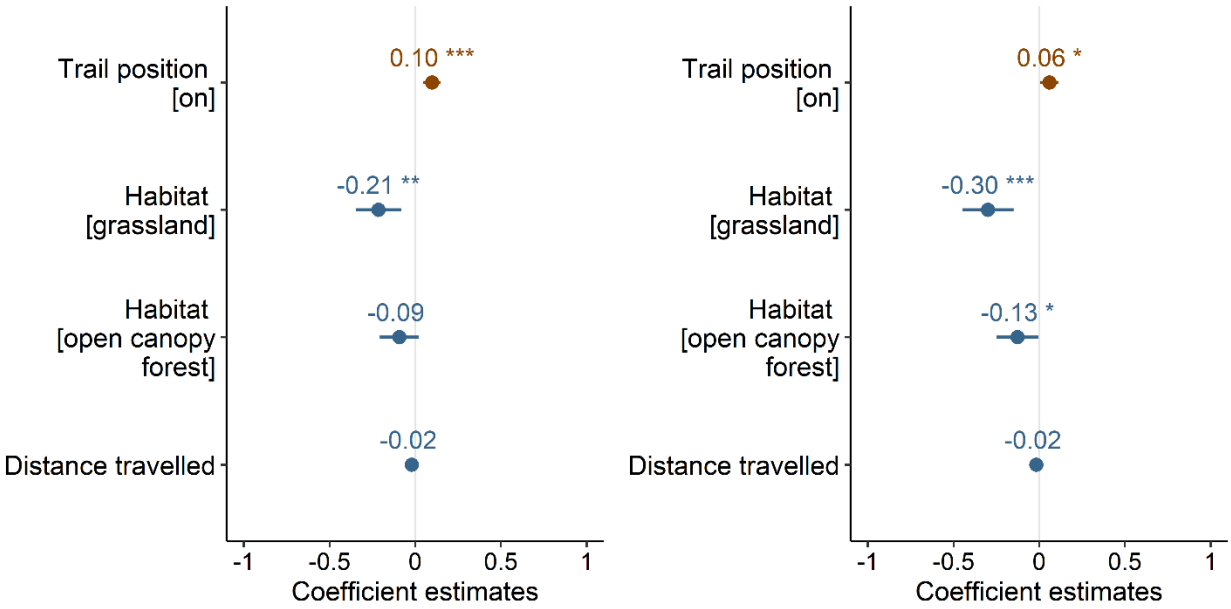



**Figure 4.2** Standardized coefficient estimates with standard error bars for variables in the top model examining taxonomic richness by trail position, habitat type, and straight-line distance traveled by observer. Trail position is "on" or "off". The three habitat types were "grassland", "open-canopy forest" and "closed-canopy forest". * = $p < 0.05$, ** = $p < 0.01$, and *** = $p < 0.001$. Blue indicates a negative relationship between the covariate and taxonomic richness, while brown indicates a positive relationship. **a)** Total taxonomic richness model. **b)** Native taxonomic richness.

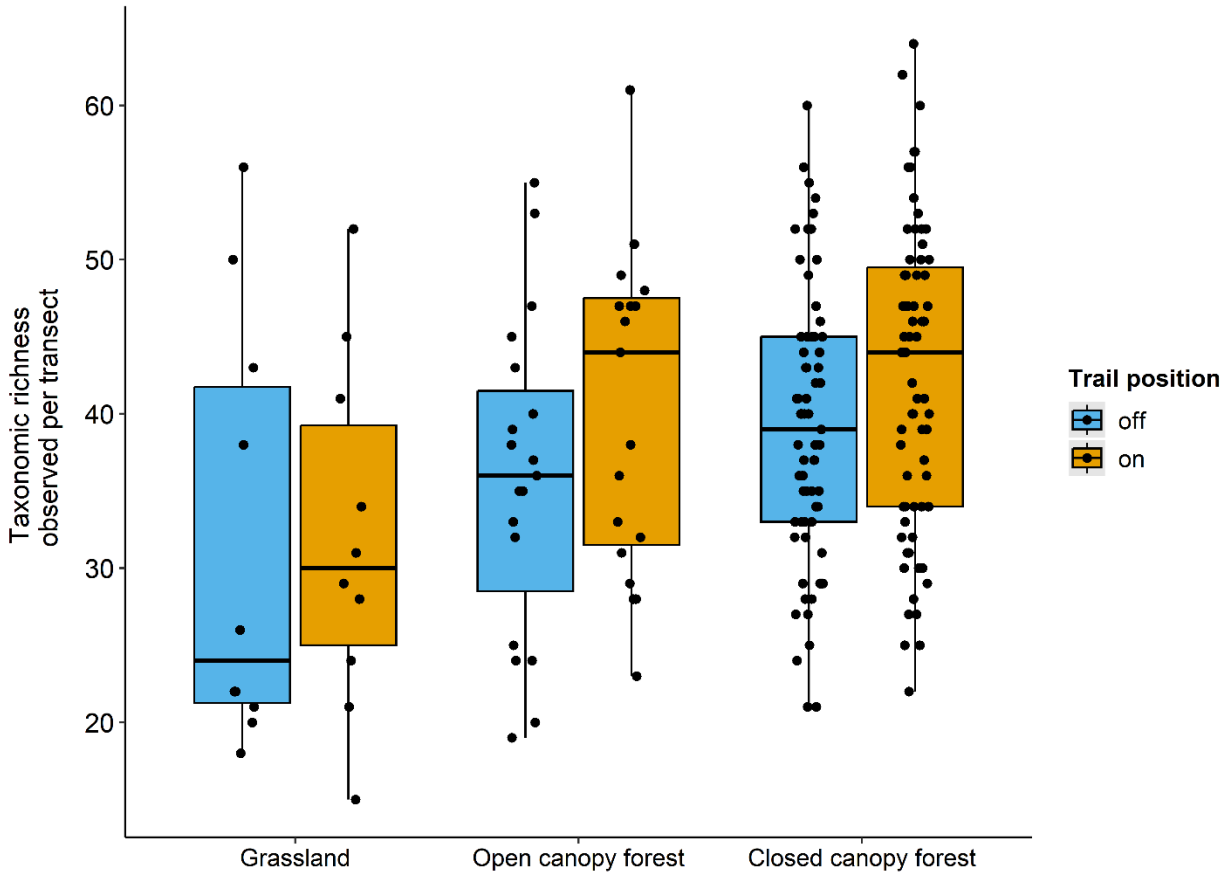



**Figure 4.3** Boxplot of taxonomic richness observed per transect across the three habitat types. Taxonomic richness is the number of unique taxa on iNaturalist per transect. Black bar = median. Box = interquartile range. See Fig. 4.2 for the model coefficient results.

### 4.4.2 Trail bias - vulnerable, rare, and exotic species

Seven of the 192 individual transects contained at least one vulnerable species (on-trail = 6 transects, off-trail = 1 transect). Only one transect had more than one vulnerable species (n = 2). Examples of vulnerable species are shown in Fig. 4.4. I observed four vulnerable insect species: Bunch Grass Locust (*Pseudopomala brachyptera*), a potter wasp (*Odynerus dilectus*), Kiowa Grasshopper (*Trachyrhachys kiowa*), and Huron Short-winged Locust (*Melanoplus huroni*). I also observed one bird species (White-throated Swift - *Aeronautes saxatalis*) and one

plant species (Large-flowered Triteleia (*Triteleia grandiflora*). There was no difference in number of rare species between on and off trail (Wilcoxon signed rank test with continuity correction: V = 73.5, p = 0.40).

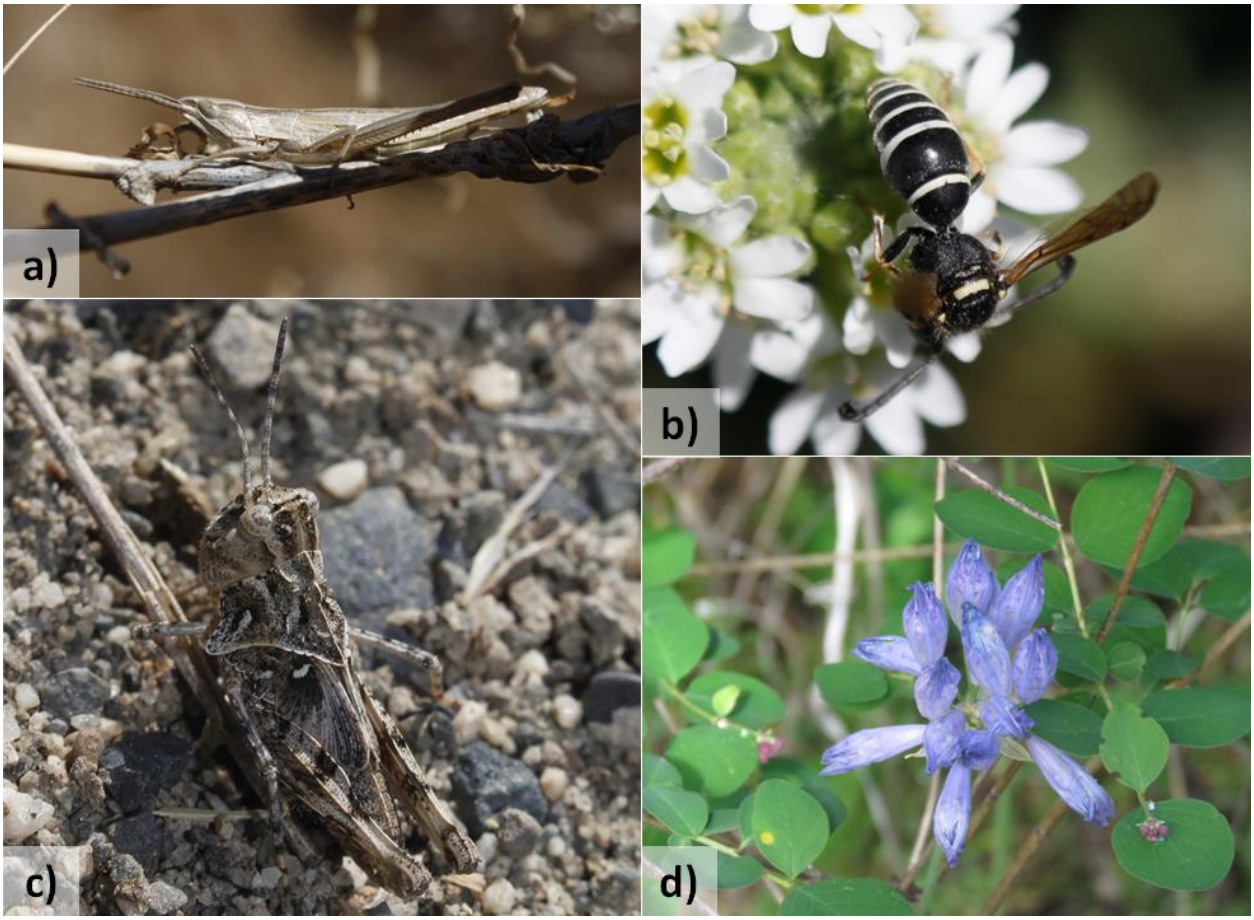



Figure 4.4 Vulnerable species observed during transects on-trail in British Columbia, Canada. Vulnerable species range from "special concern" (S3) to "historical species or possibly extirpated communities" (SH) in British Columbia. a) Bunch Grass Locust (*Pseudopomala brachyptera*) observed in Steelhead Provincial Park. b) *Odynerus dilectus* observed in Fintry Provincial Park and Protected Area. c) Kiowa Grasshopper (*Trachyrhachys kiowa*) observed in Steelhead Provincial Park. d) Large-flowered Triteleia (*Triteleia grandiflora*) observed in Kalamalka Lake Provincial Park. Photos taken by Lena Dietz Chiasson (a-c) and Erin Springinotic (d).

For exotic species, 110 of the 192 individual transects had at least one exotic species (Table 4.2). As predicted, I observed more exotic species on than off trail (Wilcoxon signed rank test with continuity correction: V = 2.5, p < 0.001). Exotic species appeared to make up a higher

proportion of the observations for transects in grassland habitats than in forest habitats (Fig. 4.5). See Table 4.2 for mean and range of number of exotic species observed on and off trails.

Table 4.2 Summary statistics for the number of exotic observations and species recorded by trail position for all transects (n = 96 paired transects).

| Trail position | Number of transects that encountered at least one exotic species | Mean number of exotic observations | Range of number of exotic observations | Mean number of exotic species | Range of number of exotic species |
|---|---|---|---|---|---|
| Off | 38 | 1.10 | 0 - 15 | 0.89 | 0 - 11 |
| On | 72 | 2.97 | 0 - 15 | 2.43 | 0 - 10 |

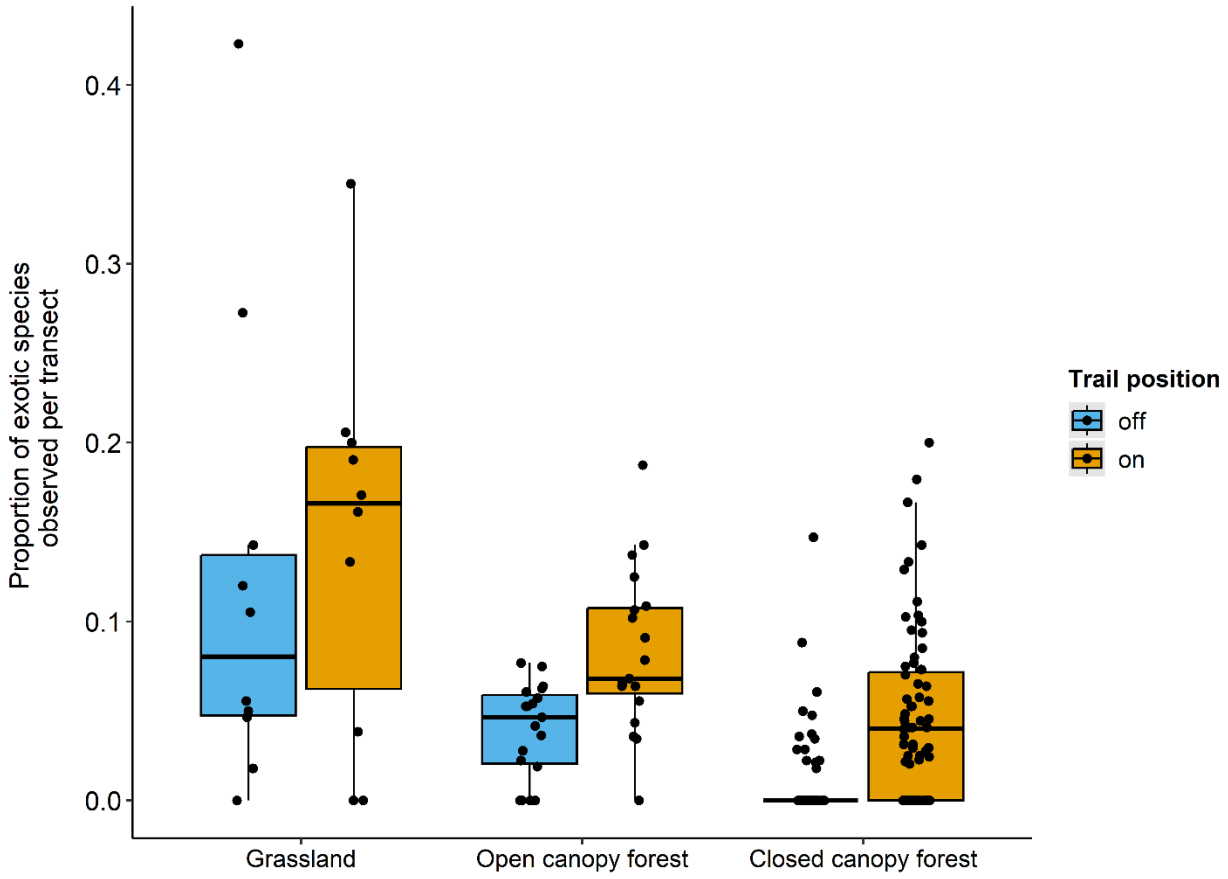



Figure 4.5 Boxplot of the proportion of exotic species observed per transect across the three habitat types. Proportion is calculated out of the total taxonomic richness observed per transect. Black bar = median. Box = interquartile range.

*4.5 Discussion*

Higher taxonomic richness observed on-trails than off-trails during the 30-minute biodiversity transects suggests that fine-scale bias (*i.e.,* trail bias) does not lead to reduced biodiversity measurements from opportunistic community science. The lack of difference between on and off trail transects for rare (*i.e.,* uncommonly recorded) species also supports the inference that trail bias does not limit the quality of biodiversity observations derived strictly from trails. These results are encouraging evidence for park managers and conservation researchers: they show that spatially biased data from opportunistic community science can provide comparable or stronger taxonomic richness estimates on-trails compared to off-trails in a variety of habitats.

### 4.5.1 Trail bias - taxonomic richness

Our test for effects of trails on estimates of taxonomic richness is novel because it includes all taxa in a variety of habitats in time-controlled transects where observers mimicked typical behaviour of people using community science. Previous results using standardized surveys have been mixed regarding trail impacts on species diversity (Ballantyne & Pickering 2015). Some studies have found higher species richness on-trails for vascular plants and ground-dwelling arthropods (Queiroz et al. 2014; Root-Bernstein & Svenning 2018; Swart et al. 2019; Wedegärtner et al. 2022), while other studies found no difference *e.g.,* plants (Jägerbrand & Alatalo 2015) or even higher species richness away from trails *e.g.,* lichens, flower-visiting insects, and gall-inducing insects (Jägerbrand & Alatalo 2015; Kamel 2020, 2021; De Almeida et al. 2022).

There are a few potential explanations for why I observed higher taxonomic richness on-trail than off-trail. It is possible that there were indeed more species on-trails, perhaps because

of more light and disturbance, which could favour ruderal species (Queiroz et al. 2014; Ballantyne & Pickering 2015; Root-Bernstein & Svenning 2018). However, species may also have been more readily detected from trails due to increased visibility and greater ease of observers moving through the habitat. As the number of taxa observed was highly correlated with number of observations, it is likely that observing from trails imparts advantages for biodiversity surveying, allowing for more species to be recorded than off-trail. Either way, if people want to see as many species as possible (and most naturalists do), the message is that they will do so by following trail etiquette.

### 4.5.2 Trail bias - species composition differences

My finding of more exotic species along trails than off-trails agrees with other studies (Ngugi et al. 2014; Liedtke et al. 2020) and provides reassuring evidence for potential rapid detection of new invasives in regions (Tiralongo et al. 2020; Werenkraut et al. 2020). This result might misleadingly give the impression that the higher taxonomic richness on-trails versus off-trails is due to exotic species locally boosting trailside taxonomic richness. However, when I removed exotic species from the models, I still found evidence of increased taxonomic richness of native taxa on trails.

The lack of difference in detection of rare species on versus off trails also suggests that observing from trails does not impose a disadvantage on biodiversity measurements. In addition, I found more vulnerable species along trails than off-trails, though the sample sizes were too small to make strong inferences. It is interesting to note that most of the observed vulnerable species were winged *i.e.,* mobile species (Fig. 4.4). It is thus possible that vulnerable

plants and lichens could be less commonly observed on-trail and consequently be underrepresented on iNaturalist (Jägerbrand & Alatalo 2015).

### *4.6 Conclusion*

In summary, my finding of higher taxonomic richness observed along trails than away from trails by iNaturalist observers is a promising outcome for the value of the platform for surveying biodiversity. Of course, there are still going to be strong biases due to the lack of sampling guidelines on iNaturalist, and spatial coverage will be imperfect. For example, some microhabitats (*e.g.,* logs, rocks, stumps) and mesohabitats (*e.g.,* seeps, cliffs, wetlands) and their inhabitants may be underrepresented by observers who stay on trails. Targeted surveys would be needed to complement community science for species in such places. In addition, the higher number of exotic species documented along trails and lack of difference detected for rare species supports the value of these data for tracking invasive species and for locating places where rare and vulnerable species occur. I hope that the results of more experiments like this study can further illuminate biases and enhance our ability to use the data effectively.

## Chapter 5: General discussion

### *5.1 Overview of results*

My research examines the spatial biases on the popular biodiversity community science platform, iNaturalist. This work contributes to our understanding of biases associated with opportunistically collected biodiversity datasets – a rapidly growing data type – allowing researchers to more effectively use these enormous datasets to answer large-scale ecological and conservation questions.

In Chapter Three, I investigated broad-scale spatial biases and their associated environmental drivers on iNaturalist, using British Columbia, Canada, as a case study. I predicted that distance to roads and human population density were the most important variables for explaining spatial bias on iNaturalist. I found that distance to roads was indeed the most important landscape feature but discovered a surprising result; the biogeoclimatic zone variable – a broad ecosystem classification system used in British Columbia – was more important than human population density in predicting observer spatial bias. I also modelled where high and low iNaturalist observer activity occurs across British Columbia, providing an example workflow and sampling effort spatial file that can be used for researchers interested in creating species distribution models using BC iNaturalist data. This study simultaneously considers multiple environmental variables when modeling the spatial bias on iNaturalist; an important step in furthering our knowledge of spatial bias on this popular community science platform.

In Chapter Four, I tested the effect of trail bias (*i.e.,* fine-scale spatial bias) on taxonomic richness estimates on and off trail in British Columbia using timed biodiversity surveys with

trained iNaturalist observers. I also tested for differences in species compositions between on and off trail using the number of exotic and rare (*i.e.,* uncommonly recorded) species recorded. I predicted that there would be higher overall taxonomic richness and higher number of exotic species on-trail compared to off-trail due to edge effects of trails often locally boosting biodiversity and exotic species (Ngugi et al. 2014; Liedtke et al. 2020; Wedegärtner et al. 2022). I found higher taxonomic richness and number of exotic species on-trails than off-trails. In addition, broad habitat type was an important factor for explaining taxonomic richness, with lower taxonomic richness in grasslands compared to closed-canopy forests. I detected no difference in number of rare species on and off trails. The results provide reassuring evidence for conservation managers and researchers interested in using iNaturalist data for biodiversity surveying but are concerned that trail bias negatively impacts data quality. The results suggest there will be no loss of biodiversity information by users sticking to trails for observations.

*5.2 Challenges and limitations*

For Chapter Three, I recognize that by limiting the analyses to British Columbia, it made spatial analysis smoother by collecting spatial data within a single political boundary. This allowed for relatively fine-scale cell resolution analysis, but may limit how much I can confidently apply our knowledge of each variable's importance to other locations around the world. Another limitation of the analyses in this chapter is that I did not consider different spatial resolutions (*e.g.,* 250 $m^2$ versus 5 $km^2$ grid cell size) when modeling variable importance, thus reducing the applicability of my results for species distribution models at different spatial resolutions. By only testing the finest resolution at my disposal, I followed best practices for species distribution models (Manzoor et al. 2018; Chauvier et al. 2022), but it is uncertain whether the

same variable, *e.g.,* distance to roads, is important across different resolutions. There is a possibility that different variables may have become important at coarser resolutions if other resolutions were tested since Maxent is sensitive to grain size (Martin et al. 2013; Manzoor et al. 2018). This is important to consider as many freely available environmental datasets have coarser resolutions such as the "WorldClim's" bioclimatic dataset with ~1 $km^2$ resolution (https://www.worldclim.org/) and many management programs work with resolutions greater than 250 $m^2$ (Shackelford et al. 2018). Thus, applicability of my results should be considered in the context of similar resolutions, as it is uncertain if the same trends hold at larger grain resolutions. That said, my analyses were done at relatively fine scales, which should reduce these problems, and the workflow is easily reproduced to support analyses at other resolutions.

In Chapter Four, there were challenges with sampling habitats evenly due to field experiments being limited by the team's schedules, trail availability, and weather conditions. Generalized linear mixed models can handle uneven test groups, but it would make a stronger study if I had a higher number of experiments conducted in grasslands to match the forest habitat types. In addition, the habitat type variable proved a challenge to accurately include in the model as it had to be categorized and simplified after the field season due to uncertainty of which parks the BC Parks iNaturalist team would visit that summer. It was difficult to plan a full field season when road and weather conditions so frequently affected access to parks, and thus required frequent changing of plans, a common problem in field ecology! I thus chose to work with only three habitat types, though more are available and could be the subject of future studies.

There was also the challenge of getting the true distance traveled by observers due to forest canopy interference with satellites. Thus, a proxy of distance travelled (*i.e.,* straight line) had to be used instead. Lastly, due to the high correlation of number of observations with taxonomic richness, there may be two interpretations of the results. It is unclear if higher taxonomic richness observed on-trails is due to higher number of taxa on-trails or because it is easier to observe more species while surveying on-trails.

*5.3 Future directions*

This research provides insight into the broad-scale spatial biases on iNaturalist, knowledge that can be used to improve species distribution models. It also demonstrates where community science likely contributes sufficient data for biodiversity modeling and where data poor regions could use targeted surveys to improve spatial coverage in British Columbia. In addition, my research provides novel experimental evidence that fine-scale spatial bias has little impact on biodiversity measurements from opportunistic community science. However, this research is just the beginning, and much work is needed to fully grasp the spatial biases present on iNaturalist. From the modelling perspective, there is more to explore with spatial bias in relation to different taxonomic groups on iNaturalist. For instance, Geldmann et al. (2016) found differences among fungi, birds, and common indicator species in how they were related and impacted by different landscape features. Another interesting avenue is to investigate whether different road types *e.g.,* highways, logging roads, residential streets, etc., have different effects on spatial sampling bias on iNaturalist (Geldmann et al. 2016). Given the low predicted probability of iNaturalist activity in the northeast of BC despite high road density (Fig.

3.2 c-d), it would be interesting to investigate if this is due to the high industry presence and public road access restrictions (Shackelford et al. 2018).

Another question that could be pursued is whether casual and active observers have different landscape features that influence where they observe. I know casual users (*i.e.,* users with fewer than five observations) on iNaturalist tend to observe more in developed areas than active users (Di Cecco et al. 2021). This idea could be expanded by examining multiple landscape features simultaneously.

There is also much to expand on in the fine-scale spatial bias study. It would be valuable to know if my results hold across specific taxonomic groups, where observers focus on particular taxa instead of sampling all taxa. In addition, more work is needed to determine whether the higher taxonomic richness observed on-trails is due to efficient sampling from observing on-trail or because it truly reflects higher taxonomic richness. One method to address this would be to re-do the experiments, but pair them with standardized surveys in the same locations to provide baseline data. This method would likely require limiting the surveys to one taxon or a few select taxa.

This thesis provides important options for dealing with some of the biases inherent in community science data, but there is no doubt that more study of biases is needed to allow for more effective use of these data. Nevertheless, I would like to end this discussion by briefly highlighting some of the amazing discoveries that iNaturalist has contributed. New discoveries are constantly being made by iNaturalist users in urban regions and roadsides. For instance, a large range expansion was documented for the Larger Pygmy Mole Grasshopper

(*Neotridactylus apicialis*) in summer 2022. The grasshopper was found inside a BC provincial park beside a road pull-out, expanding its known range substantially. Another example is a presumed extirpated (in BC) vascular plant, Munro's Globemallow (*Sphaeralcea munroana*), which was rediscovered near a road in 2022. In 2016, a new species record for Canada was found in an urban neighbourhood, the Paintedhand Mudbug (*Lacunicambarus polychromatus*). In 2019, a rare isopod, the Small-eyed Venezillo Pill Woodlouse (*Venezillo microphthalmos*), was rediscovered in the Bay Area. These are just a few of the many, many discoveries that iNaturalist community members have made, including in developed regions and near roads where I may have presumed there was no important biodiversity left to find. It is clear that opportunistic community science data can significantly help bolster our knowledge of biodiversity and enable the study of large-scale ecological questions. As shown in this thesis, these data are most effective when the biases are fully understood and explicitly accounted for in models.

## References


Arazy O, Malkinson D. 2021. A Framework of Observer-Based Biases in Citizen Science Biodiversity Monitoring: Semi-Structuring Unstructured Biodiversity Monitoring Protocols. Frontiers in Ecology and Evolution **9**:1–13.

Austin M, Buffett D, Nicolson D, Scudder G, Steven V. 2008. Taking nature's pulse: the status of biodiversity in British Columbia. Page (Austin M, Buffett D, Nicolson D, Scudder G, Steven V, editors). Biodiversity BC, Victoria, British Columbia. Available from http://www.biodiversitybc.org/assets/pressReleases/BBC_StatusReport_Web_final.pdf.

Avon C, Bergès L, Dumas Y, Dupouey JL. 2010. Does the effect of forest roads extend a few meters or more into the adjacent forest? A study on understory plant diversity in managed oak stands. Forest Ecology and Management **259**:1546–1555.

B.C. Conservation Data Centre. 2022. BC Species and Ecosystems Explorer.

Ballantyne M, Pickering CM. 2015. The impacts of trail infrastructure on vegetation and soils: Current literature and future directions. Journal of Environmental Management **164**:53–64. Elsevier Ltd. Available from http://dx.doi.org/10.1016/j.jenvman.2015.08.032.

Ballesteros-Mejia L, Kitching IJ, Jetz W, Nagel P, Beck J. 2013. Mapping the biodiversity of tropical insects: species richness and inventory completeness of African sphingid moths. Global Ecology and Biogeography **22**:586–595.

Barber RA, Ball SG, Morris RKA, Gilbert F. 2022. Target-group backgrounds prove effective at correcting sampling bias in Maxent models. Diversity and Distributions **28**:128–141.

Barve V V. et al. 2020a. Methods for broad-scale plant phenology assessments using citizen scientists' photographs. Applications in Plant Sciences **8**:1–10.

Barve V V. et al. 2020b. Methods for broad-scale plant phenology assessments using citizen scientists' photographs. Applications in Plant Sciences **8**:1–10.

Bates D, Maechler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software **67**:1–48.

BC iNaturalist Program. 2021. BC iNaturalist. Available from https://www.bcinat.com/ (accessed May 18, 2022).

BC Parks. 2018. BC Parks 2017/18 statistics report. Available from https://bcparks.ca/research/statistic_report/statistic-report-2017-2018.pdf?v=1611181148003.

BC Parks. 2020. Summary of the parks and protected areas system. Available from https://bcparks.ca/about/park-designations.html (accessed January 19, 2021).

BC Parks Foundation. 2020. Home | BC Parks Foundation. Available from https://bcparksfoundation.ca/ (accessed December 8, 2020).

Betts MG, Mitchell D, Diamond AW, Bety J. 2007. Uneven rates of landscape change as a source of bias in roadside wildlife surveys. Journal of Wildlife Management **71**:2266.

Bivand R, Keitt T, Rowlingson B. 2021. rgdal: Bindings for the "Geospatial" Data Abstraction Library. Available from https://cran.r-project.org/package=rgdal.

Boakes EH, McGowan PJK, Fuller RA, Chang-Qing D, Clark NE, O'Connor K, Mace GM. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biology **8**.

Bohl CL, Kass JM, Anderson RP. 2019. A new null model approach to quantify performance and significance for ecological niche models of species distributions. Journal of Biogeography **46**:1101–1111.

Bois ST, Silander JA, Mehrhoff LJ. 2011. Invasive plant atlas of new England: The role of citizens in the science of invasive alien species detection. BioScience **61**:763–770.

Bolker B, R Development Core Team. 2022. bbmle: Tools for General Maximum Likelihood Estimation. R package version 1.0.25. Available from https://cran.r-project.org/package=bbmle.

Bonney R, Ballard H, Jordan R, McCallie E, Phillips T, Shirk J, Wilderman CC. 2009. Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. A CAISE Inquiry Group Report. Washington, D.C.

Brown ED, Williams BK. 2019. The potential for citizen science to produce reliable and useful information in ecology. Conservation Biology **33**:561–569.

Busetto L, Ranghetti L. 2016. MODIStsp: an R package for preprocessing of MODIS Land Products time series. Computers & Geosciences **97**:40–48. Available from https://github.com/ropensci/MODIStsp.

Calenge C. 2006. The package adehabitat for the R software: tool for the analysis of space and habitat use by animals. Ecological Modelling **197**:516–519.

Callaghan CT, Ozeroff I, Hitchcock C, Chandler M. 2020. Capitalizing on opportunistic citizen science data to monitor urban biodiversity: A multi-taxa framework. Biological Conservation **251**:108753. Elsevier. Available from https://doi.org/10.1016/j.biocon.2020.108753.

Callaghan CT, Poore AGB, Hofmann M, Roberts CJ, Pereira HM. 2021. Large-bodied birds are over-represented in unstructured citizen science data. Scientific Reports **11**:1–11. Nature Publishing Group UK. Available from https://doi.org/10.1038/s41598-021-98584-7.

Canadian Endangered Species Conservation Council. 2022. Wild Species 2020: The General Status of Species in Canada.

Chauvier Y, Descombes P, Guéguen M, Boulangeat L, Thuiller W, Zimmermann NE. 2022. Resolution in species distribution models shapes spatial patterns of plant multifaceted diversity. Ecography **2022**:e05973.

Cooper CB, Shirk J, Zuckerberg B. 2014. The invisible prevalence of citizen science in global research: Migratory birds and climate change. PLoS ONE **9**.

Cornwell WK, Pearse WD, Dalrymple RL, Zanne AE. 2019. What we (don't) know about global plant diversity. Ecography **42**:1819–1831.

Courter JR, Johnson RJ, Stuyck CM, Lang BA, Kaiser EW. 2013. Weekend bias in citizen science data reporting: implications for phenology studies. International Journal of Biometeorology **57**:715–720.

Crimmins TM, Crimmins MA. 2022. Large-scale citizen science programs can support ecological and climate change assessments. Environmental Research Letters **17**.

Cull B. 2021. Potential for online crowdsourced biological recording data to complement surveillance for arthropod vectors. PLoS ONE **16**:1–25. Available from http://dx.doi.org/10.1371/journal.pone.0250382.

Damien Sulla-Menashe, Friedl MA. 2018. User guide to collection 6 MODIS Land Cover (MCD12Q1 and MCD12C1) Product. Available from https://lpdaac.usgs.gov/documents/101/MCD12_User_Guide_V6.pdf.

De Almeida ES, Sartori RA, Zaú AS. 2022. Trail Impacts in a Tropical Rainforest National Park. Geography, Environment, Sustainability **15**:5–12.

de Sherbinin A et al. 2021. The Critical Importance of Citizen Science Data. Frontiers in Climate **3**:1–7.

Di Cecco GJ, Barve V, Belitz MW, Stucky BJ, Guralnick RP, Hurlbert AH. 2021. Observing the observers: how participants contribute data to iNaturalist and implications for biodiversity science. BioScience **71**:1179–1188.

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

Dickinson JL, Zuckerberg B, Bonter DN. 2010. Citizen science as an ecological research tool: challenges and benefits. Annual Review of Ecology, Evolution, and Systematics **41**:149–172.

Drury JP, Barnes M, Finneran AE, Harris M, Grether GF. 2019. Continent-scale phenotype mapping using citizen scientists' photographs. Ecography **42**:1436–1445.

Egan B. 1997. The ecology of the mountain hemlock zone. Available from https://www.for.gov.bc.ca/hfd/pubs/docs/bro/bro51.pdf.

El-Gabbas A, Dormann CF. 2018. Improved species-occurrence predictions in data-poor regions: using large-scale data and bias correction with down-weighted Poisson regression and Maxent. Ecography **41**:1161–1172.

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ. 2011. A statistical explanation of MaxEnt

for ecologists. Diversity and Distributions **17**:43–57.

Environmental Reporting BC. 2018a. Trends in B.C.'s population size & distribution. State of Environment Reporting. British Columbia, Canada. Available from https://www.env.gov.bc.ca/soe/indicators/sustainability/bc-population.html.

Environmental Reporting BC. 2018b. Roads & Roadless Areas in British Columbia. Available from https://www.env.gov.bc.ca/soe/indicators/land/roads.html (accessed June 14, 2022).

Feldman MJ, Imbeau L, Marchand P, Mazerolle MJ, Darveau M, Fenton NJ. 2021. Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. PLoS ONE **16**:1–21. Available from http://dx.doi.org/10.1371/journal.pone.0234587.

Fernández D, Nakamura M. 2015. Estimation of spatial sampling effort based on presence-only data and accessibility. Ecological Modelling **299**:147–155. Elsevier B.V. Available from http://dx.doi.org/10.1016/j.ecolmodel.2014.12.017.

Filazzola A, Xie G, Barrett K, Dunn A, Johnson MTJ, Maclvor JS. 2022. Using smartphone-GPS data to quantify human activity in green spaces. PLoS Computational Biology **18**:e1010725. Available from http://dx.doi.org/10.1371/journal.pcbi.1010725.

Fithian W, Elith J, Hastie T, Keith DA. 2015. Bias correction in species distribution models: Pooling survey and collection data for multiple species. Methods in Ecology and Evolution **6**:424–438.

Fourcade Y, Engler JO, Rödder D, Secondi J. 2014. Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias. PLoS ONE **9**:1–13.

Geldmann J, Heilmann-Clausen J, Holm TE, Levinsky I, Markussen B, Olsen K, Rahbek C, Tøttrup AP. 2016. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. Diversity and Distributions **22**:1139–1149.

Government of Canada. 2017. British Columbia's provincial symbols. Available from https://www.canada.ca/en/canadian-heritage/services/provincial-territorial-symbols-canada/british-columbia.html (accessed December 6, 2020).

Greenwood JJD. 2007. Citizens, science and bird conservation. Journal of Ornithology **148**.

Grolemund G, Wickham H. 2011. Dates and Times Made Easy with lubridate. Journal of Statistical Software **40**:1–25. Available from https://www.jstatsoft.org/v40/i03/.

Hamilton SL et al. 2021. Disease-driven mass mortality event leads to widespread extirpation and variable recovery potential of a marine predator across the eastern Pacific.

Harris JBC, Haskell DG. 2007. Land cover sampling biases associated with roadside bird surveys. Avian Conservation and Ecology **2**.

Hausdorf B, Parr M, Shappell LJ, Oldeland J, Robinson DG. 2021. The introduction of the European Caucasotachea vindobonensis (Gastropoda: Helicidae) in North America, its origin and its potential range. Biological Invasions **23**:3281–3289. Springer International Publishing. Available from https://doi.org/10.1007/s10530-021-02579-4.

Hijmans RJ. 2021a. raster: geographic data analysis and modeling. R package version 3.5-2. Available from https://cran.r-project.org/package=raster.

Hijmans RJ. 2021b. terra: spatial data analysis. R package version 1.4-11. Available from https://cran.r-project.org/package=terra.

Hijmans RJ, Phillips S, Leathwick J, Elith J. 2021. dismo: species distribution modeling. R package version 1.3-5. Available from https://cran.r-project.org/package=dismo%0A.

iNaturalist. 2022. iNaturalist. Available from https://www.inaturalist.org (accessed May 11, 2022).

Isaac NJB, Pocock MJO. 2015. Bias and information in biological records. Biological Journal of the Linnean Society **115**:522–531.

Isaac NJB, van Strien AJ, August TA, de Zeeuw MP, Roy DB. 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. Methods in Ecology and Evolution **5**:1052–1060.

Jackson MM, Gergel SE, Martin K. 2015. Citizen science and field survey observations provide comparable results for mapping Vancouver Island White-tailed Ptarmigan (Lagopus leucura saxatilis) distributions. Biological Conservation **181**:162–172. Elsevier Ltd. Available from http://dx.doi.org/10.1016/j.biocon.2014.11.010.

Jägerbrand AK, Alatalo JM. 2015. Effects of human trampling on abundance and diversity of vascular plants, bryophytes and lichens in alpine heath vegetation, Northern Sweden. SpringerPlus **4**.

Jain P, Forbes H, Esposito LA. 2022. Two new alkali-sink specialist species of Paruroctonus Werner 1934 ( Scorpiones , Vaejovidae ) from central California **188**:139–188.

Johnston A, Fink D, Hochachka WM, Kelling S. 2018. Estimates of observer expertise improve species distributions from citizen science data. Methods in Ecology and Evolution **9**:88–97.

Johnston A, Hochachka WM, Strimas-Mackey ME, Ruiz Gutierrez V, Robinson OJ, Miller ET, Auer T, Kelling ST, Fink D. 2021. Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. Diversity and Distributions **27**:1265–1277.

Kadmon R, Farber O, Danin A. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecological Applications **14**:401–413.

Kamel M. 2020. Impact of hiking trails on the diversity of flower-visiting insects in Wadi Telah, St. Katherine protectorate, Egypt. The Journal of Basic and Applied Zoology **81**. The Journal of Basic and Applied Zoology.

Kamel M. 2021. Hiking trails effects on the diversity of gall-inducing insects in high altitude ecosystem, St. Katherine Protectorate, Egypt. Zoology in the Middle East **67**:48–56.

Kass JM, Muscarella R, Galante PJ, Bohl CL, Pinilla-Buitrago GE, Boria RA, Soley-Guardia M, Anderson RP. 2021. ENMeval 2.0: Redesigned for customizable and reproducible modeling of species' niches and distributions.

Kass JM, Muscarella R, Pinilla-Buitrago GE, Galante PJ. 2022. ENMeval 2.0 vignette. Available from https://jamiemkass.github.io/ENMeval/articles/ENMeval-2.0-vignette.html#resources.

Kelling S et al. 2015. Can observation skills of citizen scientists be estimated using species accumulation curves? PLoS ONE **10**:1–20.

Kobori H et al. 2016. Citizen science: a new approach to advance ecology, education, and conservation. Ecological Research **31**:1–19. Springer Japan.

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

Kullenberg C, Kasperowski D. 2016. What is citizen science? - A scientometric meta-analysis. PLoS ONE **11**:1–16.

La Sorte FA, Somveille M. 2020. Survey completeness of a global citizen-science database of bird occurrence. Ecography **43**:34–43.

Lehtinen RM, Carlson BM, Hamm AR, Riley AG, Mullin MM, Gray WJ. 2020. Dispatches from the neighborhood watch: Using citizen science and field survey data to document color morph frequency in space and time. Ecology and Evolution **10**:1526–1538.

Liedtke R, Barros A, Essl F, Lembrechts JJ, Wedegärtner REM, Pauchard A, Dullinger S. 2020. Hiking trails as conduits for the spread of non-native species in mountain areas. Biological Invasions **22**:1121–1134.

Loarie S. 2020. We've reached 1,000,000 observers! · iNaturalist. Available from https://www.inaturalist.org/blog/35758-we-ve-reached-1-000-000-observers (accessed March 10, 2021).

Loarie S. 2022a. We've passed 100,000,000 verifiable observations on iNaturalist! Available from https://www.inaturalist.org/blog/66531-we-ve-passed-100-000-000-verifiable-observations-on-inaturalist (accessed August 24, 2022).

Loarie S. 2022b. iNaturalist: What is it. Available from https://www.inaturalist.org/pages/what+is+it (accessed January 28, 2023).

Loss SR, Loss SS, Will T, Marra PP. 2015. Linking place-based citizen science with large-scale conservation research: A case study of bird-building collisions and the role of professional scientists. Biological Conservation **184**:439–445. Elsevier Ltd. Available from http://dx.doi.org/10.1016/j.biocon.2015.02.023.

MacPhail VJ, Gibson SD, Colla SR. 2020. Community science participants gain environmental awareness and contribute high quality data but improvements are needed: Insights from Bumble Bee Watch. PeerJ **2020**.

Mahajan S, Chung MK, Martinez J, Olaya Y, Helbing D, Chen LJ. 2022. Translating citizen-generated air quality data into evidence for shaping policy. Humanities and Social Sciences Communications **9**:1–18. Springer US.

Mair L, Ruete A. 2016. Explaining spatial variation in the recording effort of citizen science data across multiple taxa.

Mangiafico S. 2022. rcompanion: functions to support extension education program evaluation. R package version 2.4.15. Available from https://cran.r-project.org/package=rcompanion.

Manzoor SA, Griffiths G, Lukac M. 2018. Species distribution model transferability and model grain size-finer may not always be better. Scientific Reports **8**:1–9. Springer US. Available from http://dx.doi.org/10.1038/s41598-018-25437-1.

Marshall PJ, Lintott CJ, Fletcher LN. 2015. Ideas for citizen science in astronomy. Annual Review of Astronomy and Astrophysics **53**:247–278.

Martin Y, Van Dyck H, Dendoncker N, Titeux N. 2013. Testing instead of assuming the importance of land use change scenarios to model species distributions under climate change. Global Ecology and Biogeography **22**:1204–1216.

Matutini F, Baudry J, Pain G, Sineau M, Pithon J. 2021. How citizen science could improve species distribution models and their independent assessment. Ecology and Evolution **11**:3028–3039.

McMullin RT, Allen JL. 2022. An assessment of data accuracy and best practice recommendations for observations of lichens and other taxonomically difficult taxa on iNaturalist. Botany **100**:491–497.

McMullin RT, Wiersma YF. 2017. Lichens and allied fungi of Salmonier Nature Park, Newfoundland. Journal of the Torrey Botanical Society **144**:357–369.

Meidinger D, Pojar J. 1991. Ecosystems of British Columbia. Victoria, B.C.

Merow C, Smith MJ, Silander Jr JA. 2013. A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. Ecography **36**:1058–1069.

Mesaglio T, Callaghan CT. 2021. An overview of the history, current contributions and future outlook of iNaturalist in Australia. Wildlife Research **48**:289–303.

Mesaglio T, Callaghan CT, Samonte F, Gorta SBZ, Cornwell WK. 2023. Recognition and completeness: two key metrics for judging the utility of citizen science data. Frontiers in Ecology and the Environment:1–8.

Meyer C, Weigelt P, Kreft H. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecology letters **19**:992–1006.

Miller-Rushing A, Primack R, Bonney R. 2012. The history of public participation in ecological research. Frontiers in Ecology and the Environment **10**:285–290.

Neate-Clegg MHC, Horns JJ, Adler FR, Kemahlı Aytekin MÇ, Şekercioğlu ÇH. 2020. Monitoring the world's bird populations with community science data. Biological Conservation **248**:108653. Elsevier. Available from https://doi.org/10.1016/j.biocon.2020.108653.

Newmaster SG, Belland RJ, Arsenault A, Vitt DH, Stephens TR. 2005. The ones we left behind: comparing plot sampling and floristic habitat sampling for estimating bryophyte diversity. Diversity and Distributions **11**:57–72.

Ngugi MR, Neldner VJ, Dowling R. 2014. Non-native plant species richness adjacent to a horse trail network in seven National Parks in southeast Queensland, Australia. Australasian Journal of Environmental Management **21**:413–428. Taylor & Francis. Available from http://dx.doi.org/10.1080/14486563.2014.952788.

Nichols JD, Williams BK. 2006. Monitoring for conservation. Trends in Ecology and Evolution **21**:668–673.

Nowak K, Berger J, Panikowski A, Reid DG, Jacob AL, Newman G, Young NE, Beckmann JP, Richards SA. 2020. Using community photography to investigate phenology: A case study of coat molt in the mountain goat (Oreamnos americanus) with missing data. Ecology and Evolution **10**:13488–13499.

Oliveira U et al. 2016. The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. Diversity and Distributions **22**:1232–1244.

Parks Canada. 2022. Citizen science - Science and conservation. Available from https://www.pc.gc.ca/en/nature/science/impliquez-involved/science (accessed May 18, 2022).

Pebesma E. 2018. Simple features for R: standardized support for spatial vector Data. The R Journal **10**:439–446. Available from https://doi.org/10.32614/RJ-2018-009.

Pebesma EJ, Bivand RS. (n.d.). Classes and methods for spatial data in R. R News **5**:9–13. Available from https://cran.r-project.org/doc/Rnews/.

Petersen TK, Speed JDM, Grøtan V, Austrheim G. 2021. Species data for understanding biodiversity dynamics: The what, where and when of species occurrence data collection. Ecological Solutions and Evidence **2**:1–17.

Phillips S. 2017. A brief tutorial on Maxent. Available from http://biodiversityinformatics.amnh.org/open_source/maxent/ (accessed March 1, 2022).

Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S. 2009. Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. Ecological Applications **19**:181–197.

Phillips SJ, Dudík M, Schapire RE. 2020. [Internet] Maxent software for modeling species niches and distributions (Version 3.4.4.). Available from

https://biodiversityinformatics.amnh.org/open_source/maxent/ (accessed April 1, 2022).

Pocock MJO et al. 2018. A Vision for Global Biodiversity Monitoring With Citizen Science. Page Advances in Ecological Research, 1st edition. Elsevier Ltd. Available from http://dx.doi.org/10.1016/bs.aecr.2018.06.003.

Ponciano L, Brasileiro F, Simpson R, Smith A. 2014. Volunteers' engagement in human computation for astronomy projects. Computing in Science and Engineering **16**:52–59. IEEE.

Prudic KL, Oliver JC, Brown B V., Long EC. 2018. Comparisons of citizen science data-gathering approaches to evaluate Urban butterfly diversity. Insects **9**:1–10.

Queiroz RE, Ventura MA, Silva L. 2014. Plant diversity in hiking trails crossing Natura 2000 areas in the Azores: Implications for tourism and nature conservation. Biodiversity and Conservation **23**:1347–1365.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.r-project.org/.

Reddy S, Dávalos LM. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography **30**:1719–1727.

Redlands CESRI. 2017. ArcMap 10.6.1.

Reynolds JA, Lowman MD. 2013. Promoting ecoliteracy through research service-learning and citizen science. Frontiers in Ecology and the Environment **11**:565–566.

Ries L, Oberhauser K. 2015. A citizen army for science: Quantifying the contributions of citizen scientists to our understanding of monarch butterfly biology. BioScience **65**:419–430.

Roberts CJ, Vergés A, Callaghan CT, Poore AGB. 2022. Many cameras make light work: opportunistic photographs of rare species in iNaturalist complement structured surveys of reef fish to better understand species richness. Biodiversity and Conservation **31**:1407–1425.

Rocchini D, Hortal J, Lengyel S, Lobo JM, Jiménez-Valverde A, Ricotta C, Bacaro G, Chiarucci A. 2011. Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. Progress in Physical Geography **35**:211–226.

Root-Bernstein M, Svenning JC. 2018. Human paths have positive impacts on plant richness and diversity: A meta-analysis. Ecology and Evolution **8**:11111–11121.

Ross N. 2020. fasterize: fast polygon to raster conversion. R package version 1.0.3. Available from https://cran.r-project.org/package=fasterize%0A.

RStudio Team. 2021. RStudio: integrated development environment for R. RSudio, PBC, Boston, MA. Available from http://www.rstudio.com/.

Ruete A. 2015. Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. Biodiversity Data Journal **3**.

Ryan SF et al. 2018. The role of citizen science in addressing grand challenges in food and agriculture research. Proceedings of the Royal Society B: Biological Sciences **285**.

Saldivar JA, Romero AN, Wilson Rankin EE. 2022. Community science reveals high diversity of nectaring plants visited by painted lady butterflies (Lepidoptera: Nymphalidae) in California sage scrub. Environmental Entomology:1–9.

Scher CL, Clark JS. 2023. Species traits and observer behaviors that bias data assimilation and how to accommodate them. Ecological Applications:e2815.

Schoener TW. 1968. The Anolis lizards of Bimini: resource partitioning in a complex fauna. Ecology **49**:704–726.

Selva SB. 2003. Using calicioid lichens and fungi to assess ecological continuity in the Acadian Forest Ecoregion of the Canadian Maritimes. Forestry Chronicle **79**:550–558.

Shackelford N, Standish RJ, Ripple W, Starzomski BM. 2018. Threats to biodiversity from cumulative human impacts in one of North America's last wildlife frontiers. Conservation Biology **32**:672–684.

Shirey V, Belitz MW, Barve V, Guralnick R. 2021. A complete inventory of North American butterfly occurrence data: narrowing data gaps, but increasing bias. Ecography **44**:537–547.

Shirk JL et al. 2012. Public participation in scientific research: A framework for deliberate design. Ecology and Society **17**.

Smith AC, Hudson M-AR, Aponte VI, Francis CM. 2020. North American Breeding Bird Survey - Canadian Trends Website, Data-version 2019. Environment and Climate Change Canada, Gatineau, Quebec, K1A 0H3.

Speed JDM, Bendiksby M, Finstad AG, Hassel K, Kolstad AL, Prestø T. 2018. Contrasting spatial, temporal and environmental patterns in observation and specimen based species occurrence data. PLoS ONE **13**:1–17.

Stolar J, Nielsen SE. 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. Diversity and Distributions **21**:595–608.

Stoudt S, Goldstein BR, de Valpine P. 2022. Identifying engaging bird species and traits with community science observations. Proceedings of the National Academy of Sciences of the United States of America **119**:1–7.

Strathcona Wilderness Institute. 2022. iNaturalist. Available from https://strathconapark.org/swi-research/inaturalist-data/ (accessed May 18, 2022).

Sullivan BL, Wood CL, Iliff MJ, Bonney RE, Fink D, Kelling S. 2009. eBird: a citizen-based bird observation network in the biological sciences. Biological Conservation **142**:2282–2292. Elsevier Ltd. Available from http://dx.doi.org/10.1016/j.biocon.2009.05.006.

Surmacki A. 2005. What do data from birdwatchers notepads tell us? The case of the Bearded

Tit ( Panurus biarmicus ) occurrence in western Poland . Ring **27**:79–85.

Swart RC, Pryke JS, Roets F. 2019. The intermediate disturbance hypothesis explains arthropod beta-diversity responses to roads that cut through natural forests. Biological Conservation **236**:243–251. Elsevier. Available from https://doi.org/10.1016/j.biocon.2019.03.045.

Tang B, Clark JS, Gelfand AE. 2021. Modeling spatially biased citizen science effort through the eBird database. Environmental and Ecological Statistics **28**:609–630. Springer US. Available from https://doi.org/10.1007/s10651-021-00508-1.

Teucher A, Hazlitt S, Albers S. 2021. bcmaps: map layers and spatial utilities for British Columbia. R package version 1.0.2. Available from https://cran.r-project.org/package=bcmaps.

The Ministry of Forests Lands and Natural Resource Operations. 2013. Trails strategy for British Columbia. Available from https://www2.gov.bc.ca/assets/gov/sports-recreation-arts-and-culture/outdoor-recreation/camping-and-hiking/recreation-sites-and-trails/trail-strategy.pdf.

Theobald EJ et al. 2015. Global change and local solutions: tapping the unrealized potential of citizen science for biodiversity research. Biological Conservation **181**:236–244. Elsevier Ltd. Available from http://dx.doi.org/10.1016/j.biocon.2014.10.021.

Tiralongo F et al. 2020. Snapshot of rare, exotic and overlooked fish species in the Italian seas: A citizen science survey. Journal of Sea Research **164**:101930. Elsevier. Available from https://doi.org/10.1016/j.seares.2020.101930.

Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F. 2017. Taxonomic bias in biodiversity data and societal preferences. Scientific Reports **7**:1–14.

Tulloch AIT, Szabo JK. 2012. A behavioural ecology approach to understand volunteer surveying for citizen science datasets. Emu **112**:313–325. Available from https://doi.org/10.1071/MU12009.

Tye CA, McCleery RA, Fletcher RJ, Greene DU, Butryn RS. 2017. Evaluating citizen vs. professional data for modelling distributions of a rare squirrel.

Valavi R, Guillera-Arroita G, Lahoz-Monfort JJ, Elith J. 2022. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code.

Van Eupen C, Maes D, Herremans M, Swinnen KRR, Somers B, Luca S. 2022. Species profiles support recommendations for quality filtering of opportunistic citizen science data. Ecological Modelling **467**:109910. Elsevier B.V. Available from https://doi.org/10.1016/j.ecolmodel.2022.109910.

van Wilgenburg SL, Beck EM, Obermayer B, Joyce T, Weddle B. 2015. Biased representation of disturbance rates in the roadside sampling frame in boreal forests: implications for monitoring design. Avian Conservation and Ecology **10**.

Vetter J. 2011. Introduction: Lay participation in the history of scientific observation. Science in

Context **24**:127–141.

Walker DW, Smigaj M, Tani M. 2021. The benefits and negative impacts of citizen science applications to water as experienced by participants and communities. Wiley Interdisciplinary Reviews: Water **8**:1–32.

Walker J, Taylor PD. 2017. Using eBird data to model population change of migratory bird species. Avian Conservation and Ecology **12**.

Warren D, Dinnage R. 2022. ENMTools: analysis of niche evolution using niche and distribution models. R package version 1.0.6. Available from https://cran.r-project.org/package=ENMTools.

Wedegärtner REM, Lembrechts JJ, van der Wal R, Barros A, Chauvin A, Janssens I, Graae BJ. 2022. Hiking trails shift plant species' realized climatic niches and locally increase species richness. Diversity and Distributions **28**:1416–1429.

Werenkraut V, Baudino F, Roy HE. 2020. Citizen science reveals the distribution of the invasive harlequin ladybird (Harmonia axyridis Pallas) in Argentina. Biological Invasions **22**:2915–2921. Springer International Publishing. Available from https://doi.org/10.1007/s10530-020-02312-7.

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York. Available from https://ggplot2.tidyverse.org.

Wickham H, François R, Henry L, Müller K. 2021. dplyr: a grammar of data manipulation. R package version 1.0.7. Available from https://cran.r-project.org/package=dplyr.

Wolf S, Mahecha MD, Sabatini FM, Wirth C, Bruelheide H, Kattge J, Martínez ÁM, Mora K, Kattenborn T. 2022. Citizen science plant observations encode global trait patterns. Nature Ecology & Evolution **6**:1850–1859.

Zhang G. 2020. Spatial and temporal patterns in volunteer data contribution activities: a case study of eBird. ISPRS International Journal of Geo-Information **9**.

Zizka A, Antonelli A, Silvestro D. 2020. sampbias, a method for quantifying geographic sampling biases in species distribution data. Ecography **43**:1–8.

## Appendices

*Appendix S1*

Data preparation

Table S1 Descriptions of the MODIS land cover types based on the International Geosphere Biosphere Programme's global vegetation classification system (Damien Sulla-Menashe & Friedl 2018).

| ID | Name | Description |
|---|---|---|
| 1 | Evergreen Needleleaf Forests | Dominated by evergreen conifer trees (canopy >2m). Tree cover >60%. |
| 2 | Evergreen Broadleaf Forests | Dominated by evergreen broadleaf and palmate trees (canopy >2m). Tree cover >60%. |
| 3 | Deciduous Needleleaf Forest | Dominated by deciduous needleleaf (larch) trees (canopy >2m). Tree cover >60%. |
| 4 | Deciduous Broadleaf Forest | Dominated by deciduous broadleaf trees (canopy >2m). Tree cover >60%. |
| 5 | Mixed Forests | Dominated by neither deciduous nor evergreen (40-60% of each) tree type (canopy >2m). Tree cover >60% |
| 6 | Closed Shrublands | Dominated by woody perennials (1-2m height) >60% cover. |
| 7 | Open Shrublands | Dominated by woody perennials (1-2m height) 10-60% cover. |
| 8 | Woody Savannas | Tree cover 30-60% (canopy >2m). |
| 9 | Savannas | Tree cover 10-30% (canopy >2m). |
| 10 | Grasslands | Dominated by herbaceous annuals (<2m). |
| 11 | Permanent Wetlands | Permanently inundated lands with 30-60% water cover and >10% vegetated cover. |
| 12 | Croplands | At least 60% of area is cultivated cropland. |
| 13 | Urban and Built-up Lands | At least 30% impervious surface area including building materials, asphalt, and vehicles. |
| 14 | Cropland/Natural vegetation mosaics | Mosaics of small-scale cultivation 40-60% with natural tree, shrub, or herbaceous vegetation. |
| 15 | Permanent Snow and Ice | At least 60% of area is covered by snow and ice for at least 10 months of the year. |
| 16 | Barren or sparsely vegetated | At least 60% of area is non-vegetated barren (sand, rock, soil) areas with less than 10% vegetation. |
| 17 | Water Bodies | At least 60% of area is covered by permanent water bodies. |

Table S2 Summary of the variables used in the broad spatial pattern analysis of iNaturalist observation in British Columbia, Canada, with their definitions, qualities (*i.e.*, temporal and spatial resolutions), data types (*e.g.*, point, line, polygon, raster), and data sources.

| Datasets | Definition | Resolution | Type | Date produced yyyy-mm-dd | Source |
|---|---|---|---|---|---|
| **iNaturalist observations** | Georeferenced observations of animals, plants, fungi, and protozoans in BC created on iNaturalist (n = 1,769,501). Includes marine and terrestrial observations. | NA | Point | 2021-12-02 | iNaturalist |
| **Roads** | All transport lines within BC. Includes gravel, paved, decommissioned, overgrown, and seasonal roads, and boat routes. | NA | Line | 2020-11-06 | FLNRORD - GeoBC |
| **Provincial parks** | Spatial representation of BC parks, ecological reserves, and protected areas. | NA | Polygon | 2019-03-19 | BC Parks (ArcGIS Hub) |
| **National parks** | Administrative boundaries of Canada Lands. Includes National Parks, National Park Reserves, National Marine Conservation Areas and Aboriginal Land Claims Settlement Areas | NA | Polygon | 2019-08-08 | Natural Resources Canada |
| **MODIS Land Product data – MCD12Q1** | Land cover data of British Columbia based on the IGBP vegetation classification scheme. Downloaded the 500 m layer and reprojected using the MODIStsp package which resulted with cell resolution of 277 m | 277 meters | Raster | 2020-01-01 | NASA |

| | | | | | |
|---|---|---|---|---|---|
| **Biogeoclimatic Ecosystem Classification (BEC)** | Map of the different biogeoclimatic zones in British Columbia | NA | Polygon | 2021-09-02 | Forest Analysis and Inventory |
| **Canadian Digital Elevation Model** | Digital Elevation Model for British Columbia. Data is the TRIM DEM converted to the Canadian Digital Elevation Data format | 25 meters | Raster | 2014-12-10 | GeoBC |
| **Human population** | Global population distribution using a combination of GIS, remote sensing technology, and machine learning algorithms. | 100 meters | Raster | 2020-07 | Oak Ridge National Laboratory - LandScan |
| **BC terrestrial boundary** | Spatial representation of BC | NA | Polygon, line | 2020-06-25 | FLNRORD – GeoBC |

**Measure of Association – Categorical Variables**

Cramer's V measure of association between the environmental variables, BEC (Biogeoclimatic Ecosystem Classification) and land cover type was 0.166.

Data analysis outputs

Table S3 ENMeval null model outputs. AUC = Area Under Curve for training occurrences. CBI = Continuous Boyce Index for training occurrences. AUC Val = Area Under Curve for validation occurrences. AUC Diff = Minimum difference between training and test data. CBI Val = Continuous Boyce Index for validation occurrences. OR.MTP = Minimum Training Presence' omission rate. OR.10p = 10% training omission rate.

| Statistic | AUC train | CBI train | AUC Val | AUC Diff | CBI Val | OR.MTP | OR.10p |
|---|---|---|---|---|---|---|---|
| **Empirical mean** | 0.91 | 0.9990 | 0.906 | 0.000701 | 0.999 | 0.00000675 | 0.10 |
| **Empirical standard deviation** | NA | NA | 0.000685 | 0.000382 | 0.000447 | 0.00001501 | 0.0017 |
| **Null mean** | 0.63 | 0.9995 | 0.904 | 0.2670 | 1.000 | 0.000000203 | 0.0039 |
| **Null standard deviation** | 0.00082 | 0.000745 | 0.000132 | 0.000228 | < 0.00001 | 0.00000259 | 0.00012 |
| **Z score** | 332.6 | -0.7 | 15.3 | -1181.2 | < -0.00001 | 2.5 | 821.5 |
| **P value** | < 0.00001 | 0.75 | < 0.00001 | < 0.00001 | 1.00 | 0.99 | 1.0 |

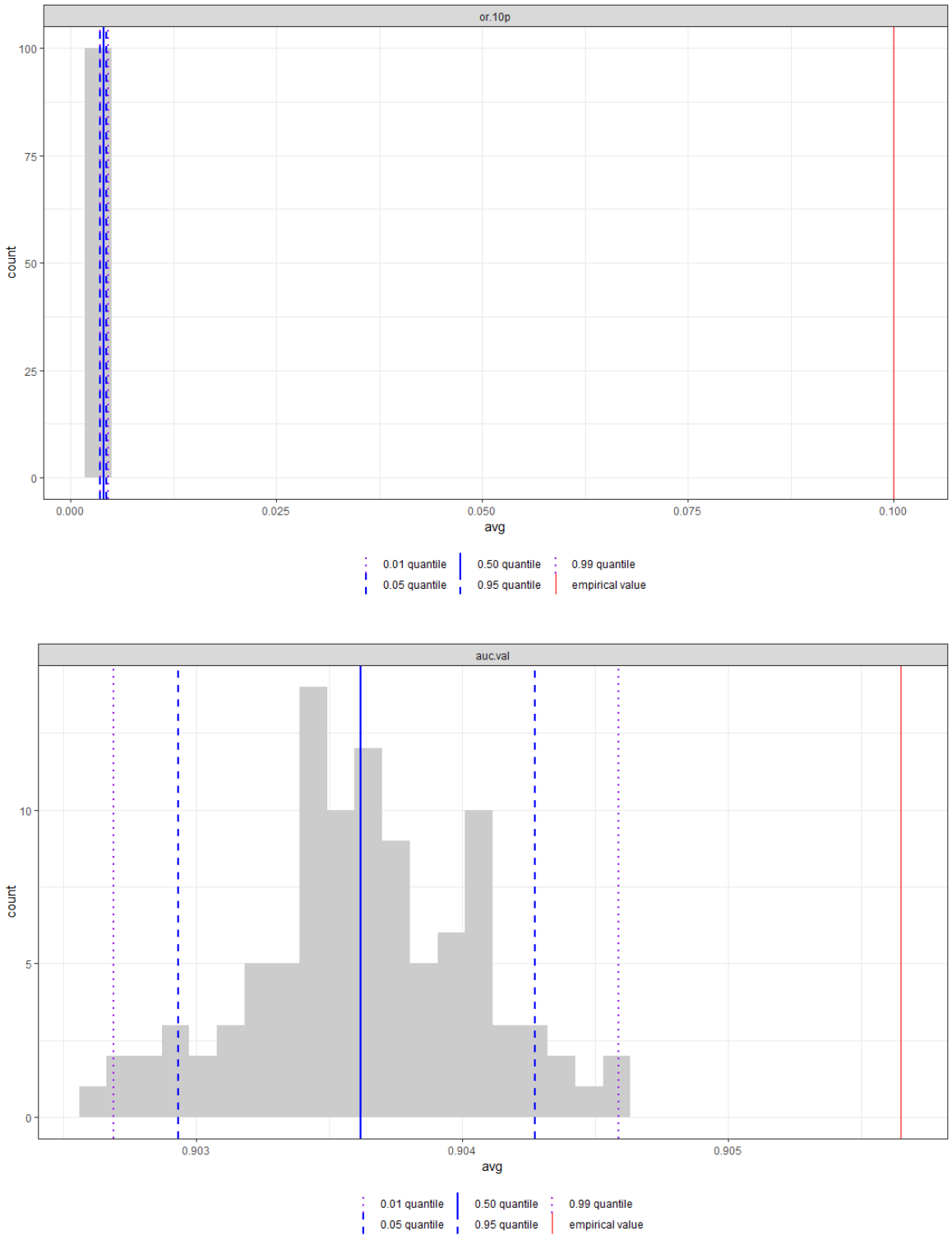



**Figure S1** Two metrics to compare null Maxent model with the top Maxent model; 10% training omission rate and validation AUC.

The jackknife test of variable importance showed the environmental variable BEC (Biogeoclimatic Ecosystem Classification) had the highest model gain when used in isolation, which suggests it contains the most useful information out of all the variables (Fig. S2). In contrast, the park land variable had the lowest gain when used in isolation, indicating it contains the least amount of useful information for predicting probability occurrence *i.e.,* the park land variable was the worst environmental variable to use on its own to predict whether an iNaturalist observer is likely to observe in a particular location (Fig. S2). The distance to road environmental variable decreased the model gain the most when omitted, which suggests it contains the most unique information out of all the variables. The Maxent model gain decreased the least when the MODIS land cover variable was excluded, suggesting it contained the least amount of unique information out of the six variables (Fig. S2).

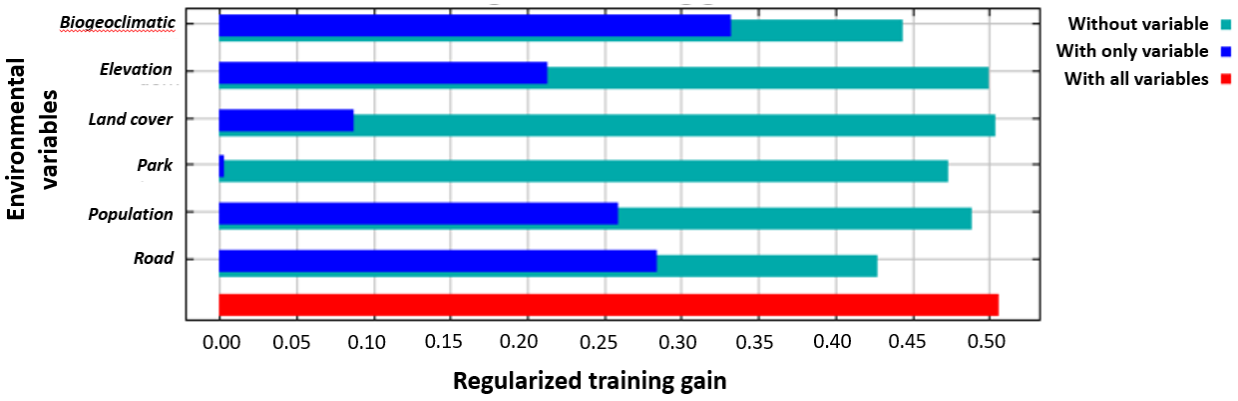



Figure S2 Jackknife test of variable importance using the regularized training gain for the top Maxent model.

The predicted probabilities for the Biogeoclimatic Ecosystem Classification zones from the isolated response curves were correlated with the mean population densities for those zones (Kendall's Tau = 0.71, z = 3.83, p-value = 0.000127).

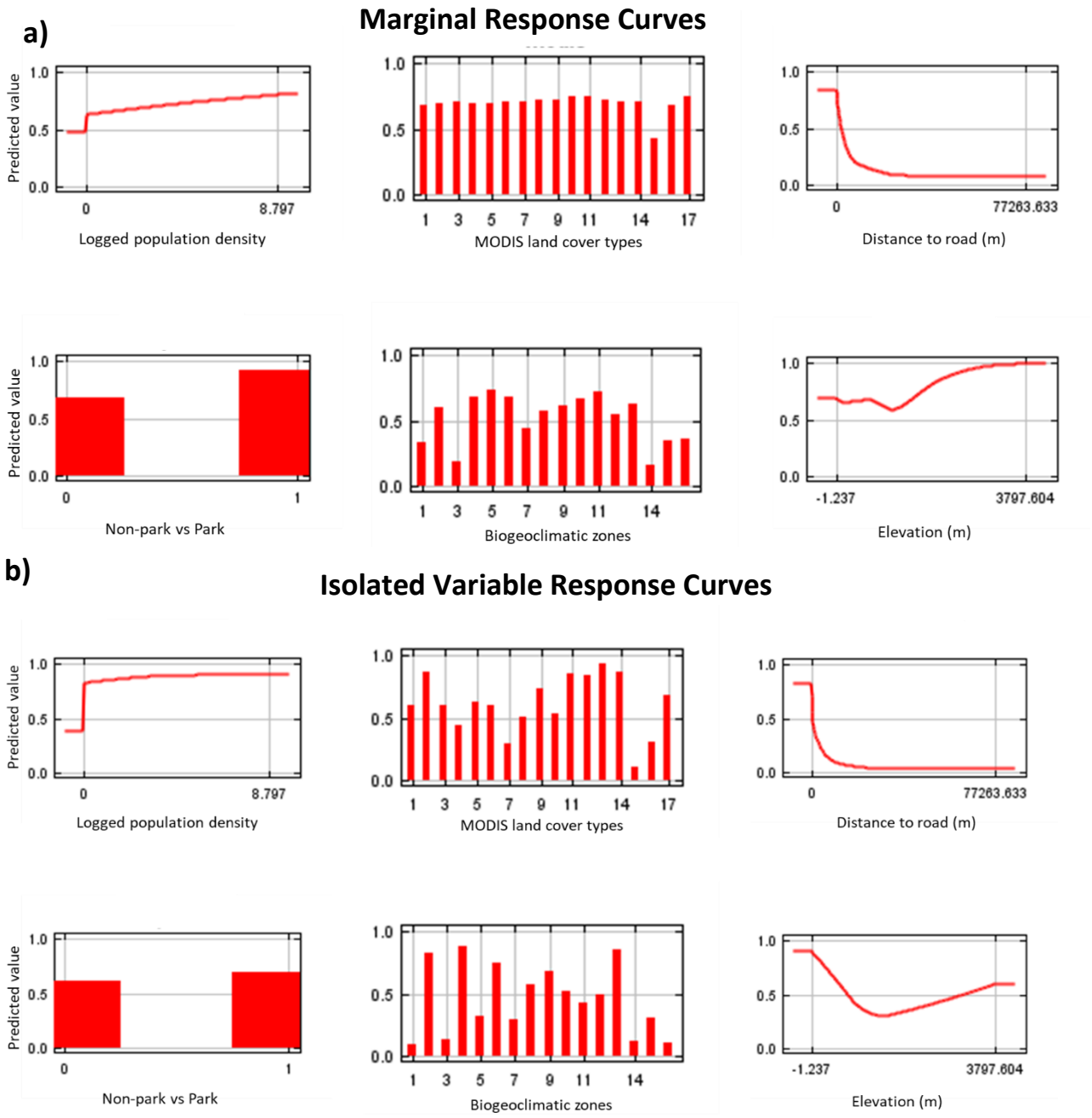



Figure S3 a) Marginal response curves for the top Maxent model. b) Response curves of the environmental variables used in isolation in the top Maxent model. Biogeoclimatic zones: 1 = Boreal Altai Fescue Alpine, 2 = Bunchgrass, 3 = Boreal White and Black Spruce, 4 = Coastal Douglas-fir, 5 = Coastal Mountain-heather Alpine, 6 = Coastal Western Hemlock, 7 = Engelmann Spruce – Subalpine Fir, 8 = Interior Cedar – Hemlock, 9 = Interior Douglas-fir, 10 = Interior Mountain-heather Alpine, 11 = Mountain Hemlock, 12 = Montane Spruce, 13 = Ponderosa Pine, 14 = Sub-Boreal Pine – Spruce, 15 = Sub-Boreal Spruce, and 16 = Spruce – Willow – Birch. See Appendix S1 Table S1 for MODIS land cover type descriptions.

*Appendix S2*

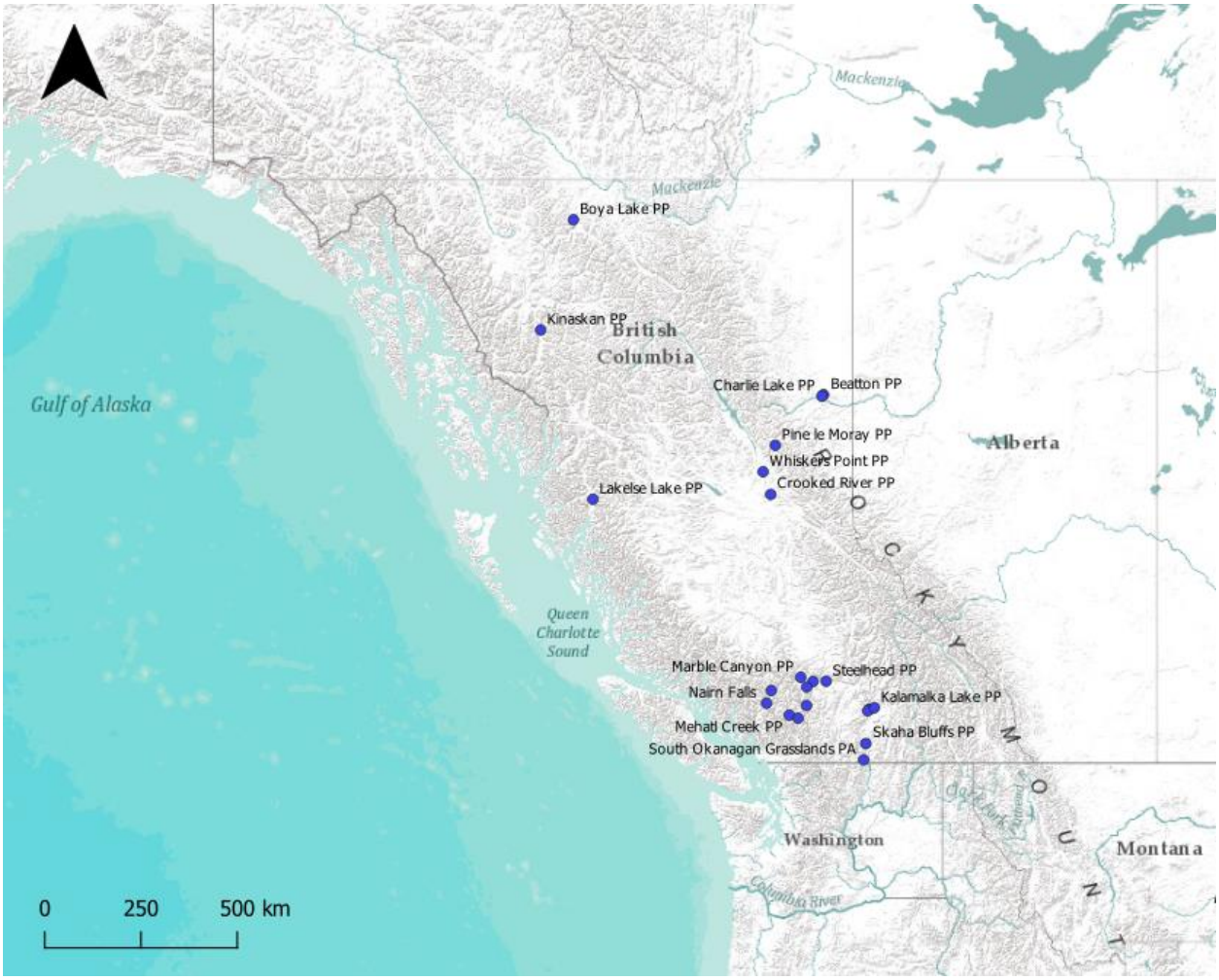



**Figure S1** The 22 provincial parks and protected areas where I conducted transects between May and August 2021, in British Columbia, Canada. I used the ESRI Terrain and Reference Overlay basemaps from QGIS.

Table S1 Provincial parks and protected areas visited along with the number of paired transects conducted and the dominant vegetation encountered for each site. Dominant vegetation types are the tallest trees and shrubs with the most cover. PP = Provincial Park. PA = Protected Area.

| Park name | Number of paired transects | Dominant vegetation #1 | Dominant vegetation #2 | Dominant vegetation #3 |
|---|---|---|---|---|
| **Beatton PP** | 14 | *Populus* sp. | *Picea* sp. | *Populus* sp. & *Picea* sp. |
| **Birkenhead Lake PP** | 8 | *Tsuga heterophylla*. & *Thuja plicata* | *Thuja heterophylla* & *Pseudotsuga menziesii* | *Acer macrophyllum* |
| **Boya Lake PP** | 2 | *Populus* sp. | *Populus* sp. & *Picea* sp. | - |
| **Charlie Lake PP** | 8 | *Populus* sp. | - | - |
| **Crooked River PP** | 6 | *Pinus contorta* | *Alnus* sp. & *Abies* sp. | *Pseudotsuga menziesii* |
| **Elephant Hill PP** | 1 | *Artemisia* sp. | - | - |
| **Ellison PP** | 6 | *Pseudotsuga menziesii* & *Pinus ponderosa* | *Pseudotsuga menziesii* | - |
| **Fintry PP** | 4 | *Populus* sp. | *Pseudotsuga menziesii* & *Pinus ponderosa* | *Medicago sativa* |
| **Kalamalka Lake PP** | 8 | *Symphoricarpos* sp. or *Amelanchier* sp. | *Pinus ponderosa* | *Pseudotsuga menziesii* |
| **Kinaskan PP** | 3 | *Picea* sp. & *Abies* sp. | *Picea* sp. & *Pinus contorta* | - |
| **Lakelse Lake PP** | 3 | *Tsuga heterophylla* | *Tsuga heterophylla* & *Populus* sp. | - |
| **Marble Canyon PP** | 2 | *Pseudotsuga menziesii* | - | - |
| **Mehatl Creek PP** | 3 | *Thuja plicata* & *Pseudotsuga menziesii* | *Pseudotsuga menziesii* | *Tsuga heterophylla* |
| **Nahatlatch PP** | 3 | *Pseudotsuga menziesii* | *Thuja* sp. & *Pseudotsuga menziesii* | - |
| **Nairn Falls PP** | 4 | *Thuja plicata* & *Tsuga heterophylla* | *Tsuga heterophylla* | *Thuja plicata* |
| **Oregon Jack PP** | 4 | *Pseudotsuga menziesii* | - | - |
| **Pine le Moray PP** | 4 | *Picea* sp. | *Picea* sp. & *Betula* sp. | *Pseudotsuga menziesii* |
| **Skaha Bluffs PP** | 4 | *Pinus ponderosa* | *Salix* sp. | *Amelanchier alnifolia* or *Philadelphus lewisii* |
| **Skihist PP** | 4 | *Pseudotsuga menziesii* & *Pinus ponderosa* | - | - |
| **South Okanagan Grasslands PA** | 2 | *Pseudotsuga menziesii* | *Pseudotsuga menziesii* & *Pinus ponderosa* | - |
| **Steelhead PP** | 1 | *Ericameria nauseosa* | - | - |
| **Whiskers Point PP** | 2 | *Pseudotsuga menziesii* | - | - |
| **Total** | **96** | | | |