Increasing Access to Biodiversity Data in metrovancouver

SERVICES AND SOLUTIONS FOR A LIVABLE REGION

Gabriel Smith
Lesley Miller
Raghav Aggarwal

2019

UBC

UBC
Data Science
Institute

Table of Contents

1. Introduction

1.1 What is Data Science for Social Good?

The Data Science for Social Good program, administered through the University of British Columbia Data Science Institute, is a 14 week fellowship program aimed at developing solutions to problems with social impact while providing data science training to UBC graduate and undergraduate students. UBC students partner with organizations in the community to work on data science projects that contribute to the public good. To greatly enrich the research experience, the program aims to recruit fellows from a diversity of fields. This particular project partners with Metro Vancouver to determine how biodiversity data can be made easily accessible to regional planners to assist in making land-use decisions.

1.2 Defining Biodiversity

Biodiversity is defined to be all living things on earth and includes the environmental context in which they live. Biodiversity encompasses multiple scales, from the molecular level all the way to entire ecosystems; it not only encompasses differences between and within species but also differences between habitats such as soil composition, temperature, elevation and tree cover.

1.3 Ecosystem Services and Biodiversity as a Social Good

Broadly defined, ecosystem services are the benefits that humanity receives from properly functioning ecosystems (Fig. 1). While some of these services such as the provision of water, wood, natural gas are quite apparent, others such as the regulation of air composition and the disposal of waste products are not obvious (Bouma, & Van Beukering, 2015; Costanza et al., 1997). Humans rely on microorganisms to generate healthy soils for plants, insects to pollinate food crops and wetlands to purify water (Bouma, & Van Beukering, 2015; Costanza et al., 1997, WWF, 2018). Almost all pinnacles of human development and progress have relied on the extraction and use of natural resources. From the diversity of medicines generated by thousands of plant species to mental and physical health benefits derived from inhabiting green spaces, the preservation of human wellbeing requires a diversity of species that remain healthy and intact (WWF, 2018). It is therefore of vital importance to recognize the economic as well as spiritual and aesthetic value that ecosystems provide and do everything possible to protect it from ongoing threats. In a conservative attempt to put a value on their contribution to the totality of the human economy, Costanza et al. (1997) estimated these combined ecosystem services to be over 33 trillion USD annually (in 1994 dollars) and made it clear that as ecosystems are degraded with time, the services they provide become far scarcer and more valuable.

Biodiversity is the foundation of any ecosystem, and is thus a cornerstone of all ecosystem services (Habib, 2016; Ninan, 2009). Ecosystems are only as strong as the species and habitats that compose them; for instance, loss of soil microbes interferes with an ecosystem's ability to replenish nutrients and regulate the chemical composition of the air. Loss of pollinators can prevent plants from flourishing which in turn has downstream impacts on the species that rely on these plants. Documenting biodiversity thus becomes imperative for ensuring the sustainability of an ecosystem and the services it provides.

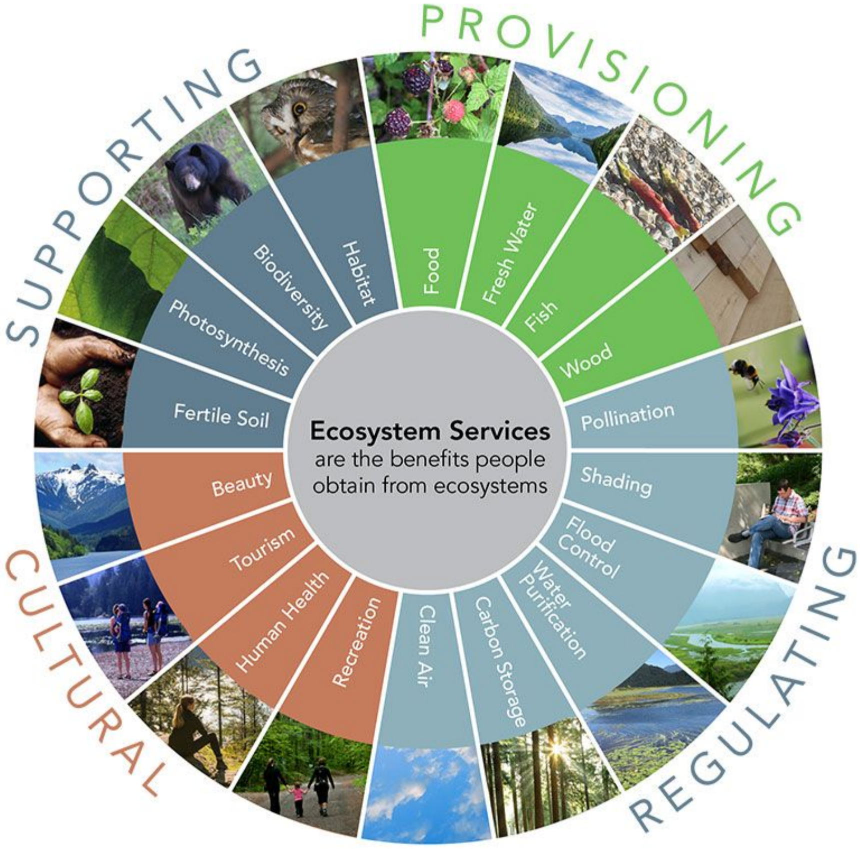


Figure 1. Ecosystem services supported by biodiversity.

1.4 Threats to Biodiversity

There are numerous threats to sustaining healthy biodiversity, chiefly among them are unsustainable agricultural practices, overuse of species (ie overfishing) and habitat loss through development decisions (Isbell, 2010; Martinez-Ramos, Ortiz-Rodriguez, Pinero, Dirzo, & Sarukhan, 2016; WWF, 2018). All of these pressures stem from unsustainable levels of human consumption. The World Wildlife Foundation's Living Planet Index (LPI) that tracks vertebrate species populations over time has shown a 60% species decline since 1970, a trend that seems likely to continue (Fig. 2). The removal of sensitive ecosystems has a multitude of negative impacts, including habitat loss, reduced rain coverage, and increased soil erosion, all of which threaten biodiversity in the regions around human habitation.

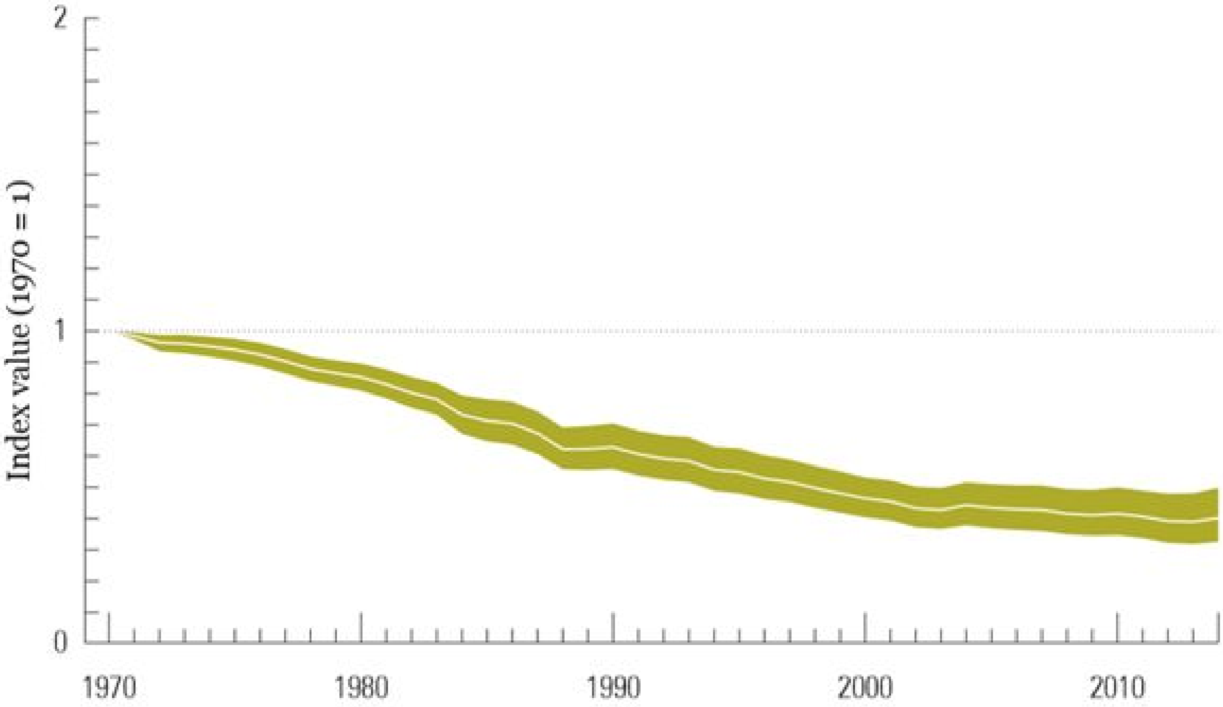


Figure 2. The global Living Planet Index shows a 60% decline between 1970 and 2014. The white line shows the index values and the shaded areas represent the 95% confidence limits surrounding the trend (WWF, 2018)

1.5 Barriers to Biodiversity Integration in Land Use Decisions

Since habitat loss due to land use decisions is a key threat to maintaining biodiversity, it is of critical importance that urban and regional planners have access to up-to-date, easily accessible data on biodiversity. At present in Metro Vancouver, easy access to biodiversity data has been an ongoing challenge. Planners must consult multiple websites on numerous platforms to get the information they need. Some biodiversity tools were built for the national and international scale leaving regional and municipal planners with data at too coarse a scale for local decision making. Protecting biodiversity in a land use context is incredibly complex and requires multiple components that no single source provides. A planner needs access to not only species occurrence data but also ecosystem classification information and endangered species lists at International, National and Provincial levels. Due to these various data access barriers, this leaves the inclusion of biodiversity in decision making to be tenuous at best.

1.6 Citizen Science

Citizen Science refers to the participation of non-expert members of the public in the process of scientific inquiry. In the field of biodiversity, data collection is the chief contribution of citizen scientists. By documenting observations of organisms with applications such as iNaturalist and e-Bird and having their reports later verified by experts, the public can produce a volume of data that would otherwise be virtually impossible to obtain. While there will always be concerns about the validity of this methodology, citizen science initiatives provide a vast and rich source of data with greater benefits than drawbacks.

1.7 Project Aims

This project constitutes the first attempt to survey the state of biodiversity data available for the entire Metro Vancouver region. The five main goals of the project include: 1) Discover the types of biodiversity data that are available (ie species occurrence, habit information), 2) Assess what data should be included in a map viewer, 3) Assess the source and quality of the data, and 4) Explore what kinds of questions that could be answered with the web viewer, 5) Identify data gaps by establishing a baseline inventory of Metro Vancouver species.

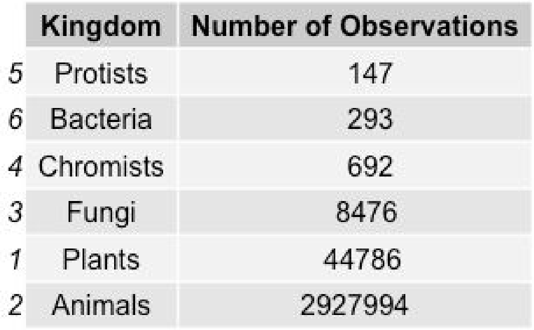
2. Data Sources & Description

A central goal of the project was to gather biodiversity data from a variety of sources. The data descriptions below detail the various sources of data that were integrated together. All data sources are publicly accessible.

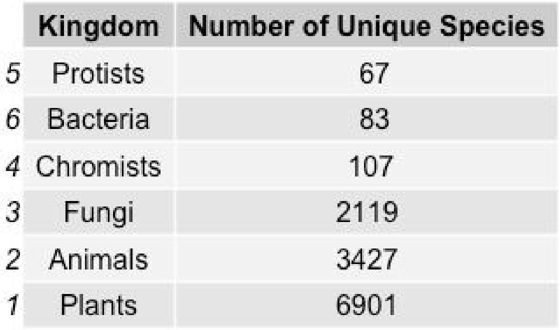
2.1 Global Biodiversity Information Facility

The Global Biodiversity Information Facility (GBIF) is an international web database that houses raw species occurrence data. GBIF integrates the data from a multitude of sources such as citizen science projects, natural history collections and academic research institutions all while making them accessible to the public. It is a valuable resource that provides species occurrence records in a format that is easy to analyze and manipulate. As of 2019, there are 235 data sources for Metro Vancouver with about 3 million species records spanning from 1700 to the present day. Each species record provides a variety of details such as the institutional source of the data, latitude and longitude, year the organism was observed as well as full taxonomic information. Though GBIF integrates data from multiple biodiversity projects, for Metro Vancouver, 95% of the data derives from citizen science projects such as eBIRD and iNaturalist with the remaining 5% originating from academic institutions such as the Royal BC Museum and the UBC Herbarium. Though animals and more specifically birds compose the greatest number of records, plants actually have a greater number of unique species observations, comprising a little over 50% of the total unique species (Table 1).

Table 1. *Comparison of Kingdom Observation Counts vs. Unique Species*



	Kingdom	Number of Observations
5	Protists	147
6	Bacteria	293
4	Chromists	692
3	Fungi	8476
1	Plants	44786
2	Animals	2927994



	Kingdom	Number of Unique Species
5	Protists	67
6	Bacteria	83
4	Chromists	107
3	Fungi	2119
2	Animals	3427
1	Plants	6901

The GBIF data constitutes ~ 10,500 unique species observed in ~30,000 unique localities across 319 years for the Metro Vancouver region. Though the data spans over 300 years, 75% of the species have been recorded within the last 10 years and 89% of the data has been recorded in the last 20 (Figure 3).

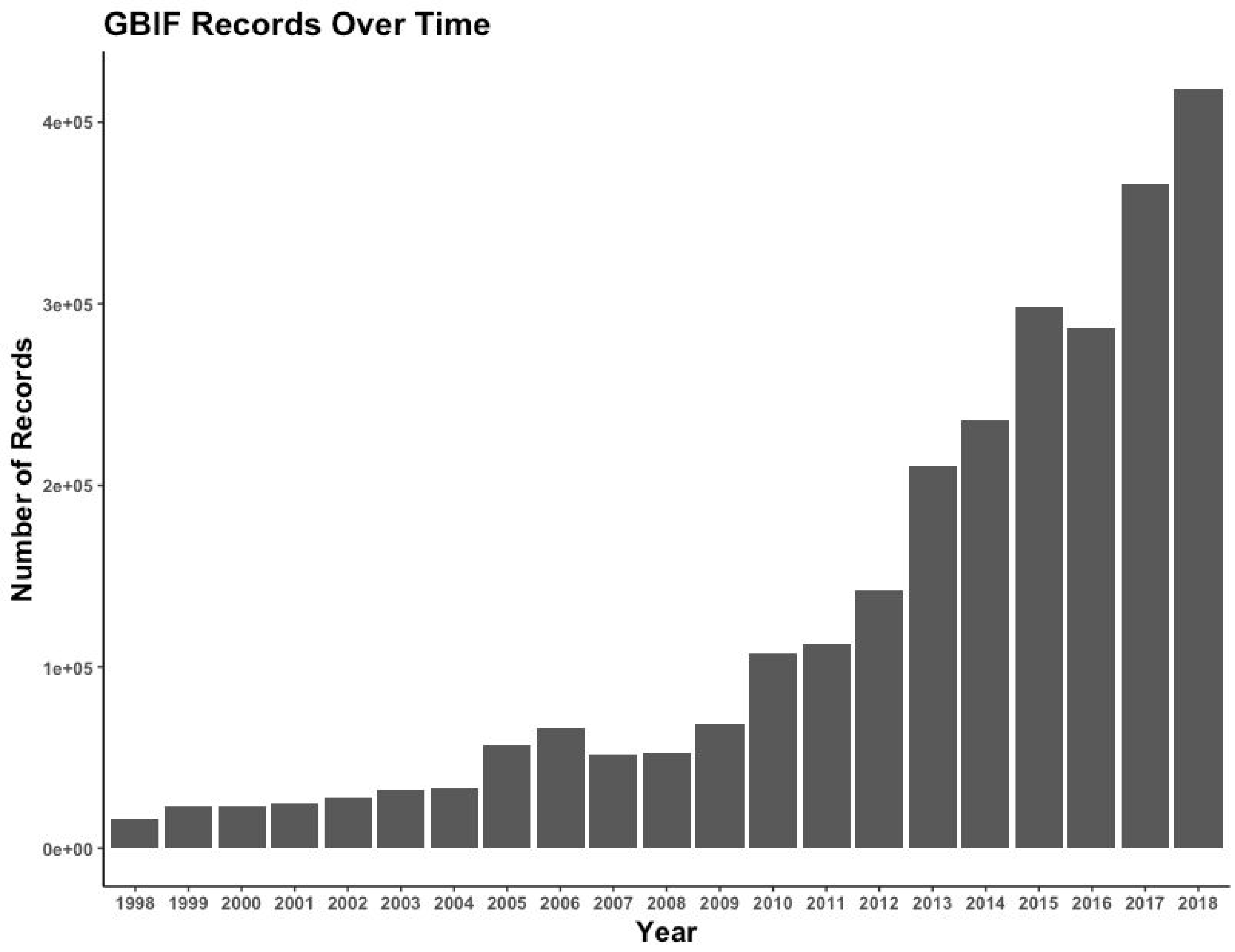


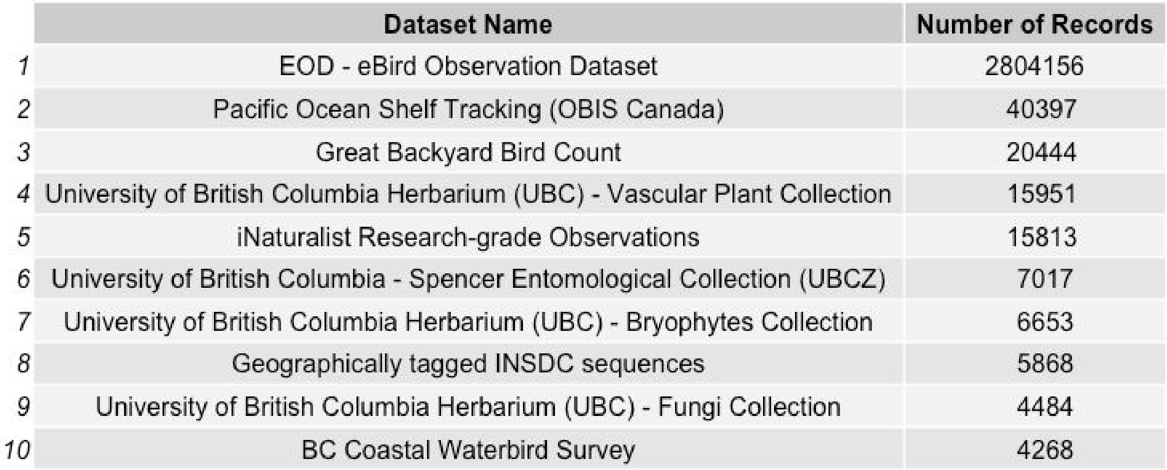
Figure 3. The rise of GBIF records over the 20 year period between 1998-2018

2.1.1 GBIF Data Biases

An exploration of the GBIF data reveals that 94% of the records come from eBIRD, a citizen science platform for collecting bird species observations (Table 2). This creates a distinct bias in the data and demonstrates how most species will be either underrepresented or not present in the data at all. This implicit bias towards birds has a number of consequences on subsequent data analyses and their interpretations.

First, the bias towards birds means that for some of the most diverse organismal groups such as microorganisms and fungi there is almost no information. This places a limitation on getting the full scope of biodiversity as it operates from the micro to the macroscopic level. Secondly, the SEI species predictions, (see section 5.5.1) are strongest for birds since they are sampled more heavily. SEI predictions cannot be made for the large majority of species since they are either not observed in any SEI polygon or have been observed in too few polygons to make reliable predictions. Thirdly, the same issue also arises for Species Distribution Modelling (see section 5.5.2) as the algorithms required to do such modelling require at least 30-90 observations. This precludes almost all fungi and a great number of plants from being analyzed.

Table 2. *Top 10 GBIF Data Contributors*



	Dataset Name	Number of Records
1	EOD - eBird Observation Dataset	2804156
2	Pacific Ocean Shelf Tracking (OBIS Canada)	40397
3	Great Backyard Bird Count	20444
4	University of British Columbia Herbarium (UBC) - Vascular Plant Collection	15951
5	iNaturalist Research-grade Observations	15813
6	University of British Columbia - Spencer Entomological Collection (UBCZ)	7017
7	University of British Columbia Herbarium (UBC) - Bryophytes Collection	6653
8	Geographically tagged INSDC sequences	5868
9	University of British Columbia Herbarium (UBC) - Fungi Collection	4484
10	BC Coastal Waterbird Survey	4268

2.2 International Union for Conservation of Nature

The International Union for Conservation of Nature (IUCN) is an organization tasked with monitoring species at risk on a global scale as well as supporting public and private institutions in conservation efforts. The IUCN conducts assessments of species and determines in which of nine classifications it falls. A species’ status can fall anywhere between Least Concern to Critically Endangered with each level of classification increasing the organism’s risk of extinction. Currently there are 79 species in the Metro Vancouver region that are on the IUCN Red List with 7 considered to be Critically Endangered, 9 Endangered, 24 Vulnerable and 39 Near Threatened.

2.3 BC Species and Ecosystem Explorer

In addition to IUCN, the government of British Columbia also monitors species at risk within its provincial borders. The BC Species and Ecosystem Explorer openly provides raw data on species in BC that have been Red, Blue or Yellow listed, with Red and Blue being the most sensitive or vulnerable categories and the Yellow list including species that are considered to be common and relatively secure. There are currently 120 red listed species in Metro Van, along with 180 Blue listed.

2.4 Sensitive Ecosystem Inventory

The Sensitive Ecosystem Inventory (SEI) is a spatial dataset created by Metro Vancouver between 2010 and 2012 and then subsequently updated in 2014. (Metro Vancouver, 2014). It classifies regions within Metro Vancouver into 12 broad ecosystem categories such as Old forests (trees > 200 years old), Wetlands or Woodlands. These ecosystems have been identified as at-risk, fragile or are important habitats for biodiversity. In addition to Sensitive Ecosystems, the dataset also maps Modified Ecosystems; though they have been younger or more disturbed by human activity, Modified Ecosystems are natural ecosystems that provide important habitat for a variety of species.

2.5 International Taxonomic Information System

The International Taxonomic Information System (ITIS) provides standardized taxonomy information for species identification. This dataset was integrated to obtain the common names of species since this information is not provided by GBIF.

2.6 Data BC

The provincial government under the Ministry of Forests, Lands, Natural Resource Operations and Rural Development published 3 spatial datasets that were used to inform the degree of proximity a specific SEI polygon is to water (fresh and saltwater). The three spatial datasets utilized were the Freshwater Atlas for Rivers, Freshwater Atlas for Lakes and the NTS BC Coastline Polygons. Each dataset was used when making species predictions in SEI polygons.

2.7 WorldClim Version2

WorldClim is a dataset of 19 spatially interpolated bioclimatic variables (i.e., temperature, precipitation) obtained from weather stations between 1970 and 2000 at a resolution of ~ 1 km^2. The climate data were used as environmental predictors in species distribution models.

2.8 Elevation-API

Elevation-API is a product that offers a coarse-resolution (5km) free version of their API for retrieving altitude information. This API was used to associate each SEI with an elevation by retrieving the elevation at its centroid.

2.9 Canada Digital Surface Model

The Canada Digital Surface Model (CDSM) is a publically-available .75 second resolution altimetry map of Canada’s land surface created by Natural Resources Canada. The elevation data was retrieved in .tif format from the coordinates that corresponded to Metro Vancouver to use with species distribution modelling (NRC, 2015).

3. Gap Analysis

An ongoing interest for project partners was to identify the gaps in the data by finding the species or groups that are poorly represented or are not observed at all when there is an expectation that they would be. Identifying data gaps would aid in potentially harnessing citizen science projects to target the species that are the most data poor.

3.1 Gap Analysis Challenges

Identifying the species or higher level taxon groups that are poorly represented is not a straightforward task. For example, a species having few observations in the data could mean that it is poorly sampled despite being common in the landscape, or the number of observations could reflect the truth, that the species is in fact rare in nature. Without manual literature searches of every poorly represented species, the answer to this question is impossible to determine from observing the raw data alone.

Another significant challenge is identifying species that are not in the GBIF records but are expected to be in Metro Vancouver based on prior knowledge of the region. Identifying a data source representing prior knowledge of the region in a format that lends itself to analytics tools proved to be only partially fruitful.

The E-Fauna and E-Flora websites have endeavored to document the biodiversity of BC; however, these platforms were not specific to the Metro Vancouver region and did not contain species lists in a format that lends itself easily to high-throughput data analysis (e.g., PDFs).

The BC Yellow list is in an accessible format and it was hoped that it would constitute those species that are common or relatively secure in the region, encompassing every local species that is not on the BC Red or Blue lists. However, the BC list is a targeted list and is not meant to be an exhaustive for the region. The BC Yellow list does include common species that can expect to be found in Metro Van but it would not include all species you'd expect to find, for instance, the list is only several thousand entries long despite ranging across BC, whereas GBIF alone contains over ten-thousand unique species for Metro Vancouver.

In short, a comprehensive database of Metro Van species in a useable format does not currently exist. A great deal of time and resources would need to be spent to create a baseline database using the existing platforms. See section 6.3.3 on future opportunities for exploring species tagged as Metro Van on the BC List.

3.2 Data Poor Taxon Groups: Plants, Fungi, Microorganisms

Even though it is not clear what specific species are the most poorly represented due to lack of background knowledge, it is abundantly clear that plants as a group are undersampled as a whole; plants compose only 1.5% of the observation records on GBIF for the region of interest. Another indicator of undersampling is while the GBIF records are rising overall, they generally have not risen for plants over time. Non-vascular plants like mosses compose very little of the plant data with most of the plant data coming from vascular plants like trees or flowering plants. Fungi compose an even smaller share of the data at less than half of a percent and most of the fungi data comes from only two phyla. Microorganism data are the poorest still, including bacteria and protist groups that collectively provide about 1000 observations. Any future citizen science or bioblitz efforts would do well to purposely sample from these high level taxon groups.

4. Methodology

4.1 Software Development

The majority of code was written in Python and JavaScript. In addition to those languages, the time-series plot was coded in R and communicates with the otherwise separate python map-viewer through a URL.

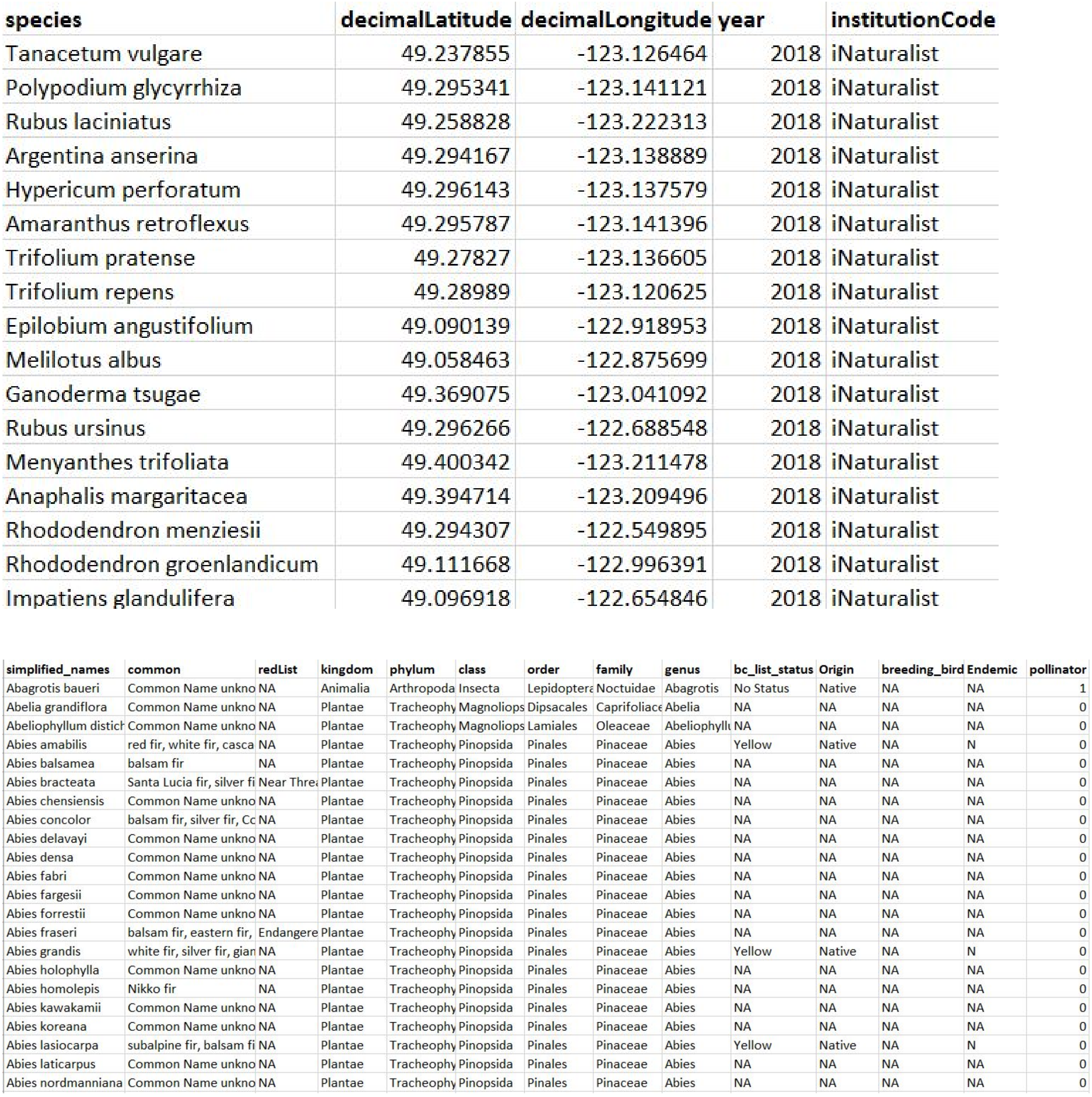
4.2 Data Curation and Reduction

In order to make it useful, all data sources needed to be aggregated together. Since different sources contained data on multiple levels, several data structures were created, each containing all pertinent data from a given level. The two principles levels in question for constructing the map viewer and summary plots were observations (GBIF, administrative boundaries) and species (GBIF, ITIS common names, IUCN Red List, BCSEE, custom lists; see Figure 4). Aside from these two organism data levels, there were three levels of environment data: SEI polygons (SEI, BC Freshwater Atlas, NTS BC Coastline Polygons, PCIC, Elevation-API), municipalities (administrative boundaries only), and raster information (Canadian Digital Surface Model, WorldClim, SEI).

4.2.1 Organism Data Structures

The foundational dataset was raw species occurrence data extracted from GBIF for Metro Vancouver with all species-specific variables (i.e., variables that described the species observed, not features of the observation itself) stripped away. This left only the identifying species name itself, the location in the form of latitude and longitude, and the dataset from which the observation came (year information was discarded, although a version with temporal information was retained for use with the time-series plot; see section 3.2.3 below). While the municipality in which an observation is located is redundant with location, a numeric variable was added for municipality (each number representing a different administrative unit) for the sake of simplicity of filtering the data based on this variable.

The species-specific data frame was created by reducing the raw data from GBIF to a single observation per unique species and removing all observation-specific variables, leaving only the full taxonomy (species, genus, family, order, class, phyla, and kingdom). To add to this, the number of raw data observations for each species was added as the column ‘freq’. Following this, the species names were merged with the ITIS common name variable, the IUCN Red List, and the BCSEE (BC endangered status, endemic status, and breeding bird status). Lastly, the multi-level variables (e.g., IUCN has several possible designations and is thus not an all-or-nothing status) were transformed into binary variables. This was done for ease of use by categorizing all species with a particular status together (e.g., group together all species that are either designated as Red List or Blue List by the BC Ministry of the Environment), these simple classifications could be used to filter the data in the time-series plot, and the binarized IUCN status in particular was used in the map-viewer to highlight clusters of observations containing IUCN status species as well as taxons containing such species in the accompanying hierarchical tree.



species	decimalLatitude	decimalLongitude	year	institutionCode
Tanacetum vulgare	49.237855	-123.126464	2018	iNaturalist
Polypodium glycyrrhiza	49.295341	-123.141121	2018	iNaturalist
Rubus laciniatus	49.258828	-123.222313	2018	iNaturalist
Argentina anserina	49.294167	-123.138889	2018	iNaturalist
Hypericum perforatum	49.296143	-123.137579	2018	iNaturalist
Amaranthus retroflexus	49.295787	-123.141396	2018	iNaturalist
Trifolium pratense	49.27827	-123.136605	2018	iNaturalist
Trifolium repens	49.28989	-123.120625	2018	iNaturalist
Epilobium angustifolium	49.090139	-122.918953	2018	iNaturalist
Melilotus albus	49.058463	-122.875699	2018	iNaturalist
Ganoderma tsugae	49.369075	-123.041092	2018	iNaturalist
Rubus ursinus	49.296266	-122.688548	2018	iNaturalist
Menyanthes trifoliata	49.400342	-123.211478	2018	iNaturalist
Anaphalis margaritacea	49.394714	-123.209496	2018	iNaturalist
Rhododendron menziesii	49.294307	-122.549895	2018	iNaturalist
Rhododendron groenlandicum	49.111668	-122.996391	2018	iNaturalist
Impatiens glandulifera	49.096918	-122.654846	2018	iNaturalist

simplified_names	common	redList	kingdom	phylum	class	order	family	genus	bc_list_status	Origin	breeding_bird	Endemic	pollinator
Abagrotis baueri	Common Name unkno	NA	Animalia	Arthropoda	Insecta	Lepidoptera	Noctuidae	Abagrotis	No Status	Native	NA	NA	1
Abelia grandiflora	Common Name unkno	NA	Plantae	Tracheophy	Magnoliops	Dipsacales	Caprifoliace	Abelia	NA	NA	NA	NA	0
Abeliophyllum distich	Common Name unkno	NA	Plantae	Tracheophy	Magnoliops	Lamiales	Oleaceae	Abeliophyllu	NA	NA	NA	NA	0
Abies amabilis	red fir, white fir, casca	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	Yellow	Native	NA	N	0
Abies balsamea	balsam fir	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies bracteata	Santa Lucia fir, silver fi	Near Threa	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies chensiensis	Common Name unkno	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies concolor	balsam fir, silver fir, Cc	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies delavayi	Common Name unkno	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies densa	Common Name unkno	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies fabri	Common Name unkno	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies fargesii	Common Name unkno	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies forrestii	Common Name unkno	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies fraseri	balsam fir, eastern fir,	Endangere	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies grandis	white fir, silver fir, gian	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	Yellow	Native	NA	N	0
Abies holophylla	Common Name unkno	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies homolepis	Nikko fir	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies kawakamii	Common Name unkno	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies koreana	Common Name unkno	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies lasiocarpa	subalpine fir, balsam fi	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	Yellow	Native	NA	N	0
Abies laticarpus	Common Name unkno	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0
Abies nordmanniana	Common Name unkno	NA	Plantae	Tracheophy	Pinopsida	Pinales	Pinaceae	Abies	NA	NA	NA	NA	0

Figure 4. Combined organism data structure example. a) the observation dataframe; b) the species dataframe.

4.2.2 Environment Data Structures

The shapefiles containing SEI polygons were extended by four additional data sources. For each polygon, a centroid was calculated and thus the centroid was used in conjunction with the Elevation-API to determine the elevation at that particular point; this elevation was assigned to the polygon as a whole. Shapefiles corresponding to rivers, lakes, and the coastline were taken from the Freshwater Atlas and the NTS Coastline project; the distance from each polygon to each source of freshwater (river or lake) and salt water (coast) was calculated. Finally, each polygon was given the average temperature and rainfall variables from the closest station in the PCIC dataset. In the event that the closest station did not contain both variables, the next closest station with this data available was chosen.

The municipality shapefiles were used only to filter observation data and were thus unaltered with one exception. In the shapefiles of administrative boundaries, the western contiguous regions of Electoral Area A, the land north of North Vancouver and West Vancouver, the University Endowment Lands, and the Georgia Straight are not distinguished. In order to facilitate the examination of each of these areas separate from one another, the shapefile was altered such that each is treated as its own administrative unit.

Uniquely, because they were used exclusively with species distribution modelling (SDM) which does not require them to be combined, the raster files used in this project were not combined with one another, but were stored separately. The elevation data rasters obtained from the CDSM were retrieved in eight separate TIFF files; these rasters were combined and saved as a single file.

4.2.3 Collapsing Redundant Observations

Certain coordinate points within Metro Vancouver were associated with over one-hundred separate observations, possibly due to citizen-scientists reporting their sightings to the nearest landmark as opposed to using precise coordinates (e.g., sightings reported in different parts of the same park may each be recorded simply as being in the park, in which case they will be given the exact same coordinates). In early prototypes, our team found it cumbersome to use the map viewer as such landmarks and common reference points often contained enough separate observations to make themselves unwieldy, and many of the observations were of the same species.

Since such information was redundant, the data was collapsed across species and location, removing ‘duplicate’ data points that reported the same organism at the same location as another observation. This new dataset contained far fewer data points (approximately one sixth the size), indicating that the majority of the original dataset was indeed redundant. All data collapse was performed in R. However, since the time-series plot relies on the 'year' variable and the spatial component is not a concern, a version of the non-collapsed data with 'year' information included was retained to be used with this plot.

During this collapse, the *'month'* variable was transformed into four binary season variables: *'winter', 'spring', 'summer',* and *'fall'* to record in which season a given observation took place.. This allowed preservation of the seasonality of the data when collapsing across time, when multiple observations of the same species in the same location were collapsed, the new entry was given for each season variable equal to '1' if at least one constituent observation occurred in that season and a '0' otherwise. While this seasonality was not used in the final product, it is still included in the data if there is a desire to filter based on season in the future.

4.2.4 Removal of Observations Outside of Metro Vancouver

Because the GBIF spatial selection tool used when retrieving data was manual, a polygon was drawn that superseded the borders of Metro Vancouver in order to ensure that no Metro Vancouver observations were missed. After the municipality shapefiles were used to assign a municipality label to each observation, those observations with no associated municipality (i.e., those that fell outside of the borders of Metro Vancouver) were removed from the data.

4.3 SEI Occurrence Modelling

Given the uneven sampling of organisms in the GBIF occurrence data that is largely inherent to citizen science work, it was deemed important to model the predicted distributions and preferences of species and taxa across the entire area of Metro Vancouver, to better evaluate the importance of the land outside of the main data collection area to biodiversity. The first technique focused specifically on sensitive ecosystems, using the presence or absence of organisms in SEI polygons of specific classes and with specific characteristic as evidence of the likelihood of finding such species in similar environments.

4.3.1 Data Setup

As discussed in Section 2, environmental data sources were used to characterise each SEI in terms of its average temperature, precipitation, elevation, and distance to both fresh and saltwater; these data were coupled with intrinsic data on each polygon such SEI class, size, condition, and surrounding land context (Metro Vancouver 2014). Before proceeding, polygons with no observations were removed in order to avoid negatively biasing predictions. In addition, species with observations in fewer than 20 distinct polygons were removed; while technically the analyses for these species would have the same ‘n’ values as all others (i.e., the same amount of polygons factored in), very low numbers of positive/present polygons can cause the equation to fail to converge and produce unreliable results, particularly with large numbers of predictive factors. Subsequently after the removal, this left 536 species remaining.

4.3.2 Prediction

A logistic regression equation predicting the presence/absence of organisms of a given species was produced for the 536 species mentioned above. Due to a number of the equations failing to converge, the species were further restricted with those producing equations with extreme and unrealistic parameters removed. This left us with a final group of 241 species. The parameters of these equations were stored in a dataframe which was given as input to the Python backend of the main application, which was then able to use them to compute an expected probability of finding each species for any given polygon.

As an optional feature for the application, the procedure mentioned above was also conducted on the class taxonomy level. The intention of this was to produce predictions that, while less precise due to the heterogeneity of species within a class, take advantage of a greater proportion of the data since the requirement of the sum total of organisms within a class to have been seen in at least 10 polygons is far less restrictive. Of the 25 classes present in the full dataset, 19 classes both met this restriction and did not produce equations with extreme parameters. This feature can been included in parallel to the species-level predictions mentioned above. It should be noted that the choice of class was arbitrary. The same procedure could be applied to Kingdom, Family, or any other taxonomic level in the future if desired.

4.4 Species Distribution Modelling

Why Use Species Distribution Modelling

While the SEI-based analysis in Section 4.3 takes advantage of the SEI classification system, it is limited both by the resolution of predictions (i.e., some polygons represent very large areas) and by its applicable area (i.e., no prediction is possible outside of sensitive ecosystems, which compose a great deal of Metro Vancouver). For this reason, a pipeline was established for Species Distribution Modelling (SDM), a statistical framework that would allow the assignment of probabilities of seeing an organism across the entire landscape and on a uniform scale (i.e., in consistent blocks as opposed to heterogenous polygons).

What is Species Distribution Modelling

The purpose of performing SDM is to predict suitable environments across the Metro Van landscape for the organism(s) being modelled as well as identify regions that are predicted to be unsuitable (Beauvais et al., 2006). The analysis produces a heat map indicating a suitability gradient with deep colored areas indicating landscapes that are predicted to be highly suitable. The modelling requires only two sources of data, species occurrence and environmental information . The modelling approach uses the values of the environmental variables at locations of known occurrence to find other geographic regions within the study area that are similar; locations with similar environmental values are then predicted to be a suitable environment for the organism though it may not have been actually observed there. The model has no temporal component and assumes that the relationship between the organism and its environment remains fixed over time.

Model Implementation

The species occurrence data was taken from GBIF with only species that had 40 or more observations included; environmental predictors included climate variables obtained from WorldClim2 and altitude data from the CDSM. The modelling was implemented in R using the "SSDM" package (Schmitt et al., 2017). The package allows for the modelling of a single species distribution or a collection of different species. When multiple species are modelled together, Stacked Species Distribution Modelling (SSDM), the map output predicts the locations that would accommodate the greatest number of species that are being modelled (Schmitt et al., 2017). Locating these areas that are predicted to have high species richness could provide a clue towards identifying biodiversity hotspots (areas that support a wide range of species).

The model output was visualized using an R shiny interactive application which allows the user to run different models with a choice of algorithms. See Schmitt et al (Schmitt et al., 2017). for more algorithm details. Changing the choice of algorithm may lead to more or less accurate results; since each species has its own unique distribution, certain algorithms will do a better job of predicting that species. The model map includes an accuracy measure which indicates the overall percentage of correct predictions when the model was tested on unseen data. These output maps represent untested hypotheses about how species are distributed across Metro Van and contain uncertainty. They provide clues as to what environmental factors may be most important for driving species distribution and indicate areas of high suitability for the species.

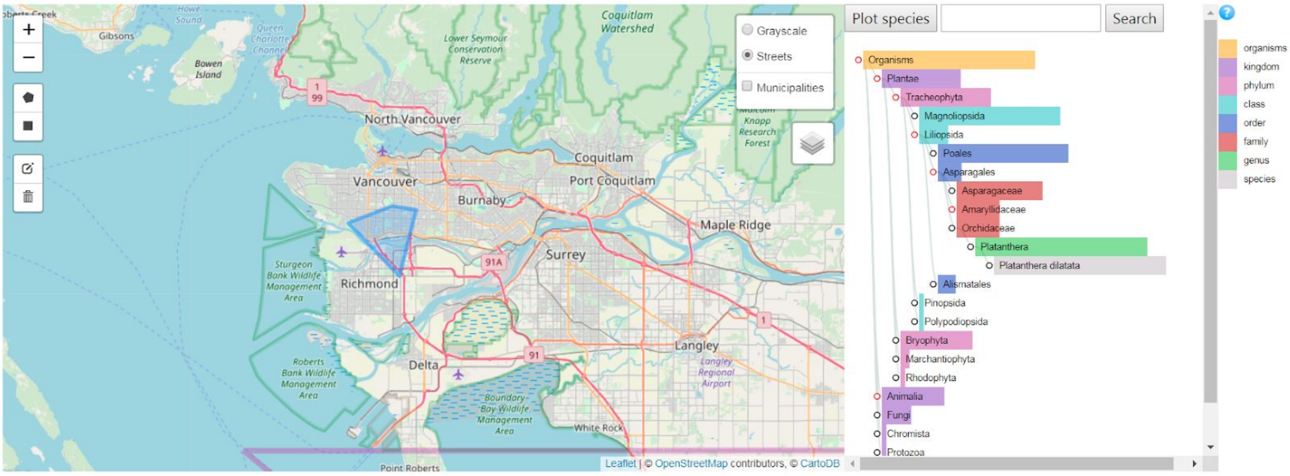
5. Map Application Features

The final deliverable of the project is a mapping application that facilitates not only spatial exploration of species occurrence, but a suite of additional features such as endangered species information, interactive summary charts, SEI species predictions, interactive time-series plots, and species distribution models.

5.1 Map Viewer and Data Exploration

The main utility of the map feature is to visually explore the GBIF species occurrences (Fig 5). A user must first begin by drawing an area of interest with either of two drawing tools (free-selection or rectangle). To plot species occurrences within the selected region, a user may choose from two approaches; either use the search feature to type a specific taxonomic group (i.e., kingdom, genus, species) or navigate through the taxon tree to select multiple taxa of interest. A user may then explore individual clusters of species until zooming in on one specific organism. Each species point provides a pop-up with scientific name, common name (if available), link to its wikipedia webpage and its IUCN red list status if applicable.

Figure 5. Screenshot of the base map viewer.



○ Grayscale
● Streets
□ Municipalities

The map viewer contains two additional layers of information that may be turned on or off by the user. A user may turn on the municipalities feature to automatically draw the boundaries around each municipality found in Metro

Van. Clicking within the boundaries of any municipality produces a pop-up label for the municipality.

A second layer allows a user to select and plot ecosystems of interest (i.e., Mature Forest, Wetlands) based off of SEI data . Clicking on any ecosystem polygon produces a pop-up label with more specific information about that polygon such as ecosystem sub-classifications and a quality score. Though the Modified Ecosystem category has 5 sub-classes, the map viewer clusters all these polygons together as one color; however, when the user clicks on a Modified Ecosystem polygon, all the specific information for that polygon (ie sub-class, quality score, % composition) can be seen in the pop-up label.

5.1.1 Interactive Summary Charts

When a user draws a region and plots species, summary charts are automatically generated for that region. The first chart provides the count of unique species for each taxon level observed. Clicking on any taxon bar will automatically generate a new unique species distribution plot for that particular group selected. For example, if a user selects the kingdom "Animalia", then a unique species distribution of the phyla within the animal kingdom will be generated.

The second summary chart displays the number of observations of that taxon group found in the data. A user may interactively select any taxon bar and generate a new plot based on that taxon group in the same way as with the unique species distribution chart.

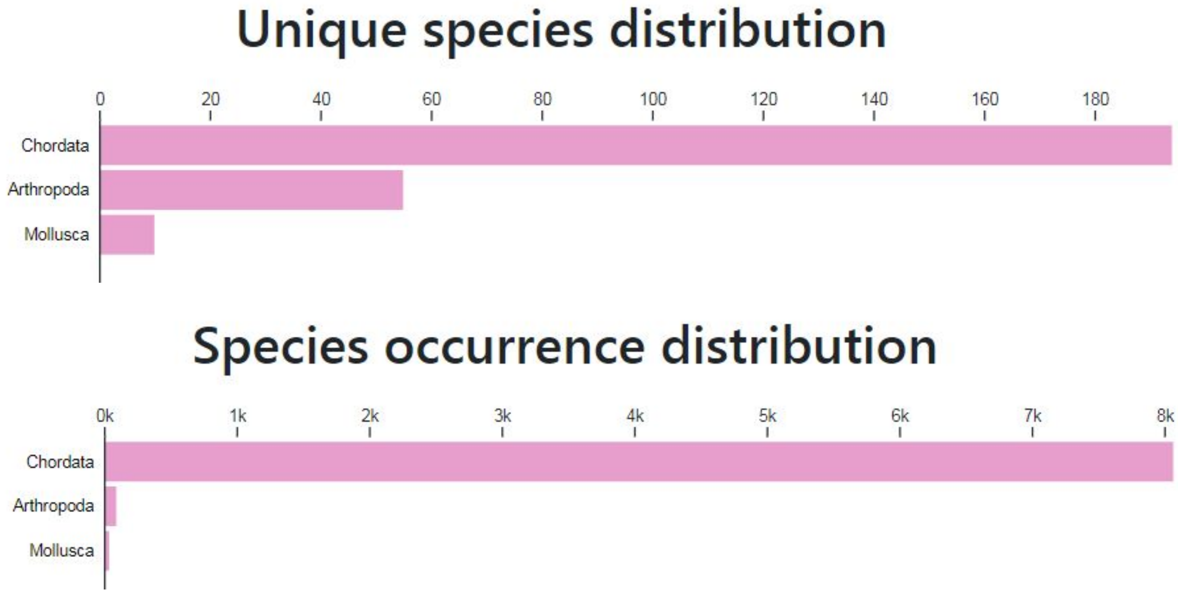
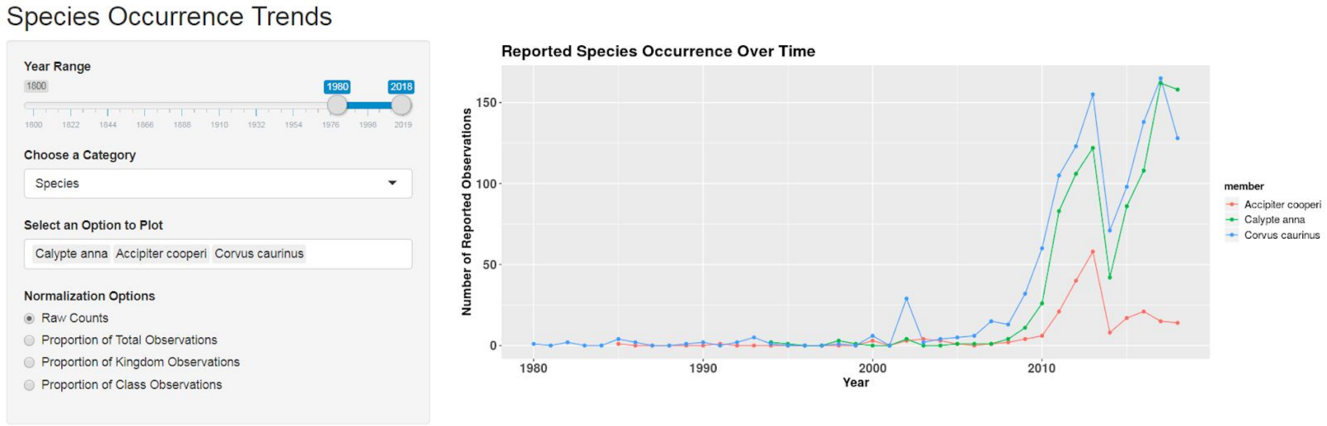


Figure 6. Screenshot of the interactive summary charts.

5.1.2 Time-Series Plots

The time-series feature is also interactive in that the user may adjust the year interval as well as how to aggregate and normalize the data. The plot visualizes the number of observations for a particular category over time, providing a rough approximation for species abundance. A user has a number of categories to choose from to aggregate the observations; they may choose to view individual species records over time or entire kingdoms. There are also a number of additional custom aggregations that highlight species of special concern like pollinators, species that are endemic to the area, and IUCN red listed species.

Viewing raw observation records over time may be somewhat misleading for some species since the GBIF records overall have been steadily climbing for the last decade (Fig 3). To remedy this, a user can choose to normalize the data by Year, Kingdom or Class. Normalizing by year assumes that if a species remains constant in nature, even if the number of new records climbs every year, the relative proportion the species contributes to the data overall will remain constant. Thus, while it's expected that there are more observations of a given species for a given year than the year before, there is no *a priori* reason to believe that the percentage of data that belongs to the species should change. Normalizing by Kingdom and Class recognizes that citizen science initiatives can cause spikes in the sampling of certain taxa (even though those taxa may remain constant in nature), violating the assumptions of the normalization by Year. This can be true especially for birds due to the majority of all records that are attributed to e-Bird (Table 1). Normalizing by class essentially equates a bird that represents 2% of the total "Aves" data with, for example, a mammal that represents 2% of the total "Mammalia" data.



Species Occurrence Trends

Year Range: 1980 – 2018

Choose a Category: Species

Select an Option to Plot: Calypte anna, Accipiter cooperi, Corvus caurinus

Normalization Options

● Raw Counts
○ Proportion of Total Observations
○ Proportion of Kingdom Observations
○ Proportion of Class Observations

Figure 7. Screenshot of the time-series plot.

5.2 Predictions

5.2.1 SEI-Based Prediction

The SEI-based prediction module has been designed to be an effortless extension of the SEI layer of the viewer. Provided the 'SEI Prediction' switch has been turned on from the menu, whenever the user selects a polygon from said layer, the prediction table automatically computes probability of finding each species/class with associated equations in that polygon and displays the results in a table. The table presents the rank of likelihood and whether or not the species/class was actually observed there in addition to the name and the probability itself; it is both searchable and sortable based on any of the mentioned variables.

5.2.2 Species Distribution Modelling (SDM)

The SDM module allows the user to run SDM analyses in real time and display the prediction results in a map output. A user must first decide if they will model single species or multiple species together. Next, the user can select from a range of different algorithms. Details of the algorithms included in the "SSDM" package can be found in Schmitt et al. Lastly, the user can select the species they want to model. The stacked modelling works by modelling each species separately and then overlying the results together, therefore, modelling multiple species can take a great deal more processing time. When modelling a single species the darker colored areas have higher probability of being suitable environment for the species. When modelling multiple species, the darker colored regions indicate higher species richness.

6. Discussion

6.1 Occurrence Data Sources: GBIF and Alternatives

Having the entirety of the observation data come from a single source, GBIF, streamlines the data collection process and makes the application far easier to update in the future; but this does open up the possibility that there are potentially other useful data sources that are not utilized. However, this possibility should be slight based on extensive research into alternate data sources at the beginning of the project. By and large, it was discovered that all sources of biodiversity occurrence information either already fed into GBIF, were not publically accessible, or did not meet set standards. In regards to data that are not publicly accessible, such sources would greatly limit the ability to distribute the application even in the cases where the data could eventually be obtained.. As for data that did not meet set standards, it was found that many sources that at first appeared as though they might have been useful (e.g., E-Flora and E-Fauna) either did not have organized data that was extractable in a useable format (e.g., data would need to have been extracted on a species-by-species basis), or did not have georeferenced data which would have completely prevented its use in the application. In sum, it was determined that GBIF constituted the best data source for the project.

6.2 Limitations

6.2.1 Limitations of Citizen Science

As indicated by Table 1, the majority of occurrence reports in the GBIF data came from the citizen science-based sources iNaturalist and especially e-Bird. While citizen science represents a cost-effective and community-minded method of aggregating biodiversity information across large regions, it comes with a number of inherent limitations.The two primary limitations that affected every aspect of our work were bias in what organisms were recorded and bias in spatial sampling.

Organism Bias

In the first case, the data were hugely biased by the relative size of the e-Bird data source, which contains only organisms from a very specific class (Aves). Even if none of the data sources were this specific, bias across the spectrum of life is inevitable, the concept of 'plant blindness' is well documented (Allen, 2003 for example (and likely applies to fungi as well). Even though ecosystems virtually require more plant life than animal life in order to function, our data contained almost all animal reports and less than 2% plant reports. Another issue is that microscopic life cannot be sampled at all through citizen science, when it is known that the microbiome of an ecosystem is hugely important to its health (Garcia-Palacios et al., 2014, Schnitzer & Klironomos, 2011).

Spatial Bias

The second limitation is in regards to spatial bias. There is no way to control for the places that citizen scientists frequent and are thus more likely to make observations. This problem can clearly be seen from the viewer when examining the number of observations in places such as the unincorporated land of Electoral Area A north of West Vancouver and the municipality of North Vancouver. In general, it appears that the more remote an area is the fewer observations will be recorded; in reality, the quantity and diversity of life is likely to be just as high if not substantially higher in these undeveloped areas. Another illuminating example is the high number of observations in the University Endowment Lands, around which many people live and work.

Overcoming Bias

When it comes to organism bias, reducing redundant data helped to reduce the bias (since most of the overlapping observations of the same species that were removed were birds and hardly any were non-animals), but eliminating the bias without purposefully and arbitrarily removing data is simply impossible. This information is still highly useful as long as users keep this bias in mind, and understand that this is a tool for tracking human observations of wildlife, not for accurately sampling relative proportions of wildlife.

Spatial bias is similarly hard to remove, and has also been addressed on the viewer home page. The two modelling techniques used in the project are essentially both attempts to address it by extending our ability to make judgements on the presence of a species over a wider area than just that area in which they have been recorded. SEI-based predictions have the obvious spatial limitation of being limited to sensitive ecosystems, though thankfully this fact is clear to the user since they access the predictions through selecting individual SEIs.These modelling techniques come with their own limitations, discussed below.

6.2.2 Lack of Abundance and Diversity Metrics

One inherent limitation to the eclectic mix of data sources included within GBIF is the exclusion of abundance data, a variable that is heavily used in ecology (Preston, 1948). The underlying assumption of the application is that the more observations of a given species in a given area, the greater that species' abundance in that area. However, when dealing with non-systematic data, the number of observations is less of a measure of real-world abundance and more of a correlated variable, as it is expected that observations for a species/location that is more abundant in actuality should produce more abundant records in the data. However, this may not be the case. As discussed above, systematic biases over both the tree of life and the area of Metro Vancouver conspire to amplify observations of certain organisms in certain places, drastically limiting the ability to make comparisons. For example, while there are more observations logged for the City of Vancouver than for the City of Coquitlam, it cannot be confidently concluded that there is more wildlife in the City of Vancouver. An effort has been made to combat organism bias in the time-series viewer by allowing the user to standardize raw counts based not only on the total number of records for the year but by the number of records in the corresponding Kingdom and Class. However, this doesn't eliminate the problem.

The same argument applies to the measure of diversity (the degree to which an environment supports many different kinds of life). Due to the same biases (particularly the undersampling of non-animal life which makes up the majority of extant species), even the multi-species output of the SDM module should be considered a rough estimate of diversity hotspots as opposed to a true metric.

The inability to achieve true abundance and diversity metrics has less to do with the presence of citizen science data in GBIF and more to do with the lack of systematically-sampled sources. In order to derive abundance for our region of interest one would need a comprehensive survey of the entire area of Metro Vancouver, and in order to arrive at diversity, this would need to be done for at least a wide-ranging 'bundle' of representative species that may or may not include microscopic organisms. Overall, such ecological metrics are preferably used in systematic academic studies that focus on small areas and limit their species scope.

6.2.3 Temporal Limitations

Depicting the element of time in our dataset proved to be a challenge throughout this project, and several methods were prototyped to allow the user to control what time periods they saw on the map viewer (including the option to display occurrences from different seasons). This feature was ultimately abandoned for two reasons: a) in most forms it proved unsightly, unwieldy, and often imposed a processing power cost that slowed down the application, and b) it was decided that it would be unlikely for a user to require such temporal control in tandem with the spatial navigation offered by the map viewer. These concerns inspired us to create the temporal summary module that exists in the final version; this module confers many benefits including the ability to standardize across time using several different methods and compare different taxa change over time.

The other concern arising from the temporal nature of our data is whether or not biodiversity information from different time periods can truly be compared. While this is not a problem on a small scale, our data span 319 years, it is quite possible that old observations logged organisms that can no longer be found in the Metro Vancouver area or vice versa. In addition, designations that were applied to the data (most notably the IUCN Red List) are also temporal in nature. At present, occurrences are flagged as being part of the Red List if the species matches the copy of the Red List from June 2019, regardless of whether or not they were endangered at the time of observation. While this is unlikely to bother most users, it does raise the question of whether or not our data should have an 'expiry date': if a large portion of the observations of an organism were made prior to a decline in their population, can the information concerning their spatial distribution be trusted?

Some of these issues are mitigated by the fact that the vast majority of the data was recorded in the 21st century (Figure 2), but this in and of itself limits our ability to make judgements about longitudinal change that the temporal summary module was created to enable.

6.2.4 General Modelling Limitations

While they are intended to address spatial bias in the data, one manner in which either SEI or SDM-based predictions are still affected by the spatial distribution biases of our data is if there are significant differences in some of the bioclimatic/topographical variables used to inform the predictions. For example, as mentioned above the mountainous region of Metro Vancouver north of West and North Vancouver is sparse in observations - if these polygons/areas have higher altitudes than the southern areas with more observations, then the model is likely to underestimate the probability of finding organisms in any area of the map with high altitude. This is especially a problem for SEI-predictions since so many of the SEI polygons are located in this area. While this bias was reduced by eliminating the SEI polygons with no observations, it was decided to not remove polygons containing low numbers of observations. The reverse could be true for SDM. Since there are few observations at sea, for example, the low altitude values of ocean areas might prompt the model to predict that low-altitude regions are less hospitable to life than they really are.

As discussed earlier, the major limitation of SEI-based predictions on the species level was that only a small proportion of species could be assigned predictions at all. While SDM was based entirely on continuous data (i.e., temperature, elevation), SEI classes themselves are multi-level categorical variables and very few species were seen in all 14 possible classes. While ideally all equations would have been based on the same input variables, inputting a predictor with zero variance (i.e., including the percentage of a polygon that is class X as a predictor, when the species has never been seen in that type of terrain) is both uninformative and can cause the model to fail to converge. Thus, for each species' equation, SEI classes were dropped in which it was never seen. This leads to the question of how to predict the probability of finding a species in one of these absent classes, and while the intuitive answer is to propose that a probability simply cannot be assigned, this becomes difficult when trying to make judgements for polygons that are only partially composed of the offending class. Assigning values of zero was not an option either, since many of the class parameters were negative, and thus a “zero” parameter might have indicated a preference for an environment that an organism was never seen in compared to an environment it was rarely seen in. It was decided to assign the lowest parameter in the equation to all “missing” parameters. What this essentially means is that with all other factors being equal, the probability of finding an organism in an environment in which it was never observed is equal to the probability of finding it in the least common class it was observed in. It was determined that these alterations to the prediction equations were necessary, but they are still problematic, and the results are not the “true” results of a regression equation.

6.2.5 SDM Specific Limitations

Limited Species

SDM can only be performed on a subset of the GBIF data. The minimum number of observations required to make a good predictive model is highly dependent on the species and the nature of the data that has been collected; the literature has offered anything from as low as 10 observations to as high as 90 as a minimum threshold (Beauvais et al., 2006). A threshold of at least 40 observations was chosen to ensure there was a reasonable amount of observations to evaluate the model after model training. The data is partitioned into 70% for training and 30% for testing. Applying this cut-off meant that only 361 unique species out of our ~10,000 could be modelled using SDM.

Presence Only Data

The GBIF data is presence-only data and contains no information on species absence or potential absence. This reduces the statistical power and prediction performance of the models. Since in practice absence data is rarely available, pseudo-absence data is commonly used when building SDM models and this method is implemented automatically in the SSDM package (Beauvais et al., 2006)(Elith et al., 2006). Though the models are evaluated quantitatively, the accuracy metric should still be approached with caution because the evaluation data set is built using this presence-only data which is bound to contain spatial biases (Elith et al., 2006).

6.3 Future Directions

6.3.1 Updating the Application

At present the application does not update automatically to accommodate new information being absorbed by GBIF (for reference, each data source on GBIF updates according to a different schedule, some as frequently as every two weeks). An effort has been made to ensure that by downloading the data pertaining to Metro Vancouver and processing it with certain scripts, the administrator of the application can manually update the database at whatever regularity suits them. In the future, it would be helpful to allow the application to automatically update itself through a GBIF API.

Furthermore, as mentioned in section 6.2.3 endangered statuses for species are inherently temporal; the species that are endangered today may not be the same five years from now. While it would be useful to automatically update endangered statuses, it should be noted that this task may be more difficult than updating the base GBIF data since endangered statuses come from multiple sources. These sources are also not necessarily as accessible as GBIF. Both the IUCN and BC/SARA endangered species lists had to be obtained in .csv format by manually selecting options and downloading files. It is possible that there are APIs or other ways to scrape data from these sources without using their website user interfaces but this was not uncovered during the project.

In addition to fetching the newest data from the sources, it is also necessary to aggregate the new data and update the prediction metrics. Currently, a separate file stores all the data related to the species in GBIF. That file needs to be updated with incoming data. The models for SEI prediction also need to be re-run.

Continued work on the app is not limited to the addition of new data, however. Minor maintenance will likely be needed in the next few years to ensure that as the packages that the application relies on develop and change, these changes do not create downstream problems. This is particularly true of D3, the JavaScript library that underlies the hierarchical organization of both the taxonomic tree and the species distribution bar charts. Because of its recency and rapid development, it is likely to force a change before the other elements of the app.

The app can also be modified to add new features. One of the simplest ways to add functionality is to add geographic information in the form of layers like polygons similar to the SEI or municipality data currently being used. Just as municipalities and polygons can currently be overlaid on the basemap and selected to allow the viewer to examine species occurrence within that region, new layers such as watersheds or university land boundaries can be added. Since the application is already setup to accommodate these geojson layers and transmit their coordinates to its exploratory modules, integrating further layers can a simple process.

6.3.2 Extensions of Modelling

Both the SEI and SDM modelling techniques used in this project were developed towards the tail end of the project, and while they represent a beginning point for modelling taxon presence across space, they were not the focus of the project and were not refined to the extent of the other modules. Future work could extend these models in a number of ways.

The simplest would be including additional data in each modelling technique. For example, distance from fresh and saltwater was calculated only for SEI polygons and was not factored into SDM - similarly, SDM contains many climate-relevant factors that were not used in the SEI-based predictions. Another possibility is rasterizing the SEI polygon vectors and including them as a predictive layer in the SDM analysis so that SEI classification can benefit both modelling techniques - alternatively, Metro Vancouver possesses a land classification raster dataset that is less precise in some of its categorizations than the SEI but covers the entirely of the Metro Vancouver area at a precision of 5m (more than enough to be used by the current SDM model). One suggested candidate for completely novel data that could be added to both models is air quality. While current air quality data is provided by Metro Vancouver, historical data that would be needed to characterize a tract of land was not found during the project. If future work could obtain such data it might be useful, especially for modelling the prevalence of organisms that depend heavily on the molecular composition of the air, such as many species of plants. In addition, including soil quality rasters would be useful when modelling plant species.

Alternatively, future work could explore the inverse of the above: removing uninformative variables. The SDM model especially contained a suite of bioclimatic variables many of which are semi-redundant with one another (e.g., “average temperature”, “average temperature of the hottest month”, “average temperature of the coldest month”). By comparing their explanatory power across different species, it is possible to determine the best candidates for removal.

6.3.3 Gap Analysis

Though the BC species list was not an exhaustive baseline for the species that live in Metro Van, it did identify at least some species that are expected to be found in Metro Vancouver. When comparing the Metro Van tagged species to the GBIF data, ~ 70% of the Metro Van BC List species were found in GBIF. However, this leaves ~ 30% (about 100 species) that are expected to be found in Metro Van but were not observed on GBIF. This list of 100 species is a good first step towards identifying a data gap. In future, these species would be candidates for an organized Bioblitz or other local Citizen Science initiative.

References

Allen, W. (2003). Plant blindness. *BioScience, 53*(10), 926-926. doi:10.1641/0006-3568(2003)053[0926:pb]2.0.co;2

Beauvais, G., Keinath, D., Hernandez, P., Master, L. and Thurston, R. (2006). ELEMENT DISTRIBUTION MODELING: A PRIMER. *NatureServe*, [online] pp.1-42. Available at: https://www.natureserve.org/biodiversity-science/publications/element-distribution-modeling-primer [Accessed 23 Aug. 2019].

Bouma, J., and Van Beukering, P. (Eds.). (2015). *Ecosystem Services: From Concept to Practice*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781107477612

Costanza, R., d'Arge, R., de Groot, R., Farber, S., Grasso, M., Hannon, B., Limburg K., Naeem, S., O'Neill, R.V., Paruelo, J., Raskin, R.G., Sutton, P., and van den Belt, M. (1997). The value of the world's ecosystem services and natural capital. *Nature. 387.* 253-260. https://doi.org/10.1038/387253a0

Elevation-API.io (2018). *Free elevation API.* Retrieved from https://elevation-api.io/

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., J. Phillips, S., Richardson, K., Scachetti-Pereira, R., E. Schapire, R., Soberón, J., Williams, S., S. Wisz, M. and E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2), pp.129-151.

Fick, S.E. and Hijmans R.J. (2017). Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology.* https://doi.org/10.1002/joc.5086

García‐Palacios, P. , Vandegehuchte, M. L., Shaw, E. A., Dam, M. , Post, K. H., Ramirez, K. S., Sylvain, Z. A., Tomasel, C. M. and Wall, D. H. (2015). Are there links between responses of soil microbes and ecosystem functioning to elevated CO2, N deposition and warming? A global perspective. *Global Change Biology*, 21: 1590-1600. doi:10.1111/gcb.12788

Habib, T.J., (2016). Impacts of Land-Use Management on Ecosystem Services and Biodiversity: An Agent-Based Modelling Approach. Edmonton, AB, CA: Alberta Biodiversity Monitoring Institute.

Isbell, F. (2010). Causes and consequences of biodiversity declines. *Nature Education Knowledge, 3*(10):54. Retrieved from https://www.nature.com/scitable/knowledge/library/causes-and-consequences-of-biodiversity-declines-16132475/

Martinez-Ramos, M., Ortiz-Rodriguez, I.A., Pinero, D., Dirzo, R., and Sarukhan, J. (2016). Human disturbances affect biodiversity in reserves. *Proceedings of the National Academy of Sciences, 133*(19) 5323-5328; doi:10.1073/pnas.1602893113

Metro Vancouver (2014). *Sensitive Ecosystem Inventory for Metro Vacouver and Abbotsford*. Meidinger, D., Clark, J., and Adamoski, D. Metro Vancouver, Burnaby, Canada. Retrieved from http://www.metrovancouver.org/services/regional-planning/PlanningPublications/SEITechnicalReport.pdf

National Resources Canada (2015). *Canada Digital Surface Model.* Retrieved from https://open.canada.ca/data/en/dataset/768570f8-5761-498a-bd6a-315eb6cc023d

Ninan, K. (Ed.), Steiner, A. (2009). Conserving and Valuing Ecosystem Services and Biodiversity. London: Routledge, https://doi.org/10.4324/9781849770859

Preston, F.W. (1948). The commonness, and rarity, of species. *Ecology 29*(3), 254-283. Retrieved from http://www.bgu.ac.il/desert_agriculture/Vegecology/Papers/Preston48.pdf

Schmitt, S., Pouteau, R., Justeau, D., de Boissieu, F. and Birnbaum, P. (2017). ssdm: An r package to predict distribution of species richness and composition based on stacked species distribution models. *Methods in Ecology and Evolution*, 8(12), pp.1795-1803.

Schnitzer, S. A., and Klironomos, J. (2011). Soil microbes regulate ecosystem productivity and maintain species diversity. *Plant signaling & behavior*, *6*(8), 1240–1243. doi:10.4161/psb.6.8.16455

WWF (2018). *Living Plant Report - 2018: Aiming Higher.* Grooten, M. and Almond, R.E.A. (Eds.). WWF, Glands, Switzerland. Retrieved from https://wwf.panda.org/knowledge_hub/all_publications/living_planet_report_2018/