


Fall 12-14-2018

# Employing Natural History Collections in the Aid of Conservation: Streamlining an Approach to Model Species Distributions En Masse for the Preservation of Biodiversity

Alice Fornari  
fornariae@gmail.com

Follow this and additional works at: <https://repository.usfca.edu/capstone>

 Part of the [Biology Commons](#), [Ecology and Evolutionary Biology Commons](#), [Geographic Information Sciences Commons](#), [Museum Studies Commons](#), and the [Zoology Commons](#)

---

## Recommended Citation

Fornari, Alice, "Employing Natural History Collections in the Aid of Conservation: Streamlining an Approach to Model Species Distributions En Masse for the Preservation of Biodiversity" (2018). *Master's Projects and Capstones*. 857.  
<https://repository.usfca.edu/capstone/857>

This Project/Capstone is brought to you for free and open access by the Theses, Dissertations, Capstones and Projects at USF Scholarship: a digital repository @ Gleeson Library | Geschke Center. It has been accepted for inclusion in Master's Projects and Capstones by an authorized administrator of USF Scholarship: a digital repository @ Gleeson Library | Geschke Center. For more information, please contact [repository@usfca.edu](mailto:repository@usfca.edu).

# **Employing Natural History Collections in the Aid of Conservation: Streamlining an Approach to Model Species Distributions En Masse for the Preservation of Biodiversity**

Keywords: Museum Studies, Natural History Collections (NHCs), Geographic Information System (GIS), Species Distribution Model (SDM), Maxent, Presence-Only Data, Biodiversity Crisis, Conservation Biology.

by

Alice Elizabeth Fornari

Capstone project submitted in partial fulfillment of the requirements for the  
Degree of Master of Arts in Museum Studies

Department of Art + Architecture  
University of San Francisco

Faculty Advisor: Marjorie Schwarzer

Academic Director: Paula Birnbaum

December 14, 2018

## Abstract

Using **species distribution models** (SDMs) in **Natural History Collections** (NHCs) can influence how humans implement **conservation** changes in flora and fauna communities and ecosystems. Through the use of legacy data (old NHCs and their associated locality/collection information), data correction (background data or pseudo absences added to **presence-only data**), and the SDM software, **Maxent** (and its associated **geographic information systems** or GIS projected models), it has been shown that it is feasible to create a low budget protocol/setup to project the past, present and future of species population changes. This has been done in the past few decades as more collections and their locality data have become digitized, potentially allowing more natural history collecting institutions and scientists to participate in more conservation projects. We can learn from how past and present population ranges have changed due to climate change, urbanization, and deforestation (among other changes) to be able to project where species ranges could exist in the future. The ultimate goal of this project is to provide both a streamlined protocol to input NHC data into Maxent in order to share the results of the Maxent models and associated statistics of NHC data, even if not publication worthy, to larger stakeholders, environmental policy makers and non-profits. Additionally, this project can allow scientists to follow up on the methods and results of the models to see if there really are possible conservation concerns. Interns, citizen sciences, collections workers (non-PhD scientists) can do this in smaller NHCs, and report their findings from their collections. This project has the potential to have a broader impact on rare species housed in smaller collections. Further, it has the capacity to be able allow for specific species and biota to be conserved with the help of precise small grants for specified flora and fauna to be modelled. While this type of project is not the end all be all cure for the **biodiversity crisis**, it can be a way to use available resources and technology for the advancement of our planet and its inhabitants.

## **Acknowledgements**

This capstone is dedicated to my mother, Carol Amsterdam Fornari, who lived with a passion for individuality, creativity, kindness, and was devoted to the conservation of the ecology of the St. Lawrence River. The spirit of her initiative and character has not only driven this work to be realized, but it has pushed me to integrate her unique qualities into my own passions and career.

## Table of Contents

Chapter 1 .....	Page 5
Chapter 2 .....	Page 6
Chapter 3 .....	Page 21
Chapter 4 .....	Page 27
Appendix A .....	Page 28
Appendix B .....	Page 35
Appendix C .....	Page 37
Bibliography .....	Page 41

## Chapter 1: Introduction

There are at least five hundred million specimens of animals and plants housed in US museums, and worldwide, there are estimated to be over two to three billion. According to evolutionary biologists in the year 2000, it was approximated that we have 50 years to answer the challenge of rapid extinction caused by the larger biodiversity crisis. While the technology exists to make these collections useful for biodiversity conservation, there is a concern that scientists are acting far too slowly to enact any real change. As one way to address this issue, this capstone examines the use of species distribution models that use presence-only data, their accuracy, and the types of data available to input into models from natural history collections. In the first section of this capstone, in an analysis of the literature, I have reviewed case studies that use the species distribution software, Maxent, with historical natural history collection data in order to understand and project various species ranges over time. The review done by Yackulic et al. 2012, as well as the study done by El Gabbas et al. 2018 were valuable to this analysis as both studies conclude that not only is Maxent the most reliable modeling software for Presence-only data, but that scientists were not using Maxent appropriately for their datasets even though the software is user friendly and streamlined. In the second section I created a streamlined, zero-cost protocol for collections managers of natural history museums to implement to help figure out what type of data they have related to a species, what type of additional background data they need to complete their sample, where they can find this data, all in order to input into Maxent to create species distribution models that can be reported to scientists and conservation policy makers and funders worldwide as to report potential shifts in species that otherwise may have not been known to have a conservation issue or concern. My hope in exploring the potential that species distribution modeling software like Maxent has in natural history collections, is to be able to provide evidence to larger stakeholders like the IUCN of more species and biodiversity that is at risk so that funding and resources can be used towards their conservation before it is too late.

## Chapter 2: Literature Review

### Introduction: The Value of Natural History Collections

There are at least five hundred million specimens of animals and plants housed in US museums, and worldwide, there are estimated to be over two to three billion. What exactly is the value of these specimens? According to Leonard Krishtalka, “These specimens document the global composition, identity, spatial distribution, ecology, systematics, and history of known life forms (approximately 1.8 million species)”. Further, they allow for analyses and research to be conducted on evolution and ecology on specimens that have been collected over the past three centuries. Thus, natural history specimens offer us invaluable tools which can aid in the conservancy of the natural world. While only approximately 10% of flora and fauna have been identified and described, and in the year 2000, roughly only “5% of collections were captured in electronic databases,” natural history collections (NHCs) are faced with an immense dilemma. Collections hold secrets to the understanding of the past, present, and future of biodiversity, but this data has minimally been digitized and shared to be able to analyze (Krishtalka et al. 2000). There is valuable data hiding in storage, gathering dust, while the planet and its inhabitants experience the impacts of urbanization, climate change, vast species declines, and extinctions.

Krishtalka et al. have identified four major challenges in which natural history museums can participate in making change: the biodiversity crisis, education, public programs, and management and leadership. The same authors, in the year 2000, claimed that we have 50 years to answer the challenge of rapid extinction caused by the larger biodiversity crisis (Krishtalka et al. 2000). While the technology exists to make these collections useful for biodiversity conservation, there is a concern that scientists are acting far too slowly to enact any real change.

In 1994, the Committee on Environment and Natural Resources and National Science and Technology Council’s Strategic Planning Document stated:

“Enhance access to information on the nation's plants and animals. Existing collections of data for millions of specimens will be computerized and made more accessible to the nation's scientists and the public. Increased information...on...geographical occurrence and associated environmental conditions would greatly increase the ability to sustain terrestrial and aquatic ecosystems and to conserve biodiversity in harmony with land use.” (CENR, NSTC, 1994, Chapter 3, pp 6).

Subsequently, with the advent of the Geographic Information System (GIS), and other mapping and computational analysis tools, this is exactly what larger, well funded NHCs began to do, responding to myriad of biologically found crises. According to Suarez et al. 2004, "... [The] importance of these collections and their contributions to society have increased in recent years, particularly following acts of terrorism in the United States and abroad". The bottom line here, is that NHCs play a critical role in public health and safety, environmental health, epidemiology, can act as homeland security tools against biological terrorism, and not just for the purpose of studying the evolution and ecology of animal and plant specimens (Suarez et al. 2004). An example of this can be seen through a study that utilized the Smithsonian National Museum of Natural History's bird study skins that were infected with strains of influenza from the turn of the twentieth century, providing further evidence for the cause of outbreaks, and what may cause future outbreaks (Suarez et al. 2004).

Suarez et al. 2004 lists three major categories as potential applications for natural history specimens: *habitat loss*, *biological invasions*, and *global climate change*. Habitat loss, "... is widely considered to be the greatest threat to biodiversity, and museum collections allow researchers to document the pace of these changes and their ecological consequences". An example of museum collections being used to combat *habitat loss*, can be seen in a study where eighteen museums' collections data were used to show that a decrease in prairie habitats in the midwestern United States, "have led to the decline or local extinction of small mammals that require this habitat to survive" (Suarez et al. 2004). When it comes to *biological invasions*, "museum collections have been used to determine the current distributions of invaders, identify the source of introduced populations, reconstruct rates of spread, and gauge the ecological impact of invaders" (Suarez et al. 2004). In a recent study, Suarez and colleagues used museum collections to reconstruct the spread of the invasive Argentine ant (*Linepithema humile*) throughout the United States during the past 100 years (Suarez et al. 2004). Lastly, *global climate change* has been investigated, "By examining museum specimens, researchers have documented the effects of climate change on a variety of organisms and furnished a glimpse of future impacts" (Suarez et al. 2004). Studies that were done in the 1990's on butterfly species, by analyzing and matching historical data from NHCs (large and small), "... Showed that southern populations (in Mexico) were four times more likely than northern populations (in Canada) to have gone extinct, resulting in a significant northward range shift" (Suarez et al. 2004). These studies each provide practical insight to historical changes, each which can provide valuable evidence to enact conservation changes, having the potential to allocate new or specified funding to the management, maintenance, and preservation of the ecosystems in question.

Shaffer et al. 1998 categorizes locality data types in NHCs as either, "sites as fixed effects," or, "sites as random effects". The former assumes: "... the same sampling techniques were used, the expertise



of both teams was equal, the sampling effort was the same in both surveys, normal biotic and abiotic factors regulating population fluctuations were the same during the sampling periods, and detectability (the detection threshold) of the target species has remained the same” (Shaffer et al. 1998). The latter assumes, “... historical and current sampling should be sufficient so that a lack of occurrence in a region is meaningful, the size of the region over which the sampling sites are pooled should include enough sites to be statistically rigorous, but should not be so large as to be biologically trivial..., the size of the sampling unit over which sites are pooled should be larger than the scale of the biotic and abiotic forces affecting population fluctuations” (Shaffer et al. 1998). These two categories can assist in organizing types of herbaria data and therefore these assumptions are valuable and can be integrated into a protocol/methodology of creating species distribution models of small herbaria specimens.

A variety of computer based data analyzing programs have been used to map, database, model, and analyze NHC data. Most of these programs and studies have been completed by larger institutions, such as the Smithsonian National Museum of Natural History, the California Academy of Sciences, the American Museum of Natural History, and The New York Botanical Garden, to name just a few. Smaller institutions may not have the funding, volunteer hours, and training available to digitize, map, and model much of their collections, even though studies have found that the data from smaller collections are valuable and should be combined with those of the world's large collections, in the hopes to fill gaps in data. Some programs used for these efforts in the past, include various platforms and iterations of GIS software, various collaborative databases such as NABIN (the North American Biodiversity Information Network), GBIF (the Global Biodiversity Information Facility), and gazetteers, which are geographic indexes and references (Funk et al. 1999, Krishtalka et al. 2000, and El Gabbas et al. 2018).

Ward (2012) graphically organized the data from NHCs in New Zealand, with the goal of helping scientists by suggesting conservation implementations in various ecological niches, as well as creating a more streamlined, and cost efficient collecting strategy for biologists. Ultimately, the researchers found that there are gaps in the data, temporally, spatially, and in the collecting methods used. In the end, they were able to come to a variety of conclusions related to their analyses, some of which are represented as figures, charts, and tables (Ward, 2012). For example, a figure titled, “Proportions of NHC records collected at different time periods from selected area codes” shows trends in density proportions of how much was collected in each region, showing how skewed the data is towards certain time periods (Ward, 2012). This type of trend showing a decrease in collections over time can potentially be explained by declining populations, population ranges shifting, and a lack of funding and internal and external lobbying for collecting efforts (both institutional and governmental). Potential solutions to remedy the skewness of this data, would be to digitize more museum collections to fill in gaps in the existing records,

as well as to conduct more specified collecting efforts in areas that seem to have limited records available. The figure titled, “Frequency distribution of the number of repeat visits at specific locations” can further assist in focusing limited funds and grants on collecting to make sure even collecting is occurring (Ward, 2012). Another figure of interest, titled, “Number of records of introduced species from urban (diamond) and non-urban (square) locations”, shows a double line graph of introduced species (specimens) collected over time from urban and non-urban areas (Ward, 2012). This type of information can help us to understand the impact of invasive or non-native species on various regions over time, and has the potential to influence where and what to look for when trying to create a conservation intervention. Further, this type of data can reveal more about gaps in the current record, and can be improved with the collaboration and digitization of more NHCs of any size. Whereas the first part of this study looked at existing records from the compiled databases, the next part focused on combining the NHC data with public records of land cover and climate data (mean annual temperature and rainfall were used in this study) specific to the regions that corresponded to the NHC data (Ward, 2012). A figure titled, “Comparison of NHC records (black) and background data (white) from different land-cover categories” compares the percentage of records that come from NHC collections and existing data in various bioregions, showing where there are gaps in the data (Ward, 2012). This figure provides yet another example of a ‘cry for help’ to digitize and add existing data from more NHCs to databases worldwide. Lastly, a figure titled, “A principal components analysis (PCA) plot comparing NHC records (red) and background data (black),” compares background data on mean annual temperature and rainfall from areas similar and dissimilar to those where the NHCs have data from (Ward, 2012). This again shows where the gaps are, and where there may be an excess in collections from certain areas, allowing for streamlining. This can help scientists to understand which datasets are able to be analyzed using various models due to sample size and biases.

In Funk et al. 1999, the authors conclude that by increasing available NHC data by 10-15%, these datasets can more accurately represent the populations of concern. Moreover, the costs of improving these datasets can be minimal. By figuring out from where we need more collections and surveys, we can focus efforts, using grant funds more efficiently. Beyond using available resources as wisely as possible, it is vital to combine NHC data from large and small institutions and again combine the datum with open access environmental, climate, and geospatial datasets in order to have a whole understanding of populations and how they have changed and will change over time. Krishtalka et al. 2000 suggest two ways to solve the biodiversity crisis: 1. Deploy the information, making the collections digitally available, and 2. Biodiversity Informatics, which, “...integrates biological research, computational science, and software engineering to deal with biotic data - their storage, integration, retrieval, and use in analysis,

prediction, and decision-making.” Decisively, the predicament of there not being enough data available stems from the obstruction of institutions collaborating more, and limited resources for these bioinformatic efforts. Funk et al. 1999 states that, “A community enterprise can mobilize biocollections information for institutions large and small that, alone, could not finance, develop, or support the essential elements of a biodiversity informatics infrastructure.”

Ultimately, the big questions with respect to a community collaboration of the mapping and modelling of the past, present, and future of NHC specimens are:

- What are the distributions of plants and animals?
- What are the areas of greatest species richness and rarity?
- How well do the data explain the biodiversity of specific regions?
- What areas are in most need of additional collecting efforts?
- Can these data be used in conservation decision-making?
- How can we correct for unequal sample size when dealing with NHCs in specific regions?
- At what spatial resolution can we look at the data?

(Adapted from Funk et al. 1999 and Steege et al. 2000)

## **Issues to Consider When Modelling NHC Collections**

While a good argument can be made to model NHC collections for conservation of flora and fauna, there remains to be significant concerns that must be taken into account before interpreting and inputting NHC locality data. First and foremost, this is not the end all be all cure for the biodiversity crisis, and this type of project should not limit any future collecting and surveying. According to Ward 2012, “... The extent to which NHCs can provide information is often uncertain.” Even more so, Macdougall et al. 1998 suggests that these limitations of NHC data, “lead some to question the value of such information for directing ecologically-based conservation work”.

From a policy standpoint, there are issues regarding museums having enough funding for this type of project, and digitization in general, which would have to be completed before mapping and modelling occurs (Suarez et al. 2004). Moreover, in Funk et al. 1999, Stork et al. 1995 suggests that, “The data are not presented in a format that policy-makers and managers can use”. These two areas are likely related. If policy-makers and managers cannot easily interpret or approach the hard science reports of these studies, how are they able to support future funding, let alone advocate for policies that combat the biodiversity crisis?

From a data standpoint, a wealth of concerns exist regarding the mapping and modelling of NHCs. This list covers most of what is discussed in the literature, speaking more generally about the modelling of NHC specimens. More specific issues will be analyzed in the case studies discussed in the next section. There is no question that the scientific community uses or wants to use NHC data and are trying to develop strategies for their use towards the study of biology. (Adapted from Ward 2012, Macdougall et al. 1998, Syfert et al. 2013, Funk et al. 1999, Steege et al. 2000, Ponder et al. 2001, and Graham et al. 2004).

- The personal interests and curatorial techniques of collectors (e.g. discarding damaged individuals, only accessioning a certain number of individuals, targeting rare or unusual over common taxa);
- The spatial biases where areas have been under-sampled, or where samples are biased towards easily collected localities (e.g. near towns/cities and/or along roadsides);
- Information is often restricted only to the presence of a species (i.e. there is no information on where a species is absent);
- The difficulty of getting information on other taxa from the same location (e.g. NHCs are organized taxonomically, not geographically) rather than systematically or randomly, so their sampled localities may not be representative of the true range of environmental conditions in which the species occurs;
- Accounting for the effects of geographical sampling bias in the acquisition of data can be critical to the accuracy of Species Distribution Models (SDMs) generated from presence-only datasets, but options to correct for sampling bias are not always applied. Failure to correct for geographical sampling bias can result in a SDM that reflects sampling effort rather than the true distribution of a species;
- The locality information accompanying specimens, such as region and habitat descriptions, are sometimes imprecise, especially for older records;
- Coordinates of the collection sites were not always available and were estimated from descriptions on the herbarium labels (or collector trip reports);
- Most species are rare and provided insufficient data for (statistical) analysis;
- A significant proportion of available records for specimens do not have recorded locality information;
- There is reduced confidence in predictive power with small numbers of records and when the data are highly clustered;

- The background-sampling method will not be able to verify the distributions of isolated allopatric taxa, or populations occurring in different geographic regions, surrounded by large areas of habitat unsuitable for members of their background group(s);
- The fundamental premise of the background-sampling analysis (that if the intensity of background sampling is “adequate,” then the mapped distribution probably reflects the true distribution) involves some fairly major biological assumptions—that seasonality, habitat fidelity, and annual population variation are understood. These assumptions cannot always be made, especially for insects and other taxa with high seasonality and host specificity; and
- Inaccurate species-level taxonomic identifications.

### **Suggested Solutions from the Literature**

Graham et al. 2004 suggests that the databasing of fieldnotes, “... will contribute significantly to the detection and correction of errors, improve the interpretation of data, add significant new information (habitats, local conditions, etc.) about the time of collection and, together with increased density of specimen data, and enhance the use of NHC evidence to detect changes in historical versus current properties of the distributions of species.” This may resolve many of the aforementioned list of concerns. Graham et al. 2004 also suggests associating morphological and locality records from NHC specimens to genomic databases (with DNA sequences, and other nucleotide records) in order to keep up with taxonomy changes, which can assist in the correct naming of both genomic records and locality records.

Ponder et al. 2001 suggests a solution to background sampling of isolated allopatric taxa that are irregularly distributed. The authors state that, “This problem can be partially overcome through use of appropriate environmental overlays, in an a posteriori fashion, that exclude the unsuitable areas.” Ponder et al. 2001 also provides some suggestions to alleviate concerns regarding assumptions being met for background-sampling with NHC data., “... Through application of appropriate filters to test particular biological assumptions, such as seasonality and host-plant preference. In the same way, data can also be filtered to include or exclude particular sampling methods.” Ultimately, Ponder et al. 2001 states that, “... Even if the outcomes may be less reliable, [the] methodology can still indicate sampling gaps and give indications of the reliability of distributions by means of the statistical procedures and/or density surfaces.” This means that even if results or inputted data have biases, the results can still be immensely valuable for understanding how to more wholly and fully collect and survey specimens, they also can tell us with how much certainty we can accept our results.

Syfert et al. 2013 discusses a procedure that Phillips et al. 2009 created for modeling presence-only data with sampling bias. This procedure included making pseudo-absences, sometimes referred to as “background data” which, “has a similar geographical sampling bias to that of the presence data (the background data is the set of geographical locations that will be used to train the SDM)” (Syfert et al. 2013). While the process adds a few additional steps to set up the SDM before the data is inputted, it does allow for certain assumptions to be met that otherwise wouldn’t have been just inputting the raw NHC locality data. As Phillips et al. 2009 states, “This is achieved by creating a sampling bias grid representing relative survey effort across the landscape, using the presence localities of a broader group of species within the region of interest (e.g. all bird species if modeling a single bird species), which is used in the SDM training algorithm”. This ultimately turned basic presence-only data into presence-absence data, in order to limit the bias that is inherent when using raw presence-only data, and further, “they demonstrated that predictive accuracy improved when using this approach” (Syfert et al. 2013). This procedure is not the only one used to limit sampling bias concerns with NHC data, however this method does a good job fixing sampling bias to allow for increased power of the outputted SDM with minimal extra steps added to the procedure.

Steege et al. 2000 dealt with resolving the concern that, “Most species were rare and provided insufficient data for (statistical) analysis”. Their procedure, unlike most of the case studies discussed, followed a different approach. They, “overlaid soil and climate maps with species distributions to come up with probable relationships” and did this by linking their database with GIS to extract specified regional data (Steege et al. 2000). They then (mostly) used a simple Chi-squared test to, “assess if plant distributions were non-random with regard to abiotic factors (mainly rainfall and major soil type)” (Steege et al. 2000). Here we see an approach that involves only two additional steps to create background-data and an analysis, combine existing GIS datasets with their NHC datasets, and do a simple statistical analysis by doing a Chi-squared test.

### **Case Studies: Modelling NHCs with Maxent**

“The Maxent software is based on the maximum-entropy approach for modeling species niches and distributions. From a set of environmental (e.g., climatic) grids and georeferenced occurrence localities (e.g. mediated by GBIF), the model expresses a probability distribution where each grid cell has a predicted suitability of conditions for the species” (Phillips et al.).

Why does it matter which model is used? Why can't each model be used on all datatypes? Why is a model like Maxent useful for NHC locality data? These are all questions that can be answered through the review of the following case studies. Graham et al. 2004 has produced this figure, “Box 1”, where the basics of Ecological Niche Modeling are laid out.

#### **Box 1. Strategies and methods for ecological niche modeling**

The ecological niche of a species can be defined as the set of conditions and resources necessary for an organism to maintain a viable population [57]. By integrating known occurrences of species with environmental GIS data layers that summarize meaningful niche dimensions, it is possible to determine the key suites of environmental conditions for that species (and, therefore, its approximate niche). Statistical models are used to develop relationships between environmental values and species presence (and absence, in some cases). This relationship can then be mapped spatially to predict potential geographical distributions [58,59].

##### **Modeling methods**

The methods for distributional modeling (reviewed in [60]) vary in their applicability to natural history collections (NHC) data, and should be selected based on the nature of the question and the data [61,62], as well as statistical issues [60]. Some modeling methods (e.g. BIOCLIM, [63] and DOMAIN, [64]) require only the records of species presence; others incorporate multiple predictive approaches with varying requirements (e.g. GARP, [65]); whereas others require both the presence and absence data (e.g. general linear and additive models,

and decision trees; [66]). Bayesian approaches are currently being revived because they formally and explicitly combine estimates of sampling bias, ‘expert opinion’, or other previous information, with observations to build posterior distributions for predictor variables and, in turn, predictions of geographical range [67].

##### **Model evaluation**

Substantial progress has been made recently with methods to evaluate models [60,61], although, again, researchers should consider the nature of the question as a guide for the choice of methodologies [61,62,68]. Some assessment has been made of the effects of sample size on the performance of NHC-based models [69]. Few studies have evaluated the predictive performance of models based on NHC data using independent, high-quality presence-absence data sets [70] and little attention has been given to the effects of data error and bias on performance (Box 3). Filling this gap is necessary to inform users of NHC data about which of the many approaches to modeling and evaluation should be applied in relation to specific questions and data sets or, perhaps, when such data should not be used for predictive modeling.

A study conducted by El Gabbas et al. 2018, looked at how three different species distribution model analysis methods modelled biased bat survey data from Egypt (by using presence-only data). The three models that were used include: Generalized Linear Models (GLMs) with subset selection (Poisson Regression), GLMs fitted with an elastic-net penalty (Poisson Regression, and Lasso and Ridge Regularization), and Maxent (El Gabbas et al. 2018). A process called “block cross validation” was used in this study, as it, “is commonly used for evaluating model performance when no independent data are available” (El Gabbas et al. 2018). This is done, “To maintain independence between folds and improve transferability of models” (El Gabbas et al. 2018). In order to do this, spatial cross-validation was implemented by creating a grid with randomly placed “blocks” over the region in which was being looked at (various parts of Egypt) (El Gabbas et al. 2018). In regards to sampling bias, this study evaluated the difference between models that corrected for biases (two separate methods) and those that did not (El Gabbas et al. 2018). The first bias that the authors integrated covariates for in their SDMs, were related to, “distances to nearest cities, roads and protected areas,” which they call the “Accessibility Model” (El Gabbas et al. 2018). The next source of bias that they created a covariate for, was related to, “relative intensity of sightings of all bat species,” which they call the “Effort Model” (El Gabbas et al. 2018). Their results show that, “sampling bias, if not corrected for effectively, can substantially affect the predicted intensity and model evaluation of SDMs” (El Gabbas et al. 2018). More specifically, the covariates related to accessibility bias led to higher validation scores than that of the covariates related to sampling

efforts (El Gabbas et al. 2018). However they found that removing bias with covariates may not alleviate enough of the sampling bias that exists in presence-only or bias-free presence-absence data (El Gabbas et al. 2018). The results of the models do not vary too much, but show a slight advantage for using the Maxent software (El Gabbas et al. 2018). Ultimately, El-Gabbas et al. 2018 found that, “Augmenting local records with data from across the species’ range allowed [them] to make consistently high-quality predictions to hold-out data from an entire country (in this case Egypt). Bias-free predictions can enhance future conservation planning and target future surveys when limited resources are available to cover large study areas. However, due to possible lower certainty at unsurveyed locations, they should be used cautiously (maps including bias are of use only during model cross-validation).” The conclusion that is made here, is that ideally unbiased presence-absence data should be used (El Gabbas et al. 2018). However, in most cases the only available geographic data comes from biased, spotty records, which can be resolved by integrating covariates related to sampling efforts and accessibility (depending on which are more biased in the specific dataset) (El Gabbas et al. 2018). Through this study we can see that without a doubt, through these simple data-fixes and Maxent, conservation planning is possible to model and implement. [Figure 1 describes a flow chart of El-Gabbas et al. 2018’s workflow and protocol, it can be found in Appendix C].

Phillips et al. 2009 wrote a paper titled “Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data” where the authors essentially have created a loose protocol for modelling NHCs through the use of Maxent. Through both “target-group background” data and “randomly sampled background” data of 226 species from various bioregions, overall they found that, “... target-group background improves average performance for all the modeling methods we consider, with the choice of background data having as large an effect on predictive performance as the choice of modeling method” (Phillips et al. 2009). Like other studies discussed in this literature review, Phillips et al. 2009 finds that increasing background data and pseudo-absences through available online databases, filling in gaps in the original datasets, can significantly affect the accuracy and reliability of the resulting SDMs. Like other studies in this review, Phillips et al. 2009, also used GAM (Generalized Additive Models), MARS, and Maxent (with default settings). Again, cross validation was used in this study, like in El Gabbas et al. 2018. Models were made for: “Presence-absence with random background, presence-absence with biased background, Maxent with unbiased samples, [and] Maxent for biased samples” (Phillips et al. 2009). Each of these models has two additional background datasets, also called “random background” in this study (because random samples are an assumption of the model) (Phillips et al. 2009). Area under the curve (AUC) was used to test how well the data fit the predictive models. Phillips et al. 2009 defines AUC as:



“... The probability that the model correctly ranks a random presence site vs. a random absence site, i.e., the probability that it scores the presence site higher than the absence site. It is thus dependent only on the ranking of test data by the model. It provides an indication of the usefulness of a model for prioritizing areas in terms of their relative importance as habitat for a particular species. AUC ranges from 0 to 1, where a score of 1 indicates perfect discrimination, a score of 0.5 implies random predictive discrimination, and values less than 0.5 indicate performance worse than random.”

Many of the other studies discussed in this review as well as studies by other ecologists use AUC as a metric to determine how data agrees with a model's predictions. After this, Spearman's rank coefficient, “which is a nonparametric measure of correlation” was used to test, “whether there is a monotone relationship between two variables” (Phillips et al. 2009). To be able to quantify how much bias that existed in each target group, the authors of this study estimated, “... How well we can discriminate target-group sites from the background, by using Maxent to make a model of target-group sites and using the AUC of the target-group sites vs. background as a measure of discrimination” (Phillips et al. 2009). The authors call this AUCTG for short (Phillips et al. 2009). A high AUCTG value means, “environmental variables can be used to distinguish the spatial distribution of target-group presences from random background, and therefore target-group presences sample environmental space in very different proportions from the proportions present in the study area, i.e., the target-group presences are biased both in environmental and geographic space” (Phillips et al. 2009). This metric, AUCTG, is suggested for future studies to integrate into their analyses, as it can help to determine how reliable a model's outputs are based on how biased the inputted data originally was (Phillips et al. 2009). Phillips et al. 2009 ultimately suggests that it is just as important to correct for biases in the data, as it is to choose the most appropriate model based on the type of data. [Tables 1,2, and 3, describe various metrics, how they change when being corrected for biases, as well as different types of pseudo absences, and can all be found in Appendix C].

In a paper titled, “Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections”, Mateo et al. 2010, analyzed whether it is possible to create accurate and reliable models from NHC data through the use of pseudo-absences on five species of plants in the genus *Anthurium*. Six methods in total were used (Mateo et al. 2010). Group Discriminative techniques that were used were: MARS (Multivariate Adaptive Regression Splines), Maxent, and LMR (Logistic Multiple Regression) (Mateo et al. 2010). Profile techniques that were used were: BIOCLIM and Gower's distance index (Mateo et al.

2010). And lastly, a combination of Profile and Group Discriminative approaches, GARP (Genetic Algorithm for Rule-set Production) was used, which actually, “generates its own pseudo-absences” (Mateo et al. 2010). For the Group Discriminative techniques that were employed, “... Three types of absences were generated: (1) random pseudo-absences in equal number to presences and excluding a buffer area around presences (except for Maxent, which assumes that this background sample includes presences), (2) a large number (10,000) of random pseudo-absences, also excluding a buffer area around each presence and (3) ‘target-group absences’ (TGA), consisting of sites where other species of the group have been collected by the specialist, but not the species being modelled” (Mateo et al. 2010). Here we see three subsets of background-data being created from existing and available databases online. The conclusions made by Mateo et al. 2010 are that if there is enough data to employ these models, “... Use group discriminative techniques as they are more reliable than profile techniques”. This narrows down to the three tests, LMR, MARs and Maxent, as they conclude that the other three models tend to overpredict and are therefore not as reliable with this type of data (Mateo et al. 2010). They then suggest to, “... Use MARS or other regression techniques if the aim is to perform an in-depth analysis of each species, or Maxent when a less intensive analysis of large numbers of species is required”, which tells us that for low budgets and man hours, Maxent really is the best choice for NHC modelling (Mateo et al. 2010). Beyond this, they suggest to, “... Use TGA instead of pseudo-absences if sufficient data are available to generate them and ... if forced to use pseudo-absences, create a buffer around each presence to minimize the false-negative rate” (Mateo et al. 2010). The major takeaways from Mateo et al. 2010 are: that Maxent is the most efficient model for NHC data, and that target-group absences and pseudo-absences are integral in decreasing biases in NHC data, and can be found through open access datasets/databases.

Through the use of presence data of tree ferns in New Zealand the results of Syfert et al. 2013, show that NHC data that has bias can be modelled accurately through the use of Maxent. Further, the authors argue that biased herbaria data can be mapped for conservation purposes through Maxent (Syfert et al. 2013). They used an “... Online biodiversity data portal (GBIF, presence-only), with two sources that differed in size and geographical sampling bias: a small, widely-distributed set of herbarium specimens and a large, spatially clustered set of ecological survey records (NVS, National Vegetation Survey Bank, presence-absence)” (Syfert et al. 2013). This is a key addition to the study, because it allows the data to be “geographically and environmentally comprehensive, thus providing a more reliable evaluation of performance than obtained from subsampled data” (Syfert et al. 2013). They then added “sampling bias grids” to each of their SDMs, as well as “fitting a wide variety of environmental response curves” to try to account for some of the sampling bias present in the data (in Maxent these are called “feature types”) (Syfert et al. 2013). Additionally, Maxent’s default option, “... allows the software to

automatically select functional forms to describe species' responses to environmental conditions, but users can select from a list of functional forms (i.e. linear, quadratic, threshold, hinge, product, and categorical) and allow different functional forms for different environmental variables" (Syfert et al. 2013). This can allow for slight variations to be made for each SDM due to the variety of data types even in a single NHC database (i.e., who collected, collecting methods, rarity, etc). Syfert et al. 2013 then, "...used MaxEnt to fit ("train") a species distribution model to a random sample of 75% of the species occurrence data, with the remaining 25% of the data used to assess ("test") model performance". This was then repeated 40 times for each subsample used in the model. They then, "used MaxEnt to assess uncertainty of the SDM predictions" (Syfert et al. 2013). The authors evaluated the goodness of fit for each SDM by, "... comparing predicted range maps with tree fern presences and absences using an independent national dataset" (Syfert et al. 2013). While correcting the data to resolve some of the sampling biases did improve the goodness of fits, the results show that the issues were not fully fixed (Syfert et al. 2013). Ultimately, they suggest that fixing the NHC data with background data even in certain areas will likely help the results just as much as correcting for all of the data (because there is only so much bandaging the problem can resolve) (Syfert et al. 2013). Conclusively, Syfert et al. 2013 follows and "endorses" Phillips et al. 2009, as an easy, efficient, and reliable way to account for bias in SDMs, in order to create simple and easy to understand ecological niche models for the conservation of threatened plant species. (Tables 4 and 5 show how correcting for sampling bias effect predictive performance, false presences, and false absences, can be found in Appendix 6). (Figure 2 showing: presence/absence locations with predicted presences and absences generated from average LQ model predictions (with geographical sampling bias correction), can be found in Appendix C).

In a recent review of 108 articles published between 2008 and 2012, Yackulic et al. 2012 analyzed various recurring trends in studies that were creating SDMs through Maxent, with presence-only data. The authors were skeptical of the studies being analyzed, as to whether they were addressing, and assuming assumptions were being met, when in fact they may not have been (Yackulic et al. 2012). In fact, Yackulic et al. 2012 found that 87% of studies used samples that likely had selection bias, but biases were rarely resolved or mentioned. Further, a meager 14% of studies, "mentioned detection probability" (Yackulic et al. 2012). Here, we see biologists organizing from the literature, a real concern that their fellow science community are not using modelling tools like Maxent appropriately, and moreover, they are taking advantage of a tool without understanding its own limitations. Another fascinating statistic from this review is that 54% of the studies analyzed falsely interpreted the outputs of Maxent (Yackulic et al. 2012). [Table 6, which shows other statistical findings from this review can be found in the Appendix]. Because of these shocking mis-uses of a tool that can be valuable for bias limiting data like NHC data, the

authors of this review devised a few measures to decide what corrections certain biased data need in order to appropriately be inputted into Maxent. The first suggestion that Yackulic et al. 2012 make, revolves around the type of data that will be analyzed: is it presence-absence, if it's presence-only, how randomly were the samples collected? Here, the concern is that there is a random sample, with both presence and absence data, so if there is no absence data, background data will most likely need to be made. The second suggestion the authors make, is concerned with the method of sampling; for example, they state that if the detection probability is less than one, "could detection probability vary with the environmental covariates that determine occurrence probability?" (Yackulic et al. 2012). The authors state that, "in the event that detection probability varies with respect to covariates that also determine occupancy, data are not amenable to presence-absence analysis and data were collected through a standardized sampling scheme, it may still be possible to estimate occurrence probability if additional data (e.g. multiple visits to a subset of sites) are collected to estimate how detection probability varies with the covariates that determine occupancy" (Yackulic et al. 2012). Here we see a suggestion to make more collections or surveys, which would only be relevant to more recent collections in NHCs being studied. They then suggest creating a, "potential a priori hypothesis" which can then be compared with the results of "modelled relationships" (Yackulic et al. 2012). The main concern is that even though this seems like an obvious approach, many scientists are just accepting the results of Maxent even if these results differ significantly from their hypotheses, which means there may be issues with controlling for biases or the created covariates may need to be adjusted (Yackulic et al. 2012). Further, creating a posteriori hypothesis, "could be tested through additional data collection" (Yackulic et al. 2012). This process should continue and repeat until there are very minimal if not any concerns about the biases present in the inputted data. The last, and one of the most important takeaways from this review, is to:

"Provide readers with the necessary information to critically evaluate your results. A hallmark of scientific reporting is that future researchers should be able to compare results of their studies to yours. Maps alone do not provide sufficient output to allow for this, and inclusion of estimated response curves and parameters, either in the bodies of studies or in appendices, would greatly improve the transparency and usefulness of presence-only studies. Moreover, authors should be encouraged to make original data available, when legally appropriate, through online appendices or data repositories."

In all, Yackulic et al. 2012 provides yet another example of how complex data can be (especially presence-only data), and that even though there are serious limitations to its uses, when handled and corrected for properly, tools like Maxent can be implemented appropriately. [Table 7 titled, "Corrections

required for both presence-absence and presence-only analyses under various assumptions”, can be found in Appendix C).

## **Conclusion**

The effects of modelling NHCs can influence how humans implement conservation changes in flora and fauna communities and ecosystems. Through the use of legacy data (old NHCs and their associated locality/collection information), data correction (background data or pseudo absences), and Maxent, research has shown that it is feasible to create a low budget protocol/setup to project the past, present and future of species population changes. This has been done in the past few decades as more collections and their locality data have become digitized, potentially allowing more natural history collecting institutions and scientists to participate in more conservation projects. Larger institutions like the American Museum of Natural History in New York, the California Academy of Sciences in San Francisco, as well as the Smithsonian National Museum of Natural History in Washington, D.C., have participated in such efforts as their collections have been digitized, but this has rarely been executed by smaller NHCs, however this can and must change. We can learn from how past and present population ranges have changed due to climate change, urbanization, and deforestation (among other changes) to be able to project where species ranges could exist in the future.

## Chapter 3: Proposed Project

### Goals and Objectives

The first goal of this project is to conserve threatened or at risk species and ecosystems. This can be achieved by using SDMs, in order to focus conservation, digitization, and collecting spending budgets based on the findings of the outputted models. The second goal of this project is to model the past, present and future of species ranges. This can be attained by determining the type of data that any NHCs have (small-large collections), and what types of covariates and background data are available open source to combine with the dataset to attain presence-absence data for input into the modelling software (which in this case will be Maxent). Why is Maxent the best choice for a project like this? It has a simplified methodology, while at the same time being accurate and user friendly, it uses presence-only data (which is what is available from all NHCs), and the software is open source with a plethora of available tutorials, guides and updates that are widely used in the ecology and evolutionary biology research community (it was also made by the American Museum of Natural History). In all, with the help of this protocol, formulated models that are somewhere between publication quality and raw data pin drops will be made, and can potentially be used as evidence for people asking for funding to make conservation changes, as well as to digitize, and model more NHCs.

### Program Specific Questions and Concerns:

- What type of background data do we need to create a more random, representative sample?
- Where can we access the data online/open access resources?
- Have to remember to use the right subcategories for maxent that fit the specific datasets used. The same goes for the statistics that are used to interpret the data.

The broader questions that I have identified through the literature review that can be answered through this proposed program, are:

- What are the distributions of plants and animals?
- What are the areas of greatest species richness and rarity?
- How well do the data explain the biodiversity of specific regions?
- What areas are in most need of additional collecting efforts?
- Can these data be used in conservation decision-making?
- How can we correct for unequal sample size when dealing with NHCs in specific regions?

- At what spatial resolution can we look at the data?

(Adapted from Funk et al. 1999 and Steege et al. 2000)

At the same time, it is worth noting the problems and biases that can be limited through following the methods of this proposed program:

- The personal interests and curatorial techniques of collectors (e.g. discarding damaged individuals, only accessioning a certain number of individuals, targeting rare or unusual over common taxa);
- The spatial biases where areas have been under-sampled, or where samples are biased towards easily collected localities (e.g. near towns/cities and/or along roadsides);
- Information is often restricted only to the presence of a species (i.e. there is no information on where a species is absent);
- The difficulty of getting information on other taxa from the same location (e.g. NHCs are organized taxonomically, not geographically) rather than systematically or randomly, so their sampled localities may not be representative of the true range of environmental conditions in which the species occurs;
- Accounting for the effects of geographical sampling bias in the acquisition of data can be critical to the accuracy of Species Distribution Models (SDMs) generated from presence-only datasets, but options to correct for sampling bias are not always applied. Failure to correct for geographical sampling bias can result in a SDM that reflects sampling effort rather than the true distribution of a species;
- The locality information accompanying specimens, such as region and habitat descriptions, are sometimes imprecise, especially for older records;
- Coordinates of the collection sites were not always available and were estimated from descriptions on the herbarium labels (or collector trip reports);
- Most species are rare and provided insufficient data for (statistical) analysis;
- A significant proportion of available records for specimens do not have recorded locality information;
- There is reduced confidence in predictive power with small numbers of records and when the data are highly clustered;
- The background-sampling method will not be able to verify the distributions of isolated allopatric taxa, or populations occurring in different geographic regions, surrounded by large areas of habitat unsuitable for members of their background group(s);

- The fundamental premise of the background-sampling analysis (that if the intensity of background sampling is “adequate,” then the mapped distribution probably reflects the true distribution) involves some fairly major biological assumptions—that seasonality, habitat fidelity, and annual population variation are understood. These assumptions cannot always be made, especially for insects and other taxa with high seasonality and host specificity; and
- Inaccurate species-level taxonomic identifications.

(Adapted from Ward 2012, Macdougall et al. 1998, Syfert et al. 2013, Funk et al. 1999, Steege et al. 2000, Ponder et al. 2001, and Graham et al. 2004).

This project has three major groups of stakeholders:

- IUCN (International Union for the Conservation of Nature)
  - At the highest level, the IUCN could be impacted by a project like this, as the IUCN are the people that list species as threatened, endangered, etc. If this project can succinctly model species, it can help to focus spending on implementations to conservation efforts worldwide.
- SPNHC (Society for the Preservation of Natural History Collections)
  - Small to medium sized non-profits and societies can help to advertise this type of project and get feedback from the academic and museum community to help streamline the protocol and methodology (keep it technologically updated).
- Collections managers, museum workers
  - The actual museum staff who will be in charge of this type of project, are the collections managers and curators, IT department staff, as well as interns, so for this reason they are arguably the most immediate stakeholders for this project. They are the ones that are responsible for the dispersal of the results of this type of project, and therefore have the potential for the outputted models and results to be spread to researchers and conservation nonprofits who can actually make environmental changes based on the data.

The individuals who will help to make this program technically function are:

- Maxent, GIS, and other SDM and geographic analysis community members
  - These software developers can provide invaluable insight as to how to update the protocols for this project, as well as suggest other methodologies for museums that have the budgets to implement more expensive protocols.
- Biologists



- Researchers can provide insight to conservation nonprofits and the IUCN by interpreting the results of the outputted models, without which a project like this could not fully affect the implementation of environmental policy and conservation.
- Citizen scientists and museum volunteers

The fundamental intention with this project is that a simplified, open source modeling software will be used, meaning that minimal training to citizen scientists and volunteers could allow this to be very low cost if not cost-free to realize. If collections managers, research assistants, curators and other higher up staff in NHCs were to become more well versed in Maxent (which would not take more than a week of simple training), and they were to then train lower level staff including volunteers, this type of project could be seamlessly integrated into the daily tasks of curatorial assistants, technicians and volunteer duties. Similarly, there are open source sites that exist for citizen scientists to enter label data from specimens from their own personal computers at home. If NHC locality data were to be cleaned and ready to be uploaded to Maxent, there could be a similar platform and protocol for museum workers to upload data to Maxent and report the results and related statistics. This not only gets the interest of this project beyond the walls of NHCs, but it also allows non-academic citizens (citizen scientists) to participate in potential conservation related work, if they are unable to do physical surveys or other laborious tasks usually associated with doing work in the field (accessible to those with disabilities).

The resources that are needed to complete this project, are:

- Team members: volunteers, citizen scientists, collections managers, GIS and Maxent training
- Software licensing (free for maxent, statistics software “R”, and arcGIS).
- Digitize collection prior to this (need to have labels cataloged and available to be cleaned up and organized to be uploaded to Maxent).

## **The Proposed Program**

The computer program that I propose to be developed should be designed to address all of the issues and needs of the stakeholders listed above. It will allow for the following questions to be answered (listed in the checklist below), and will use the answers to compile available background data that is open source. The program will then clean the data so that it can be simply uploaded to Maxent, and will also provide suggestions based on the types of data and the results of the Maxent models, of what type of statistics should be calculated in a program like R. These questions and their various choices to answer from are a

compilation of the types of data, variables, models, and statistics that were discussed in the articles analyzed in the literature review. The protocol and backbone of the proposed program can be seen in Figure 3 [see Appendix C, or the checklist below]. Additionally, Table 7, which is described in full in the literature review, is recommended to be used as a stand alone, separate reference from the suggested program, as an alternative to further understand specific corrections that need to be done to various types of data based on the assumptions that are used in particular analyses. Ideally, the resulting models and statistics will be made available on some sort of collaborative web source that will allow any type of collection (private, or museum) to be uploaded and discussed among researchers, policy makers, and conservationists. All that a volunteer, private collector, or museum staff member will have to do is use this program, upload their data to it, download Maxent, upload their cleaned data from this program to Maxent, calculate statistics, and then upload models, maps and statistics to the collaborative site. No computer coding or database management training is needed, meaning really anyone could assist with this project no matter their prior background in computing, natural history, biology, and statistics.

<p><b>Data:</b></p> <p>What kind of data is available?</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> Presence - absence</li> <li><input type="checkbox"/> Presence - only</li> </ul> <p>How much data do we have?</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> Enough for a representative sample</li> <li><input type="checkbox"/> Not enough</li> <li><input type="checkbox"/> If this data is from a recent study can more samples be taken (ONLY FOR CURRENT STUDIES)</li> </ul> <p>Are the data:</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> "sites as fixed effects" or</li> <li><input type="checkbox"/> "sites as random effects"</li> </ul> <p>Which biases are more prevalent in the data?</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> Accessibility Model (distance to cities, roads, protected areas)</li> <li><input type="checkbox"/> Effort Model (relative intensity of sightings)</li> </ul> <p>What covariates would best "fix" the dataset?</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> Environmental (soil, temp, rain etc)</li> <li><input type="checkbox"/> Target group background - sites where other species of the group have been collected by the specialist but not the exact same species</li> <li><input type="checkbox"/> Randomly sampled background             <ul style="list-style-type: none"> <li><input type="checkbox"/> Random pseudo-absences in equal number to presences</li> <li><input type="checkbox"/> Large number of pseudo-absences</li> </ul> </li> </ul> <p>*if have to use pseudo-absences - create a buffer around each presence to minimize the false negative rate</p> <p>Where can we access this background data online?</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> USGS, NOAA (climate and land cover)</li> <li><input type="checkbox"/> GBIF (presence-only)</li> <li><input type="checkbox"/> NVSB (presence-absence)</li> <li><input type="checkbox"/> Any med-large NHC will have locality data open access (some with coordinates)</li> </ul>	<p><b>Modelling and Analysis:</b></p> <p>Which Maxent functional form works best with the data?</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> Linear</li> <li><input type="checkbox"/> Quadratic</li> <li><input type="checkbox"/> Threshold</li> <li><input type="checkbox"/> Hings</li> <li><input type="checkbox"/> Product</li> <li><input type="checkbox"/> Categorical</li> <li><input type="checkbox"/> Default settings (allows software to automatically select functional forms to describe species' responses to environmental conditions)</li> </ul> <p>What is your a priori hypothesis for the model?</p> <p>What is your posteriori hypothesis for the model?</p> <p>What other tests should be done to assess the model?</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> AUC or AUCtg (area under the receiver operating characteristic curve)</li> <li><input type="checkbox"/> Spearman's rank</li> <li><input type="checkbox"/> Chi square</li> <li><input type="checkbox"/> Block cross validation             <ul style="list-style-type: none"> <li><input type="checkbox"/> Spatial cross validation</li> </ul> </li> <li><input type="checkbox"/> Sampling bias grids</li> <li><input type="checkbox"/> Environmental response curve</li> </ul>
---	---

## Conclusion

The ultimate goal of this project is to provide a way to share the results of the Maxent models and associated statistics of NHC data, even if not publication worthy, i.e. to larger stakeholders,

environmental policy makers or non-profits, to allow scientists to follow up on methods and results to see if there really are possible conservation concerns. Interns, citizen sciences, collections workers (non-phd scientists) can do this in smaller NHCs, and report their findings from their collections. This project has the potential to have a broader impact on rare species housed in smaller collections. Further, it has the capacity to be able to allow for specific species and biota to be conserved with the help of precise small grants for specified flora and fauna to be modelled. Budgeting for conservation implementations and policy need to be backed by research and data showing that species are at risk or threatened in an environment. A project like this can be the type of evidence a policy maker can use in order to receive grants and funding, even more so, grants could be written to fund people to get paid to work solely on projects like this. Beyond this, while digitization is important, it takes a very long time, especially for large collections. Usually mapping is the last phase of digitization, but in the amount of time that the first phases of the project are done, certain species and ecosystems may have shifted and have become threatened due to the effects of climate change, urbanization, and deforestation, among other causes. And while this type of project is not the end all be all cure for the biodiversity crisis, it can be a possible way to use available resources and technology for the advancement of our planet and its inhabitants.

## Chapter 4: Conclusions

This capstone examines the use of species distribution models that use presence-only data, their accuracy, and the types of data available to input into models from natural history collections. In the first section of this capstone, in an analysis of the literature, I have reviewed case studies that use the species distribution software, Maxent, with historical natural history collection data in order to understand and project various species changing ranges over time. Though I was able to conceive of a very feasible protocol, if I were to pursue this project further, I would like to create and test an open-source program that is a precursor to Maxent that compiles internal presence-only data with various options for external, background data depending on the needs of datasets, that finally cleans the data and readies it to be inputted into Maxent, and R for statistical analysis. By creating a simplified, streamlined program, I envision that nearly any collection type, small, large, and on any budget could easily participate with minimal training. It would also be valuable to be able to have feedback and input from the GIS and biological academic community to tweak this protocol so that it is viable and realistic to implement. In all, my hope in exploring the potential that a species distribution modeling software like Maxent has in natural history collections, is to be able to provide evidence to larger stakeholders like the IUCN of more species and biodiversity that are at risk so that funding and resources can be used towards their conservation before it is too late. Ultimately this capstone speaks to an issue of the lack of digitization and availability of open access data at the deepest level, so with a project like this, we can participate in a new level of widespread information sharing and collaboration.

## Appendix A: Annotated Bibliography

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

This study looks at how three different species distribution model analysis methods modelled biased bat survey data from Egypt (presence-only data). The three models that are used are: GLMs with subset selection, GLMs fitted with an elastic-net penalty, and Maxent. The results show that the models do not vary too much, but show a slight advantage for using the Maxent software. This paper will be used as a case study to show why Maxent should be used for biased presence-only data, like that which exists in herbaria, to be able to most accurately model species distributions.

Funk, VA, Fernanda Zermoglio, M, Nasir, N. 1999. Testing the use of specimen collection data and GIS in biodiversity exploration and conservation decision making in Guyana. *Biodiversity & Conservation*, Vol. 8, 727-751.

This extensive study shows how highly funded governmental resources and large institutions (The Smithsonian Institution) can use their combined datasets and models to conserve various areas and species. ARC/INFO, ArcView, a database, and a gazetteer were compiled for the analysis and modelling of the species in collections from Guyana. Besides this paper being used as a case study in my literature review, this paper asks questions that I will be integrating more broadly into my thesis: “What are the distributions of plants and animals in Guyana?”, “What are the areas of greatest species richness and rarity?”, “How well do the data explain the biodiversity of Guyana?”, “What areas of Guyana are in most need of additional collecting efforts?”, and, “Can these data be used in conservation decision-making?” Of additional interest, is the ability to convert multiple database formats into a streamlined single platform, as well as lists of concerns when using these types of datasets (presence-only) in drawing accurate, significant conclusions. The authors also discuss that by increasing the available data by 10-15%, this will allow for these datasets to fully represent the populations of concern, and that the costs of improving these datasets can be minimal. By figuring out where we need more collections and surveys from, we can influence streamlined grants to be written in the future (as to not survey places that don’t need to be surveyed as much as other places). The authors also suggest to combine this data with environmental/geospatial datasets (all of which are open access).

Gaubert, P, Papes, M, Peterson, T. 2006. Natural history collections and the conservation of poorly known taxa: Ecological niche modeling in central African rainforests genets (*Genetta* spp.). *Biological Conservation*, Vol. 130, no. 1, 106-117.

This paper models three species of rainforest genets in central Africa, through the use of natural history data. The model used in this paper is an Ecological Niche Model, through the use of GARP (Genetic Algorithm for Rule-Set Prediction), environmental data layers, and georeferenced localities from the literature and museum databases. Their findings show that most of the localities analyzed fit in the species suitability guidelines, however those found at the outskirts are recommended to be surveyed again, in which case could indicate some unusual patterns of geographic ranges otherwise unknown. This will be used as a case study (although looking at animals), showing that combined datasets from natural history collections and databases (not just NHC data, but also literature and environmental), can yield the most well rounded and accurate models of species.

Graham, CH, Ferrier, S, Huettman, F, Moritz, C, and Townsend Peterson, A. 2004. New developments in museum based informatics and applications in biodiversity analysis. *TRENDS in Ecology and Evolution*, Vol. 19, no. 9, 497-503.

This article largely discusses the pros and cons of using informatics to analyze metadata contained in museum (natural history collections) databases. They list models which can be used to project species population spread, among other analyses like DNA analyses that are not going to be used in my literature review. An integral area of this paper that I will utilize in my discussion, is all of the potential downsides to using NHC metadata in analyses, as well as a list of potential errors that can arise from using this type of data in various modelling methods. Lastly, this paper discusses implications for using museum collection metadata towards conservation of biodiversity, while keeping in mind that many of these projections are not as accurate as if the data were taken from more thorough studies. In all, this paper can be used in my discussion of modelling platforms that are appropriate to use with NHC locality data, their associated conservation implications, as well as the serious concerns and biases that need to be understood while analyzing the results of these models.

Krishtalka, L, and Humphrey, PS. 2000. Can natural history museums capture the future? *BioScience*, Vol. 50, no. 7.

Natural history museums here have four major challenges in which they can participate in making change in: the biodiversity crisis, education, public programs, and management and leadership. I am focused on their discussion of the biodiversity crisis, and will integrate their arguments in my introduction

where I will analyze the value of natural history collections and museums, and why it is imperative for them to be actively involved in all four of the aforementioned categories. The subcategories of the “biodiversity crisis” section of this paper include: deploy the information, biodiversity informatics, barriers, and solutions. In these categories the authors discuss the types of data present, and the possible uses for these datum. So, additionally, I will be using this article to discuss the barriers to this type of data, as well as how some larger inter-museum databases exist, and how they combine data to analyze species distributions (ie: NABIN, the North American Biodiversity Information Network).

MacDougall, AS, Loo, JA, Clayden, SR, Goltz, JG, and Hinds, HR. 1998. Defining conservation priorities for plant taxa in southeastern New Brunswick, Canada, using herbarium records. *Biological Conservation* , Vol. 86., 325-338.

This paper looks at internationally held herbarium records from New Brunswick, Canada, with the ultimate goal of creating land management and conservation changes in the region. Specifically, habitat types and regions were sectioned off, and the authors were ultimately able to identify previously unknown habitats in New Brunswick. Models were not used in this study, instead the focus was on rare and uncommon plant species located in the herbarium records, in order to understand the habitats and their conservation needs. This will be used as a case study showing that even without models, underutilized herbarium data can affect our understanding of regional ecologies. What this study efficiently does, is outlines habitats with rare or uncommon species, and suggests that Land Managers in Canada should survey them to maintain their ecologies.

Ponder, WF, Carter, GA, Flemons, P, and Chapman, RR. 2001. Evaluation of museum collection data for use in biodiversity. *Conservation Biology* , Vol. 15, no. 3, 648-657.

This paper, while twenty years old, shows an effective and feasible methodology for organizing geographical metadata held in natural history collection databases. They specifically organize data by how accurate to the coordinate the recorded locality information is (by categorizing distance in meters). They also suggest combining this presence data from museum collections with open access environmental data, to overlay and cross examine with (ie: temperature, rainfall, and seasonality data), to assist in the prediction of species distributions. This study uses BIOCLIM models, not MaxEnt, however their approach helps us to understand why certain models do not work for this type of data even if it is combined with other existing environmental survey data. Further, this study clearly lists the concerns and assumptions that should be cautioned to researchers and collections managers before accepted all results as 100% true, which will be used in my discussion of the downside to using these data types with these

modelling softwares. Ultimately, this paper suggests guidelines to follow in order to have the minimum amount of bias in analyzing this type of data.

Reutter, BA, Helfer, V, Hirzel, AH, and Vogel, P. 2003. Modelling habitat suitability using museums collections: an example with three sympatric *Apodemus* species from the alps. *Journal of Biogeography*, Vol. 30, 581-590.

This study uses species distribution modelling software to create distribution maps of three alpine mouse species from Switzerland. The data for which these models were created from, were all presence-only from museum collection. This study does not use MaxEnt, it instead uses a software called “Ecological-Niche Factor Analysis (ENFA)”. This paper will be used as a case study for using natural history collection data to model habitat suitability, and species distribution of species, however it will be in a more broad discussion of modelling as it is not specifically using MaxEnt software.

Rhoads, AF, and Thompson, L. 1992. Integrating herbarium data into a geographic information system: requirements for spatial analysis. *Taxon*, Vol 41, no. 1, 43-49.

This paper is an early example (1992) of plant biologists attempting to create a standard protocol for databasing locality information for plant specimens in herbaria. This study uses the Pennsylvania Flora Database to show how different recording practices by naturalists over time make it difficult to standardize how to create comparable database records in herbaria. They ultimately recommend that this is something that must be standardized in order for the valuable data to actually be used by GIS and other mapping and SDM softwares as an eventual, “... tool in studies of endangered species and conservation efforts.” This study will be used to help argue that a coding system like the one that is suggested in this paper, can assist in the streamlined procedure of databasing herbaria records that will be ultimately mapped and used for conservation and research.

Shaffer, HB, Fisher, RN, and Davidson, C. 1998. The role of natural history collections in documenting species declines. *TREE*, Vol. 13, no. 1, 27-30.

This paper describes the wholehearted value of natural history collections and how they can specifically be used to document the decline of species and what that entails in complex ecological structures and species. This paper was written in 1998, however I will be using this paper to further my argument of what information natural history collections hold. While the technologies discussed in this paper are largely outdated, at the core, they discuss how the specific types of biased data found in natural history collections can be categorized based on the information available. The categories they assign, are,



“sites as fixed effects”, which assumes: “the same sampling techniques were used, the expertise of both teams was equal, the sampling effort was the same in both surveys, normal biotic and abiotic factors regulating population fluctuations were the same during the sampling periods, and detectability (the detection threshold) of the target species has remained the same,” and, “sites as random effects”, which assumes, “historical and current sampling should be sufficient so that a lack of occurrence in a region is meaningful, the size of the region over which the sampling sites are pooled should include enough sites to be statistically rigorous, but should not be so large as to be biologically trivial..., the size of the sampling unit over which sites are pooled should be larger than the scale of the biotic and abiotic forces affecting population fluctuations.” I feel that these two categories can assist in organizing the types of herbaria data I will be basing my capstone proposal on, and therefore these assumptions are valuable and should be integrated into the protocol/methodology of creating species distribution models of specimens from small herbaria.

Steege, HT, Jansen-Jacobs, J, and Datadin, VK. 2000. Can botanical collections assist in national protected area strategy in Guayana? *Biodiversity and Conservation* , Vol 9., 215-240.

This paper used a consortium of international herbaria locality data to analyze the species distributions of five tree taxa in Guyana. This was specifically done with the goal in mind of creating a protected area of Guyana. The two models that were calculated were: non-asymptotic model (species area curve) and an asymptotic model, each of which were calculated to understand each respective taxa's species richnesses. The geospatial models that were used in this study were done with GIS (Arcview 3.1). This study will be used as a case study depicting the mapping/use of herbaria data for conservation purposes.

Suarez, AV, and Tsutsui, ND. 2004. The value of museum collections for research and society. *BioScience*, Vol 54, no. 1, 66-74.

This article discusses the importance of the data held in natural history research collections, and their implications for understanding biodiversity, evolution and ecology. They also provide a significant discussion towards the implications that natural history collections have had towards public health and safety, and pathogens. They further discuss how natural history collections are not being utilized even close to their potential capacity, and give recommendations to increase their use. In all, I will use this paper to help argue the value of natural history collections in my introduction as well as how natural history museums can make their data more accessible and used towards making public health and conservation changes.

Syfert, MM, Smith MJ, and Coomes DA. 2013. Correction: The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. PLOS ONE, Vol. 8, no. 7, 10.

Species distribution models are predictive algorithms that use inputted locality data usually from biological surveys, to show the geographical reach of a population of a species. There are many different species distribution models, each with a set of assumptions that must be met in order for the outputted model to be confidently plausible. Through the use of presence data of tree ferns in New Zealand, this study shows through the use of MaxEnt species distribution model analysis, and ultimately shows how data that has bias can be modelled accurately. This paper will be used as a case study of how biased herbaria data can be mapped for conservation purposes through the species distribution modelling software, MaxEnt.

Ward, Darren F. 2012. More than just records, analysing natural history collection for biodiversity planning. Plos One , Vol. 7, no. 11, 1-8.

This paper is not only a case study of using natural history collections in New Zealand for Hymenoptera conservation planning, but it also provides valuable figures that describe temporally and spatially how natural history collections have quantifiably changed. These charts will be used in my introduction where I explain the value of natural history collections, how there are gaps in data, and why these geospatial analyses of NHC data are so valuable. Some of the figures to be discussed in this section, include (but are not limited to), “Summary of the spatial coverage of NHC locations across New Zealand,” “The average number of records per location from different area codes,” “Sampling effort across New Zealand,” “The number of NHC records at different time periods,” and, “Proportion of records of introduced species from urban areas over time”. Ultimately, what the authors have found from graphically organizing the data from NHCs in New Zealand, is that they can help scientists to suggest conservation implementations in various ecological niches, as well as to create a more streamlined, efficient collecting strategy for biologists.

Willis, F, Moat, J, Paton, A. 2003. Defining a role for herbarium data in Red List assessments: a case study of *Plectranthus* from eastern and southern tropical Africa. Biodiversity & Conservation , Vol. 12, no. 7, 1537-1552.

A genus of mint is used in this study to analyze how worldwide accessible herbarium data can be used to determine IUCN Red List assignments in the mint family in the southern tropical region of Africa.

This paper will be used as a case study to describe how mapping herbaria from available data can assist in conservation efforts, and more specifically to those species to which are categorized as concerns or higher. This paper uses an older version of ArcView 3.2 GIS, but can show how this type of mapping can be done with biased data, but larger consortiums of combined international herbaria data. There are also IUCN terminology and measurements described in this paper that are valuable towards creating an accurate protocol, some of which include, “Extent of Occurrence”, “Area of Occupancy”, “Fragmentation”, “Projected Continuing Decline”, and “Subpopulation”. Each of these categories have corresponding equations and definitions that will support the technical explanations in my literature review.

Yackulic, CB, Chandler, R, Zipkin, EF, Royle, A, Nichols, JD, Campbell Grant, EH, and Veran, S. 2013. Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, Vol. 4, 236-243.

This paper discusses the pros, cons, and considerations researchers need to take in order to use MaxEnt with presence-absence and presence-only data. There are a few tables in the paper that will be valuable towards my discussion of why one must use caution when analyzing biased data like that found in natural history collections. A table of interest, is, “Questions and summary responses based on 78 articles published between 2008 and 2010 and 30 articles published in the first half of 2012”, which shows frequency statistics on the types of data used in published literature, and whether answers to the questions listed that these articles actually answered. Some questions include, “is it likely that the presence only data suffers from sampling selection bias (non-random sampling)?” and, “Were absence data available and discarded?” The other table of interest in this paper, is titled, “Corrections required for both presence-absence and presence-only analyses under various assumptions”, which is a table that says which type of data is preferable based on the “detection probability”, “equal to one”, “less than one and constant”, and, “varies”, and “sampling probability” is either “constant” or “varies”.

## Appendix B: Glossary

**Bootstrapping** -The bootstrap method is a resampling technique used to estimate statistics on a population by sampling a dataset with replacement. It can be used to estimate summary statistics such as the mean or standard deviation.

(<https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/>)

**Chi square test** - a statistical method assessing the goodness of fit between observed values and those expected theoretically

**Ecological Niche Modeling** - Environmental niche modelling, alternatively known as species distribution modelling, niche modelling, predictive habitat distribution modelling, and climate envelope modelling.

**Ecological niche models** utilize associations between environmental variables and known species' occurrence localities to define abiotic conditions within which populations can be maintained (Guisan and Thuiller, 2005).

**GIS** - Geographic Information System, A geographic information system (**GIS**) is a system designed to capture, store, manipulate, analyze, manage, and present all types of geographical data.

<https://researchguides.library.wisc.edu/GIS>

**Goodness of Fit** - the extent to which observed data match the values expected by theory.

**MAXENT** - The Maxent software is based on the maximum-entropy approach for modeling species niches and distributions. From a set of environmental (e.g., climatic) grids and georeferenced occurrence localities (e.g. mediated by GBIF), the model expresses a probability distribution where each grid cell has a predicted suitability of conditions for the species. (<https://www.gbif.org/tool/81279/maxent>)

**NHC** - natural history collection

**Lasso and Ridge Regularization** -

- In statistics and machine learning, lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. (Wikipedia).
- Ridge regularization is the most commonly used statistical method of regularization of ill-posed problems. in machine learning, it is known as weight decay, and with multiple independent discoveries, it is also variously known as the Tikhonov–Miller method, the Phillips–Twomey method, the constrained linear inversion method, and the method of linear regularization. It is related to the Levenberg-Marquardt algorithm for non-linear least-squares problems. (Wikipedia).

**Poisson Regression** - In statistics, Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables. (Wikipedia).

**Presence-Absence Data** - species occurrence data that lists both where the species were and were not found

**Presence only data** - species occurrence data that only lists where the species were found

**Pseudo Absences** - Using background (or available datasets) to create Absence data for an SDM. Ie: if a study has Presence only data, they can create Pseudo absences to have Presence-Absence data.

**SDM**- species distribution model

## Appendix C: Figures and Tables

**Figure 1:**

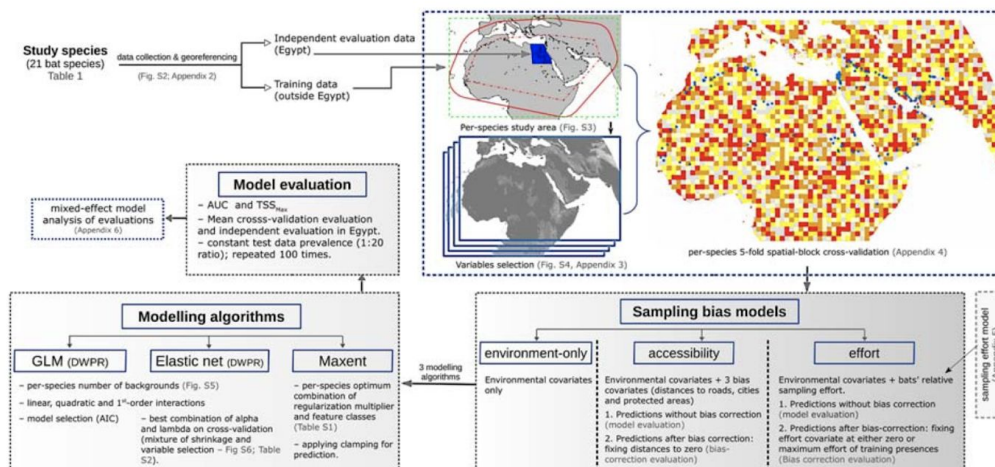
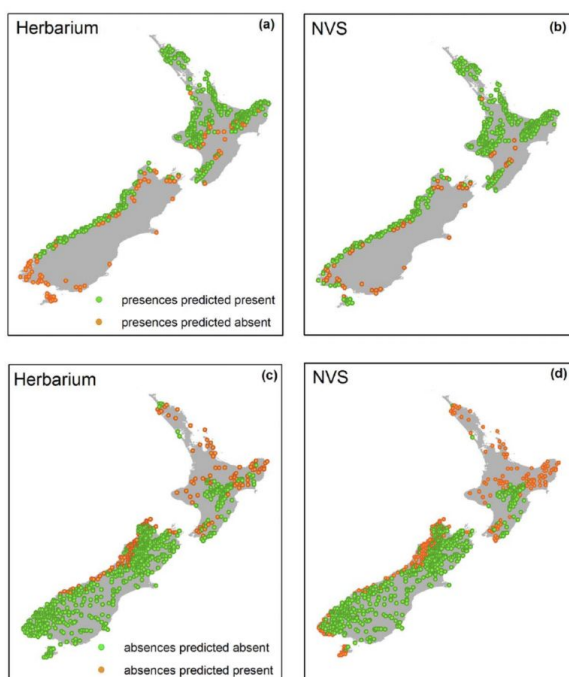


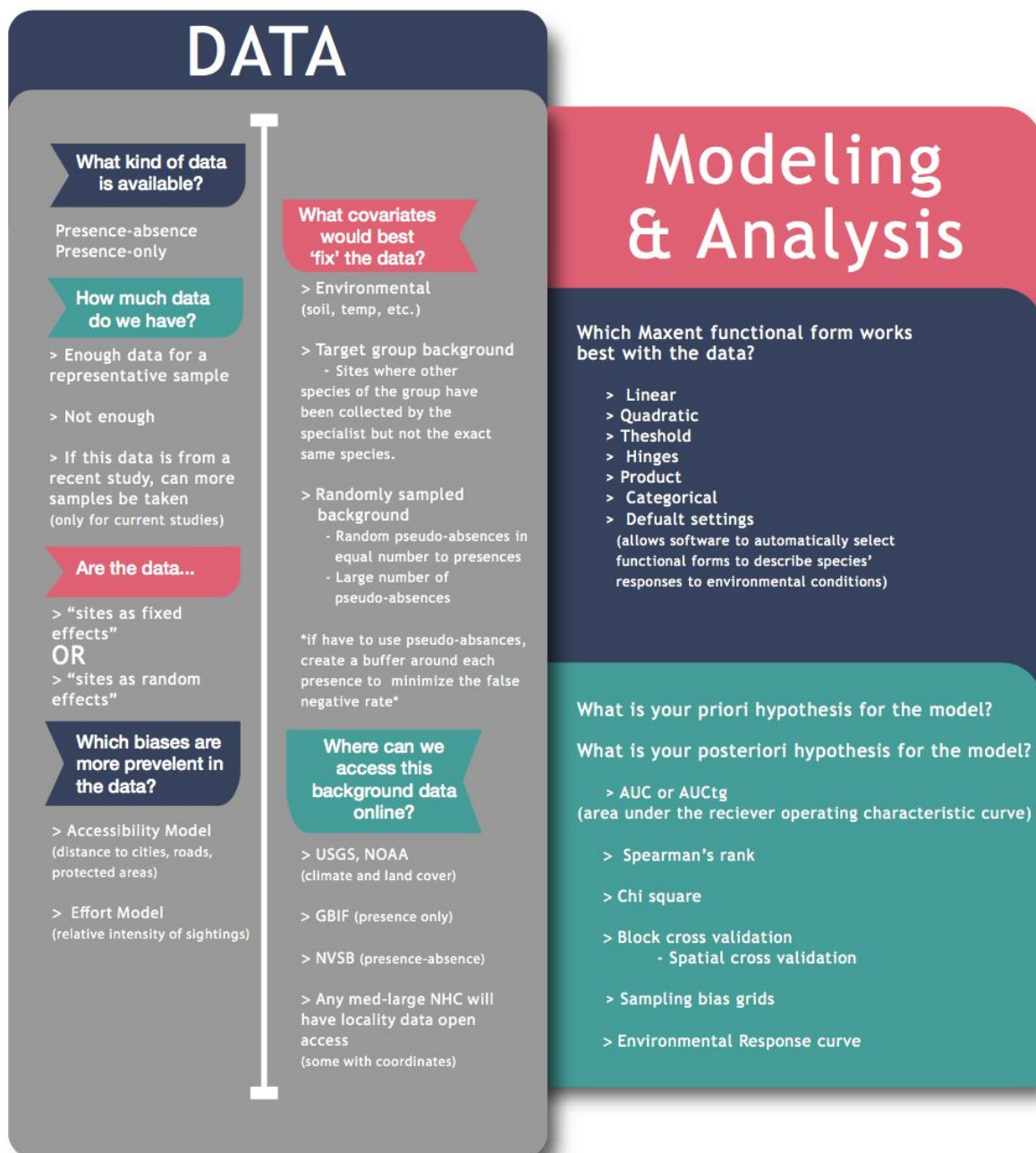
Figure 1. Flowchart of analyses in this study, illustrated with data for *Asellia tridens*. Sampling-bias models and modelling algorithms were combined factorially. Only results for validation with AUC are presented in the manuscript, while TSS-results are given in the Supplementary material.

**Figure 2:**



LUCAS presence/absence locations with predicted presences and absences generated from average LQ model predictions (with geographical sampling bias correction). Correct agreement between predicted presences/absences and LUCAS presences/absences are shown in green and incorrect agreements are shown in orange. LUCAS presence locations are shown with predictions from (a) herbarium dataset and (b) NVS dataset, LUCAS absence locations are shown with predictions from (c) herbarium dataset and (d) NVS dataset.  
doi:10.1371/journal.pone.0055158.g005

Figure 3:



### Tables 1, 2, and 3:

**Table 1** Area under the receiver operating characteristic curve (AUC) and correlation between predictions and 0–1 test data (COR) for the methods considered; values shown are averages over all 226 species.

Model	Random background		Target-group background	
	AUC	COR	AUC	COR
BRT	0.7275	0.2130	0.7544	0.2435
Maxent	0.7276	0.2100	0.7569	0.2446
MARS	0.6964	0.1787	0.7260	0.2145
GAM	0.6993	0.1765	0.7368	0.2196

*Notes:* For random-background models, background data were chosen uniformly at random from the study area. For target-group background, background data are the sites with presence records for any species from the same target group. Models are boosted regression trees (BRT), maximum entropy (Maxent), multivariate adaptive regression splines (MARS), and generalized additive models (GAM).

**Table 2** Spearman rank correlations of improvement in AUC when using target-group background instead of random background.

Model	Correlation with training bias		Correlation with test bias	
	Spearman's $\rho$	$P$	Spearman's $\rho$	$P$
Maxent	0.87	0.002	0.81	0.008
GAM	0.90	<0.001	0.93	<0.001
BRT	0.75	0.017	0.87	0.002
MARS	0.84	0.004	0.95	<0.001

*Notes:* The improvement is correlated against the degree of bias in the training data for each target group ("training bias") and a measure of how well the training data for each target group predict the test sites ("test bias"). In each case, we give Spearman's rank correlation coefficient ( $\rho$ ) and the two-sided  $P$  value for the null hypotheses that  $\rho = 0$ .

**Table 3** Coefficients for an analysis of variance for AUC and COR evaluated on independent presence-absence test data for models of 226 species.

Measure	Algorithm				Background		Effect SE		
	BRT	GAM	MARS	Maxent	Random	Target group	Species	Algorithm	Background
AUC	0.0128	−0.0101	−0.0169	0.0141	−0.0154	0.0154	0.0228	0.0030	0.0021
COR	0.0157	−0.0146	−0.0160	0.0149	−0.0180	0.0180	0.0241	0.0032	0.0023

*Note:* Factors were species (per-species effects not shown), algorithm used to make the model (BRT, GAM, MARS, or Maxent), and background data used for the model (random or target group).

### Tables 4 and 5:

**Table 4** Effects of correcting for geographical sampling bias on the predictive performance of New Zealand tree fern distribution models trained on herbarium and NVS datasets.

	Not correcting for sampling bias				Correcting for sampling bias			
	AUC	COR			AUC	COR		
Herbarium dataset	0.787	±0.012	0.474	±0.020	0.851	±0.004	0.588	±0.008
NVS dataset	0.587	±0.003	0.165	±0.005	0.837	±0.004	0.549	±0.005

MaxEnt was used to fit the models (feature type = LQ) and model performance indicated by mean (±1 standard deviation) AUC and COR values, evaluated by using the independent LUCAS dataset.  
doi:10.1371/journal.pone.0055158.t001

**Table 5** Effects of correcting for geographical sampling bias on the rates of false presences and absences, and on the predicted extent of tree ferns (as a percentage of the total land area of New Zealand).

	Not correcting for sampling bias			Correcting for sampling bias		
	False absences <sup>†</sup> (%)	False presences <sup>‡</sup> (%)	Percentage of NZ predicted to be occupied	False absences (%)	False presences (%)	Percentage of NZ predicted to be occupied
Herbarium dataset	12.4	41.2	45.4	20.3	19.5	30.9
NVS dataset	19.8	64.1	34.5	12.2	30.0	35.9

Models were fitted to two datasets (herbarium and NVS) using MaxEnt with the feature type set as "LQ". Model predictions were based on average predictions from the 40 runs and evaluated by using the LUCAS dataset.

<sup>†</sup>False presences occur when a model predicts a species as present whilst observed data indicate it is absent.

<sup>‡</sup>False absences occur when a model predicts a species as absent whilst observed data indicate it is present.

doi:10.1371/journal.pone.0055158.t002



## Tables 6 and 7:

**Table 6.** Questions and summary responses based on 78 articles published between 2008 and 2010 and 30 articles published in the first half of 2012 [Correction added after online publication 6 December 2012: responses for question 1 have been changed]

Questions	Response	Frequency (% of clear responses)		
		2008–2010	2012	Total
1. Is it likely that the presence-only data suffers from sample selection bias (nonrandom sampling)?	Yes (Y)	57 (92%)	19 (76%)	76 (87%)
	Unclear (–)	16	5	21
	No (N)	5 (8%)	6 (24%)	11 (13%)
2. Does article acknowledge detectability and/or heterogeneity in detectability? (No articles discussed heterogeneity in detectability.)	Mentioned detectability (Y)	12 (15%)	3 (10%)	15 (14%)
3. Were absence data available and discarded (i.e. could they have done a PA analysis)?	Yes (Y)	27 (36%)	9 (35%)	36 (36%)
	No absence data (N)	47 (64%)	17 (65%)	64 (64%)
	Unclear/Used for comparison (–)	4	4	8
4. Was MAXENT's output interpreted as an occurrence probability? (Possible answers: (a) Yes and interpretation of results relied heavily on this assumption, (b) Yes but results not dependent on assumption, (c) No.)	Yes (a or b)	34 (44%)	24 (83%)	58 (54%)
	(a)	20 (26%)	15 (52%)	35 (33%)
	(b)	14 (18%)	9 (31%)	23 (21%)
	No (N)	44 (56%)	5 (17%)	49 (46%)
	Unclear		1	1
5. Were response curves or betas reported? (Possible answers: (a) Response curves, (b) Beta values, (c) Signs of betas, (d) No.)	(a)	11 (14%)	4 (13%)	15 (14%)
	(b)	0	1 (3%)	1 (1%)
	(c)	0	2 (7%)	2 (2%)
	No (N)	67 (86%)	23 (77%)	90 (83%)
6. How many presences were used?	See Fig. 1			
7. How many covariates were tested?	See Fig. 1			

**Table 7** Corrections required for both presence–absence and presence-only analyses under various assumptions

Detection probability			
Equal to one		Less than one and constant	Varies*
<i>Sampling probability</i>			
Constant	Presence–absence analysis preferable. Presence-only allowable, but many methods only yield relative occurrence probability	Relative measures of occurrence possible using both presence–absence and presence-only; Royle <i>et al.</i> (2012) allows estimation of occurrence probability provided that there is a relationship between occurrence and covariates. Presence–absence methods yield occurrence probability when provided with information on detection probability	Presence–absence analysis only; requires estimating relationship between detection probability and environmental covariates [e.g. through multiple visits to some sites and use of programs such as PRESENCE (freely available online)]
Varies*	Presence-only modelling requires that sampling intensity can be standardized objectively through modelling or subsampling of data. Presence–absence analysis provides unbiased estimates of occupancy conditional on sampled areas without covariates, but requires covariates and a reasonably well-specified model for unbiased estimates of occurrence probability across a landscape	Presence-only modelling requires that sampling intensity can be standardized objectively through modelling or subsampling of data. Requires correction for detection probability in addition to covariates for unbiased estimates across a landscape	Presence-absence analysis only; requires estimating relationship between detection probability and environmental covariates. If users want to make inferences about average occupancy across a landscape (as opposed to inferences about relationship to covariates), this estimate must be based on covariate values across the landscape. In other words, average occupancy from a non-representative sample will not be equal to average occupancy across the landscape without additional steps

\*Varies here is shorthand for varies with respect to environmental covariates that are also related to occupancy patterns.

## Bibliography

- [CENR, NSTC} Committee on Environment and Natural Resources, National Science and Technology Council. 1994. *Strategic Planning Document—Environment and Natural Resources*. [www.whitehouse.gov/WH/EOP/OSTP/NSTC/html/enr/enr-plan.html](http://www.whitehouse.gov/WH/EOP/OSTP/NSTC/html/enr/enr-plan.html).
- El-Gabbas, A, and Dormann, CF. 2018. Improved species-occurrence predictions in data poor regions: using large-scale data bias correction with down weighted poisson regression and maxent. *Ecography*, Vol 41., 1161-1172.
- Funk, VA, Fernanda Zermoglio, M, Nasir, N. 1999. Testing the use of specimen collection data and GIS in biodiversity exploration and conservation decision making in Guyana. *Biodiversity & Conservation*, Vol. 8, 727-751.
- Graham, CH, Ferrier, S, Huettman, F, Moritz, C, and Townsend Peterson, A. 2004. New developments in museum based informatics and applications in biodiversity analysis. *TRENDS in Ecology and Evolution*, Vol. 19, no. 9, 497-503.
- Krishtalka, L, and Humphrey, PS. 2000. Can natural history museums capture the future? *BioScience*, Vol. 50, no. 7.
- MacDougall, AS, Loo, JA, Clayden, SR, Goltz, JG, and Hinds, HR. 1998. Defining conservation priorities for plant taxa in southeastern New Brunswick, Canada, using herbarium records. *Biological Conservation*, Vol. 86., 325-338.
- Mateo, RG, Croat, TB, Felicísimo, AM, and Muñoz, J. 2010. Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. *Diversity and Distributions*, Vol. 16, no. 1, 84-94.
- Ponder, WF, Carter, GA, Flemons, P, and Chapman, RR. 2001. Evaluation of museum collection data for use in biodiversity. *Conservation Biology*, Vol. 15, no. 3, 648-657.
- Phillips, SJ, Dudík, M, and Schapire, RE. [Internet] Maxent software for modeling species niches and distributions (Version 3.4.1). Available from url: [http://biodiversityinformatics.amnh.org/open\\_source/maxent/](http://biodiversityinformatics.amnh.org/open_source/maxent/).
- Phillips, SJ, et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, Vol. 19, no. 1, 181-197.
- Shaffer, HB, Fisher, RN, and Davidson, C. 1998. The role of natural history collections in documenting species declines. *TREE*, Vol. 13, no. 1, 27-30.
- Steege, HT, Jansen-Jacobs, J, and Datadin, VK. 2000. Can botanical collections assist in national protected area strategy in Guayana? *Biodiversity and Conservation* , Vol. 9, 215-240.
- Suarez, AV, and Tsutsui, ND. 2004. The value of museum collections for research and society. *BioScience*, Vol. 54, no. 1, 66-74.
- Syfert, MM, Smith MJ, and Coomes DA. 2013. Correction: The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. *PLOS ONE*, Vol. 8, no. 7, 10.
- Ward, Darren F. 2012. More than just records, analysing natural history collection for biodiversity planning. *PLOS ONE* , Vol. 7, no. 11, 1-8.

Yackulic, CB, Chandler, R, Zipkin, EF, Royle, A, Nichols, JD, Campbell Grant, EH, and Veran, S. 2013. Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, Vol. 4, 236-243.