Master's Thesis of Hyun Yong Chung

# Co-occurrence Pattern Learning Species Distribution Model (SDM) Quantifies Annual Reduction of American Coccinellids—Overcoming Data Inconsistencies and Insufficiencies

# 공동출현패턴 학습 종분포모델(SDM)을 이용한 북미 무당벌레의 연 단위 감소율 추정—데이터의 비일관성과 불충분성 극복을 중심으로

August 2023

Graduate School of Education
Seoul National University
Environmental Education Major

Hyun Yong Chung

# Co-occurrence Pattern Learning Species Distribution Model (SDM) Quantifies Annual Reduction of American Coccinellids—Overcoming Data Inconsistencies and Insufficiencies

Young Keun Song

Submitting a master's thesis of Art in Education

August 2023

Graduate School of Education
Seoul National University
Environmental Education Major

Hyun Yong Chung

Confirming the master's thesis written by
Hyun Yong Chung
August 2023

Chair Bongwon Suh (Seal)
Vice Chair Young Keun Song (Seal)
Examiner John E. Losey (Seal)

# Abstract

**Aim:** To predict annual distribution patterns and reduction rates of insufficiently observed species by using co-occurrence pattern learning and devising filling-in strategy to overcome structural and temporal inconsistencies in multi-source noisy data.

**Idea:** Although more than 10% of insects will face extinction in the coming decades, studies on their reduction rates that will form the basis for conservation strategies are still limited. This limitation is first due to the dominance of unstructured records available for invertebrates, secondly, to the inconsistencies among them, and thirdly, to the insufficiencies of them. While compelling to gather data across multiple sources, the small amount of data precludes deep filtering to handle structural and temporal inconsistencies among sources for time-series comparison. This is the first study to estimate annual reductions with machine learning from multi-sourced, presence-only, and small data, by overcoming its inconsistencies and insufficiencies. This study proposes and validates the following two novel strategies. (1) Co-occurrence pattern learning: By grouping low-quality, unreliable individual occurrence records into patterns, I validate that structural and temporal inconsistencies can be overcome without deep filtering. (2) Filling-in strategy: I propose a procedure for estimating population trends by filling in the prediction into the deficiencies of the collected yearly data to be evenly compared.

**Location:** 51 states of the USA and 6 provinces of Canada

**Taxa:** four ladybugs native to North America

**Methods:** In chapter 2, seven performance scores were used to evaluate the predictions on presence versus absence in the following three situations: (1) learning unstructured data to predict structured data or low-efficiecy data to high-efficiency data; (2) learning data before a particular year to predict after that year and vice versa; (3) learning 70% of multi-source data to predict the rest. During both the evaluation and generalization phases, a comparison was made between the performance of the co-occurence pattern using models and the environmental information using models, as well as with the commonly accepted benchmark.

In chapter 3, reduction rates and extinction status were estimated by ML's predicting the occupancy of species annually at all coordinates where species have appeared since 2007. In addition to that, the newly suggested approach's methodological reliability was verified, in comparison with pre-established methods. Furthermore, the reliability of the newly proposed method was validated by examining discrepancies in estimations under the following scenarios: variances in data extraction for pseudo-absence data points, variances in variable selection techniques, and the stochastic incorporation of missing or false information within the presence data.

**Results:** 1) The COP models' performance surpassed acceptable criteria for all validation steps and all species. They also ouperformed over the ENV models. 2) Reduction rates were 36.4% for *H. parenthesis* (2007-2021; VU), 29.7% for *A. bipunctata* (2010-2019; NT), 23.7% for *C. novemnotata* (2009-2018; NT), and 14% for C. trasversoguettata (2007-2018; LC). Additionally, the newly proposed

approach was confirmed to possess strong methodological validity when compared to pre-established methods. In terms of reliability tests, the range of estimations from the new method did not misrepresent IUCN conservation status to a significant extent.

Conclusion: The combination of using co-occurrence patterns as variables and filling-in strategy enabled SDM to predict distribution patterns and reduction rates of insufficiently observed species by overcoming structural and temporal inconsistencies in multi-source data integrating considerable citizen science data. As a result, it revealed that four native ladybug species have been declining in North America. This study suggests that ML developed with the new method can integrate multiple-source data without filtering, allowing for the acquisition of more data, and that COP-based SDMs would be more advantageous for predictions at finer temporal scale population changes than commonly used SDMs developed with environmental variables usually spanning over decades. This can aid in tackling the challenge in global conservation initiatives posed by rare and invertebrate taxa, which frequently face restricted data availability and are often underrepresented in conservation lists.

Keyword : conservation status, annual reduction rate, citizen scienece, presence-only, speceis distribution model, co-occurrence pattern

Student Number : 2018-25558

# Table of Contents

# Chapter 1. Introduction

## 1.1. General background of the study

The estimation of year-to-year decline in low-data-availability taxa poses a significant challenge for entomology in the face of the ongoing mass extinction. It is projected that about 9-16% (500,000-900,000 species) of insects will become extinct within the next 40 years (IPBES 2019, Whalsh *et al.* 2013), with scientists warning about the resulting loss of crucial ecosystem services and the subsequent ecological collapse and social costs (Cadosso *et al.* 2020, Losey and Vaughan 2006). However, the limited availability of methods to understand temporal changes in individual species' distribution and population trends hampers scientists ability to develop a basis for conservation strategies to address their rapid decline (Montgomery *et al.* 2020, Jönsson *et al.* 2021). Consequently, only a mere 0.2% of insect species have been evaluated on the IUCN Red List, in distinction from 100% of birds and amphibians and 93% of mammals.

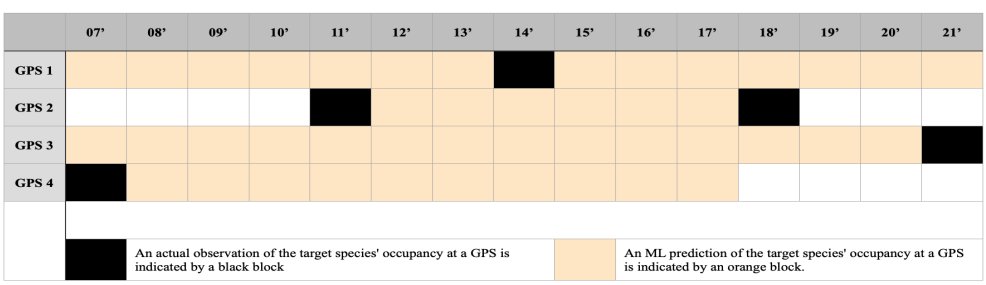
Previously available methods to estimate distribution and population size changes have heavily relied on monitoring results fulfilling high hurdles, such as (1) directly observed ordinal abundance data, (2) consistent survey protocols allowing mathematical assumptions to estimate abundance from occurrence data (ex. checklist, effort unit for each survey, periodi revisit), or (3) the availability of highly dense, independent observations over space enough to filter them to emulate consist survey protocols. However, these cost-intensive data have been rarely available for many taxa and regions, which in turn left these taxa overlooked in global conservation efforts. Even well-studied species like North American ladybugs, with three well-organized national citizen science programs, fail to meet the stringent data requirements for the previously developed estimation methods. To address this issue, there is a pressing need for the active development of methodologies that can estimate the time-series variation of species with the combination of small volume of and presence-only dataset, which is lacking consistent survey protocols or ordinal abundance, concurrently.

In this regards, for any population estimation method to be widely applicable to insects and other understudied taxa, it must

address two key challenges: data insufficiency and inconsistency. Firstly, the lack of data on these taxa hinders the application of conventional filtering techniques (Steen *et al.* 2019) to quantitatively compare regions or periods consistently. The filtering technique aims to improve the statistical analysis of data collected opportunistically by selecting points where data collection efforts are evenly distributed in space and time (Serra-Diaz et al. 2017, Kallimanis et al. 2017, Rutten et al. 2019). However, in cases where data is scarce for certain taxonomic groups, filtering will only retain a small fraction of their overall distribution due to limited survey efforts. It is important to note that if the filtering process reduces the total number of data points too significantly, the resulting model's predictive power may decrease, irrespective of the quality of remained data (Wisz *et al.* 2019, Van Eupen *et al.* 2021). Moreover, it would reduce the spatio-temporal scope of a study (Outhwaite *et al.* 2019, Outhwaite *et al.* 2020).

Secondly, compounding the problem, datasets for less data-rich taxa are often the result of integrating different survey techniques and individual surveys conducted over a wide spatiotemporal range to maximize data quantity, leading to inherent inconsistencies in the data. Although the recent increase of unstructured citizen science data is contributing to increase the volume of data, it produces an additional burden in the inconsistent rates of observation across years and locations (Figure 1-1). Furthermore, most of the multi-source data available does not provide details regarding the objectives and methods employed during the conducted investigations. Consequently, it is challenging to ascertain whether the absence of further observation reports from a location implies the genuine vanishing of the species in that area or if it is simply due to the researcher's discontinuation of reporting after a certain period of time (Table 1-1).

Table 1-1. A conceptual summary of each area's historical records for a target species.




In conclusion, for a population estimation method to be

applicable to a broader range of insect species, it must address the paradox of utilizing low-level data—presence-only data with unknown observation procedure—from various sources to compensate for data deficiencies, while overcoming temporal and structural inconsistencies among them without deep filtering (Figure 2-3 and 2-4).

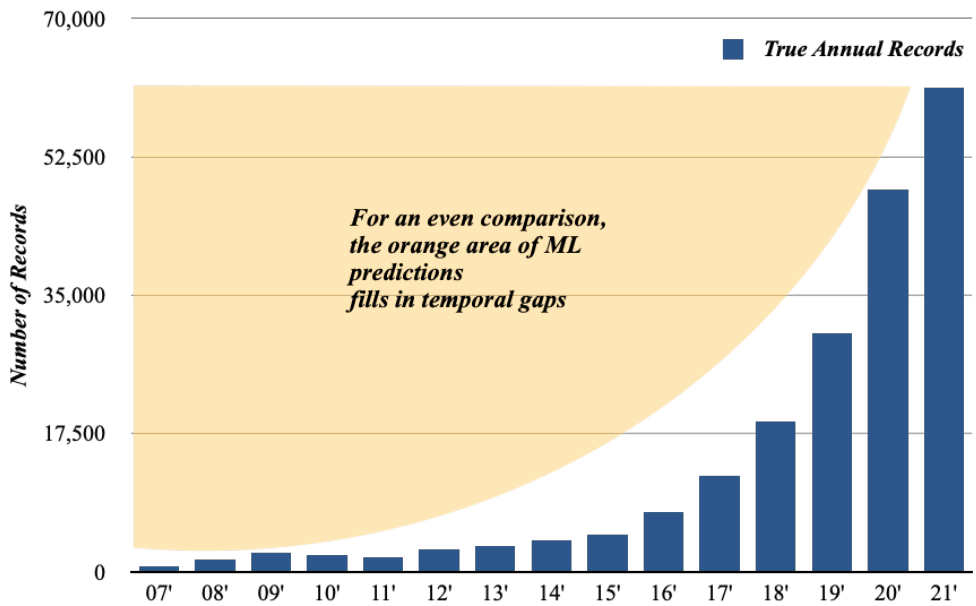



**Figure 1-1. This conceptual graph displays how Machine Learning trained with known historical observations can dilute temporal bias by producing predictions.**

This study proposes a filling-in approach that allows for quantitative comparisons between time points by standardizing them with "filling" predictions in periods and locations with insufficient data, as the alternative to filtering (Figure 1 and Table 1). In this study, ensemble-based species distribution models (SDM), a machine learning algorithms based on decision trees, are preferred for generating predictions. SDM models learn the relationship between predictors and species occupancy information, enabling the prediction of species occurrence at undocumented locations (Olden *et al.*, 2008). Machine learning-based SDM models have demonstrated their effectiveness in inferring species habitats using limited data volumes, integrated multisource data, and citizen science presence-only data (Radomski *et al.* 2022).

However, there is a lack of research on estimating the yearly

trend of taxa with small data volumes and presence-only data points. While ML-based SDM models have been extensively studied for predicting future or past trends as atlases spanning decades (Tingley and Beissinger 2009), studies focusing on finer time units have been abandoned. This is because SDM models cannot account for detection biases, making them vulnerable to datasets with changing detection rates over time. Also, models developed for a particular context may result in unrealistic predictions when utilized beyond their original spatio-temporal scope. For example, Tyler *et al.* (2019) developed independent SDM models for each annual dataset to estimate distribution per each year, but lacking a unified criterion for comparable predictions across time points prevented establishing a predicative trend. By developing a single model that can predict individual time points relatively independently of detection biases with small number of presence only data points, it can provide robust estimates of species occurrence over time, thereby expanding the taxonomic range in conservation efforts.

To increase the temporal generalizability of SDM prediction, this study also proposes employing a co-occurrence pattern (COP) as a variable. COP, which incorporates co-occurring species as variables, is gaining attention and has been validated in recent studies, including the Joint Species Distribution Model (JSDM). In this research, I adopt COP as a variable and highlight three advantages it offers over commonly used environmental variables when predicting at fine temporal scales. Firstly, biological variables can capture more immediate and dynamic changes compared to environmental variables. Secondly, unlike environmental variables, co-occurrence-related variables can encompass dynamic relationships among organisms, such as competition. Thirdly, the co-occurrence pattern reflects organisms' responses to environmental changes; thus, the effect of these changes is internalized in the model.

In sum, this study introduces a novel approach for estimating the annual population of species with limited and unstructured data by utilizing (1) the co-occurrence pattern learning model and (2) the filling-in approach by demonstrating the independence of this model from temporal-structural biases.

## 1.2. Purpose of the study

Chapter two focuses on ensuring that SDM models trained with COP can generate applicable predictions to fill in data gaps over time, given the importance of verifying the model's ability to overcome temporal and structural biases over time. Firstly, it is necessary to determine if the model can be generalized rather than only valid on patterns observed within specific time frame. Secondly, the chapter cross-examines the model for structural inconsistencies to mitigate data mixing and to regulate biases driven by prevailing sources across time. In sum, this chapter bolsters the model's generalizability. Therefore, this study verifies the following three hypotheses. First, ML classifiers can distinguish between the presence versus absence of target species, based on co-occurring species from a multi-source data pool. Second, training patterns of co-occurring species can allow ML classifiers' prediction to be generalized across survey structures and protocols, defined as structural generalization. Third, training patterns of co-occurring species can allow ML classifiers' prediction to be generalized across time periods, defined as temporal generalization.

Chapter three focuses on utilizing a methodology newly proposed by this thesis, the "filling-in" approach, to estimate population size and trends over years. This study estimates reduction rates of four native ladybugs by "filling-in" the prediction generated by our ML classifiers into the deficiencies of the collected yearly data to be evenly compared. Subsequently, this study evaluates the validity and reliability of this approach, which is critical due to the scientific and social costs of incorrect estimations of extinction risk levels. One of the key factors contributing to inaccurate population trend estimations is the temporal variation in data availability itself. Therefore, to assess validity, this study compares the filling-in approach with the Relative abundance, the Historical records accumulation), and the trends of raw data itself to determine the extent to which each method is independent of the temporal variability in the quantity of available data. Additionally, to assess reliability, the study evaluates the degree of variation in estimations among COP models developed using (a) different pseudo-absence datasets, (b) different variables, and (c) presence datasets that is partially lost and partially replaced with false data points.

## 1.3. Study history

Species Distribution Modeling (SDM) is a powerful tool in ecological research and conservation management. Initially, in the late 20th century, early attempts to model species distributions had begun in response to the broadening interest in ecology studying the relationship between a species' occurrence and the factors that explained it. Scientists relied on statistical techniques like logistic regression and discriminant analysis. These initial methods were data hungry and required high cost monitoring efforts (ex. ordinal abundance for all seasons and locations, checklist to presume true absence) of experts as with common statistical, ecological models.

Over time, advancements in computational power and data availability opened up new avenues for SDM. In the early 2000s, machine learning algorithms, such as MaxEnt (Maximum Entropy), gained prominence. It is based on the principle of maximum entropy, which fundamentally seeks to find the most unbiased probability distribution given a set of constraints. MaxEnt introduced an innovative approach by incorporating information theory principles to estimate species distributions using only presence data. This breakthrough overcame the unavailability of absence data. This model has also allowed scientists to predict far-future or past distributions with significant changes in environmental variables, such as climate change, in terms of the mean value of 47 years between each predicted atlas from 23 papers.

Moreover, the advent of ensemble modeling approaches, including Random Forest and XGboost, has added new powerful options to SDM researchers. Ensemble models combine predictions from multiple algorithms to improve model performance and account for uncertainties. Techniques like spatial undersampling and integrating different sources of dataset have been successfully applied to SDM, enabling robust performance. Also, these models have demonstrated their effectiveness in inferring species habitats using limited data volumes, integrated multisource data, and citizen science presence-only data.

Meanwhile, there is a relatively recent development in the SDM field, called Joint Species Distribution Modeling (JSDM), which originally began to emerge and gain attention as a statistical approach to studying species co-occurrence patterns in the early 2000s. In its early stages, the concept of species distribution model's primary

approach was to focus on understanding individual species' distributions based on environmental variables like SDM. Over time, scholars have recognized the need to consider species interactions and community-level processes in addition to environmental factors. This recognition led to the development of methods that could model multiple species simultaneously.

The application of Bayesian hierarchical modeling techniques and advances in computational power facilitated the development and implementation of JSDM approaches. By jointly modeling multiple species' distributions, JSDM allows for the detection of non-random patterns in species co-occurrence, such as positive or negative associations. Efforts are being made to incorporate additional factors such as species traits, spatial autocorrelation, and temporal dynamics into JSDM frameworks.

Concurrently, the improvements in the SDM field have been amplified by the increased availability of the following two types of data. Firstly, the integration of remote sensing data and Geographic Information Systems (GIS) has further strengthened SDM. High-resolution environmental data, including climate variables and land cover maps, have enhanced the accuracy and precision of predicting species' responses to environmental changes and identifying conservation areas of interest.

The second driving factor is the on-going increase in opportunistic observations from citizen science. Citizen science engages the public in scientific research, allowing non-professional individuals to participate in data collection and analysis. In the context of SDM, citizen science initiatives often involve volunteers observing and recording species occurrences more than any other survey methods. These contributions significantly increase the spatial and temporal coverage of data.

However, citizen science has the following limitations that prevent it from being used to its full potential by current SDM. Most notably, most citizen science data is collected through unknown processes, and the observer records their accidental encounters without following a specific investigation protocol. Therefore, these features make it difficult to make mathematical assumptions about the results of citizen science data, hindering quantitative analysis of population dynamics such as temporal fluctuations.

Specifically, there are limitations in controlling data quality. Participants in citizen science projects may have varying levels of expertise and training, leading to inaccuracies or inconsistencies in the collected data. Even with quality control measures and protocols, data accuracy and reliability remains in question. These characteristics deter scientists from integrating citizen science data with existing expert-derived data sources or even internally integrating among different citizen science projects.

Another limitation is spatial and temporal coverage, which can also be restricted. Participants in citizen science projects are dispersed unevenly across different locations, making it difficult to cover all areas of interest. Additionally, data collection may be inconsistent over time, resulting in gaps in monitoring and hindering the data's ability to demonstrate long-term trends or changes accurately.

Yet another limitation is the presence of sampling bias. Citizen science projects rely on volunteers who may choose to participate based on personal interest or accessibility to certain areas. This can result in the overrepresentation of certain regions or habitats, while others may be underrepresented. Biases in data collection can affect the generalizability and representativeness of the findings.

Therefore, it is important to devise a method to overcome these limitations and maximize the quantitative utilization of citizen science data, while ensuring the validity and reliability of the scientific conclusions drawn from such data. However, there is no research to adopt presence-only citizen science data to generate population trends with SDM or JSDM.

Especially, there is a lack of research on estimating the yearly trends of taxa with small data volumes and presence-only data points with SDM. While ML-based SDM models have been extensively studied for predicting future or past trends on the basis of 20 to 50-year atlases, studies focusing on finer time units have been abandoned. This is because SDM models cannot account for detection biases, making them vulnerable to datasets with changing detection rates over time. Therefore, models developed for a particular context may result in unrealistic predictions when utilized beyond their original spatio-temporal scope.

For example, Tyler *et al.* (2019) developed independent SDM

models for each annual dataset to estimate distribution per each year, but the lack of a unified criterion for comparable predictions across time points prevented establishing a predicative trend. By developing a single model that can predict individual time points relatively independently of detection biases with small number of presence only data points, it can provide robust estimates of species occurrence over time, thereby expanding the taxonomic range in conservation efforts.

Meanwhile, it is important to note occurrence Modeling (OM) as a dominant mathematical modeling-based approach in species distribution modeling (SDM) for estimating temporal variations in species populations. This method effectively incorporates investigative efforts and species detection possibilities, making it valuable for data collection by citizen scientists (Isaac *et al.,* 2014). OM assumes that changes in observed detection probabilities reflect changes in species abundance, providing insights into species dynamics. However, typical occupancy models used in OM require substantial data, including information on investigation efforts, checklists for inferring non-detection, and repeated revisits (Perkins-Taylor & Jennifer, 2020). Although attempts have been made to replace repetitive surveys with presence-only data, there are four disadvantages compared to machine learning-based models:

(1) Data density poses challenges for applying OM in many species and regions. Previous OM studies on European insects, even with "relatively small" data volumes, have 30 to 100 times higher density (number of observations per area per year) compared to data-rich groups like ladybugs in North America. In this region, even with compromises on temporal and spatial resolution, achieving the necessary data density per grid cell for bees and dragonflies required a rough time (two 20-years Atlases) resolution that is not suitable for the IUCN Red List or spatial resolution (100km² per a grid cell) exceeding the appropriate landscape scale resolution (1-25 km²) needed for conservation planning.

(2) Opportunistic presence-only observational data can violate the statistical assumptions of the OM model regarding observation efforts and revisit periods. Previous studies have used random observation probability models and average detection rates to assume observation efforts for presence-only data collected with varying efforts. However, there are concerns that these 'flexibility' neglect OM's susceptibility to temporal variations in unmodeled detection probabilities, introducing biases in trends over time (Merow and

Silander, 2014). In contrast, machine learning models do not rely on mathematical assumptions about data collection processes, reducing the risk of presence-only data violating fundamental model assumptions.

(3) Numerous studies demonstrate that machine learning (ML) models perform comparably to OM models for presence-only data. ML models exhibit superior performance, particularly for rare species and even with small sample sizes, when compared to occupancy models (Rota *et al.,* 2011; Gomley *et al.,* 2011; Lahoz-Monfort *et al.,* 2014; Perkins-Taylor & Prey, 2020).

(4) Assessing the reliability of OM predictions for estimating species trends is challenging due to the lack of 'ground truth' on the subject matter. In contrast, ML models provide their own strong performance assessment results during development.

Considering these factors, it is worthwhile to develop machine learning-based SDMs as a method to detect changes in species distribution over time in a more effective manner.

Lastly, here I documented information about the algorithm adopted in this study, XGBoost (Extreme Gradient Boosting), in detail. This is a well-adopted machine learning algorithm known for its efficiency and accuracy in solving supervised learning problems. It is an implementation of the gradient boosting framework that utilizes a combination of decision trees to make predictions. In practical terms, the performance of XGBoost in SDM has been evaluated across diverse ecological systems and species, including both plant and animal distributions. Its applications range from predicting species' potential distributions under current conditions to projecting future distributions under climate change scenarios. The computation process of XGBoost involves the following steps:

Data Preparation: XGBoost requires the input data to be in a specific format. The features and the corresponding target values need to be transformed into a structured format that XGBoost can process.

Initialization: XGBoost initializes with an initial prediction value for all instances in the dataset. This initial prediction is typically the mean or median of the target values.

Building Decision Trees: XGBoost builds decision trees sequentially in an iterative manner. Each decision tree is built to correct the mistakes made by the previous trees. It uses a technique called gradient boosting, where each subsequent tree focuses on reducing the errors of the previous trees.

Calculating Loss: XGBoost uses a loss function to quantify the errors made by the model. Commonly used loss functions include regression loss functions like mean squared error (MSE) and classification loss functions like logistic loss or softmax loss.

Gradient Calculation: XGBoost calculates the gradient of the loss function with respect to the predictions made by the previous trees. This gradient provides information on how to update the predictions to reduce loss.

Iterated Tree Building: XGBoost constructs decision trees by recursively partitioning the data based on selected features and their respective thresholds. It uses an algorithm called the "greedy" algorithm, where it selectively, or otherwise "greedily," chooses the best split points to minimize the loss function.

Regularization: XGBoost applies regularization techniques to prevent overfitting. It adds penalties to the loss function for having complex models or large coefficients, encouraging the model to be simpler and more generalized.

Update Predictions: After each decision tree is built, XGBoost updates the predictions by adding the predictions of the new tree, multiplied by a learning rate. The learning rate controls the contribution of each tree to the final prediction.

Repeat Steps 4-8: The process of calculating loss, gradients, building trees, and updating predictions is repeated for a specified number of iterations or until a convergence criterion is met.

Final Prediction: Once all the iterations are completed, the final prediction is obtained by summing up the predictions from all the individual trees.

By iteratively improving the predictions of the weak models (decision trees), XGBoost creates a strong ensemble model that can capture complex patterns and make accurate predictions.

As a result of these contributions toward ensemble learning approaches, handling missing data, nonlinear relationships, and so forth, XGBoost has made significant contributions in the realm of SDM and has been widely recognized for its effectiveness in this field.

As researchers began exploring the application of XGBoost in SDM, several studies were conducted to evaluate its performance and compare it with other algorithms commonly used in this domain, such as Random Forest and Support Vector Machines. These studies consistently demonstrated the superior predictive capabilities of XGBoost, highlighting its ability to capture complex relationships between environmental variables and species distributions.

# Chapter 2. Co-Occurrence Patterns Overcome Structural and Temporal Inconsistencies in a Multi-Source Dataset, Outperforming Environmental Variables.

## 2.1. Materials and methods

### 2.1.1. Summary of materials and methods

First, by co-occurrence pattern learning, I examined whether structural and temporal inconsistencies could be overcome without deep filtering. In this validation process, six performance scores were used to evaluate the ML predictions of presence versus absence in the following three situations. First, learning unstructured data was used to predict structured data. Second, learning data before a particular year was used to predict after that year and vice versa. Third, learning 70% of all valid data was used to predict the rest. Next in the prediction process, as a filling-in strategy, reduction rates and extinction status were estimated. ML was used to predict the occupancy of species annually at all coordinates where species have appeared since 2007. This was to fill the prediction into the deficiencies of the collected yearly data to be evenly compared.

### 2.1.2. Target species

In a variety of coccinellid complexes in North America, four native ladybug species once emerged as dominant species. These include the *Coccinella novemnotata*, *Coccinella transversoguttata*, *Adalia bipunctata*, and *Hippodamia parenthesis*. Covering a wide range of prey species and habitat types, they comprised a considerable portion of the collection (Losey 2007, 2012). Research after the mid-1980s, however, found them to be rare, with a drop estimated at 0.009-0.05 based on relative abundance in the collection (Jason *et al.* 2006). It has

been posited that the cause of this rapid disappearance was the introduction and establishment of two adventive species, *Coccinella septempunctata* and *Harmonia axyridis* (Wheeler and Hoebeke 1995, Harmon 2007). After outbreaks, it is common to find the coccinellid complex utterly dominated by these adventive species in traditional landscapes. This results in the loss of diversity and abundance of native species over the continent (Peterson and Losey 2022). The vastness of the decline range has, however, often prevented such reports from estimating the overall reduction rate and the quantitative risk of extinction (Wheeler and Hoebeke 1995, Hesler *et al*. 2004, Jason *et al*. 2006). These reports have temporal and spatial constraints due to the low density of the target species, combined with vast distribution. These factors necessitate integrating reports from different periods, regions, and methodologies while overcoming biases inherent in multi-source data.

### 2.1.3. Occurrence data

Ladybug records were collected from multiple online databases, including two types of unstructured citizen science (CS) platforms, a university collection website, and three metadata platforms (Table 2-1). One type of CS platform was for users to verify species identifications (I-Naturalist and bugguide.net), and the other for experts to verify (The Lost Ladybug Project). I confirmed the target species' identification in I-Naturalist and bugguide.net. A minimal degree of preprocessing was applied to the raw collection to confirm how co-occurrence patterns itself overcome target biases. First, the scope of multi-source data collection was limited to two areas between January 2007-December 2021. These included the territories of the US (except Alaska and Hawaii) and the border area of Canada (Manitoba, Ontario, Saskatchewan, British Columbia, Alberta, and Quebec). Second, I sorted a subset of data points identified at the species level and observed in adult forms—with relatively less error in identification. Third, GPS pinpoint accuracy was limited to 1 km, if available (89% of total). Finally, the data points with simultaneously

matching species-year-GPS were eliminated. After preprocessing, the examined dataset accounted for 188,644 data points for 353 species (including sp.) of ladybugs' occurrence data that were stored in 85 sources. To reveal the temporal and structural inconsistencies inherent in our collection, I used descriptive statistics. The collection contained 324 data points to *C. novemnotata*, 510 to C. transvuersoguettata, 732 to *H. parenthesis*, and 1,426 to *A. bipunctata*. The target species' data points were labeled as the 'presence'.

### 2.1.4. Pseudo-absence

When a species absence record is nonexistent, the presence record of other species is borrowed as a pseudo-absence point. Generally, pseudo-absence GPS is randomly sampled from a GPS mixture of all other species (Robinson *et al*. 2018). For two reasons, I specified the two alien species' GPSs (*C. septempunctata* and *H. axyridis*) as the pool of pseudo-absence points in this study. (1) The target species and two exotic species compete exclusively within communities. This is the main trigger for their reduction across the continent that this study's ML aimed to predict. The adventive species' occupation in conjunction with the target species absence within an 18km radius was regarded in two ways. It was interpreted as (i) general evidence of absence and (ii) an altered co-occurrence pattern following the local extinction of the target species. (2) Accounting for 61% of the total observations, the adventive species would have dominated the pool, although if I had followed the general rule. This ensures methodological coherence.

By sampling at regular intervals, I pooled 10,000 pseudo-absence points from each state and province. This was done to match the rate of the presence data pool for four target species. To minimize human intervention, a subset of absence points from the pseudo-absence pool was randomly sampled multiple times for practice. The model showed higher accuracy in sampling the pseudo-absence pool without considering the state ratio. Even so, the variable analysis indicated a

stronger dependence on geographically characteristic variables such as Coleomegilla maculate, which was concentrated in the east. Biological interactions such as competition were considered more suitable variables to predict temporal changes rather than static distributions. The 'matched state ratio' could address this issue.

### 2.1.5. Variables

(1) Variables for COP models: Active direct and indirect competition structures underlie ladybug assemblages, as adventive species dominance creates native species niche differentiation (Peterson and Losey 2022) and avoidance (Mukwevho *et al.* 2017, Hesler and Kieckhefer 2008, Elliott 1996). In this study, as variables representing co-occurrence patterns, I used the number of records of each ladybug species within an 18 km radius of the presence and absence points. This exact distance stems from a typical assumption of ladybug dispersal ability (Jeffries *et al.* 2013, COWISE 2017, 2018, 2020). For example, a group of predators displays high (ex. *H. axyridis*, 442km per year; McCorquodale 1998) and active (ex. moving between habitats when foraging; Woltz and Landis 2013) mobility in long-distance flight. A count of the target species was excluded from its own variables to avoid self-guidance. Counts in each variable (= a ladybug species) are min-max-scaled within each year. This is done to treat temporal discrepancy in the annual volumes of the total data points. There were no other covariates (ex. survey efforts) besides the numbers of each co-occurring ladybug species. To avoid multicollinearity in our testing (Kissling 2012), I did not use environmental variables. For each ML development, variables were filtered through the minimum number of co-occurrences (>30; 85 species remain) and 'sp.' data were excluded. Multiple regressions were implemented in a forward way ($p<0.05$) to sort predictable factors. Other parameters were variance inflation factor (<10) and absolute correlation coefficients (<50%) to reduce multicollinearity among variables. SHAP (SHapley Additive exPlanations) values (1st to 15th) were used to sort and rank a compact list of variables. SHAP is

a value to rank input variables in a model's computing. According to classic Shapley values from game theory and typical implementations, a model is defined as the linear addition of input variables. It links the optimal credit allocation with local explanations.

(2) Variables for ENV models: Environment values have been exclusively and dominantly used as SDM variables (Table 2-2). This is because they represent key ecological factors that influence species' habitat suitability. Climatic variables, such as temperature and precipitation, affect species' physiological tolerances and determine their ability to survive and reproduce. Land cover variables, on the other hand, reflect the availability of suitable resources and habitat structure for the species. The variables representing climate included annual mean air temperature, average temperature of the warmest month, and annual average rainfall (downloaded from: https://chelsa-climate.org/downloads/). The variables representing land cover included Evergreen/Deciduous Needleleaf Trees, Evergreen Broadleaf Trees, Deciduous Broadleaf Trees, Mixed/Other Trees, Shrubs, Herbaceous Vegetation, Cultivated and Managed Vegetation, Regularly Flooded Vegetation, Urban/Built-up, Snow/Ice, Barren and Open Water (downloaded from: https://www.earthenv.org/landcover). As vegetation indices, Evenness, Shannon, Simpson, and Coefficient of variation were included (downloaded from: https://www.earthenv.org/texture). To improve consistency among measurements, environmental data were downloaded from different platforms, and the platform with the closest distance to the measured values of ladybug data points was selected. The values of the closest measured coordinates of environmental variables were assigned to each presence/absence data point. The measurement year was not considered. Each variable was min-max scaled. For each machine learning development, variables were filtered based on the minimum ratio of ladybug data points that had a distance of less than 18km from the measured values. Variance inflation factor (<10) was considered to reduce multicollinearity among variables. SHAP (SHapley Additive exPlanations) values (1st to 15th) were used to sort and rank a

compact list of variables.

Table 2-1. collected data on coccinellids from seven sources. A minimal amount of data refinement was applied to the dataset in order to confirm how co-occurrence patterns alone overcome structural/temporal biases.

| Source | Volume | Refined with.. |
|---|---|---|
| The Lost Ladybug Project | 32,905 | • '07-'21 years<br>• GPS precision < 1km<br>• species level<br>• adult stage<br>• drop duplicated at year-gps-species |
| I-Naturalist | 197,990 | |
| bugguide.net | 27,018 | |
| GBIF | 143,000 | |
| BISON | 109,834 | |
| IdigBio | 99,723 | |
| NCUC | 5,425 | |

Table 2-2. collected data on environmental variables from several sources. All the data provided with less than 1km resolution.

| Class | Type | Source | Address |
|---|---|---|---|
| **Topography** | Elevation | USGS | https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-global-multi-resolution-terrain-elevation?qt-science_center_objects=0#qt-science_center_objects |
| **Climate** | Annual Temperature | Chelsa-climate | https://chelsa-climate.org/downloads/ |
| | Warmest month's temperature | Chelsa-climate | https://chelsa-climate.org/downloads/ |
| | Annual Precipitation | NASA Earth Data | https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=2130 |
| **Land Cover** | 9 land cover categories | Yale earthenv | https://www.earthenv.org/landcover |
| **Vegetation index** | Simpson | Yale earthenv | https://www.earthenv.org/texture |
| | Shanon | Yale earthenv | https://www.earthenv.org/texture |
| | Enveness | Yale earthenv | https://www.earthenv.org/texture |
| | Coefficient of variation | Yale earthenv | https://www.earthenv.org/texture |

### 2.1.6. Development and characterization of models

In practice, XGBoost Classifier ("Scikit-Learn" package) was used in the Python environment. XGBoost classifiers utilize an optimized gradient boosting algorithm to compute predictions and make decisions. By iteratively training a series of weak decision tree models, XGBoost optimizes a specific objective function, incorporating gradient descent and regularization techniques to enhance predictive accuracy. This ensemble classifier performs the advanced gradient boosting tree algorithm at high speeds. This tool is also known to be capable of dealing with regularization and overfitting-underfitting issues (Chen and Guestrin 2016). To ensure the generality of our approach, the default parameter settings of the package were applied, except setting objective="binary: logistic" and n_estimators=1000. According to the rule of thumb, the train-test ratio was set at 7:3 in all tests (except structural generalization). The ratio of presence and absence was set at 5:5 by undersampling the volume of pseudo-absence, for 50 independent practices. Within these combinations, training and testing in a dataset were randomly split for 50 independent practices, creating 2,500 unique practices. Eventually there were [(50 different splits of train/test) x (50 different combinations of absence data points)]. Each practice's performance scores were derived from the discrepancy between predictions and known labels of the test datasets. Their overall mean was evaluated through the following six scores. These included Accuracy (ratio of true response), Kappa (Cohen 1960; considering default chance of true response), Recall (true positive), and Precision (positive predictive) to evaluate each model's ability and bias in predicting binary presence versus absence. Brier (Brier 1950; mean squared discrepancy) and AUC (Fielding and Bell 1997; ranking of the prediction classes) are also used to evaluate the quality of predicted probabilities. These treatments in development and characterization were commonly applied to the following procedures.

### 2.1.7. Generalization

Generalization tests the capacity of the developed model's application to a new pool of independent data, denying autocorrelation within data (Justice *et al*. 1999). Successful generalization is important evidence of the model's when the ground truth and prediction results cannot be directly compared (Justice *et al*. 1999). In this regard, this step requires data that differs from training data in terms of temporal, geographical, or source factors (Vaughan 2005). Our tests focused on whether our approach can be generalized between structurally or timely distinct data pools (Figure 2-1).

1) Structural generalization: To test our approach's generality across survey structures, I trained 'unstructured' data points for ML to generate a prediction on 'structured' data points that the models never encountered. Additionally, the differences in co-occurrence patterns between them were quantified by ANOSIM with Manhattan distance. Four target species' presence data points were separated into unstructured data points recorded across The Lost Ladybug Project (LLP), I-Naturalist, and bugguide.net and structured data points stored in 28 institutions. In total, there were 280 unstructured versus 44 structured records for *C. novemnotata*, 485 versus 25 for *C. transversoguttata*, 626 versus 116 for *H. parenthesis*, and 1,338 versus 88 for *A. bipunctata*. The structured pseudo-absence data points available in the structural generalization ranged from 416 to 510. Moreover, I evaluated the predictive performance on the LLP data (mean efficiency = about 6.8) by the model trained with the rest of the data (=about 1.2) in terms of efficiency cross-validation.

(2) Temporal generalization: I examined two directions of testing for temporal generalization across periods. (1) forward way. From 2007 to a year when 70% of the total presence data volume had been accumulated, 70% was used for training and 30% for testing. (2) backward way. From 2021 to a preceding year when 70% of the total presence data volume had been accumulated, 70% was used for

training and 30% for testing.

### 2.1.8. Evaluation

To evaluate the general performance under our approach, I used 70% of all presence data points and matched the number of absence data in training and the remaining 30% in testing.

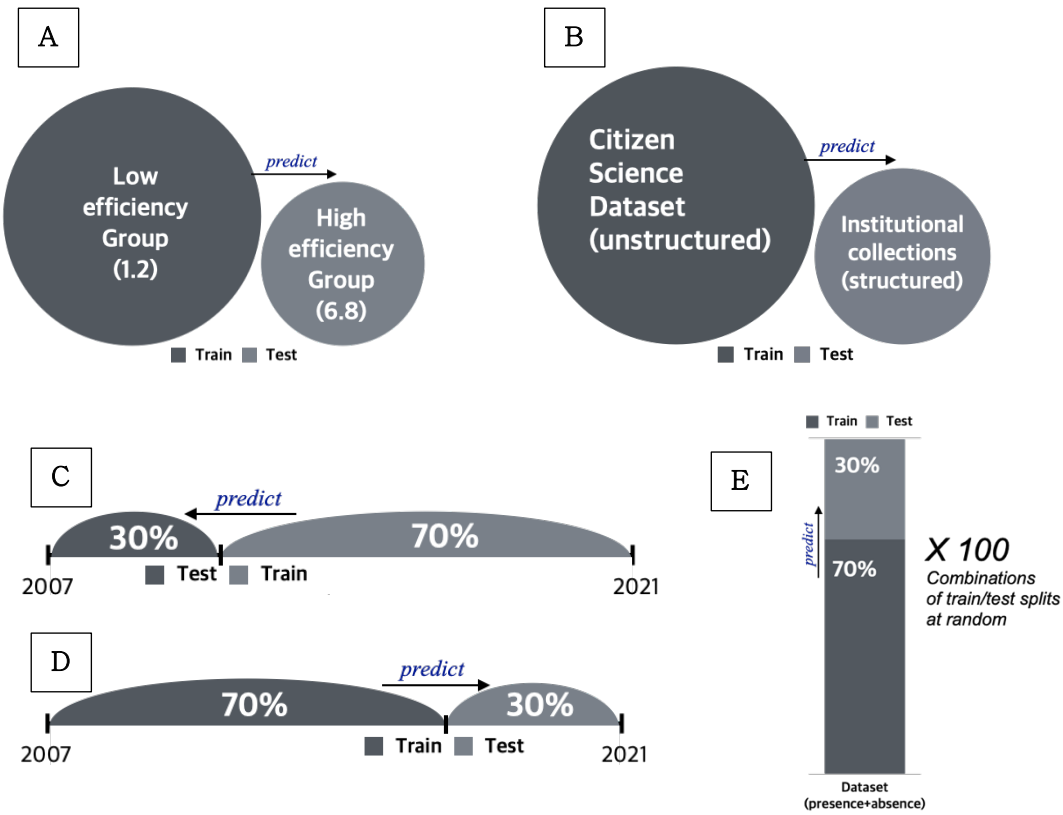



Figure 2-1. The schematic diagram illustrates the procedures employed to evaluate the machine's performance. In [A], models predict structured data using unstructured data. In [B], models predicts high-efficiency data using low-efficiency data. In [C], models train from 2007 to a year when 70% of the total presence data volume had been accumulated and predict the following years. In [D], models train from 2021 to a preceding year when 70% of the total presence data volume had been accumulated and predict the preceding years. In [E], models predict on the remaining 30% of data after training it on 70% of the available data.

## 2.2. Results

### 2.2.1. Biases in multi-source data

The dataset I compiled from multiple sources contained structural and temporal biases. The inconsistency of efficiency in finding target species for sources (quantified by each source's ratio of four target species observations divided by the total observations of the source) revealed the structural bias, which was defined as being derived from discrepant efforts and methods in producing observations (Figure 2-2). Structured data, which accounted for 3.5% of the total, recorded target species at about three (or 2.79) times the density of unstructured data, which took 96.5% of the total. In addition to the differences from institutional collections, citizen science platforms exhibited these differences among themselves as well. LLP, which accounted for 5% of the total, recorded target species at seven times the density of I-Naturalist, which took 89% of the total.

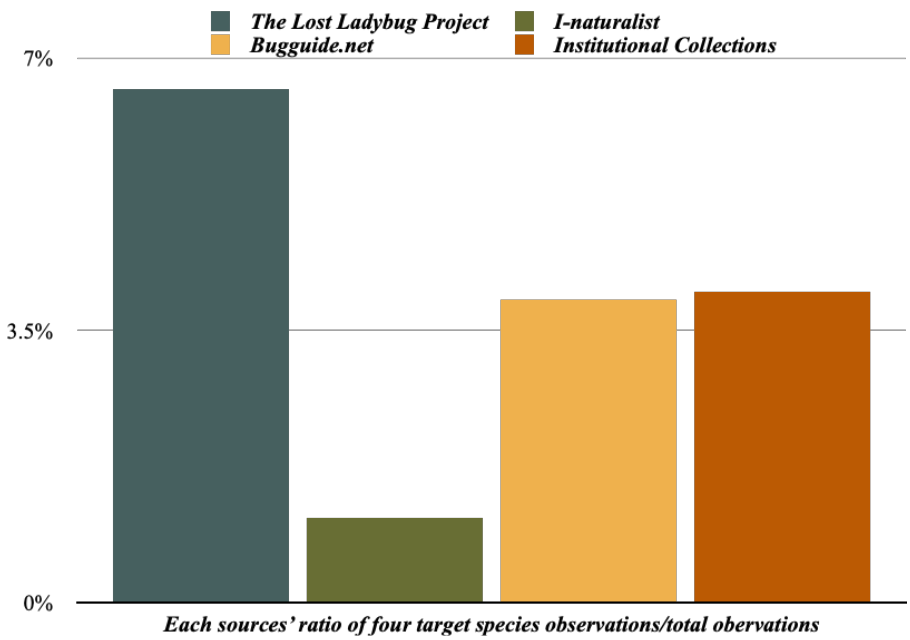



Figure 2-2. The efficiency of finding target species is inconsistent among sources. Citizen science platforms that report opportunistic observations ( "unstructured" ) also exhibit this difference (The Lost Ladybug Project, I-naturalist and the bugguide.net).

This data collection also displayed temporal bias due to the exponential increase in annual observations (Figure 2-3). Data volume until and after 2014, which is the midpoint, differed by about ten (or 9.61) times.

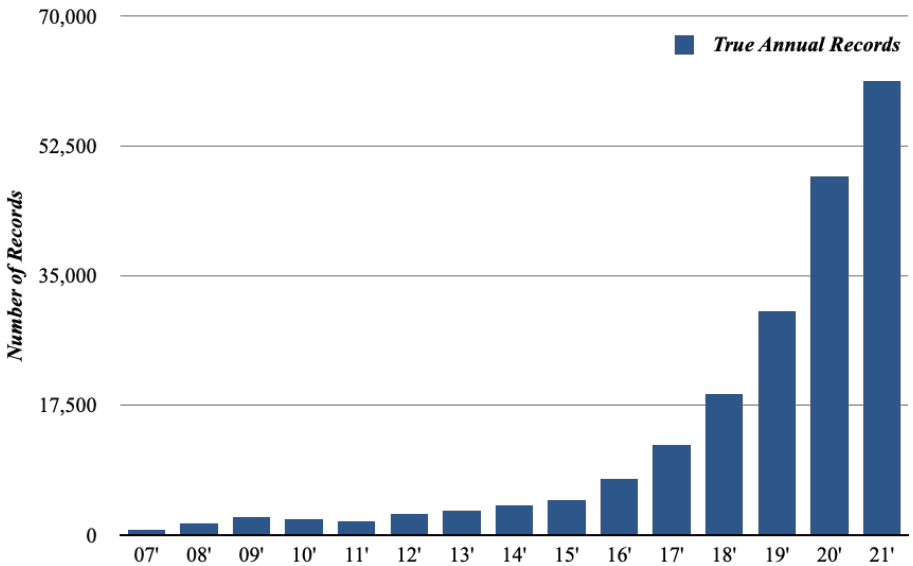



Figure 2-3. The total number of observations on coccinellids has exponentially increased.

### 2.2.2. Structural and temporal generalization

The ML classifiers developed with the co-occurrence pattern method (COP model) outperformed those developed with environmental variables (ENV model), in all types of generalization methods, including two structural validations between different survey structures and efficiencies, as well as two temporal validations conducted in both backward and forward ways. Moreover, the COP group overcame biases inherent in the dataset and were generalizable, whereas the ENV group failed. COP models trained on target species' unstructured citizen science data points (from I-Naturalist, The Lost Ladybug Project, and bugguide.net) were able to predict presence versus absence data points from structured data stored in institutions, surpassing acceptable performance standards (Accuracy, Recall,

Precision, F1 > 0.65 (as a rule of thumb); AUC > 0.70 (Mandrekar 2010), Kappa > 0.40 (Landis and Koch 1977), Brier < 0.25 (Brier 1950)). The three other validations' results also showed these levels of COP models' performance.

During COP model validation between different survey structures, *C. transversoguttata*'s model was evaluated as outstanding in AUC and excellent in Kappa (Figure 2-4). The other three models were evaluated as excellent in AUC and good in Kappa. Across the major four scores, *C. transversoguttata*'s models performed the best (Accuracy = 0.87, AUC = 0.94, Kappa = 0.75, Brier = 0.11), followed by *C. novemnotata* (0.81, 0.85, 0.61, 0.16), *H. parenthesis* (0.78, 0.84, 0.55, 0.17), and *A. bipunctata* (0.73, 0.84, 0.46, 0.19). ANOSIM results showed that there were small sizes of dissimilarity (<0.25, p=0.001) between the two types of co-occurrence patterns; one is accompanied by citizen scientific data points and the other is accompanied by institutional data points. Next, the minimum performance of COP models' predictions made by training data from a low-efficiency surveys (1.2) to points from the high-efficiency survey (6.8) is 0.72, 0.78, 0.45, 0.19, surpassing acceptable performance standards (Figure 2-4). In contrast, ENV models performed one or two ranks lower than COP models overall, and *A. bipunctata*'s model was not generalized to different efficiency surveys, or the *H. parenthesis* model to different structure surveys (Figure 2-6).

In terms of COP model's performance, *C. transversoguttata*, *C. novemnotata*, and *H. parenthesis*' models had similar levels of performance during both temporal validations, regardless of the projecting direction (Figure 2-5). In contrast, the *A. bipunctata* model had a higher performance than *H. parenthesis* in the forward temporal validation, but the lowest performance in the backward temporal validation. According to the backward validation, Recall, the ratio of correctly predicted positive values to the total true positive value, rose 11% compared to the forward validation. However, Precision, the ratio of correctly predicted positive values to all predicted positive values,

dropped by about 15%. Therefore, *A. bipunctata*'s model trained by current occupancy displayed tendencies to classify more diverse conditions as occupied habitats than it actually was in the past, unlike the other three. Meanwhile, the ENV models of *C. novemnotata* and *H. parenthesis* were not generalized in forward or backward direction, displaying a higher loss rate than the cases of structure validations, which indicated larger performance degradation (Figure 2-7).

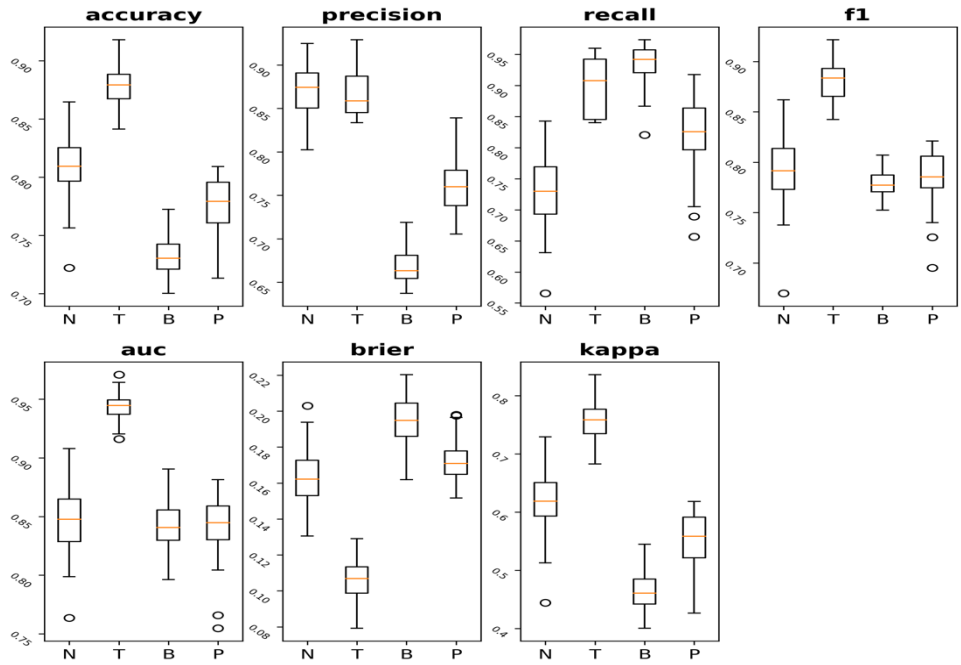
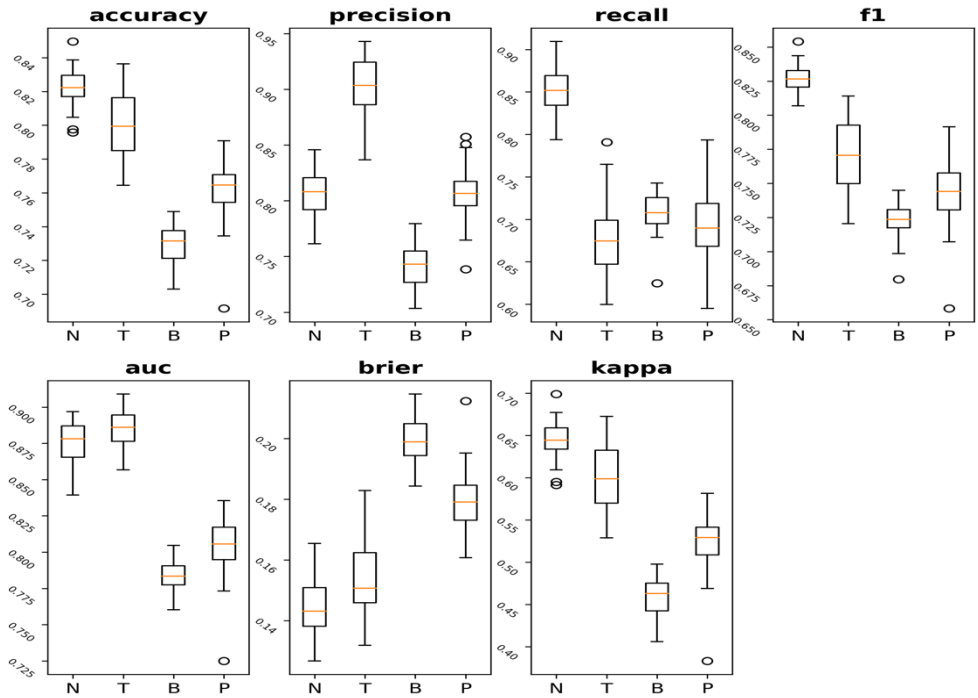



Figure 2-4. Structural generalizations of models training co-occurrence patterns (COP ML). Above: citizen science data (unstructured) trained COP ML predicted presence/absence in institutional data (structured). Below: Low efficiency data trained COP ML predicted presence/absence in high efficiency data. Box plots show the mean ML performance over 2,500 cases (combinations of 50 cases of random train/test splits and 50 cases of random absence data sampling) for each species (N = *C. novemnotata*, T = *C. transversoguttata*, B = *A. bipunctata*, P = *H. parenthesis*).



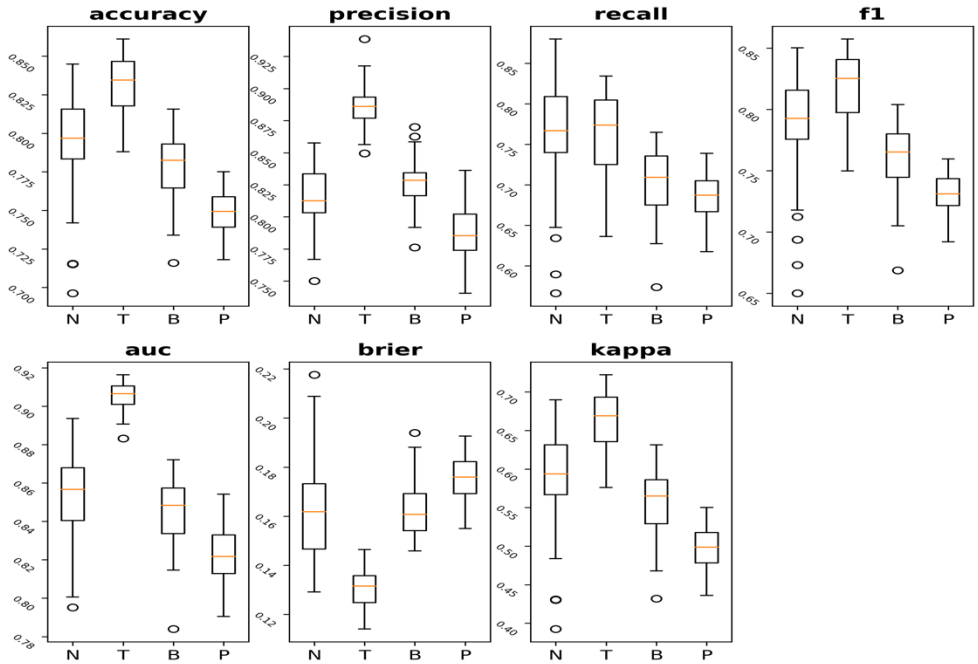
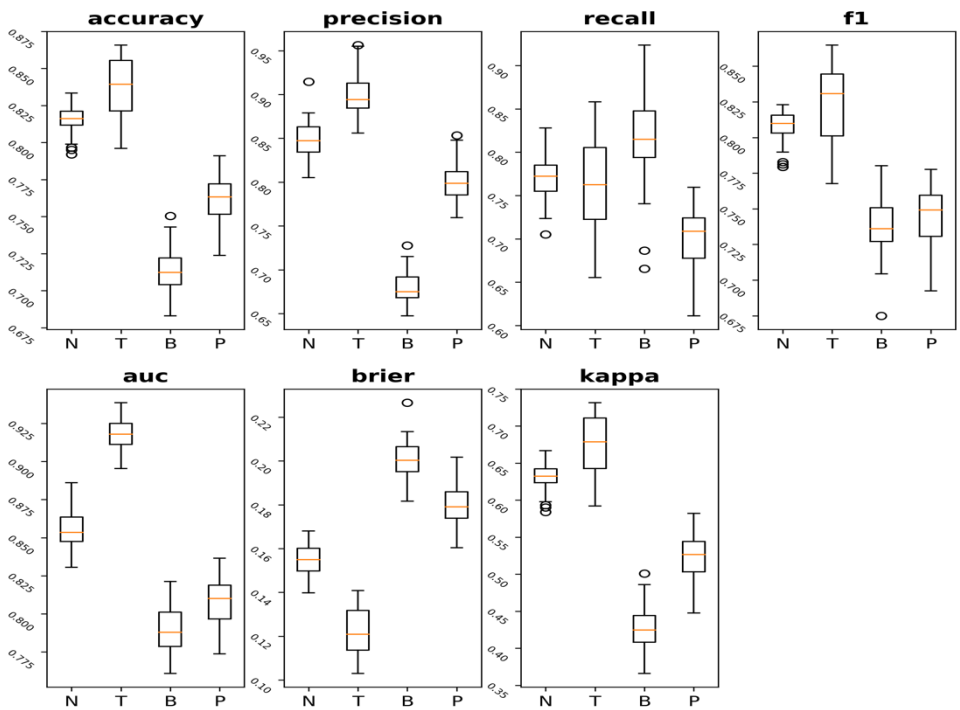



Figure 2-5. Temporal generalizations of models training co-occurrence patterns (COP ML). Above: learning data before a particular year to predict after that year. Below: vice versa. Box plots show the mean COP ML performance over 2,500 cases (combinations of 50 cases of random train/test splits and 50 cases of random absence data sampling) for each species (N = *C. novemnotata*, T = *C. transversoguttata*, B = *A. bipunctata*, P = *H. parenthesis*).



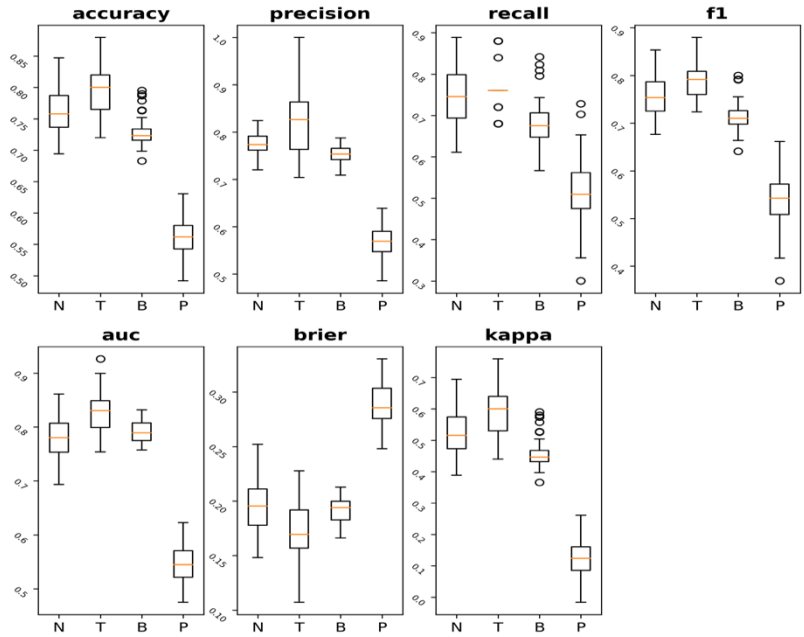
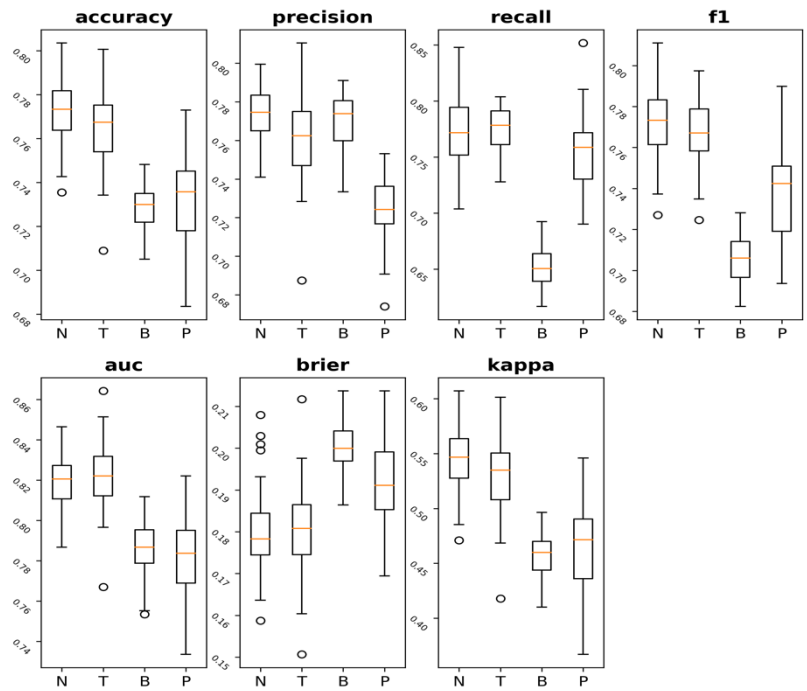



Figure 2-6. Structural generalizations of models training environemntal variables (ENV ML). Above: citizen science data (unstructured) trained ENV ML predicted presence/absence in institutional data (structured). Below: Low efficiency data trained ENV ML predicted presence/absence in high efficiency data. Box plots show the mean ENV ML performance over 2,500 cases (combinations of 50 cases of random train/test splits and 50 cases of random absence data sampling) for each species (N = *C. novemnotata*, T = *C. transversoguttata*, B = *A. bipunctata*, P = *H. parenthesis*).



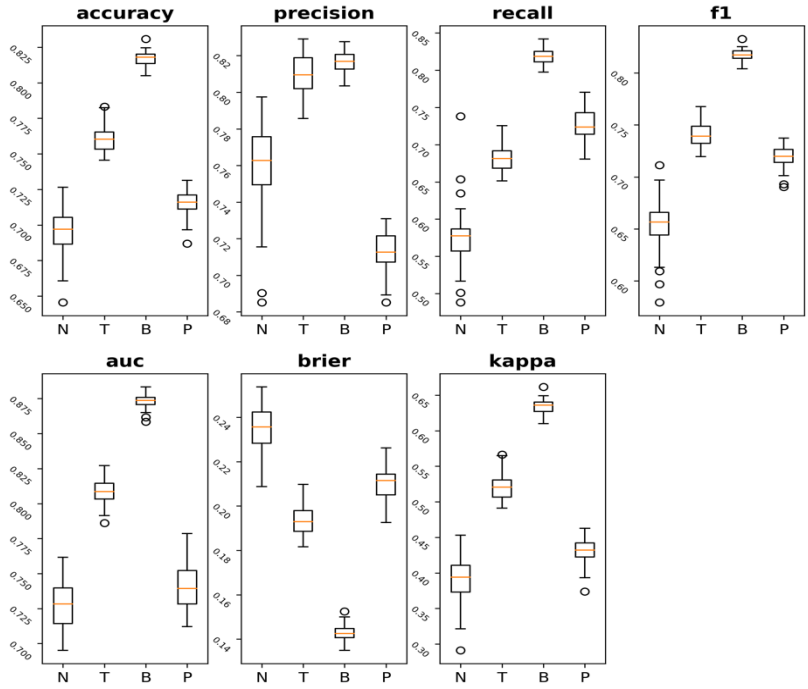
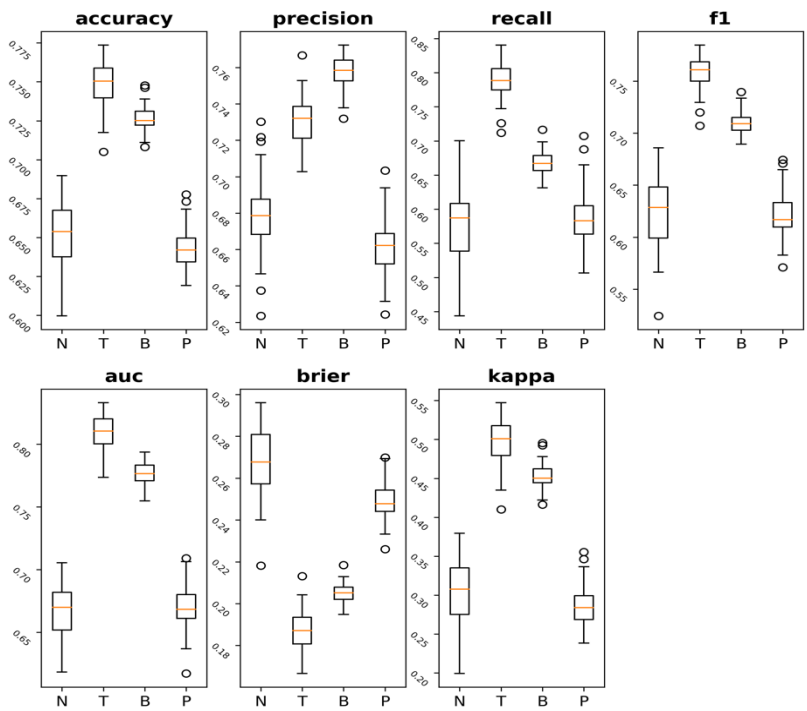



Figure 2-7. Temporal generalizations of models training environemntal variables (ENV ML). Above: learning data before a particular year to predict after that year. Below: vice versa. Box plots show the mean ENV ML performance over 2,500 cases (combinations of 50 cases of random train/test splits and 50 cases of random absence data sampling) for each species (N = *C. novemnotata*, T = *C. transversoguttata*, B = *A. bipunctata*, P = *H. parenthesis*).



### 2.2.3. Evaluation of the developed models

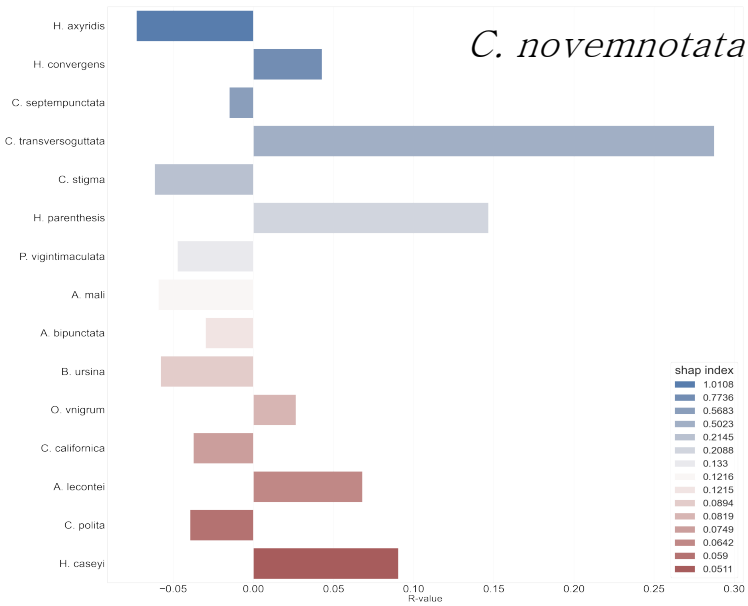
Using the entire multi-source collection, COP models relatively outperformed ENV models. Also, COP models were evaluated to be satisfactory in terms of absolute accepted standards. Among COP models, the highest overall score went to *C. transversoguttata* (number of presence data points = 510), followed by *C. novemnotata* (= 324) and then *A. bipunctata* (= 1,438), with *H. parenthesis*' (= 742) model at the lowest level. In particular, the minimum accuracy, precision, recall, and F1 of all four models were higher than 0.75 (> excellent), those AUC scores were higher than 0.87 (> outstanding) and those Kappa scores were higher than 0.57 (> substantial). These models are evaluated as being capable of predicting binary detections and non-detections relatively accurately. All models have Brier scores of less than 0.15, which confirms their consistency as prediction models.

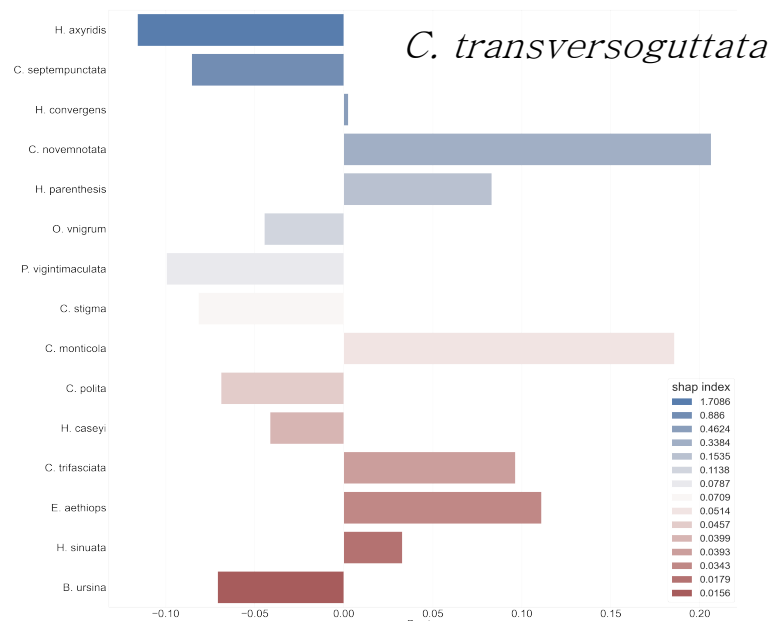
ENV models were also evaluated to be acceptable. The highest overall score went to *C. transversoguttata*, followed by *A. bipunctata* and then *C. novemnotata*, with *H. parenthesis*' model at the lowest level, similar to the COP models' order. To be specific, the minimum accuracy, precision, recall, and F1 of all four models were higher than 0.72 (> excellent), those AUC scores were higher than 0.79 (> acceptable) and those Kappa scores were higher than 0.47 (> moderate)—0.3, 0.08, and 0.1 lower than those of COP models, respectively. All models have Brier scores of less than 0.19, which is 0.04 higher than COP models, but still making them acceptable.

### 2.2.4. Importance and correlation among variables

Based on SHAP value and Pearson's correlation analysis, *C. novemnorata*, *C. transversoguttata*, and *H. parenthesis* positively correlated in our datasets and predictions of their presence depended on each other in ML calculations. *H. axyridis* and *C. septempunctata* were negatively correlated with these three target species and had the

highest significance among the variables. *A. bipunctata*, however, was an exception, showing a slight positive correlation with *H. axyridis* and *C. septempunctata*. *H. convergence*, with the third highest abundance in our collection and native to North America, displayed a positive association with three of our native species and was used as a major variable, exempt from the model of *H. parenthesis*.



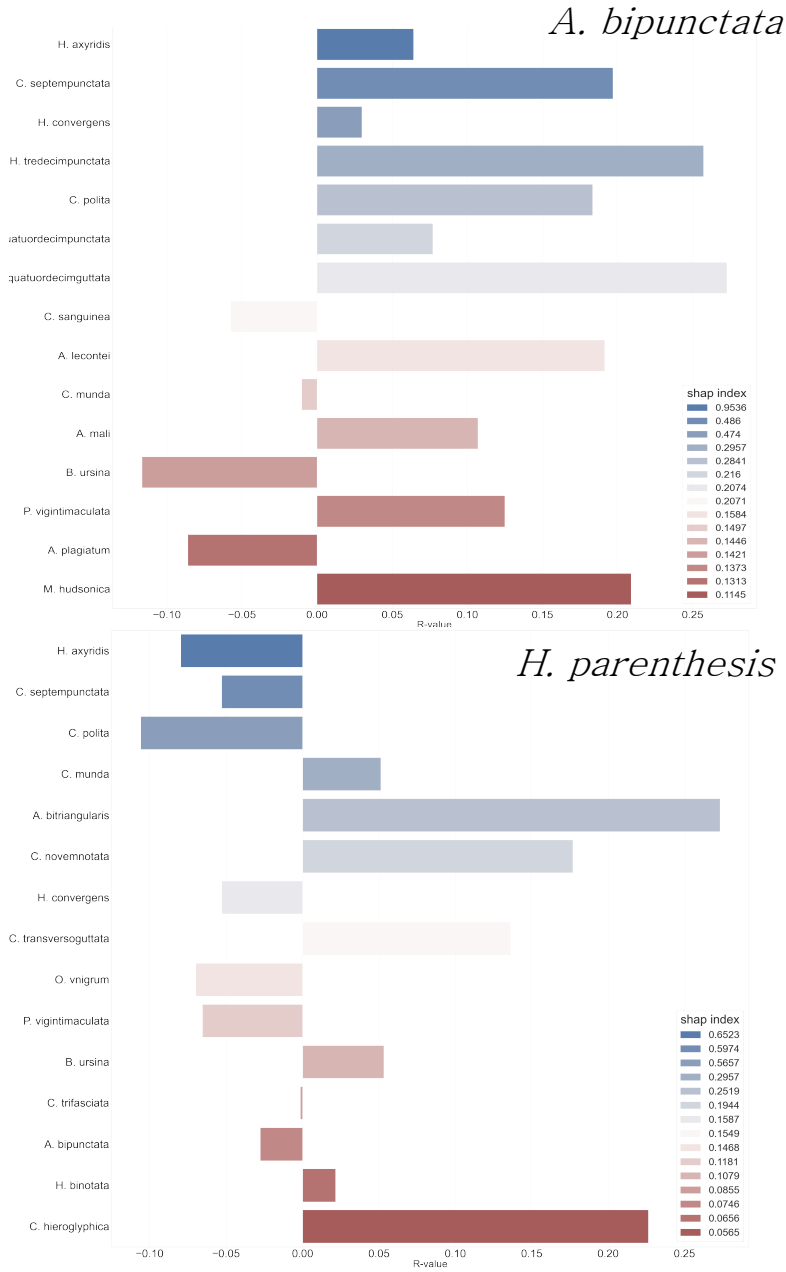



Figure 2-8. SHAP index (vertical axis) indicates the importance rank of a variable in ML calculation. R-value (horizontal axis) indicates the degree of correlations.

## 2.3. Discussion

### 2.3.1. The strength of co-occurrence pattern learning

In this study, the ensemble models based on the Co-occurrence pattern (COP) demonstrated practical performances for structurally and temporally inconsistent data.

Acceptable performances of COP models during the generalization across different data structures indicate the strength of COP to utilize datasets combined from multiple sources, including varying survey protocols. Many studies have recognized the importance of integrating data from multiple sources (Miller et al. 2019, Spear et al. 2017, Isaac *et al.* 2020, Robinson *et al.* 2020, Martino *et al.* 2021, Shirey *et al.* 2021). However, due to concerns about inherent inconsistencies between the data (Isaac and Pocock 2015), especially doubts regarding the reliability of citizen science data (Isaac and Pocock 2015), some researchers have chosen to either avoid using citizen science data or multi-source data in their SDM development, thereby sacrificing the coverage of their data (Steen *et al.* 2019).

The findings of this study demonstrate that, at least in some situations, the utilization of COP can overcome inconsistencies among data. This suggests that despite recent doubts about the true effectiveness of the expansion of citizen science (Lukyanenko *et al.* 2016, Kamp *et al.* 2016 Bayraktarov *et al.* 2019), COP can derive benefits from citizen science's quantitative expansion in SDM.

In the other hand, when trained on past data to predict future data collected on an annual basis and vice versa, the model showed acceptable performance. This suggests that COP can be utilized for more precise predictions at finer temporal scales beyond the long-term atlases that SDMs have traditionally focused on (Tingley and Beissinger 2009). Theoretically, COP has been assumed to have strength in integrating interactions among organisms (Pollock *et al.*

2014). The results of this study report that these strengths can lead to the advantage of capturing more immediate responses of organisms to micro changes compared to relatively macro changes in the environment.

Additionally, all COP models in this study outperformed the Environmental (ENV) models. Some species' ENV models exhibited methodologically unacceptable performance levels for specific generalization tasks in this research. Although environmental information is the most prevalent pool of variables in SDM (Martínez-Minaya *et al.* 2018), these findings suggest that COP can be more effective in certain tasks or species.

This study's target species are widely distributed across North America but threatened by rapid decline due to competition (Harmon *et al.* 2007, Losey *et al.* 2012). From this perspective, the following factors could have made COP stronger than ENV. Firstly, a major cause of decline was competition with adventive species (Turnipseed *et al.* 2014, Tumminello *et al.* 2015). Secondly, the dominance of adventive species altered the coccinellid community within its traditional landscape (Wheeler and hoebke 1995, Turnock et al. 2003, Harmon *et al.* 2007, Losey *et al.* 2007, Hesler and Kieckhefer 2008, Behali *et al.* 2015, Peterson and Losey 2022). Thirdly, there have been no prominent landscape changes around the presence points over the past 15 years (in yale Earth ENV data). Lastly, the target species are habitat generalists and their high dispersal abilities allow them to actively move to new habitats (McCorquodale 1998, Woltz and Landis 2013). The fact that biological change is more pronounced than environmental change might have played a role in causing the performance gap between the COP and ENV models. In other words, in these situations, it may be advantageous to utilize COP. Therefore, further research is needed to investigate the extent to which the performance differences uncovered in this study can be generalized.

### 2.3.2. The interpretation of used variables

The relationship between variables and target species revealed by the SHAP index and correlation coefficient analysis aligned with the previously known ecological relationships of the following major species. In the ML of *C. novemnotata* and *C. transversoguttata*, these species largely depended on each other. Their known preference for overlapping habitats and resources accounts for this observation (Hesler *et al.* 2009). Meanwhile, *H. axyridis* and *C. septempunctata* were negatively correlated with three target species and had the highest significance among the variables. Presumably, intense competition across North America produced this result (Wheeler and hoebke 1995, Turnock et al. 2003, Harmon *et al.* 2007, Losey *et al.* 2007, Hesler and Kieckhefer 2008, Behali *et al.* 2015, Peterson and Losey 2022). *A. bipunctata*, however, was an exception, showing a slight positive correlation with them. The species is known to share a certain degree of niche overlap with *H. axyridis* on a macro scale (Coderre *et al.* 1995, Koch, 2003, Omkar and Pervez 2005, Hentley *et al.* 2016). The overlap between *C. septempunctata* and *A. bipunctata* has also been reported in much European research, where both are native (Honěk 1985, Nedvěd 1999). However, this does not necessarily imply that *A. bipunctata* is immune to the negative effects of competition on a smaller scope (less than the 18 km radius that this study employed; Kajita *et al.* 2000, Soares and Serpa 2007, Kajita *et al.* 2006). Conversely, it appeared that a lower SHAP index corresponded to a lesser degree of known ecological interaction with the target species.

### 2.3.3. The incorporation of new variables

As variables, co-occurring species would provide information on (1) interactions (ex. competition), (2) habitat types (ex. urban garden), and (3) geographical realms (ex. West-East). In this regard, other potential organisms from other taxonomic groups of the target species (i.e. not the species' own taxonomic groups) could be conceptually

incorporated into the pool of variables.

In terms of information, a species that seems unrelated sometimes may provide more valuable information compared to a species that is directly interactive with the target species. For example, records on aphids and parasitic wasps, that directly interact with coccinellids as prey and predator, were too scarce to be employed. Meanwhile, our ML was shown to perform about 2% better when records of Passeriformes (perching birds), slightly related to the ecology of ladybugs but abundantly monitored, are included.

Conceptually, this may implies that these variables worked as indicators of (2) habitat types or (3) geographical realms. For example, it is possible that the geographical distribution of some perching birds could have aligned with the geographical realms of target species. Even though this study intentionally excluded geographical indicator variables to achieve our goal of selecting variables that can predict over time, this kind of variables (ex. coordinates) is commonly used to static prediction models with invertebrates (Gaul *et al.* 2022, Tyler *et al.* 2019), so the perching birds might be useful. In any case, to establish the validity of new potential variables, increases in performance scores should be interpreted as valid only if they result in a more descriptive model of the real world, rather just a technically improved model.

Over the past 20 years, the increase in observation data is a phenomenon shared by most taxonomic groups, but the rate varies significantly among species and even across taxonomic groups (Knape *et al.* 2022). Birds, in comparison to insects, exhibit the steepest rate of data accumulation. Therefore, if COP can harness information from species with rapid data accumulation to predict habitats for species with slower data accumulation, it may help bridge the data gap between taxonomic groups. However, an increase in overall accuracy cannot guarantee the descriptive reliability of prediction actually improved. Therefore, further research is needed to determine whether

incorporating birds or other taxa will strengthen co-occurrence patterns or if it will generate more biases and illusory patterns.

Moreover, additional studies are required to explore the balance between expanding the information content through the integration of biological variables of new organisms and retaining only high-quality correlated variables.

Although environmental variables were excluded from this study in order to reveal the effect of biological variables, the use of biological and environmental variables is not mutually exclusive, and future models may present more accurate results when combined.

### 2.3.4. The limitation in application

Biological interactions that are temporally and spatially inconsistent have been identified as a weakness of the co-occurrence pattern as a predictor in species distribution models (Tikhonov *et al.* 2017). It is noteworthy that the performance of predicting the recent distribution based on the relative past of *A. bipunctata* declined significantly compared to predicting the past based on the recent distribution. This may be derived from the changes in the habitat preferences of *A. bipunctata* as a result of the "habitat compression" following the introduction of *C. septempunctata* and *H. axyridis* over the past 15 years (Bahlai *et al.* 2015). Therefore, there may be limitations in predicting the past habitat preferences of A. bipunctata based on recent data. In terms of temporal prediction, the COP model may have limitations in applying those predictions to a distant past or future where there are rapid changes in the relationships between species over time or where such changes are inevitable.

## 2.4. Conclusion

This study developed an ensemble model to predict the

presence/absence of four native ground beetle species in North America that are known to decline rapidly due to competition with adventive species. Specifically, the performance of the ensemble model was compared when trained on co-occurrence pattern variables and environmental variables. The study assumed a scenario of developing species distribution models (SDMs) for most invertebrates and rare species, where only presence-only data is available, data quantity and density are low, data from different survey processes need to be combined, and accurate predictions at fine temporal scales are required. The study compared and derived conclusions about the structural and temporal generalization of SDMs by assuming these restrictions. The co-occurrence pattern model (COP ML) revealed two advantages over the environmental variable model (ENV ML). Firstly, co-occurrence patterns were advantageous for integrating data from multiple sources compared to environmental variables. In both structural generalization tests (trained and tested between different degrees of efficiency or structure data groups), COP ML consistently outperformed the agreed-upon benchmark, while ENV ML failed to meet the benchmark in some generalization tests. Additionally, the COP ML performed better than the ENV ML for all four species. Secondly, co-occurrence pattern variables provided more predicative models than environmental variables because they could generalize to predict the present based on the past or predict the past based on the present with more accuracy. Moreover, with respect to both temporal generalization tests, COP ML consistently outperformed the agreed-upon benchmark, while ENV ML failed to meet the benchmark in some generalization tests. Additionally, the COP ML performed better than the ENV ML for all four species. The analysis of biological variables revealed that the correlation and importance of key variables with the target generally matched known ecological patterns. Based on these results, the study suggests that ML developed with COP can integrate multiple-source data without filtering, allowing for the acquisition of more data, and that COP-based SDMs may be advantageous for predictions at finer temporal scales (and thus more precise than commonly used SDMs developed with environmental variables usually

spanning over decades), which is especially necessary for many invertebrates and rare taxa.

# Chapter 3. Filling Machine Learning Predictions In Temporal Data Gaps Can Estimate Annual Reductions Across Every Historical Distribution.

## 3.1. Materials and methods

### 3.1.1. Summary of materials and methods

Same with 2.1.1.

### 3.1.2. Target species

Same with 2.1.2.

### 3.1.3. Occurrence data

Same with 2.1.3.

### 3.1.4. Pseudo-absence

Similar with 2.1.3. but I was randomly pooling pseudo-absence datasets without considering the state ratio for reliability comparison while pooling 'matched state ratio' pseudo-absence dataset for standard models.

### 3.1.5. Variables

Same with 2.1.5. but I only used COP as variables.

### 3.1.6. Development and characterization of models

Same with 2.1.6. but I didn't evaluate developed models.

### 3.1.7. Prediction on annual distributions and reduction rates

For an even comparison of temporal changes in population size, I applied the trained ML classifiers to fill in the prediction into the deficiencies of the collected yearly data. To do this, the models predicted the annual existence of the target species in every historical coordinate after 2007.

(1) Prediction: Prediction models were developed in the same way as the Development and characterization step, except those models with 100% training of available presence data were used for predicting past annual distributions of each target species, to boost the predictive performance (Fielding 1997, Rencher 2002). If more than 50% of 50*50 numbers of the models developed with each unique combination of pseudo-absence data points were in favor, a GPS was considered as occupied at the year. This whole process was repeated 30 times by changing pseudo-absence pools to obtain confidence intervals of the annual distribution.

(2) Analysis: To evaluate and analyze distribution trends, the AOO (Area of occupancy) and EOO (Extent of occurrence) of the IUCN Red List criteria system were used. Under criteria A, the population reduction rate over a ten-year moving time window was obtained from analyzing the predicted annual distribution measured by counting 4 km^2 grid cells, a way devised to measure AOO that indicates how much area taxon occupies and presents an indirect value of population size.

### 3.1.8. Validity evaluation

Validity refers to the practical relationship between our measurement and the ability of that measurement to accurately represent a targeted object. To confirm whether the estimates of the methodology for the population over the years are independent of the temporal variation in data availability, linear regression analysis and

correlation coefficient analysis were conducted with the LLP's total annual observations. LLP was selected as an independent variable for validation because it has the highest number of data points for the target species among sources and its fluctuation is not consistently increasing or decreasing.

In addition to the filling-in approach, three traditional methods widely used for estimating the rate of decline in presence-only data were also implemented and compared: RA (Relative Abundance; described mathematically as the target species' annual observations divided by the total coccinellids' annual observations), ACC (Historical records accumulation; trends of the annual AOOs given that GPS has been continuously occupied by the target species from 2007 to the last discovery on it), and Raw Records (the simple number of annual reports).

### 3.1.9. Reliability evaluation

Reliability is the consistency and stability of measurement when the study is repeated or replicated under similar conditions. To evaluate reliability, this study evaluates the degree of variation in estimations among COP models developed with (1) different pseudo-absence datasets, (2) different variables, and (3) different presence datasets that is partially lost and partially replaced with false data points.

In terms of pooling pseudo-absence, variations between (a) randomly sampling absence data from all the research areas and (b) sampling absence data to match its ratio with presence data in each state was compared were evaluated.

In terms of pooling variables for modeling, variations between the two sets of variables, obtained by (a) [independent simple linear regression + VIF] and (b) [multiple linear regression + VIF] were evaluated.

In terms of errors inherent in the presence dataset, variations of estimates was evaluated when 10% of the presence records for the target species were randomly deleted (representing imperfect observations) and replaced with occurrence records of other species (representing misidentification errors) for 2,500 times of repetitions. Next, the mean squared error (MSE), compared to the values obtained from complete dataset, was obtained.

## 3.2. Results

### 3.2.1. Estimated reduction rates and conservation status

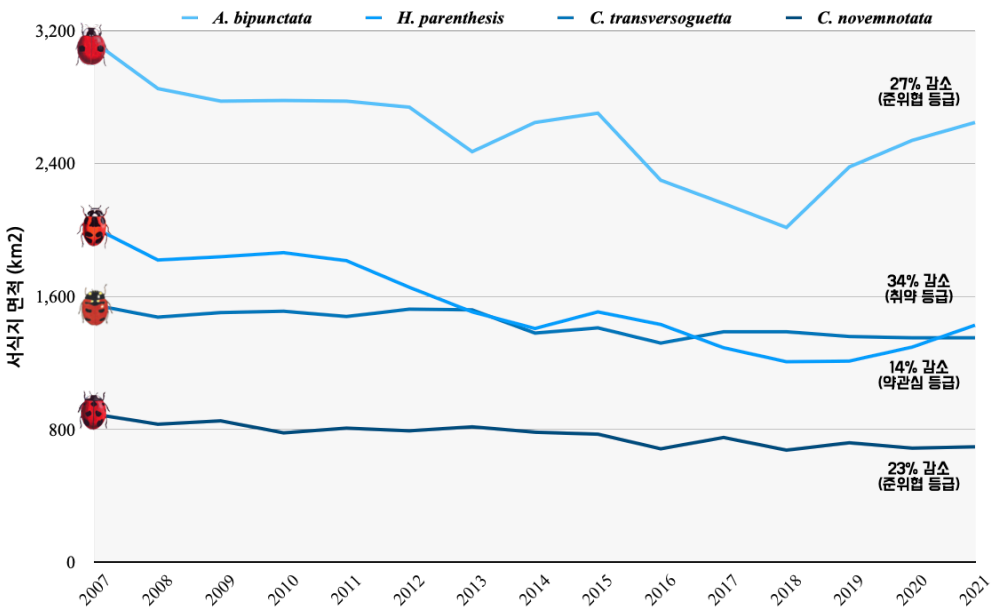



Figure 3-1. Estimated reductions of the four ladybugs native to North America.

The ML classifiers trained with co-occurrence patterns was applied to obtain annual occupancy by filling the prediction into the deficiencies of the volume of collected yearly data to be evenly compared from 2007 to 2021.

According to ML's predictions, three target species could have

been undergoing a fundamental extinction process in North America (Figure 3-1), currently. Area of occupancy (AOO), a direct indicator of the area occupied by the taxon and an indirect indicator of population size (IUCN 2022), was found to have declined since 2007 in all four species; *H. parenthesis*' AOO, which showed the largest decline, decreased by 1,962 km2, *A. bipunctata* by 584 km2, and *C. novemnotata* and *C. transversoguttata* by 480 km2 (Table 3-1).

According to IUCN Red List Criteria A, which evaluates extinction risk based on a reduction rate of a population size in recent years, a maximum 10-year decline within 2007-2021 was estimated to be 36.4% for *H. parenthesis* (2007-2021; "Vulnerable"), was 29.7% for *A. bipunctata* (2010-2019; "Near Threatened"), was 23.7% for *C. novemnotata* (2009-2018; "Near Threatened"), and was 14% for C. trasversoguettata (2007-2018; "Least Concern").

On the other hand, the extent of occurrence (EOO), a parameter of the ability of spatially spreading risks (IUCN 2022), was the most significant decrease by numbers in *C. transversoguttata*-This indicates that this LC species also has been undergoing deterioration in the ability to resist extinction as well as the progressive population reduction.

Table 3-1. Summary of distributional trends and IUCN status of the four target species.

| Species | Peak Reduction Rate (time window) | IUCN Status | AOO (km2) | | EOO (km2) | |
|---|---|---|---|---|---|---|
| | | | 2007 | 2021 | 2007 | 2021 |
| *A. bipunctata* | 27% (09'-18') | NT | 3,128 | 2,648 | 11,538,691 | 10,817,443 |
| *H. parenthesis* | 34% (10'-19') | VU | 1,548 | 1,352 | 8,450,469 | 7,749,070 |
| *C. transversoguttata* | 14% (07'-16') | LC | 892 | 696 | 9,820,525 | 9,146,848 |
| *C. novemnotata* | 23% (07'-16') | NT | 2,012 | 1,428 | 5,480,067 | 5,399,901 |

### 3.2.2. Validity comparison with pre-established methodologies

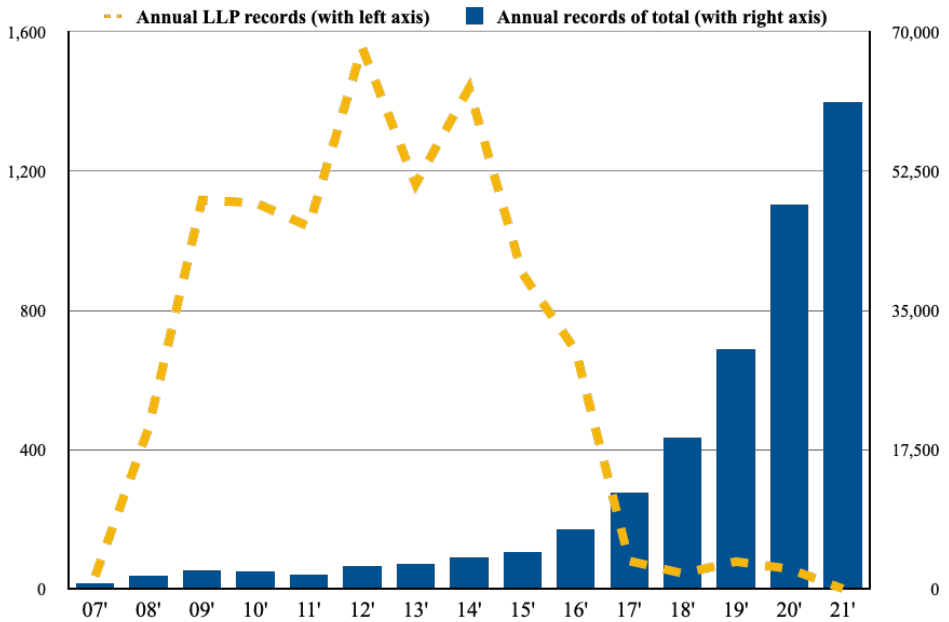



**Figure 3-4. CS's spreading has led to a continuous increase of the number of datapoints. However, LLP records with the highest efficiency in reporting native ladybugs have declined since 2014.**

This research's filling-in approach (ML), Relative Abundance (RA), Historical records accumulation (ACC) and Raw Records (Raw) estimated different predictions of population trends as well as distinct peak reduction rates over 10 years, thereby determining different IUCN Red-List conservation category for each species. Category for a species varied from LC (less than 20%: currently least concern in extinction risk) to CR (over 80%: indicating the highest level of extinction risk). When estimating population trends, the IUCN Red List category was highest for RA (*H. parenthesis*: 88%, *C. novemnotata*: 94%, *C. transversoguttata*: 90%, *A. bipunctata*: 78%), middle for ACC (71%, 84%, 76%, 65%), and lowest for ML (34%, 23%, 14%, 27%; Table 3-2). Meanwhile, the volume of our yearly data collection has increased since 2007 as the "raw records" indicated that targets had been reported more frequently (2% of increase per 10 years), while the overall density of target species' datapoints divided by all coccinellid observations had declined. Linear regression showed ML

had the most statistical independence from these time-series changes in data volume and yearly density of target species

Table 3-2. The estimated reduction rates derived from the raw data, two previously established methods, and a novel approach proposed in this study.

| | Machine Learning | Relative Abundance | Historical records accumulation | Raw Records |
|---|---|---|---|---|
| *A. bipunctata* | 27% (09-18) | 78% (11-20) | 65% (12-21) | -2% (09-18) |
| *H. parenthesis* | 34% (10-19) | 88% (10-19) | 71% (12-21) | -1% (09-18) |
| *C. transversoguttata* | 14% (07-16) | 90% (11-20) | 76% (12-21) | -1% (09-18) |
| *C. novemnotata* | 23% (07-16) | 94% (11-20) | 84% (12-21) | -1% (09-18) |

Descriptive statistics showed that there was an overturn between LLP's volumes and I-Naturalist's volumes (Figure 3-4). LLP, which accounted for only 5% of the total observations on coccinellids, reported target species at six times the density of I-Naturalist, the largest source of the dataset by containing 89% of all observations (Figure 2-2 and Figure 3-3). Nonetheless, while the total number of LLP observations had been on the decline since 2014, I-Naturalist had exponentially and steadily grown, lowering the density of the target

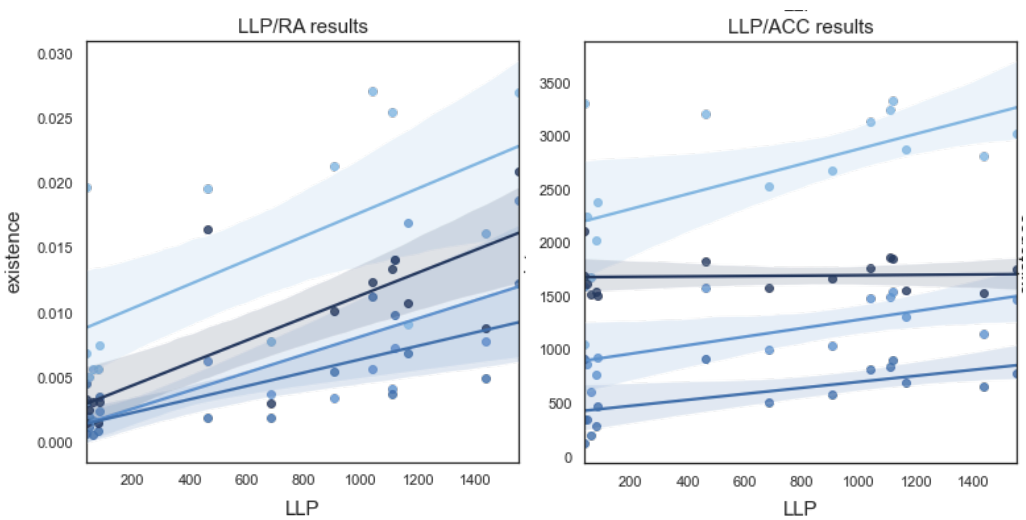



Figure 3-5. Linear Regressions between the number of observations from LLP and population size estimated by RA (Relative Abundance) and ACC (Historical Records Accumulation) methods.

species in the yearly collected coccinellid data.

To assess the validity of population trends estimated by each methods, a linear regression analysis was conducted (Figure 3-5). This analysis examined the relationship between the number of annual LLP observations and the annual estimated population size from RA, ML, and ACC. The result revealed that estimates of RA and ACC for all species' population sizes were dependent on the number of annual LLP observations (Table 3-3; RA: $F(1,13) > 8.64$, $p < 0.05$; ACC: $> 6.08$, $< 0.05$). ML results were relatively free from bias, since they kept the null hypothesis ($p > 0.05$); the only exception was *C. transversoguttata*.

Although LLP observations only took 4.9% of the multi-source dataset, its R squared value wasn't small, higher than 0.40 for all RA models and 0.31 for all ACC models. Conversely, the increase in the number of observations within the I-Naturalist dataset explained the tendency towards a decrease in the estimated population size ($R2 > 0.31$). The volume of LLP observations was positively correlated with estimations of RA and ACC (correlation coefficient $> 0.631$, $p < 0.05$), while the volume of I-Naturalist was negatively correlated with them ($< -0.563$, $< 0.05$).

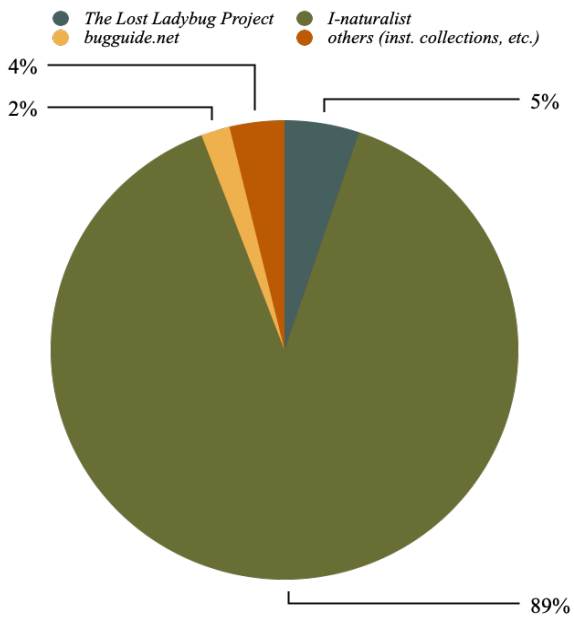



Figure 3-3. The proportion of each source from the total data collected in this study.

Table 3-3. The results from linear regression between the number of annual LLP observations and the annual estimated population size from RA, ML, and ACC.

| Species | Methods | linear regression with annual LLP observations | | | |
|---|---|---|---|---|---|
| | | $R^2$ | *correlation* | *p-value* | *F(1,13)* |
| *C. novemnotata* | Year | | | | |
| | RA | 0.57 | 0.755 | 0.001 | 17.25 |
| | ML | | | | |
| | ACC | 0.348 | 0.589 | 0.02 | 6.939 |
| *C. transversoguetta* | RA | 0.552 | 0.742 | 0.001 | 16 |
| | ML | 0.452 | 0.672 | 0.006 | 10.71 |
| | ACC | 0.333 | 0.576 | 0.02 | 6.478 |
| *A. Bipunctata* | RA | 0.4 | 0.631 | 0.01 | 8.643 |
| | ML | | | | |
| | ACC | 0.361 | 0.6 | 0.017 | 7.342 |
| *H. parenthesis* | RA | 0.627 | 0.791 | 0.0004 | 21.82 |
| | ML | | | | |
| | ACC | 0.319 | 0.564 | 0.028 | 6.08 |

### 3.2.3. Reliability analysis on filling-in approach

Using the filling-in approach resulted in less variation in the prediction of the population trend compared to aggregating different pseudo-absence data points, variables and presence data points.

In terms of aggregating pseudo-absence data points, (a) randomly sampling absence data from all the research areas and (b) sampling absence data to match its ratio with presence data in each state showed less than 10% of difference in peak reduction rates in each period (Table 3-4). This left less than one level of difference in IUCN conservation status for each species.

In terms of pooling variables for the Machine Learning modeling, the two sets of variables, obtained by (a) [independent simple linear regression + VIF] and (b) [multiple linear regression + VIF], showed marginal differences (less than 2%) in the reduction rates (Table 3-5).

Lastly, in terms of errors inherent in the presence dataset, this study went through 2,500 iterations of randomly deleting 10% of the presence of the target species or replacing the presence records with occurrence records for other species. This methodology represented imperfect observations and misidentifications errors, respectively. On average, the 2,500 iterations had an estimated range of variation that was within 8% (with an interquartile range less than 3% and a standard deviation less than 0.019) for all four species. The mean squared error (MSE), compared to the values obtained from the complete dataset, was 0.0014.

Conclusively, the filling-in approach demonstrated limited degree of variability in population trend and conservation status estimation when there are sorting variables, selecting different psuedo-absence datasets, missing occurrence records, and including misidentification data.

Table 3-4. Variations in 10 years of reduction rates for IUCN Red List between ways to sample absence data. Left: reduction rates for *H. parenthesis* and *C. novempnotata* were calculated using 10,000 pseudo-absence points, which were sampled in proportion to the presence data of each state or province. Right: reduction rates for the same species were calculated using 2,000 randomly sampled pseudo-absence points.

| | **Reduction rate in AOO of *H. parenthesis*** | |
|---|---|---|
| **The start point of the next 10 years decline** | ***Method 1*** | ***Method 2*** |
| 2007 | 0.32 | 0.29 |
| 2008 | 0.30 | 0.28 |
| 2009 | 0.37 | 0.27 |
| 2010 | 0.37 | 0.25 |
| 2011 | 0.30 | 0.25 |
| 2012 | 0.16 | 0.12 |
| | **Reduction rate in AOO of *C. novemnotata*** | |
| **The start point of the next 10 years decline** | ***Method 1*** | ***Method 2*** |
| 2007 | 0.23 | 0.18 |
| 2008 | 0.12 | 0.10 |
| 2009 | 0.20 | 0.18 |
| 2010 | 0.08 | 0.11 |
| 2011 | 0.15 | 0.25 |

Table 3-5. Variations between ways to sort variables. Left: with 15 variables chosen by simple linear regression. Right: with 9 variables chosen by multiple linear regression. These shared seven variables.

| | **Reduction rate in AOO of *H. parenthesis*** | |
|---|---|---|
| **The start point of the next 10 years decline** | ***Method 1*** | ***Method 2*** |
| 2007 | 0.25 | 0.26 |
| 2008 | 0.13 | 0.12 |
| 2009 | 0.20 | 0.21 |
| 2010 | 0.08 | 0.08 |
| 2011 | 0.13 | 0.13 |
| 2012 | 0.02 | 0.04 |

### 3.2.4. Predicted distributions

Machine learning predictions suggested that target species continued to exist in states where no recent additional records have been found (Figure 3-2). *C. novemnotata* was predicted to still be found in Washington, Wyoming, Arizona, South Dakota, Nebraska, Wisconsin, and Alberta. *C. transversoguttata* was to Saskatchewan, South Dakota, and Nebraska. *A. bipunctata* was to Manitoba, Wyoming, North Dakota, South Dakota, and Nebraska. *H. parenthesis* was to Oregon, Idaho, North Dakota, Wyoming, Utah, and Arizona. These predicted occurrence wasn't considered in any of RA and ACC's estimations and conservation status.

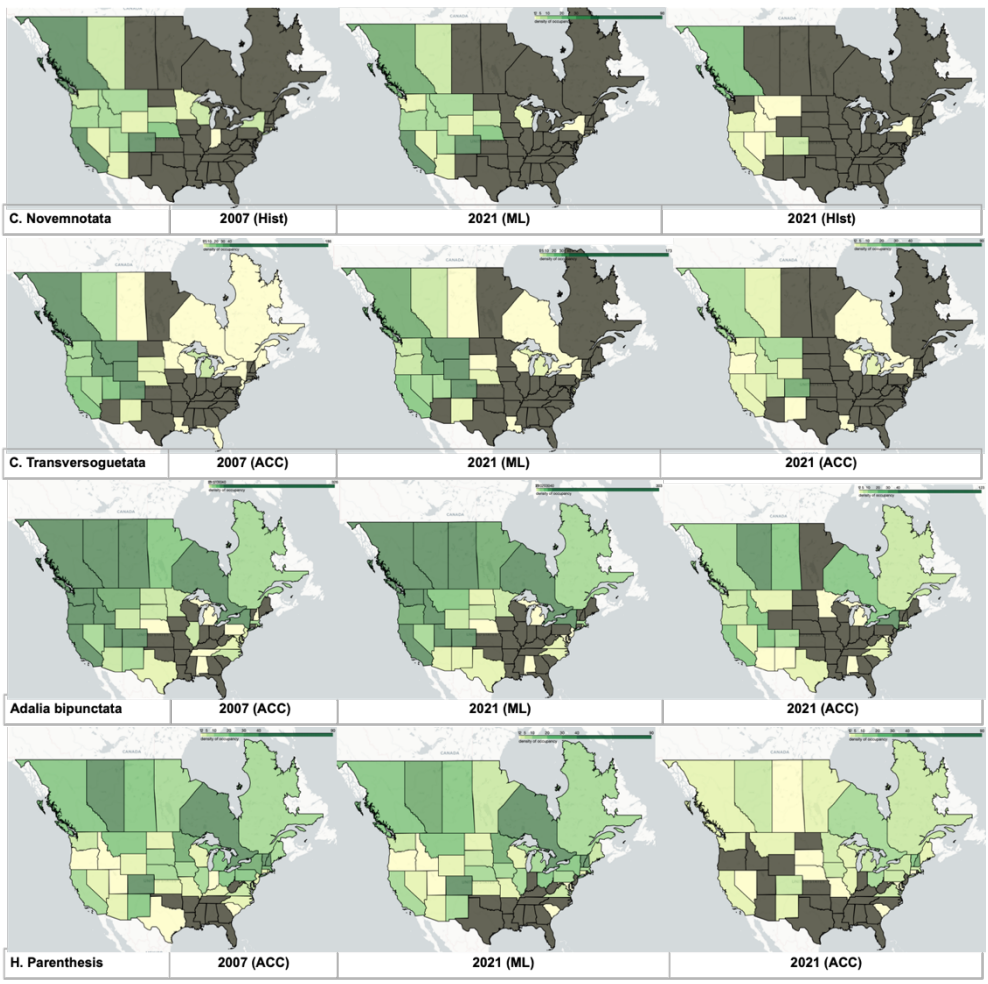



Figure 3-2. Comparision of yearly distributions between ML and ACC.

## 3.3. Discussion

### 3.3.1. The theoretical rationale for the ML reduction rates

It has been documented that the population of native species plummeted largely in recent decades. Nevertheless, there are several reasons that the scale of the current reduction rate (2007-2021) produced by ML, which is smaller than previous studies, is reliable. (1) The decline occurred mainly during the '80s and '90s soon after exotic species landed in North America (Colunga-Garcia and Gage 1998, Bahlai *et al.* 2015). (2) Several recent reports from regional habitats indicate there has been no further decrease in native species (Alyokhin and Sewell 2004, Bahlai *et al.* 2015). (3) Some researchers speculate that the coccinellid complex is likely to modulate the effects of adventive species over time (Turnock *et al.* 2003, Harmon *et al.* 2007, Hesler and Kieckhefer 2008), as many 'biological invasions' reached the chronic phase (Elton 2000, Strayer *et al.* 2006). (4) The remaining colonies might be more resistant and sustainable than other lost ones because of their metapopulation dynamics, refuge supply, etc (Evans 2000, 2004, Evans *et al.* 2011). Based on these reasons, I can assume that the current reduction amplitude should be lower than it was at the onset of the spreading.

In addition to that, (5) When I estimated the RA of *H. parenthesis* (548 number of observations) from a single study (2007-2019) in South Dakota (Bahlai *et al.* 2015; downloaded from: lter.kbs.msu.edu/datatables/67), 20 years after the advent of foreign species, the reduction rate of the trend line was 40% on a 10-year basis. This finding shows a similar intensity to the 34% in our study over the continent. Meanwhile, in the UK and Belgium, the rate of decrease in *A. bipunctata* population was 30% and 44% in a single study (2003-2008) conducted 25 years after the introduction of *H. axyridis* (Roy *et al.* 2012). This finding shows a similar level of 27% in this study.

From a policy perspective, it is crucial to emphasize that the projected rate of decline, although less severe than the well-known rate of decline, should not be interpreted as an indication that these species are no longer at risk or recovering. Despite having been heavily exploited in the past and having suffered substantial population losses, these species continue to face a decline.

### 3.3.2. Affect of temporal fluctuations of data on various models

Since the 2000s, the observations stored in crowd-source systems have shown a steep increase worldwide (Amano *et al.* 2016). The rapid expansion of citizen science has been identified as the primary driver of this trend (Dickinson *et al.* 2010, Isaac and Pocock 2015, Chandler *et al.* 2017). However, the temporal variability in the availability of citizen science data poses limitations in directly estimating species' abundance and the extent of its change (Boersch-Supan *et al.* 2019, Kamp *et al.* 2016, Bayraktarov *et al.* 2019). In particular, Knape *et al.* (2022) reported a decrease in the number of records submitted per observer, despite an overall increase in number of observations for insect taxa. They suggested that this could be due to lower knowledge and survey efforts regarding subsequent observations by new participants compared to initial participants (Knape *et al.* 2022). Nonetheless, there are studies suggesting that the population trends of semi-structured citizen science data can predict structured survey trends significantly, given appropriate treatment (Boersch-Supan *et al.* 2019, Horns *et al.* 2018). However, effective methods to control such temporal variability have not yet been agreed upon, particularly for taxa with low spatiotemporal data density, such as rare species or invertebrates, and especially in situations when data collection is entirely unstructured (Kamp *et al.* 2016).

According to our results, by accounting for the temporal discrepancy in the dataset, the estimation derived from the filling-in approach proved to be the most independent by accounting for the time-series fluctuations in the total dataset. These fluctuations were

represented by the number of LLP and I-Naturalist participants. Similar with Knape *et al.* (2022), the I-Naturalist, which constituted 89% of the total data and did not involve expert's verification and specific focus, had a six times lower ratio of target species observation compared to the LLP (1.17%), which constituted only 4.9% of the total data and was under a campaign to search for native ladybugs, aided by educational documents, and verified by experts in identification of photos.

Therefore, the more I-Naturalist data enter the model, the lower the estimated population density of the target species, even though the raw number of observed members of the target species increased. This phenomenon has been particularly noticeable since 2014 when LLPs began to decline. Consequently, an increase in the number of observations within the I-Naturalist dataset, while LLP had decreased, was associated with a tendency towards a decrease in the estimated population size of RA and ACC.

The results of linear regression revealed that RA and ACC's estimates were confounded by their dependence on the LLP data; in other words, their measurements could validate replications of LLP data, but not valid measures of actual population shifts. In addition to that, the reduction rates for RA and ACC are much sharper (as EN to CR in IUCN Red-List) than those for ML (LC to VU). To further emphasize this point, with the exception of ML, the periods of peak reduction rates from other methods tend to vary less, such as 2012 to 2021 for ACC, no matter the target.

These findings indicate that the estimates derived from RA and ACC may result from a decline in LLP or a trade-off with the I-Naturalist. As a result, using these methods to estimate population trends without addressing the inherent fluctuations in unstructured citizen science data could lead to significant errors in policy decisions.

### 3.3.3. Practical benefits of the filling-in approach

most reliable method for estimating population trends is to conduct long-term sampling at consistent locations (Elliott *et al.* 1996, Strayer *et al.* 2006, Honek *et al.* 2016). However, due to the high cost involved, most species lack this type of monitoring effort. In comparison, unstructured citizen science data is three to four times more cost-effective than structured sampling efforts (Gardiner *et al.* 2012). Nevertheless, in citizen science data, the sampling effort is not consistently distributed in space and time to cover the entire population (Bayraktarov *et al.* 2019). Consistently filtering out only the well monitored areas (ex. 5 times of revisit during the study period; Schultz *et al.* 2017) has been a long-standing dilemma for ecologists because it removes a substantial portion of the available data volume and the overall distribution of species, resulting in a trade-off between quantity and quality (Gábor *et al.* 2019, Wisz *et al.* 2008). This poses particular challenges for species with small population sizes, limited geographical ranges, or rapid population declines, which are prioritized for assessments of extinction risk, as their sample sizes are small (Hertzog *et al.* 2021).

To mitigate the inherent heterogeneity in citizen science and the high costs associated with traditional monitoring, the filling-in approach can be utilized to maximize the utility of citizen science data and enhance the cost-effectiveness of data collection. For example, alternative strategies have been proposed to fill data gaps, such as sending people to unsampled areas or providing financial incentives (Tulloch *et al.* 2013, Xue *et al.* 2016). However, employing machine learning predictions to fill these gaps can be more economical than sending individuals.

Additionally, in general, more citizen scientists' participation broaden the coverage of the surveys' taxonomic scope (Chandler *et al.* 2017, Pocock *et al.* 2019). Instead of reducing participation and volume of observation due to strict protocols (Pocock *et al.* 2017), this study's

approach is about to adjust for the inherent discrepancies in unstructured data and can cover a broader spatiotemporal range of under represented species by maximizing quantitative use of unstructured, therefore low hurdled, citizen science. In other words, this study's approach benefit from low-hurdled citizen science data's broad coverage while also demonstrating robustness.

In summary, utilizing citizen science data through the filling-in approach and leveraging machine learning techniques to address data gaps and discrepancies can enhance the economic efficiency of data collection, particularly for species at risk and with limited monitoring efforts.

### 3.4.4. The filling-in approach in conjunction with data filtering methods

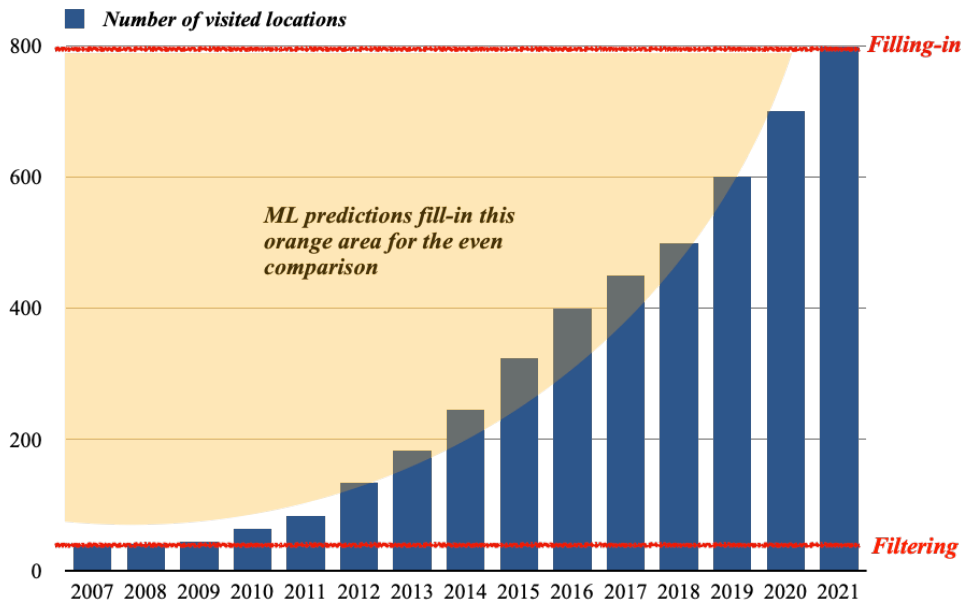



Figure 3-6. Conceptual comparison between Filling-in and Filtering methods in their coverages on locations for temporally consistent estimation.

Citizen science surveys can be used to contour population trends in three ways. The first two methods are a direct application of

abundance (Newson *et al.* 2015, Walker and Taylor 2017, Schultz *et al.* 2017) or secondly, an application identified effort inputs (ex. survey time or check list; LeCroy *et al.* 2020, Fink *et al.* 2020) to statistically assume an observational chance. The one-million Coccinellid data points this study relied on, however, did not carry these pieces of information. Thus, if I adopted methodologies (1) and (2), I would have been forced to disregard these data in holding with common academic practices.

Thirdly, 'filtering' is a procedure to remove some data points from unstructured, noisy, or multi-sourced (such as GBIF) datasets in order to make them more consistent and evenly distributed enough to compare (Steen *et al.* 2019). Filtering ecological datasets has frequently been used as a way to minimize bias to reveal a signal of biological change (Hickling *et al.* 2005, 2006, Kuussaari *et al.* 2007, Roy *et al.* 2012, Isaac *et al.* 2014, Aiello-Lammens *et al.* 2015, Galante *et al.* 2018, Robinson *et al.* 2018).

However, in the case of taxa with limited data, it is generally assumed that information obtained from a larger quantity of records surpasses the potential bias of opportunistic sampling (Boersch-Supan *et al.* 2019). Therefore, there is typically a trade-off between collecting a relatively heterogeneous (i.e., lower "quality") large volume of data and collecting a smaller quantity of higher "quality" data that adheres to a defined common structure (Boersch-Supan *et al.* 2019). The outcomes of this trade-off between quantity and quality are still not fully understood (Aceves-Bueno *et al.* 2017, Bayraktarov *et al.* 2018, Kelling *et al.* 2018, Specht and Lewandowski 2018).

In contrast with filtering, our method ‘fills-in’ the gaps that exist in inconsistent datasets using prediction (Figure 3-6). As a way to moderate the trade-off, it is possible to increase spatio-temporal coverage through filling-in while eliminating the inherent bias using filtering. Therefore, 'filling-in' and 'filtering' are not mutually exclusive and can be combined by the following procedure: (a) filling-in before

filtering or (b) filtering before filling-in. This joint methodology would be essential due to an insufficient amount of data available on the most endangered species or minor taxa to cut off some parts of them. It is necessary to test the synergistic effects of the combination in terms of the trade-off between data volume and quality.

## 3.5. Conclusion

Due to the lack of consistent surveys across their entire habitats, most invertebrates and rare species are underrepresented in international conservation efforts. The small size of available data, predominantly consisting of presence-only data collected through different survey efforts, makes it challenging to estimate temporal changes in population size. Previous studies have primarily used filtering techniques to retain only a subset of reliable data. In contrast, this study proposes and tests a method that fills in predictions using machine learning at points of temporal inconsistency, enabling consistent temporal comparisons across the entire species habitat (= filling-in strategy). The results of validity tests showed that the filling-in strategy was independent of inherent temporal variations in the data, while the traditionally used methods (RA and ACC) were not. Additionally, when there were differences in the methods of data extraction for pseudo-absence datapoints, variations in variable selection methods, and the random inclusion of missing or false information in the presence data, the range of estimates from the filling-in strategy for population trends did not misrepresent IUCN conservation status to a significant extent. These findings indicate that the filling-in strategy, despite having lower quality but cost-effective data, exhibits resistance to temporal variations in data richness across the entire habitat range of a species and can produce theoretically valid predictions. Therefore, this study suggests that the filling-in strategy is a promising approach that can include a greater number of taxonomic groups in conservation planning.

# Bibliography

Aceves-Bueno, E., Adeleye, A. S., Feraud, M., Huang, Y., Tao, M., Yang, Y., & Anderson, S. E. (2017). The accuracy of citizen science data: a quantitative review. Bulletin of the Ecological Society of America, 98(4), 278-290.

Alyokhin, A., & Sewell, G. (2004). Changes in a lady beetle community following the establishment of three alien species. Biological Invasions, 6(4), 463-471.

Amano, T., Lamming, J. D., & Sutherland, W. J. (2016). Spatial gaps in global biodiversity information and the role of citizen science. Bioscience, 66(5), 393-400.

Bahlai, C. A., Colunga-Garcia, M., Gage, S. H., & Landis, D. A. (2015). The role of exotic ladybeetles in the decline of native ladybeetle populations: evidence from long-term monitoring. Biological Invasions, 17(4), 1005-1024.

Bayraktarov, E., Ehmke, G., O'Connor, J., Burns, E. L., Nguyen, H. A., McRae, L., ... & Lindenmayer, D. B. (2019). Do big unstructured biodiversity data mean more knowledge?. Frontiers in Ecology and Evolution, 239.

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., ... Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. Biological Conservation, 173, 144–154.

Boersch-Supan, P. H., Trask, A. E., & Baillie, S. R. (2019). Robustness of simple avian population trend models for semi-structured citizen science data is species-dependent. Biological Conservation, 240, 108286.

Boersch-Supan, P. H., Trask, A. E., & Baillie, S. R. (2019). Robustness of simple avian population trend models for semi-structured citizen science data is species-dependent. Biological Conservation, 240, 108286.

Chandler, M., See, L., Copas, K., Bonde, A. M., López, B. C., Danielsen, F., ... & Turak, E. (2017). Contribution of citizen science towards international biodiversity monitoring. Biological conservation, 213, 280-294.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37–46.

Colunga-Garcia, M., & Gage, S. H. (1998). Arrival, establishment, and habitat use of the multicolored Asian lady beetle (Coleoptera: Coccinellidae) in a Michigan landscape. Environmental Entomology, 27(6), 1574-1580.

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

Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen science as an ecological research tool: challenges and benefits. Annual review of ecology, evolution, and systematics, 41, 149-172.

Elliott, N., Kieckhefer, R., & Kauffman, W. (1996). Effects of an invading coccinellid on native coccinellids in an agricultural landscape. Oecologia, 105, 537-544.

Evans EW (2000) Morphology of invasion: body size patterns associated with establishment of Coccinella septempunctata (Coleoptera: Coccinellidae) in western North America.

Eur J Entomol 97:469–474

Evans EW (2004) Habitat displacement of North American ladybirds by an introduced species. Ecology 85:637–647

Evans EW, Soares A, Yasuda H (2011) Invasions by ladybugs, ladybirds, and other predatory beetles. Biocontrol 56:597–611

Fink, D., Auer, T., Johnston, A., Ruiz-Gutierrez, V., Hochachka, W. M., & Kelling, S. (2020). Modeling avian full annual cycle distribution and population trends with citizen science data. Ecological Applications, 30(3), e02056.

Fisher, A., Saniee, K., Van der Heide, C., Griffiths, J., Meade, D., & Villablanca, F. (2018). Climatic niche model for overwintering monarch butterflies in a topographically complex region of California. Insects, 9(4), 167.

Gardiner, M. M., Allee, L. L., Brown, P. M., Losey, J. E., Roy, H. E., & Smyth, R. R. (2012). Lessons from lady beetles: accuracy of monitoring data from US and UK citizen-science programs. Frontiers in Ecology and the Environment, 10(9), 471-476.

Gaul, W., Sadykova, D., White, H. J., León-Sánchez, L., Caplat, P., Emmerson, M. C., & Yearsley, J. M. (2022). Modelling the distribution of rare invertebrates by correcting class imbalance and spatial bias. Diversity and Distributions, 28(10), 2171-2186.

Harmon, J. P., Stephens, E., & Losey, J. (2007). The decline of native coccinellids (Coleoptera: Coccinellidae) in the United States and Canada. Beetle conservation, 85-94.

Hentley, W. T., Vanbergen, A. J., Beckerman, A. P., Brien, M. N., Hails, R. S., Jones, T. H., & Johnson, S. N. (2016). Antagonistic interactions between an invasive alien and a native coccinellid species may

promote coexistence. Journal of Animal Ecology, 85(4), 1087-1097.

Hertzog, L. R., Frank, C., Klimek, S., Röder, N., Böhner, H. G., & Kamp, J. (2021). Model-based integration of citizen science data from disparate sources increases the precision of bird population trends. Diversity and Distributions, 27(6), 1106-1119.

Hesler, L. S., & Kieckhefer, R. W. (2008). Status of exotic and previously common native coccinellids (Coleoptera) in South Dakota landscapes. Journal of the Kansas Entomological Society, 81(1), 29-49.

Hesler, L. S., Catangui, M. A., Losey, J. E., Helbig, J. B., & Mesman, A. (2009). Recent records of Adalia bipunctata (L.), Coccinella transversoguttata richardsoni Brown, and Coccinella novemnotata Herbst (Coleoptera: Coccinellidae) from South Dakota and Nebraska. The Coleopterists Bulletin, 63(4), 475-484.

Hesler, L. S., Kieckhefer, R. W., & Catangui, M. A. (2004). Surveys and field observations of *Harmonia axyridis* and other Coccinellidae (Coleoptera) in eastern and central South Dakota. Transactions of the American Entomological Society, 113-133.

Hickling, R., Roy, D. B., Hill, J. K., Fox, R., & Thomas, C. D. (2006). The distributions of a wide range of taxonomic groups are expanding polewards. Global Change Biology, 12, 450–455.

Honěk, A. (1985). Habitat preferences of aphidophagous coccinellids [Coleoptera]. Entomophaga, 30, 253-264.

Horns, J. J., Adler, F. R., & Şekercioğlu, Ç. H. (2018). Using opportunistic citizen science data to estimate avian population trends. Biological conservation, 221, 151-159.

Isaac, N. J., & Pocock, M. J. (2015). Bias and information in biological

records. Biological Journal of the Linnean Society, 115(3), 522-531.

Isaac, N. J., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., ... & O'Hara, R. B. (2020). Data integration for large-scale models of species distributions. Trends in ecology & evolution, 35(1), 56-67.

Jason, L. A. (2006). Benefits and challenges of generating community participation. Professional Psychology: Research and Practice, 37(2), 132.

Jeffries, D. L., Chapman, J., Roy, H. E., Humphries, S., Harrington, R., Brown, P. M., & Handley, L. J. L. (2013). Characteristics and drivers of high-altitude ladybird flight: insights from vertical-looking entomological radar. PloS one, 8(12), e82278.

Kallimanis, A. S., Panitsa, M., & Dimopoulos, P. (2017). Quality of non-expert citizen science data collected for habitat type conservation status assessment in Natura 2000 protected areas. Scientific reports, 7(1), 8873.

Kamp, J., Oppel, S., Heldbjerg, H., Nyegaard, T., & Donald, P. F. (2016). Unstructured citizen science data fail to detect long-term population declines of common birds in Denmark. Diversity and Distributions, 22(10), 1024-1035.

Kelling, S., Fink, D., La Sorte, F. A., Johnston, A., Bruns, N. E., & Hochachka, W. M. (2015). Taking a 'Big Data'approach to data quality in a citizen science project. Ambio, 44, 601-611.

Knape, J., Coulson, S. J., van der Wal, R., & Arlt, D. (2022). Temporal trends in opportunistic citizen science reports across multiple taxa. Ambio, 1-16.

LeCroy, K. A., Savoy-Burke, G., Carr, D. E., Delaney, D. A., & Roulston,

T. A. H. (2020). Decline of six native mason bee species following the arrival of an exotic congener. Scientific reports, 10(1), 18745.

Losey, J. E., Perlman, J. E., & Hoebeke, E. R. (2007). Citizen scientist rediscovers rare nine-spotted lady beetle, Coccinella novemnotata, in eastern North America. Journal of Insect Conservation, 11, 415-417.

Losey, J., Allee, L., & Smyth, R. (2012). The Lost Ladybug Project: Citizen spotting surpasses scientist's surveys. American Entomologist, 58(1), 22-24.

Lukyanenko, R., Parsons, J., & Wiersma, Y. F. (2016). Emerging problems of data quality in citizen science. Conservation Biology, 30(3), 447-449.

Martínez-Minaya, J., Cameletti, M., Conesa, D., & Pennino, M. G. (2018). Species distribution modeling: a statistical review with focus in spatio-temporal issues. Stochastic environmental research and risk assessment, 32, 3227-3244.

Martino, S., Pace, D. S., Moro, S., Casoli, E., Ventura, D., Frachea, A., ... & Jona Lasinio, G. (2021). Integration of presence-only data from several sources: a case study on dolphins' spatial distribution. Ecography, 44(10), 1533-1543.

McCorquodale, B. (1998). Adventive lady beetles (Coleoptera: Coccinellidae) in eastern Nova Scotia, Canada. Entomological news, 109(1), 15-20.

Mukwevho, V. O., Pryke, J. S., & Roets, F. (2017). Habitat preferences of the invasive harlequin ladybeetle *Harmonia axyridis* (Coleoptera: Coccinellidae) in the Western Cape Province, South Africa. African Entomology, 25(1), 86-97.

Newson, S. E., Evans, H. E., & Gillings, S. (2015). A novel citizen

science approach for large-scale standardised monitoring of bat activity and distribution, evaluated in eastern England. Biological Conservation, 191, 38-49.

Outhwaite, C. L., Gregory, R. D., Chandler, R. E., Collen, B., & Isaac, N. J. (2020). Complex long-term biodiversity change among invertebrates, bryophytes and lichens. Nature ecology & evolution, 4(3), 384-392.

Outhwaite, C. L., Powney, G. D., August, T. A., Chandler, R. E., Rorke, S., Pescott, O. L., ... & Isaac, N. J. (2019). Annual estimates of occupancy for bryophytes, lichens and invertebrates in the UK, 1970–2015. Scientific data, 6(1), 259.

Pocock, M. J., Roy, H. E., August, T., Kuria, A., Barasa, F., Bett, J., ... & Trevelyan, R. (2019). Developing the global potential of citizen science: Assessing opportunities that benefit people, society and the environment in East Africa. Journal of applied ecology, 56(2), 274-281.

Pocock, M. J., Tweddle, J. C., Savage, J., Robinson, L. D., & Roy, H. E. (2017). The diversity and evolution of ecological and environmental citizen science. PloS one, 12(4), e0172579.

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., ... & McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology and Evolution, 5(5), 397-406.

Robinson, O. J., Ruiz-Gutierrez, V., Reynolds, M. D., Golet, G. H., Strimas-Mackey, M., & Fink, D. (2020). Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. Diversity and Distributions, 26(8), 976-986.

Roy, H. E., Adriaens, T., Isaac, N. J. B., Kenis, M., Martin, G. S., Brown,

P. M. J., et al. (2012). Invasive alien predator causes rapid declines of native European ladybirds. Diversity and Distributions, 18, 717–725.

Schultz, C. B., Brown, L. M., Pelton, E. & Crone, E. E (2017). Citizen science monitoring demonstrates dramatic declines of monarch butterflies in western north america. Biol. Cons. 214, 343–346.

Serra-Diaz, J. M., Enquist, B. J., Maitner, B., Merow, C., & Svenning, J. C. (2017). Big data of tree species distributions: how big and how good?. Forest Ecosystems, 4, 1-12.

Spear, D. M., Pauly, G. B., & Kaiser, K. (2017). Citizen science as a tool for augmenting museum collection data from urban areas. Frontiers in Ecology and Evolution, 86.

Specht, H., & Lewandowski, E. (2018). Biased assumptions and oversimplifications in evaluations of citizen science data quality. Bulletin of the Ecological Society of America, 99(2), 251-256.

Steen, V. A., Elphick, C. S., & Tingley, M. W. (2019). An evaluation of stringent filtering to improve species distribution models from citizen science data. Diversity and Distributions, 25(12), 1857-1869.

Strayer, D. L., Eviner, V. T., Jeschke, J. M., & Pace, M. L. (2006). Understanding the long-term effects of species invasions. Trends in ecology & evolution, 21(11), 645-651.

Svancara, L. K., Abatzoglou, J. T., & Waterbury, B. (2019). Modeling current and future potential distributions of milkweeds and the monarch butterfly in Idaho. Frontiers in Ecology and Evolution, 7, 168.

Tikhonov, G., Abrego, N., Dunson, D., & Ovaskainen, O. (2017). Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. Methods in Ecology and Evolution, 8(4), 443-452.

Tingley, M. W., & Beissinger, S. R. (2009). Detecting range shifts from historical species occurrences: new perspectives on old data. Trends in ecology & evolution, 24(11), 625-633.

Tumminello, G., Ugine, T. A., & Losey, J. E. (2015). Intraguild interactions of native and introduced coccinellids: the decline of a flagship species. Environmental entomology, 44(1), 64-72.

Turnipseed, R. K., Ugine, T. A., & Losey, J. E. (2014). Effect of prey limitation on competitive interactions between a native lady beetle, Coccinella novemnotata, and an invasive lady beetle, Coccinella septempunctata (Coleoptera: Coccinellidae). Environmental Entomology, 43(4), 969-976.

Turnock, W. J., Wise, I. L., & Matheson, F. O. (2003). Abundance of some native coccinellines (Coleoptera: Coccinellidae) before and after the appearance of Coccinella septempunctata1. The Canadian Entomologist, 135(3), 391-404.

Van Eupen, C., Maes, D., Herremans, M., Swinnen, K. R., Somers, B., & Luca, S. (2021). The impact of data quality filtering of opportunistic citizen science data on species distribution model performance. Ecological Modelling, 444, 109453.

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

Walker, J., & Taylor, P. (2017). Using eBird data to model population change of migratory bird species. Avian Conservation and Ecology, 12(1).

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan,

A., & NCEAS Predicting Species Distributions Working Group. (2008). Effects of sample size on the performance of species distribution models. Diversity and distributions, 14(5), 763-773.

Woltz, J. M., & Landis, D. A. (2013). Coccinellid immigration to infested host patches influences suppression of Aphis glycines in soybean. Biological Control, 64(3), 330-337.