

# **Visualizing Strengths and Limitations of Semi-Structured Versus Structured Approaches on eBird**

**Rodrigo Rolim Fialho Nunes de Freitas**

Thesis to obtain the Master of Science Degree in

## **Information Systems and Computer Engineering**

Supervisors: Prof. Ana Gualdina Almeida Matos  
Dr. Ana Patrícia Subtil da Graça Freitas Garcia

### **Examination Committee**

Chairperson: Prof. Pedro Tiago Gonçalves Monteiro  
Supervisor: Prof. Ana Gualdina Almeida Matos  
Member of the Committee: Prof. Carlos António Marques Pereira Godinho

**June 2022**



# Acknowledgments

Firstly I would like to thank my parents, sister, grandparents, and all my close family who have unconditionally supported and openly encouraged me throughout this memorable phase of my life.

I also cannot thank enough my two supervisors, Professor Ana Matos and Professor Ana Subtil, for the opportunity I was given to work with them throughout the last year on this thesis. This would not have been possible without their enviable willingness and readiness to guide me through in the work itself and in the writing of the thesis. I would like to thank Pedro Cardia as well, who has accompanied most of the work by giving invaluable insight and key feedback with his expertise in the field of ornithology and in the functioning of the platform in which this work is based, eBird. This MSc dissertation was also supported by *Instituto de Telecomunicações*.

Also, a big thank you to all the friends who have accompanied and to the friendships made along the way, with whom I was lucky enough to share unforgettable experiences throughout these years.

Last but not least, thank you to the great institutions that are *Instituto Superior Técnico* and *Universidade de Lisboa* for providing the conditions and making this possible to achieve.

Thank you to all.





# Abstract

Citizen science is gaining more participants by the day. Consequently, its broader sphere of influence has been reflected throughout the scientific community in recent years, comprising a wide variety of fields - with results not only on research regarding biological diversity and climate change, but also efforts with respect to policy-making for conservation purposes.

One of the most notable platforms created for this purpose, eBird, is our case study for this thesis. Thus, we present and discuss topics regarding the platform, and inherent challenges that come with it, focusing in the case of biased reporting. With this in mind, we explore the state of the art with the coverage of studies that try to model the species' range and abundance distributions using citizen science data, revealing the true potential of these data.

To this end we study methods for exhibiting biases that are present in the observation reporting data that is collected and made available via the eBird platform. We are in particular interested in how different levels of protocol structure in the observation data collection phase can lead to different forms of biases, and propose methods for identifying and analysing them. Hence, our approach facilitates the comparison of structured versus semi-structured approaches to data collection in eBird, and includes: mapping and graphing of differing metrics calculated as well as available data depicting each checklist's search effort; and a visual interactive tool, named Shiny eBird, developed using the programming language R's Shiny framework that allows to display these computed metrics interactively.

## Keywords

Citizen Science; Birdwatching; Sampling Effort; eBird Platform; Data Visualization; Interactive Tool.



# Resumo

A ciência cidadã tem ganho mais participantes por cada dia que passa. Consequentemente, tem ganho cada vez mais relevância nos últimos anos num cada vez maior contributo para toda a comunidade científica, abrangendo uma ampla variedade de campos - demonstrando resultados tanto em investigação acerca de diversidade biológica e mudanças climáticas, mas também esforços respeitantes à criação de políticas para fins de conservação animal.

Uma das plataformas mais notáveis usadas para este fim, eBird, é o nosso caso de estudo. Assim, apresentamos e discutimos temas relativos à plataforma e desafios inerentes que a acompanham, com o foco no enviesamento dos dados recolhidos, explorando o estado da arte e referindo estudos focando-se na modelação de espécies usando estes dados.

Envergamos neste trabalho, assim, no estudo métodos que ajudam a expôr enviesamentos nos dados de listas que estão disponibilizadas na plataforma eBird. Interessa-nos como diferentes estruturas de protocolos de observação de aves durante a fase de recolha de dados podem levar a vários tipos de enviesamento, e propomos métodos para os identificar e analisar. Assim, a nossa abordagem faz a comparação entre duas estruturas distintas, estruturadas e semi-estruturadas, para recolha de dados no eBird, que inclui: o mapeamento e a criação de gráficos de diferentes métricas escolhidas por nós, bem como outros dados do eBird que descrevem o esforço de procura de cada lista; também é proposta uma ferramenta visual interativa, chamada Shiny eBird, desenvolvida com a framework Shiny através da linguagem R, que permite exibir essas mesmas métricas de forma interativa.

## Palavras Chave

Ciência Cidadã; Observação de Aves; Esforço de Amostragem; Plataforma eBird; Visualização de Dados; Ferramenta Interativa.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Goals . . . . .	4
1.2	Contributions . . . . .	4
1.3	Document Structure . . . . .	5
<b>2</b>	<b>Background</b>	<b>7</b>
2.1	How eBird Works . . . . .	9
2.1.1	eBird Data Submission . . . . .	9
2.1.2	eBird Data Validation . . . . .	10
2.1.3	eBird Products and Services . . . . .	11
2.2	Applications of eBird . . . . .	12
2.3	Challenges in Citizen Science . . . . .	14
2.3.1	Types of Bias . . . . .	14
2.3.2	Solutions to Address Bias . . . . .	15
2.4	eBird's Survey Structure . . . . .	15
2.4.1	Semi-Structured and Structured Data in eBird . . . . .	17
<b>3</b>	<b>Related Work</b>	<b>19</b>
3.1	Approaches to Address Bias in Citizen Science . . . . .	21
3.1.1	Quantifying Bias using SampBias . . . . .	21
3.1.2	Species Distribution Model (SDM) . . . . .	21
3.1.3	Spatio-Temporal Exploratory Model (STEM) . . . . .	22
3.1.4	Uses of STEM . . . . .	23
3.1.5	Spatial Interpolation Methods . . . . .	25
3.1.6	Point Process Model (PPM) . . . . .	26
3.2	Data Visualization . . . . .	26
<b>4</b>	<b>Approach</b>	<b>29</b>
4.1	Proposed Solution and Goals . . . . .	31
4.2	Data Considered . . . . .	31

4.3	Metrics Considered . . . . .	32
4.3.1	Metrics Applied to a Single Species . . . . .	32
4.3.2	For Each Grid Cell . . . . .	35
4.3.3	Mapping representation . . . . .	38
4.3.4	Cohen's Kappa . . . . .	42
4.3.5	Observable Agreement . . . . .	44
4.3.6	Percentage of Species Reported . . . . .	45
4.3.7	Species Accumulation Curves . . . . .	46
4.3.8	Species Richness . . . . .	47
<b>5</b>	<b>The Tool Proposed</b>	<b>51</b>
5.1	R's Shiny Framework . . . . .	53
5.2	Development Workflow . . . . .	54
5.2.1	Acquiring the Data and Analysis . . . . .	54
5.2.2	Pre-processing of the data . . . . .	54
5.2.3	Grid and Coordinate Reference System Used . . . . .	55
5.2.4	Database System . . . . .	56
5.2.5	Application's User Interface . . . . .	57
5.2.6	Deployment of the Tool . . . . .	58
<b>6</b>	<b>Conclusion</b>	<b>59</b>
6.1	Concluding Remarks . . . . .	61
6.2	Limitations and Future Work . . . . .	61
	<b>Bibliography</b>	<b>63</b>

# List of Figures

2.1	Image depicting the first step in submitting a checklist - the type of protocol, along with information used to subsequently measure the observer's effort are requested. Taken from eBird's webpage [9]. . . . .	10
2.2	Circle charts illustrating the end-uses of 1100 eBird data requests. Image obtained from Sullivan et al. (2014) [6]. . . . .	12
2.3	Illustration summarizing the main characteristics of different types of citizen science project structures [26]. The size of the dots represent to which extent each structure type meets the listed sets of criteria. Obtained from Kelling et al. (2019) [26]. . . . .	16
3.1	Comparison between STEM and AdaSTEM <i>stixels</i> . <i>Left</i> : STEM's fixed-size <i>stixels</i> are shown, not offering total coverage. <i>Right</i> : AdaSTEM's <i>stixels</i> . Worth noting that the smaller <i>stixels</i> correspond to areas with higher observation density. Image (a) obtained from Fink et al. (2010) [20] and image (b) from Fink et al. (2014) [43]. . . . .	22
3.2	<i>Left</i> : using AdaSTEM, colored areas indicate bias where standardized residuals are more than twice as large as their associated standard errors. <i>Right</i> : AdaSTEM distribution estimates in the American continent for Barn Swallow during the breeding season, and winter, using eBird data. The bar on the right indicates the relative probability of occurrence. Obtained from Fink et al. (2014) [43] and Fink et al. (2013) [44], respectively. . . . .	23
3.3	Year-round predictor relative importance of each land cover class for the Wood Thrush population, calculated weekly for each base model (fit within each <i>stixel</i> ). Classes whose value is below zero are not used in the core population. Image taken from Fink et al. (2020) [46]. . . . .	24
3.4	Workflow developed for mapping the distribution of a species, using the hummingbird species Glittering Starfrontlet as an example. Obtained from Palacio et al. (2020) [47]. . .	25
4.1	Distribution of the semi-structured checklists' duration where the species <i>Turdus merula</i> was reported. . . . .	33

4.2	Distribution of the semi-structured checklists' distance where the species <i>Turdus merula</i> was reported. . . . .	33
4.3	Distribution of the semi-structured checklists' starting time where <i>Turdus merula</i> and <i>Tyto alba</i> were reported. . . . .	34
4.4	Top reported in cell 10 semi-structured, atlas timeframe. . . . .	35
4.5	Checklists reported per weekday for the grid cell 10. . . . .	35
4.6	Distribution of the checklist duration in a given grid cell per <b>duration</b> bin. . . . .	36
4.7	Average number of species submitted in a given grid cell per <b>duration</b> bin. . . . .	36
4.8	Distribution of distance bin the number of checklists submitted in a given grid cell per <b>distance</b> bin. . . . .	37
4.9	Average number of species reported per <b>distance</b> bin. . . . .	37
4.10	Maps depicting the total number of species reported by semi-structured and structured checklists, during the atlas timeframe. . . . .	38
4.11	Map showing the difference between the number of species reported by semi-structured and structured checklists, during the atlas timeframe. . . . .	39
4.12	Map depicting where species was observed between an arbitrary date. . . . .	40
4.13	Map depicting where Eurasian Blackbird have been observed during the atlas timeframe. . . . .	40
4.14	Maps depicting the logged total of semi-structured checklists in (a), and the its percentage of completeness in map (b). . . . .	41
4.15	Kappa coeficient for all species reported between 15th of March and 15th of July of the years 2015 to 2021. . . . .	43
4.16	Observable agreement between structured and semi-structured observations, expressed as a percentage. . . . .	44
4.17	Comparison between structured and semi-structured percentages of species found over the atlas timeframe. . . . .	45
4.18	SAC plot comparing structured (in orange) and semi-structured checklists (in green). . . . .	46
4.19	Comparison structured and semi-structured <i>Shannon Indexes</i> , over the atlas timeframe. . . . .	48
4.20	Comparison structured and semi-structured <i>Shannon Equitability</i> , over the atlas timeframe. . . . .	49
4.21	Breeding Bird Atlas' coverage and global search effort, following the same grid layout. Image kindly provided by Pedro Cardia. . . . .	50
5.1	Simplified shiny app layout. . . . .	53
5.2	Diagram depicting the development workflow. . . . .	54
5.3	Comparison between the grids considered. Images generated using QGIS. . . . .	56
5.4	User-Shiny server architecture. . . . .	58



# List of Tables

- 2.1 Comparison between the structured approach used for the P65 protocol and the semi-structured approach . . . . . 17
- 4.1 Table depicting confusion matrices used for the calculation of the value of Kappa, depicting the theoretical matrix, with the variables relations. The variables  $a$ ,  $b$ ,  $c$ ,  $d$  are the names given in this work's code, and are somewhat easier to visualize. . . . . 42
- 4.2 Interpretation of the Cohen's Kappa coefficient. . . . . 43
- 4.3 Table depicting confusion matrix with the values measured for the grid cell with ID 266 throughout the atlas timeframe, which has presented the highest value of  $\kappa$  (0.512). . . . 44



# Acronyms

<b>CRS</b>	Coordinate Reference System
<b>DB</b>	Database
<b>GIS</b>	Geographic Information Systems
<b>IDE</b>	Integrated Development Environment
<b>UI</b>	User Interface
<b>STEM</b>	Spatio-Temporal Exploratory Model
<b>SAC</b>	Species Accumulation Curves
<b>SDM</b>	Species Distribution Model



# 1

## Introduction

### Contents

1.1	Goals . . . . .	4
1.2	Contributions . . . . .	4
1.3	Document Structure . . . . .	5



Humans have been collecting data from nature for centuries. In the late 19<sup>th</sup> century, with the professionalization of Science, new approaches were taken in order to improve the efficiency, accuracy and precision in the process of recording ecologic information, in order to study the dynamics of spatial and temporal patterns of the distribution and abundance of organisms. The wider community, initially set aside from these activities, has ever since been steadily gaining importance in the processes of data collection, analysis, and even interpretation - in what we could consider as being a democratization of knowledge [1]. These came to improve both the scientific and educational outcomes, specially when regarding broader ecological questions at unfeasible scales through professional science alone, or simply projects in which there is no professional scientists willing to do the job on their own [2]. Although there is no clear definition of what a citizen scientists is, a fair description would be a member of the general public that voluntarily engages in scientific research, contributing to the expansion of our knowledge.

In recent years, citizen science has been seeing a steady boost on the number of new projects and participants of its projects, namely on those regarding biodiversity monitoring or conservation [3] [4]. With more than 800 projects worldwide [5], these are allowing for ecological research at unprecedented spatial and temporal scales [6]. One of the main reasons for this boom in participation being the advancements in information technology [7] from the last couple of decades - granting us the ability to access and share data in a both centralized and standardized manner, and recently to be able to do this from virtually anywhere with an internet connection, via mobile apps. Consequently, new paradigms and methods were created to the way citizen science is approached, such as the ability to review observations and provide immediate feedback to keep users engaged with the project [8].

The work focuses in a particular citizen science project, eBird [9], a growing birdwatching online network launched in 2002, run by conservationists and information scientists at Cornell University in the US, in the member-supported Cornell Lab of Ornithology unit<sup>1</sup>. It is aimed to those who wish to create and keep a record of this form of wildlife observation, while actively contributing for the enrichment of information about abundance and distribution of the world's bird populations as a reliable source that can further be used for research. eBird has since been a case study for researchers in a wide variety of fields, such as computer science, statistics, ecology and biology [6] [10].

In this thesis we study methods for exhibiting biases that are present in the observation reporting data that is collected and made available via the eBird platform. We are in particular interested in how different levels of protocol structure in the observation data collection phase can lead to different forms of biases, and propose methods for identifying and analysing them.

---

<sup>1</sup> <https://www.birds.cornell.edu/home/>

## 1.1 Goals

The work that is described in this document seek to achieve and contribute to the following goals:

1. Present a contrast between the different approaches of bird reporting on the eBird platform considered in our problem.
2. Help identify areas that seem misrepresented in the citizen science program, resorting to mapping and plotting of certain attributes of the observation process.
3. Motivate future work from conclusions taken through our analysis of the data.
4. Creating a visual and interactive tool capable of facilitating the points above.

## 1.2 Contributions

With the intent of helping the study of biases that are entailed by different data collection processes, we consider eBird's freely available dataset to describe the platform's state in Mainland Portugal. The dataset, which can be subdivided into two categories depending on how strict the procedure conducted during the observations is - structured and semi-structured - is used to contrast between both approaches, by resorting to mapping and graphing of differing metrics calculated as well as available data depicting each checklist's search effort. These include maps describing the total of species reported, its degree of agreement, species accumulation curves, among other metrics that helped describing the regular *eBirder's* semi-structured activity throughout the territory, as well as the outcome of the structured approach.

Additionally, considering all the results gathered, we present a visual interactive tool using the programming language R's Shiny framework that allows to displays these computed metrics interactively in a grid laid over the territory of Portugal, following the same spatial subdivisions as those used by the structured approach. This tool, named **Shiny eBird**, was developed to allow the user to explore the data computed in the contrasting of both approaches, but also letting the user examine semi-structured data collected by regular volunteers in eBird in other timeframes, ultimately with the goal of contributing for the of combat bias in this platform.



## 1.3 Document Structure

This document is structured in the following manner:

- Chapter 1 provides an introductory overview of citizen science, the object of study and the motivation for this work, as well as its goals.
- Chapter 2 the main context around the work is provided, illustrating some of the essential citizen science concepts, challenges, and hence the work's underlying motivation.
- Chapter 3 articles are shared that have been published regarding this topic, including methodologies together with how biased data is accounted for for estimating distributions from citizen science programs, ending on the topic of visualization.
- Chapter 4 where the solution is outlined, accompanied with different metrics that are expected to help achieve the goals in mind.
- Chapter 5 where the proposed tool's development workflow and details about the solution are described in detail.
- Chapter 6 ends with the main conclusions taken from the work, and future work relatively to both features and improvements to the tool, and possible different approaches to other data available.



# 2

## Background

### Contents

---

<b>2.1 How eBird Works</b>	<b>9</b>
2.1.1 eBird Data Submission	9
2.1.2 eBird Data Validation	10
2.1.3 eBird Products and Services	11
<b>2.2 Applications of eBird</b>	<b>12</b>
<b>2.3 Challenges in Citizen Science</b>	<b>14</b>
2.3.1 Types of Bias	14
2.3.2 Solutions to Address Bias	15
<b>2.4 eBird's Survey Structure</b>	<b>15</b>
2.4.1 Semi-Structured and Structured Data in eBird	17

---



This chapter aims to explain some of the main topics regarding citizen science in general, as well as some intrinsic to the platform at hand, eBird, starting off with its basic concepts.

## 2.1 How eBird Works

eBird is a massive collector of data partnered with hundred of organizations, thousands of regional experts, and hundreds of thousands of users who submit their lists of observations through predefined protocols, via virtual checklists [9]. The subsection below illustrates the process behind the submission of new data.

### 2.1.1 eBird Data Submission

To submit a checklist, participants follow a series of steps, starting by specifying the location where it took place. This is followed by indicating what survey protocol was being used, from a predefined selection. There are 4 main types of observation protocols: *incidental* if birding was not the primary purpose; otherwise, it can be *historical* if the date is known, but either the exact time of duration or distance traveled are unknown; and, when these are known it can be *stationary* or *travelling* whether if occurred at a fixed location or with a displacement of over 30 meters [9]. Allied to each type, more information is asked to describe the activity, and is required to answer as it is later used to compute the user's effort on its observations. Figure 1 depicts this first step. In the mobile application, some of this data can be automatically and effortlessly extracted [8] such as the trip distance or the time and place of the observations, with the user's permission.

Last but not least, the final step consists on filling the checklist itself. This is done by naming which species were observed out of a list of those that might be present in the specified location. If possible, the user should specify how many were encountered throughout the activity, otherwise simply fill the correspondent box with an "x" character. A relevant detail is the fact that users are asked to indicate whether they are reporting all the birds they were able to identify - and, if that's the case, it will be regarded as a **complete checklist**, which will add more value to the list, as we will discuss further. Also, there is the possibility to add notes detailing each sighting, and users are in fact encouraged to do so, along with media such as photographs, videos, or audio of birds' calls in order to enrich the checklist.

Even though eBird offers a predetermined list of probable species that an observer might encounter, by taking into account previously made observations, the platform allows to record a different one from the set presented, and hence it is not a closed system. This proves useful, for instance, in cases of sighting of non-native species, or migratory ones present in an unusual period.

To help users start developing their skills of birdwatching on eBird, guides<sup>1</sup> are made available on

---

<sup>1</sup> <https://support.ebird.org/>

The screenshot shows the eBird checklist submission form. At the top right, a legend indicates that an asterisk (\*) denotes a required field. The form includes the following sections:

- \* Observation Date:** A date selector showing 'Apr', '...', and '2021', with a calendar icon to the right.
- \* Observation Type:** A section with five radio button options:
  - Traveling:** Selected by default. Description: 'You traveled a specific distance — walking a trail, driving a refuge loop, field birding.' A 'More Info...' link is to the right.
  - Stationary:** Description: 'You stayed at a fixed location — watching from a window, hawkwatching, seawatching.' A 'More Info...' link is to the right.
  - Historical:** Description: 'Birding was your **primary purpose**, but you cannot estimate start time, duration, and distance; use Traveling or Stationary if you can estimate these.' A 'More Info...' link is to the right.
  - Incidental:** Description: 'Birding was not your **primary purpose** — noting a bird while driving or gardening.' A 'More Info...' link is to the right.
  - Other:** Includes a 'Choose...' dropdown menu.
- \* Start Time (24-hour):** Two input boxes for hours and minutes, with a 'Use 12-hour Clock' button.
- \* Duration:** Two input boxes for hours and minutes.
- \* Distance:** An input box followed by a 'miles' dropdown menu.
- \* Party Size:** An input box with the text 'Enter the total number of people in your birding party'.
- Checklist Comments:** A large text area at the bottom.

**Figure 2.1:** Image depicting the first step in submitting a checklist - the type of protocol, along with information used to subsequently measure the observer's effort are requested. Taken from eBird's webpage [9].

the best practices to virtually anything related to the platform, from how to document sensitive species to how to best prepare and upload media. For instance, it is not recommended for two users to be birdwatching at the same location and time while filling in their separate checklists - eBird offers the ability to share the checklist - this is to prevent duplicate observations that can potentially introduce bias in the system.

### 2.1.2 eBird Data Validation

As in many citizen science projects, eBird seeks to increase its data volume through the recruitment of new participants and engagement at a global scale. As of June 2020, eBird had collected 705 million valid observations eligible for research, in its dataset on GBIF [11], an international organisation that centralises online free and open access biodiversity data. This amount of data raises the question on what is the process behind the validation of a checklist. There is a need to improve data quality both during and after submission, as anyone regardless of their experience can submit one.

A submission to eBird is subjected to automatic filtering and in some cases only validated by experts. For each and every species that is reported in a checklist ultimately goes through a series of data quality filters, that evaluate the specie's presence and/or count. The filters are set according to previously collected data along with reliable sources of knowledge on bird distributions, and are managed by hundreds

of regional experts. They are also in charge of reviewing the records which remain flagged after contacting the user for confirmation or more details [12]. This method creates an active learning feedback loop and bigger proximity between users and the system that increases the quality of the data [13].

### 2.1.3 eBird Products and Services

It is worth noting that eBird offers a handful of services to facilitate the analysis of the gathered data:

- **eBird Basic Dataset** EBD<sup>2</sup> is the dataset containing all the observations being made to date. Anyone who wishes to can request these data, for non-commercial use, in exchange for answering a small survey to indicate for which purpose it is being used. [14]
- **Status and Trends** eBird's Status and Trends uses statistical models to provide analytical insight and predictions, by making use of the data collected on eBird while considering environmental data from NASA for its models. This allows for new ways of viewing birds across continental scales including range, abundance maps and animations, regional charts and other stats comprising plenty of species. Some of the models used are explored in this report. As of now, eBird has only made available estimates relative to the western hemisphere, relative to the year 2018 [15]. It will serve as a source of inspiration for the work ahead.
- **R Packages** Such as **auk** [57] and **ebirdst** [58]. The former, **auk**, will be of much value for this work, as we will have to extract and process subsets of data from the EBD for analysis. [16]
- **Merlin** The mobile application Merlin<sup>3</sup> makes use of computer vision and machine learning advances to look for patterns that are shared between photos and sound recordings of birds. Its goal is to assist birdwatchers in naming observed species, hence lowering the barrier to bird identification, as it takes a lot of experience to know what species are expected at a given location and date.
- **Macaulay Library** This database<sup>4</sup> serves as the scientific archive of natural history audio, video, and photographs. Also powers Merlin's models.
- **eBird API** Offers a way to allow for dynamically fetching information from the eBird's database up until 30 days ago.

---

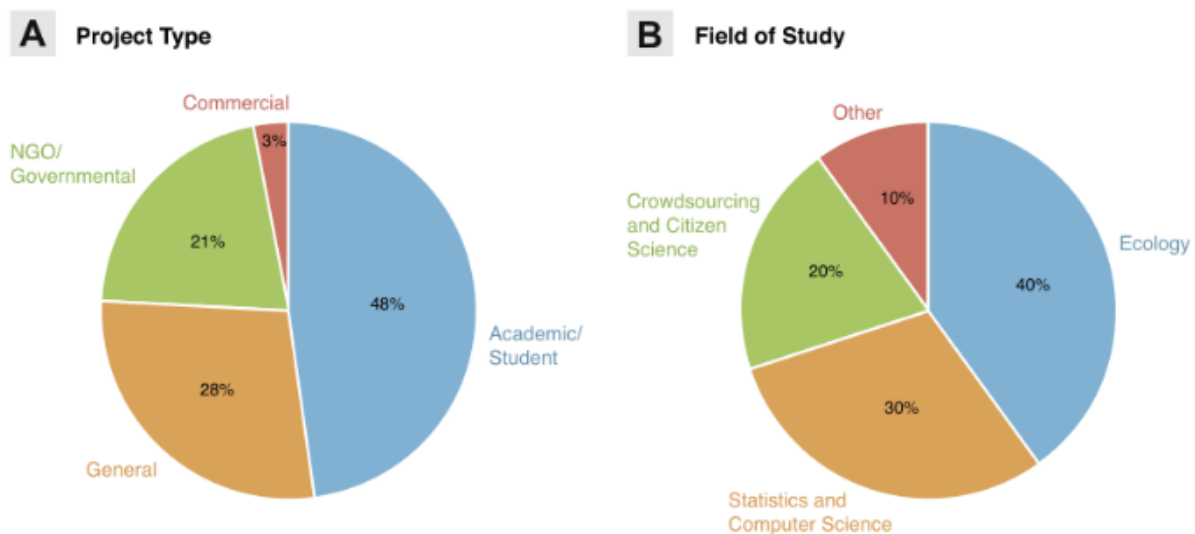
<sup>2</sup><https://ebird.org/data/download>

<sup>3</sup><https://merlin.allaboutbirds.org/>

<sup>4</sup><https://www.macaulaylibrary.org/>

## 2.2 Applications of eBird

Citizen science data has been having a relevant impact in the amount of information available for research purposes. In eBird, particularly, the vast number of different academic fields that have resorted to eBird data ultimately dictates the platform's success, and citizen science's overall [6]. Figure 2.2, below, depicts its diverse use, plotting the information taken from the survey asking how the data is going to be used once requesting access to eBird's data, from its website<sup>5</sup>.



**Figure 2.2:** Circle charts illustrating the end-uses of 1100 eBird data requests. Image obtained from Sullivan et al. (2014) [6].

eBird acts as a reliable case study for several scientific education initiatives and raises awareness among the general public [17]. eBird's collected data allows for more accurate estimates over wide spatial and temporal scales and has been making its contribution in studies and innovations across a wide span of fields [6], some of the most prominent being in computation and statistics, ecology, and in conservation policy. Below, we dive further into these:

### Computational and Statistical Advancements

Steps have had to be taken in platforms such as eBird in order to keep generating high-frequency and large-volume data streams, implying taking a Big Data approach [10]. With this, new opportunities and challenges are created - such as how to deal with noisy data - as well as new methods are demanded to process the vast amount of data, including its visualization and analysis [18].

Having this in mind, statistical models capable of taking in vast amounts of data and examine patterns involving multiple covariates across space and time are applied, which is the case of Species Distribution

<sup>5</sup><https://ebird.org/data/request>



Models (SDM) [19]. Besides these, other novel approaches have been created with the initial intent of interpreting eBird's data [20], that are considered later in Chapter 3.

Last but not least, machine learning models are also put into practice, as previously mentioned, to efficiently mine data and improve its quality at the submission phase [12], and in tools such as Merlin.

## **Avian Ecology**

As expected, knowledge gained from eBird related to bird's diversity and its environments have a direct application in ecology and biology. In particular, in phenology, an ecology branch aimed at studying periodic events in biological life cycles and how these play out with the seasonal and inter annual variations in climate and habitat - focuses on studying indicators of the specie's status in the wildlife, including potential threats. All begins by understanding patterns, monitoring and determining shifts [21] of distributions, abundance, and movements of individuals, which has been more accurately and dynamically attainable through advancements such as those mentioned in Chapter 3.

It is worth mentioning that eBird has proven to be of special value to measure as well as project medium to long-distance migration strategies [10] [6], through the analysis of sightings at large geographical scales. For instance, La Sorte et al. (2017) has used eBird data to model patterns of migrant species in the American continent to determine how they are currently associated with public protected areas and projected changes in climate and land-use [22].

## **Land-use and Conservation Policy**

With the world rapidly changing, resulting in a pressure build up on bird's habitats due to global threats such as climate change, invasive species, illegal killing or simply habitat loss from human activity [7], commitment from the authorities is increasingly required to make an effort so that potential harm is mitigated and conservation efforts are held. Thanks to more accuracy in the knowledge of spatial and temporal variation in bird occurrence over large areas, citizen science projects such as eBird have been considered as a reliable alternative to directly inform officials in policy-making of land-management and conservation planning [23].

A concrete example is a case study presented by Ruiz-Gutierrez et al. (2021) [24], about how eBird data was used by the US Fish and Wildlife Service as a source of information regarding wind energy development to define low-risk collision areas as part of the permitting process.

## 2.3 Challenges in Citizen Science

Due to citizen science's own nature, a distinguishable set of challenges are to be expected. For one, these projects often face resistance within the scientific community and decision makers [27]. This stems from concerns related to the **data veracity** and other issues such as noise accumulation [10] - which may in turn impact data quality, validity, and consistency. Other generic challenges may also be present, such as the need for specific programs or analyses to process data, depending on data variation and scope, obstacles to engage more people across less populated areas, legal impediments, lack of expertise and funding, or even barriers to participation - considering the range of different cultures and customs in larger programs [28].

With regards to eBird, being one of the largest citizen science projects, by far the biggest bird occurrence reporter on GBIF [11] and increasingly so even in data-poor regions [3], two of its main objectives of quantifying and controlling data quality issues come with a great responsibility. Taking into consideration the volume of data needed to create a reliable depiction of birds' distributions and abundance across the globe, numerous challenges have to be acknowledged and faced in order to ensure the best results. Due to its semi-structured nature, data collection done by non-professional users can regularly be erroneous, incomplete, and patchy [28] - meaning that the data collected may end up becoming **biased**, and consequently not be reliable to the point of being able to answer pressing questions. This brings us to the diverse categories of bias that are considered when dealing with citizen science projects such as eBird.

### 2.3.1 Types of Bias

Considering the wide range of non-professional users uploading their observations onto citizen science platforms, a uniform representation of large areas comprising a vast number of species can prove to be a hard task, causing an inherent bias [9]. Frequent biases that are associated with citizen science projects that may lead to **under** or **over-reporting**, are:

- **Spatial Bias** is one of the most common types of bias present where there is a tendency for people to choose certain locations. It has been showed to be prevalent on unstructured [29] and semi-structured programs such as eBird, particularly when it comes to areas with higher infrastructure and population density [30] - meaning of easier accessibility - closer to the observers' homes, or with higher biodiversity [25].
- **Temporal Bias** is, along with the former type, one of the most common sources of bias. It consists on the periodic and seasonal patterns noted by participants picking on particular dates and/or certain times of the day to carry out their observations [31], resulting in higher reporting on weekends, for instance [32].

- **Taxonomic Bias** refers to the preference - or the lack of it - by users to report specific species or subspecies during the observation process. Even though “chasing behaviour” of rare birds hasn’t been proved on eBird [33], it’s fair to consider that there is a tendency to select certain groups of birds among untrained observers [29]. Besides the sampling process, it has also been demonstrated how societal preferences correlate more steadily to taxonomic bias, in contrast to scientific research [34].

Besides these, two other challenges present in non-professional wildlife reporting programs, that may add up to the bias on the sampling process, particularly on eBird, are:

- **Class imbalance** refers to a problem with classification where the classes are not represented equally. In relation to eBird, this occurs when species harder to detect are not reported, resulting in a faulty record in case of a complete checklist, which reports non-detections [31].
- **Variation in detectability** as the name suggests, the unevenness of detectability can create flawed representation of birds real presence [25], as the observation process is heterogeneous, due to variation in time of the day, weather. It is also safe to consider the observer’s variability as a another cause for its variation, since the observer’s expertise in identifying birds and sampling effort weighs on the end result [6].

### 2.3.2 Solutions to Address Bias

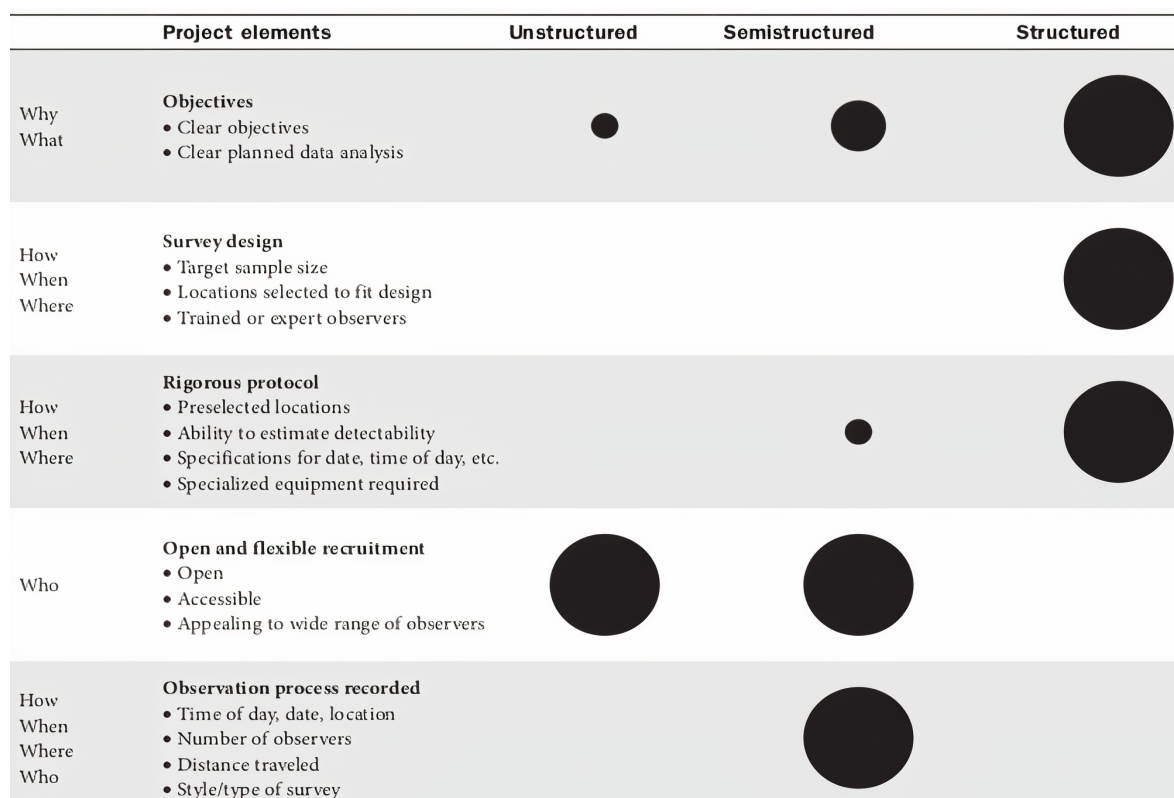
Various approaches have been suggested in order to minimize the unwanted effects of bias inherent to citizen science. From resorting to deep learning algorithms to account for spatial bias [35], introducing gamification features on the application to encourage more balanced coverage of the territory [37], to the use of local leaderboards, rankings, and suggestion systems [36] [29]. Besides these methods aimed at lowering the unevenness during the data collection process, bias can also be embraced and taken into consideration in the data analysis stage. To this end, some models are mentioned in Chapter 3, below.

## 2.4 eBird’s Survey Structure

Two characteristics named in the Section 2.1 differentiate eBird from other platforms. Firstly, the nature of the checklist structure itself allows the collection of data relative to the non-detection of species. This means that, if the user confirms to have recorded every species being found (*i.e.*, is reporting a complete checklist), information relative to those absent at the site can later be used to infer its distribution. Should it not be specified, only those spotted are considered. Secondly, information about the observation process is taken into account, to measure search effort. This data is recorded to account for the bias

resulting from variation in detection and observation [25] and can later be used to improve species distribution models, such as those from the Spatio-Temporal Exploratory Model framework, addressed in Chapter 3.

These characteristics make eBird a **semi-structured** [6] citizen science program, which seeks to collect information from both the opportunistic observations (*presence-only*) and simultaneously infer from those that are absent (known as *presence-absences*). Another related concept is that of *presence-background*, which informs presence along with environmental information. In contrast, platforms such as BioDiversity4All<sup>6</sup> document observations following lighter protocols (mainly through media), and could be considered as an **unstructured** program. Nonetheless, it has its advantages, such as lower skill level requirements to new participants. **Structured** programs, characterized by low volume, velocity, and variety, are those mostly done by professionals, following strict protocols. Though, to be fully correct, eBird also stores data of some structured surveys, such as in the case of the *Breeding Bird Atlas*, which will be payed special attention in this work, and whose protocol is explained in the subsection below. The following Figure shows different characteristics intrinsic to these three types of programs.



**Figure 2.3:** Illustration summarizing the main characteristics of different types of citizen science project structures [26]. The size of the dots represent to which extent each structure type meets the listed sets of criteria. Obtained from Kelling et al. (2019) [26].

<sup>6</sup><https://www.biodiversity4all.org/>

### 2.4.1 Semi-Structured and Structured Data in eBird

As stated above, eBird mainly focuses on collecting data from birdwatchers following semi-structured reporting. However, the data collected may vary from dozens of different types of protocols, depending on its location and purpose, as the platform also keeps stored some structured data. With regards to Portugal, besides the regular eBird protocols explained initially in subsection 2.1.1, three other protocols that follow a structured methodology can be found in the Portuguese dataset: *RAM–Iberian Seawatch Network*, a monthly seabird counts from coastal points [51]; *Common Bird Survey*<sup>7</sup>, a long term monitoring program of common birds more directed at reporting the demographics trends of those species [52]; and, lastly, *Breeding Bird Atlas* protocol which serves as the structured approach to the *III Portuguese Breeding Bird Atlas*<sup>8</sup>. Besides structured, this Atlas also takes into account semi-structured observations as uploaded to eBird by volunteers in conjunction with census directed to specific species as a complementary way to enrich the estimates of the species' distributions and abundance, which in turn contribute to the *European Breeding Bird Atlas*, among other conservation initiatives<sup>9</sup>.

The *Breeding Bird Atlas* protocol will be looked into further below and in the following chapters.

#### Breeding Bird Atlas

This protocol, which can be found on eBird's database identified by the code **P65**, seeks to collect systematically as much information about the species in Portugal throughout the breeding season. It has taken place across Portugal between the years 2015 to 2021, with observations being carried between March 15th and July 15th. The volunteers follow systematic 30 minutes counts of the detected species (both visually and aurally) and record each one's *Breeding Code* describing its breeding activity. These counts are performed inside 6 2x2 km sub-squares, referred to as tetrads, that are distributed in a given 10x10 km square out of a grid covering the territory. Ideally, each larger square should be visited twice, at different periods within the timeframe.

We can summarise eBird's semi-structured and Breeding Bird Atlas' structured approaches with the following table:

Data collected	Semi-structured Observations (eBird)	Structured Observations (P65)
Duration	Any	30 minutes count
Visits per grid cell	Any	Ideally 2
Coverage per grid cell	Any	At least 6 2x2 km sub-squares
Breeding code	Recommended	Collected
Time interval	All year round	Between March 15th and July 15th
List Completeness	Complete and non-complete	Complete

**Table 2.1:** Comparison between the structured approach used for the P65 protocol and the semi-structured approach

<sup>7</sup><https://spea.pt/censos/censo-aves-comuns/>

<sup>8</sup>[https://www.spea.pt/wp-content/uploads/2020/12/Metodologia-campo\\_v6\\_20201209.pdf](https://www.spea.pt/wp-content/uploads/2020/12/Metodologia-campo_v6_20201209.pdf)

<sup>9</sup><https://spea.pt/censos/iii-atlas-aves-nidificantes/>



# 3

## Related Work

### Contents

---

<b>3.1 Approaches to Address Bias in Citizen Science . . . . .</b>	<b>21</b>
3.1.1 Quantifying Bias using SampBias . . . . .	21
3.1.2 Species Distribution Model (SDM) . . . . .	21
3.1.3 Spatio-Temporal Exploratory Model (STEM) . . . . .	22
3.1.4 Uses of STEM . . . . .	23
3.1.5 Spatial Interpolation Methods . . . . .	25
3.1.6 Point Process Model (PPM) . . . . .	26
<b>3.2 Data Visualization . . . . .</b>	<b>26</b>

---





This chapter addresses published work that regards some of the challenges mentioned in the previous section by making use of citizen science data, as well as models used by eBird for their estimations on the Status and Trends product, which were explored at an early stage of the work. This is followed by a brief overlook at work regarding data visualization.

## 3.1 Approaches to Address Bias in Citizen Science

### 3.1.1 Quantifying Bias using SampBias

Zizka et al. (2021) [38] proposed an algorithm for quantifying the effect of biases related to accessibility in sparse species occurrence data sets, making it open-access through an R-package<sup>1</sup>. The algorithm works by assessing to which extent variation in sampling rates can be explained by distance from bias factors, whose formulas are detailed in the mentioned paper [38].

### 3.1.2 Species Distribution Model (SDM)

These models are used to predict species distributions across geographic space and time using environmental data, to understand how these variables play out in their abundance or occurrence. They also serve for the detection of potential sources of bias and prediction of the distributions in areas that have not been sampled.

Several algorithms can be used to model species distribution, using presence data or both presence and absence data. These include statistical techniques such as generalized linear models, a generalization of ordinary least squares regression, and generalized additive models. Machine Learning methods can also be used for creating an SDM such as boosted regression trees, random forests (RF), or maximum entropy, a method known as MaxEnt [39].

One example of the application of SDMs is from Yu et al. (2010) [40] who developed a probabilistic model, the Occupancy-Detection-Expertise (ODE), which finds and incorporates the expertise of birders submitting checklists to eBird. This information could potentially be used posteriorly to inform participants in order to improve the reliability of their observations.

On another front, Matutini et al. (2021) [41] proposed ways on how citizen science could better these models. They made use of both presence-only and presence-absence data to test the use of filtered citizen science data to make an independent dataset for external evaluation for comparing the performance of presence-only SDM predictions.

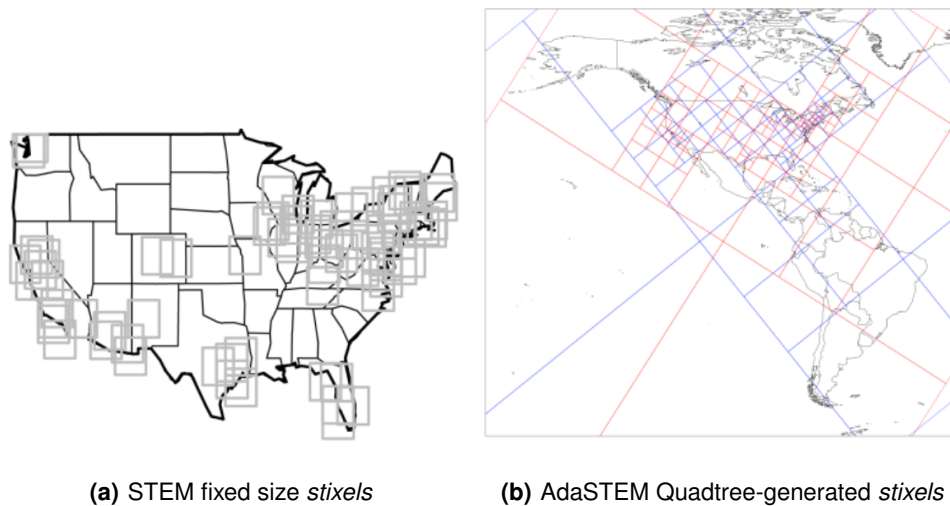
---

<sup>1</sup><https://github.com/azizka/sampbias>

### 3.1.3 Spatio-Temporal Exploratory Model (STEM)

With the growing need for sophisticated models and techniques to process and analyze eBird's data, its research group pushed for advancements in the statistical and computational fields that led to the creation of the Spatio-Temporal Exploratory Model [6]. These offer a solution for the challenge of variation through regions and time in a specie's habitat preferences throughout the year, while simultaneously controlling important sources of observation variability such as search effort [42]. Fink et al. (2010) [20] introduced the novel methodology using eBird data to estimate monthly changes in the distribution of the Tree Swallow, a migratory species present in the United States, to prove how its accuracy improved in comparison to the more conventional bagged decision tree model.

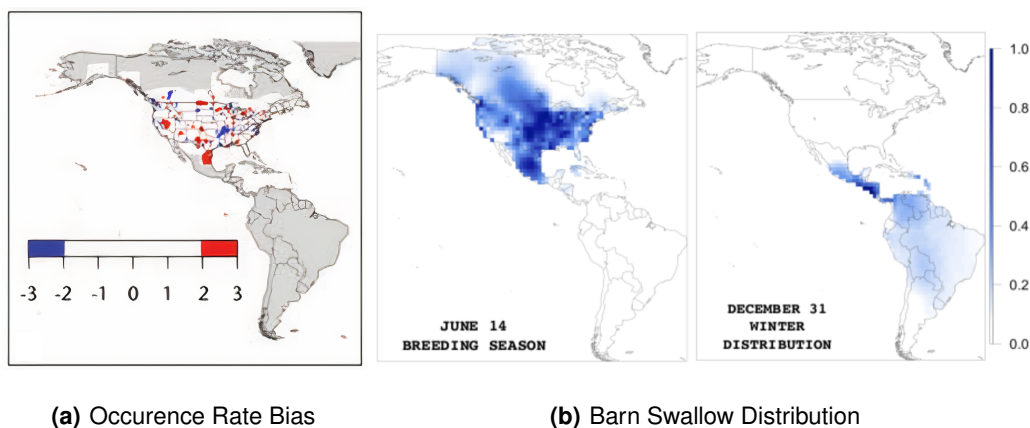
STEM work by adding spatiotemporal structure to existing SDMs without requiring specification about underlying dynamics. This is achieved through the creation of a set of multiple SDMs comprising smaller fixed areas (as in Figure 3.1 (a)), called *stixel*, short for "spatio-temporal pixel". Each one has a base model used to model the distribution that accounts for variation as a function of predictor values. In turn, these are restricted by two main parameters: the spatial, which controls the size of the region, and the time scale, controlling the time period. This is to account for patterns at a local extent, limiting the risk of extrapolation to longer distances - consequently, the size of the *stixel* results in a trade-off between the coverage of the study and that same risk. Then, these local models are combined into a single "ensemble" model to generate the end result - distributions over large regions and all year round. Because of this, the model is also referred as using a two-stage SDM approach [42].



**Figure 3.1:** Comparison between STEM and AdaSTEM *stixels*. *Left:* STEM's fixed-size *stixels* are shown, not offering total coverage. *Right:* AdaSTEM's *stixels*. Worth noting that the smaller *stixels* correspond to areas with higher observation density. Image (a) obtained from Fink et al. (2010) [20] and image (b) from Fink et al. (2014) [43].

Later, an extended version of STEM was developed by the same authors, using the same strategy of dividing and recombining models. Fink et al. (2013) [44] proposed the **Adaptive Spatio-Temporal Exploratory Model**, also called AdaSTEM. This approach, also making use of eBird checklists, aimed to geographically expand the range of the spatiotemporal estimates in order to produce a hemisphere-wide distribution of long-distance migration species at the population level. Here, the function setting the size of the *stixel* is not fixed, and varies according to the density of observations, as observable in Figure 3.1 (b). This makes it more adaptable to the continental scale, where the spatial resolution of the estimates can range from less than 100 kilometers length to more than 10.000 kilometers [43].

To identify regions with bias, while studying distributions for Barn Swallows, Fink et al. (2014) [43] determined the areas where the interpolated residuals were considerably larger than those expected by chance. This can be observed in Figure 3.2 (a), with higher estimated occurrence rates in blue, and lower for the colour red. Finally, Figure 3.2 (b) depicts the result of using AdaSTEM to estimate distribution of occurrence of the Barn Swallow.



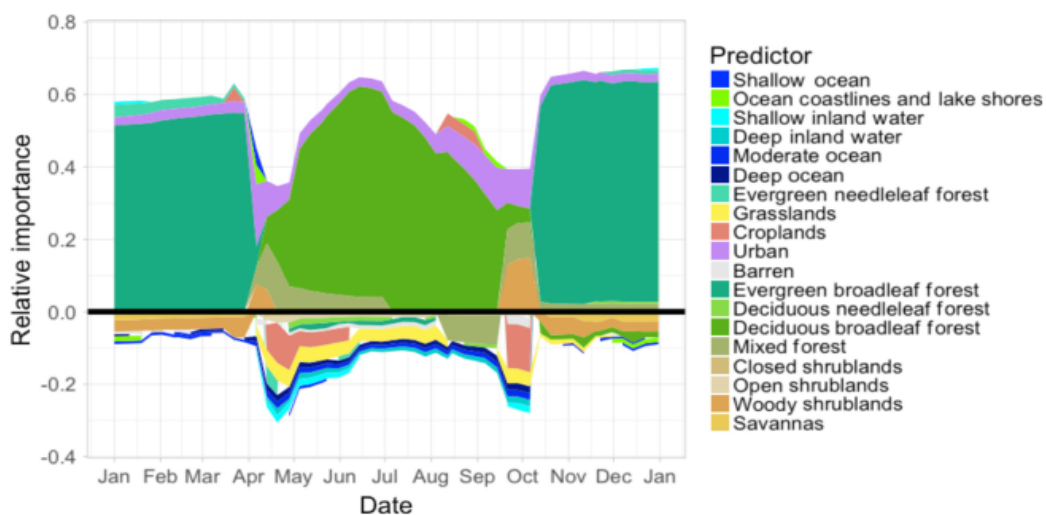
**Figure 3.2:** *Left:* using AdaSTEM, colored areas indicate bias where standardized residuals are more than twice as large as their associated standard errors. *Right:* AdaSTEM distribution estimates in the American continent for Barn Swallow during the breeding season, and winter, using eBird data. The bar on the right indicates the relative probability of occurrence. Obtained from Fink et al. (2014) [43] and Fink et al. (2013) [44], respectively.

### 3.1.4 Uses of STEM

Johnston et al. (2015) [45] put into practice a model used to estimate Californian migratory waterbird species' occurrence and relative abundance using eBird data, to help prioritizing times and locations for conservation decision-making. It details the strategy used to calculate these estimates, which are present on eBird's **Status and Trends**. The model used takes into account variation in detectability resulting from the observers' effort, using covariates such as the duration of the activity, number of people involved, distance traveled, among other information recorded in each checklist, to describe

the observation process. Additionally, environmental covariates are used to define the environmental process, such as elevation and annual land cover. It adopts a three-stage modelling strategy, with one of them being the use of the STEM framework to generate estimates throughout the year across the region for each species, while accounting for the nonstationarity of the relationships between the counts of birds and environmental variables. The base model, consisting on a **zero-inflated boosted regression tree**, is used to handle the frequent absence of observations, and estimate covariate effects.

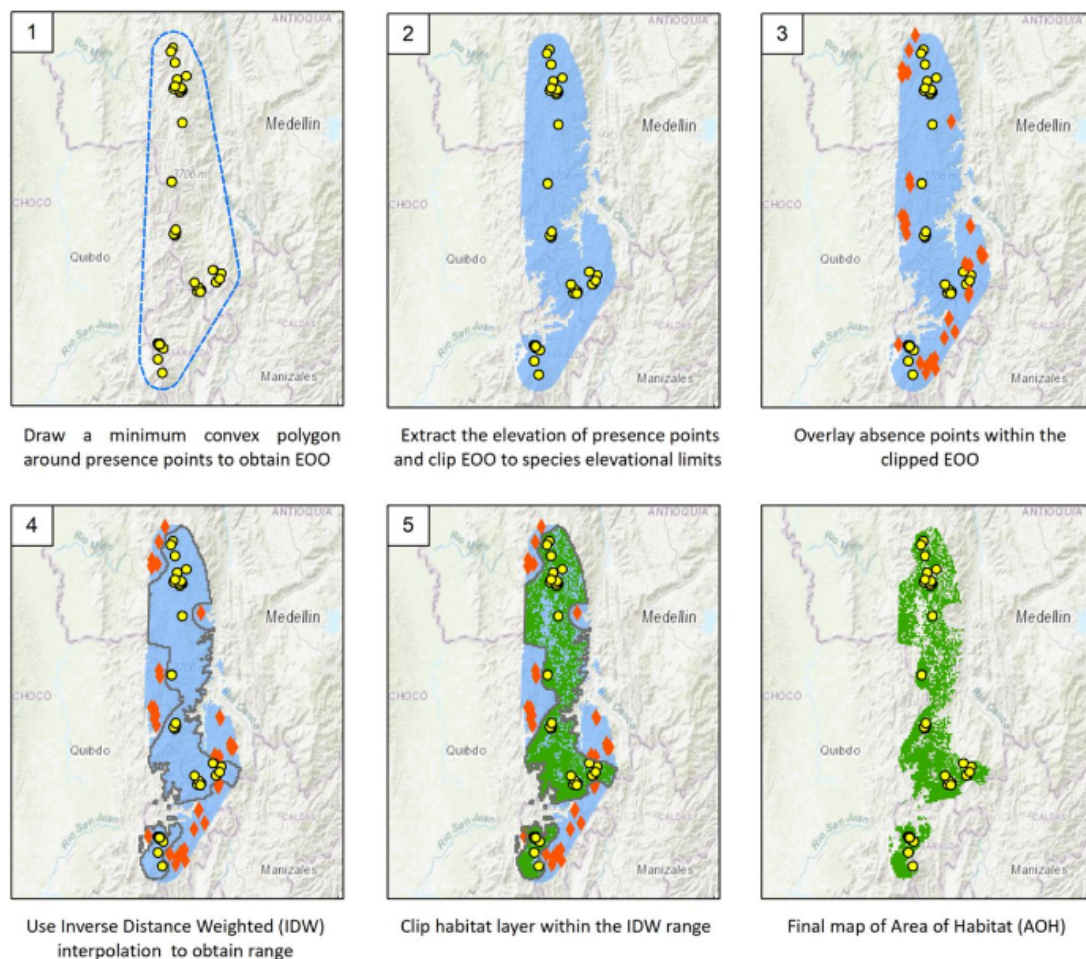
Furthermore, Fink et al. (2020) [46] provide an overview of the analytical methodology that is used for the Status and Trends data products, applied to the long-range migrant species Wood Thrush. In this case, a two-step approach based on the AdaSTEM framework is used, as the estimates' scope entails the whole western hemisphere, to tackle the challenges inherent from irregularly and sparsely distributed data. These comprise estimates of abundance, occurrence, area of occurrence (AOO), and trends to estimate the average annual rate of change in relative abundance. As STEMs can shape how the detection or non-detection of species change with a set of predictor variables, a considerable set of 87 predictors is incorporated into the model, which is divided across three categories that account for observation effort, variation across temporal scales, and a larger one related to environmental covariates. Figure 3.3 measures the contribution of the predictors describing land and water cover, and how they vary in importance throughout the year for the population concerned, in order to analyse how habitats are used.



**Figure 3.3:** Year-round predictor relative importance of each land cover class for the Wood Thrush population, calculated weekly for each base model (fit within each *stixel*). Classes whose value is below zero are not used in the core population. Image taken from Fink et al. (2020) [46].

### 3.1.5 Spatial Interpolation Methods

A different metric for modelling and mapping distributions of vulnerable bird species and assessing its extinction risk was proposed by Palacio et al. (2020) [47], citing the complexity of distribution modelling through SDMs. By making use of presence and absence points from GBIF [11] - comprising eBird observations - Palacio et al. used the interpolation procedure Inverse Distance Weighting (IDW) to map the distribution and develop a framework of a geospatial workflow that is compatible with the IUCN Red List - a well-known inventory of the global conservation status of biological species - following the procedure shown in Figure 3.4. These estimates were then compared to the expert-drawn range maps in IUCN/BirdLife.



**Figure 3.4:** Workflow developed for mapping the distribution of a species, using the hummingbird species Glittering Starfrontlet as an example. Obtained from Palacio et al. (2020) [47].

### 3.1.6 Point Process Model (PPM)

These are used to extract and understand spatial patterns from a set of random points in space. Even though less commonly used on citizen science data, Geldmann et al. (2016) [30] resorted to PPMs to map areas of hotspots on presence-only observations from four distinct citizen science programs, find potential regions of lower biodiversity and identify spatial factors that determine where people observe. This is done by determining the likelihood of each point in space a function of a series of covariates such as land cover and infrastructure. One advantage stated about PPMs for this type of analysis is that it allowed for modeling covariates in a spatial context, in contrast to a grid analysis. The other side of the coin is that these models were regarded as being limited to points representing moving objects or with high error rates. Therefore, the observations were viewed as simply being an event of observation taking place, rather than considering the environmental knowledge it may have for the species at hand - keeping the focus on the intensity of observations. The authors also conclude that additional contextual data proves useful to make further analysis.

Additionally, Hefley and Hooten (2016) show how count, presence-absence, and presence-only data can be conceptualized to a point process distribution, through Hierarchical Species Distribution Models.

## 3.2 Data Visualization

Regarding data visualization, a couple of papers are expected to be useful in the stage of developing the visual tool proposed in the following Chapters.

Firstly, Roth (2013) [49] provides a review of the current state of science regarding interactive mapping, by laying six fundamental questions. These, although some being abstract, serve as principles on the best practices for cartographic interaction, including how it should be provided, its design and interface styles, where it should be applied and to whom, among others. Likewise, Midway (2020) [50] proposes a set of 10 principles, though more focused on the visual representation of the data on scientific visuals. Both papers may prove to be advantageous to build more robust images in the work ahead.

Regarding eBird, two main types of visualization are available: visualization from the main eBird website, allowing users to interactively explore birding hotspots and where species have been observed; and the visualization tools from Status and Trends, which, as already mentioned, have available maps describing how bird populations change through time, displaying abundance animations, range maps and abundance maps. Besides eBird, one tool that may serve as a source of inspiration is a case study from the European Union's environmental program Copernicus Climate Change Service, that have created an educational storytelling map<sup>2</sup> that shows through an interactive timeline bird migration

---

<sup>2</sup><https://birdmigration.climate.copernicus.eu/the-progression-of-bird-migration>



movements of four bird species in continental Europe.

Besides eBird's visualization tools, another rather valuable source of inspiration was the R Shiny gallery<sup>3</sup>, which is comprised of contributions from the community showcasing examples of interactive tools built using the Shiny framework [63], which that will be explained in depth further. Additionally, it also contains several examples that highlight specific features of the package. These have undoubtedly given an idea of the capabilities and potential of the aforementioned framework.

---

<sup>3</sup><https://shiny.rstudio.com/gallery/>





# 4

## Approach

### Contents

---

<b>4.1</b>	<b>Proposed Solution and Goals</b>	<b>31</b>
<b>4.2</b>	<b>Data Considered</b>	<b>31</b>
<b>4.3</b>	<b>Metrics Considered</b>	<b>32</b>
4.3.1	Metrics Applied to a Single Species	32
4.3.2	For Each Grid Cell	35
4.3.3	Mapping representation	38
4.3.4	Cohen's Kappa	42
4.3.5	Observable Agreement	44
4.3.6	Percentage of Species Reported	45
4.3.7	Species Accumulation Curves	46
4.3.8	Species Richness	47

---



This next chapter focuses on the solution that is being presented, taking into account the previous information about how citizen science data is collected and utilized in eBird. The visual tool that is proposed along with the data and statistics that accompany are thoroughly described in this Section.

## 4.1 Proposed Solution and Goals

From the start, one of the main focus for this work was to develop an interactive tool capable of interactively displaying the imperfections of citizen science programs, with the focus being on eBird.

In this solution, we didn't focus on the inherent variability of the observers' skills, which is always present, but rather on the data directly available via checklists uploaded to eBird and validated. Thus, seeking to paint the picture on how the different regions are being reported and therefore help identify unreported as well as misrepresented areas and species throughout the country, for, ideally, a more accurate depiction of its distributions in the future.

We focus on two main angles for analysing the data, which is reflected on the behaviour of the tool presented:

1. Graphical and mapped analysis of the data following a semi-structured procedure applied to either all or to a single species, across an arbitrary timeframe.
2. Contrasting semi-structured and structured approaches, for a fixed timeline - that of the Breeding Bird Atlas - done by mapping and graphing different metrics that allowed to compare each one of the outcomes. Also, similarly to the previous angle, to be able to display this analysis for either a single species or for all of those reported during the aforementioned timeframe, with the latter offering additional metrics for helping to contrast the results.

## 4.2 Data Considered

As previously mentioned, the data considered for the proposed solution can be split into two categories, concerning its collection method - **semi-structured** and **structured** data.

Regarding the semi-structured data, the protocols focused on were the *Stationary* and *Travelling* ones, with protocol codes **P21** and **P22**, respectively, from 2010 up until December 2021. Checklists registered as *Incidental* protocols (protocol code **P20**), formerly was referred to as "Casual Observations", were not taken into account for our solution, since the former, as the name suggests and as described in 2.1.1, does not have birdwatching as its prime objective and due to the amount of required data fields collected being less strict for both. For this same reason, those checklists registered as *Historical* (protocol code **P62**), were also excluded from the set.

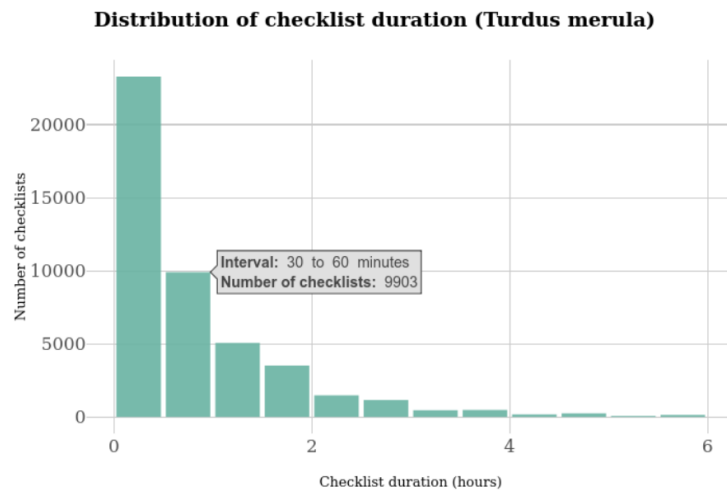
In relation to the structured data, *Breeding Bird Atlas*, was naturally selected for, just like complete checklists, following a procedure of reporting all species present during observation. On the other hand, the remaining two structured protocols referred in Section 2.4.1 were not considered - as both present differing methodologies, with the *RAM-Iberian Seawatch Network* being focused on maritime seabirds in specific spots [51], and the *Common Bird Survey* being also taxonomically biased for focusing on the reporting and population of a fixed set of common bird species, and geographically for having covered a relatively small portion of the territory in the previous years, besides making use of a grid using a separate coordinate system [52].

## 4.3 Metrics Considered

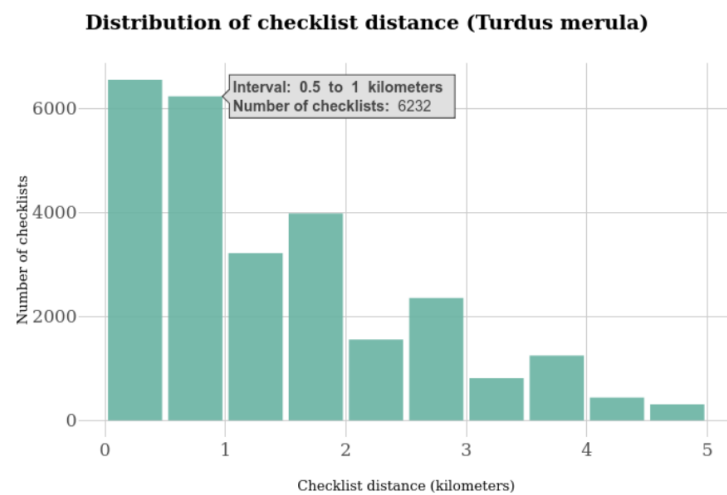
Below, metrics that were applied to the raw eBird data are listed. These can be split into two main categories: firstly, those plotted in graph form, made to analyse the search effort for either each grid cell or for the entire map; secondly, those that make use of the grid to depict a metric by means of a color. Also, it should be noted that in the proposed tool, the metrics described below are shown depending on the input, which can display one of the different angles listed in 4.1 - for instance, the map containing the total number of different species reported won't need to be mapped for the analysis of a single species. Regarding the graphs, these also may display information when one of its elements are hovered on with the cursor - which is referred to as the **tooltip**. These were purposely included in the figures of all graphs, listed below.

### 4.3.1 Metrics Applied to a Single Species

Starting with metrics illustrating the sampling effort of the checklists - these were created with the intent of describing the effort taken relatively to a single species's lists. In the graphs below the time period considered is the same as *Breeding Bird Atlas*'s, although different time periods can be visualized in the tool proposed. The species Eurasian Blackbird (*Turdus merula*) was chosen for these graphs due to being a common species in checklists throughout the territory. Naturally, these graphs require the checklists to be split into bins, *i.e.* into intervals of the data being distributed, in order to graph the data.

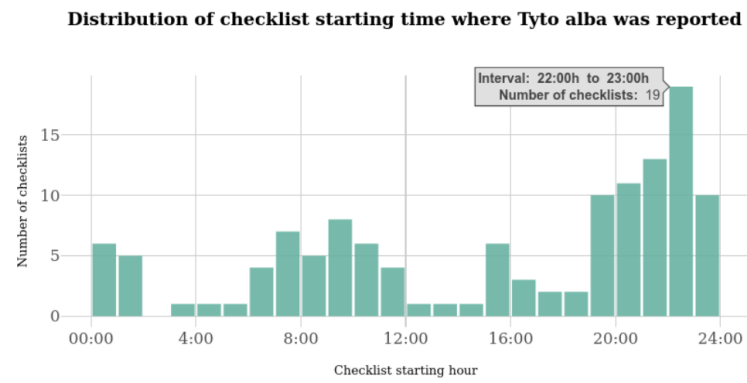
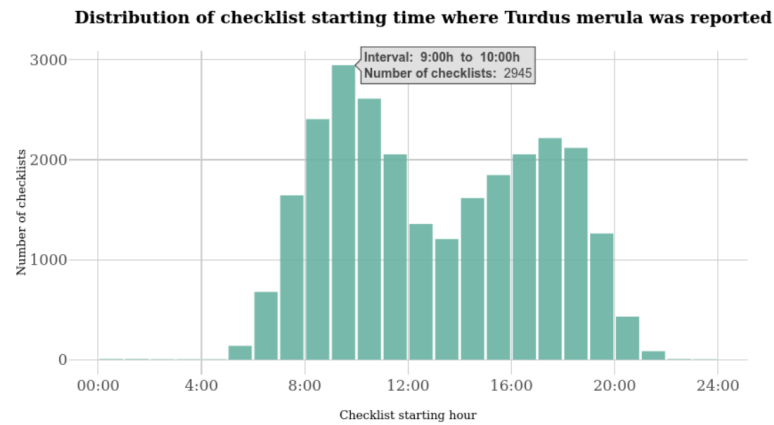


**Figure 4.1:** Distribution of the semi-structured checklists' duration where the species *Turdus merula* was reported.



**Figure 4.2:** Distribution of the semi-structured checklists' distance where the species *Turdus merula* was reported.

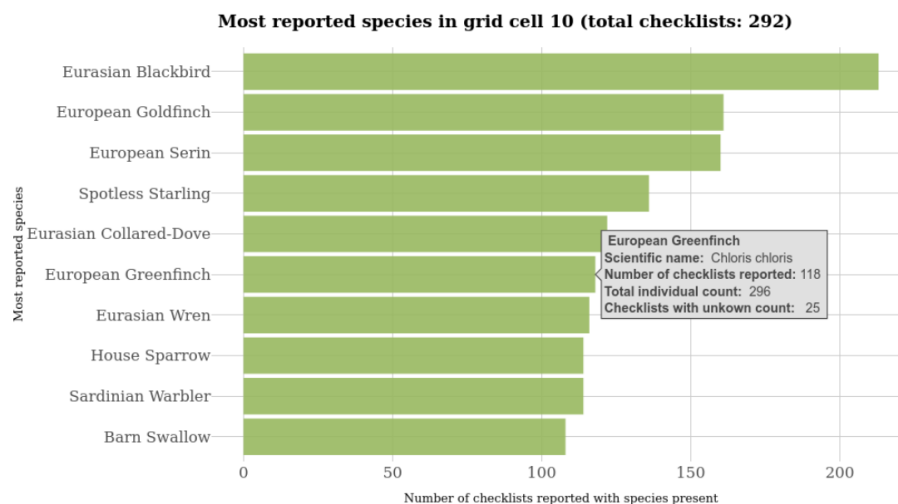
These metrics usually serve as the basic search effort measures one can consider to characterize the behaviour of the observers. Below, two additional graph plots the starting time of checklists that have observed two distinct species - *Turdus merula* and *Tyto alba*, commonly known as Barn Owl. The latter was included to simply show how it's possible with these plots to get information not only about the observer behaviour, but also about the species itself.



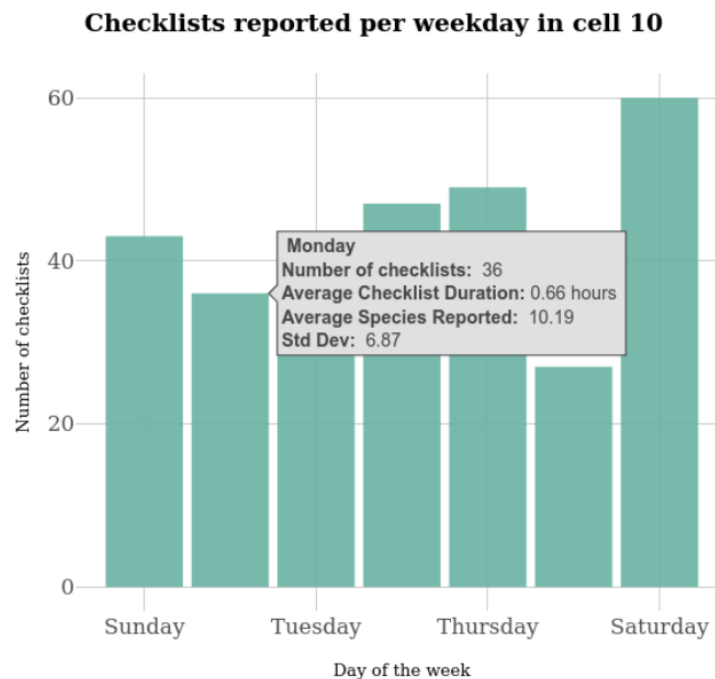
**Figure 4.3:** Distribution of the semi-structured checklists' starting time where *Turdus merula* and *Tyto alba* were reported.

### 4.3.2 For Each Grid Cell

The following graphs depicted, to be plotted with the data of a specific grid cell, are meant to be generated as the user clicks the grid presented on the map, being one of the responsive features of the tool initially proposed. The plot below reports, for a given grid cell clicked, the top species reported in that cell, displaying more info about the number of checklists the species were observed, as well as the count of individuals.



**Figure 4.4:** Top reported in cell 10 semi-structured, atlas timeframe.

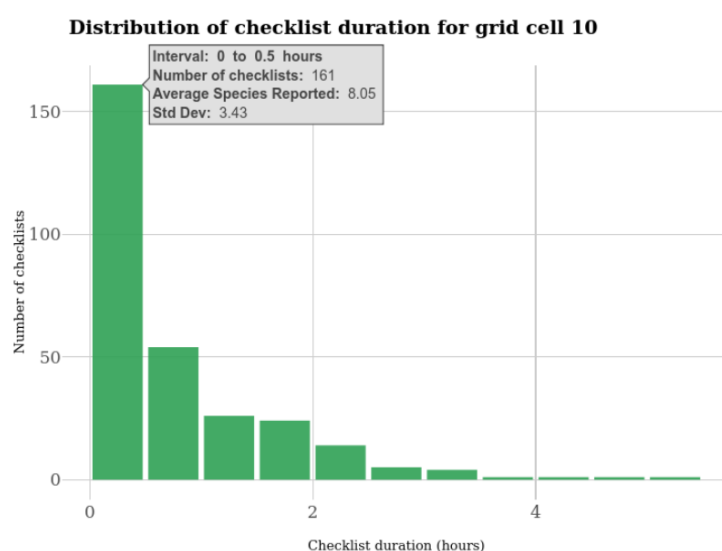


**Figure 4.5:** Checklists reported per weekday for the grid cell 10.

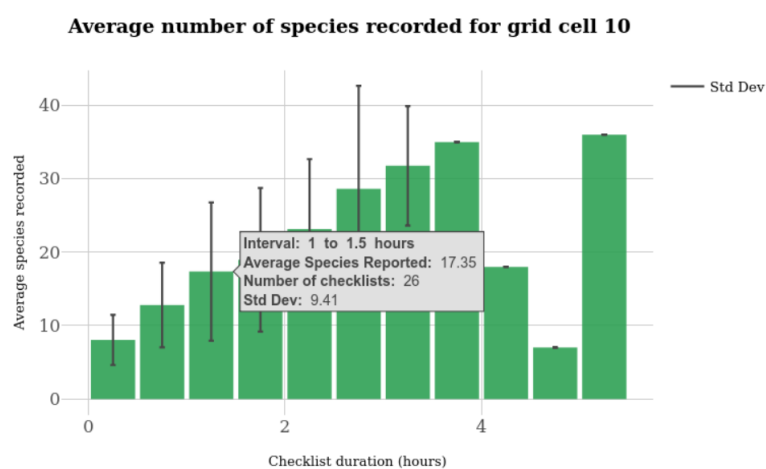
The plot above features the average number of checklists reported per weekday for the grid cell number 10 (here with a notable spike on Saturday), along with info on each days average checklists duration and species reported on the tooltip.

Below, the following two plots categories are analogous the first two graphs in 4.3.1, but since these are not directed at a single species, more data can be displayed, including the average number of different species reported per duration and distance bins, respectively.

### As a function of duration



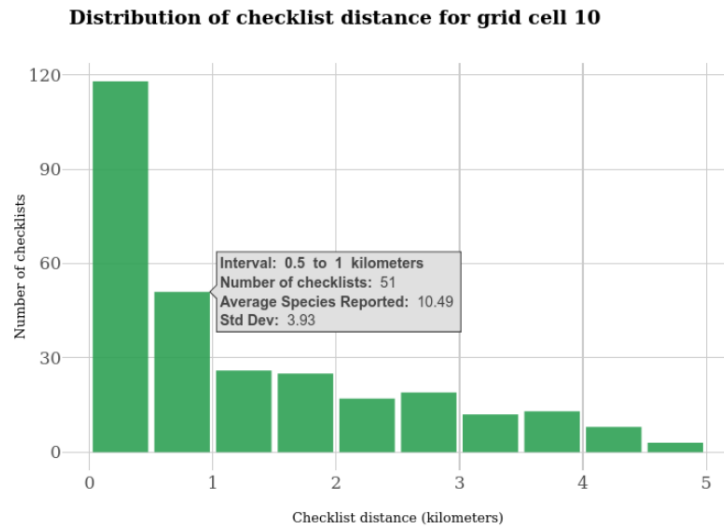
**Figure 4.6:** Distribution of the checklist duration in a given grid cell per **duration** bin.



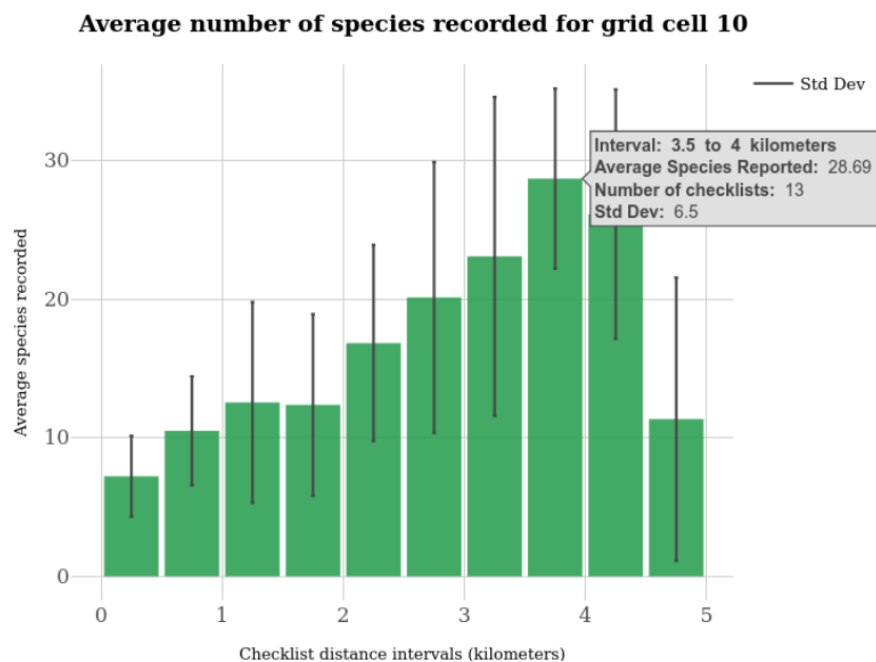
**Figure 4.7:** Average number of species submitted in a given grid cell per **duration** bin.



## As a function of distance



**Figure 4.8:** Distribution of distance bin the number of checklists submitted in a given grid cell per **distance** bin.



**Figure 4.9:** Average number of species reported per **distance** bin.

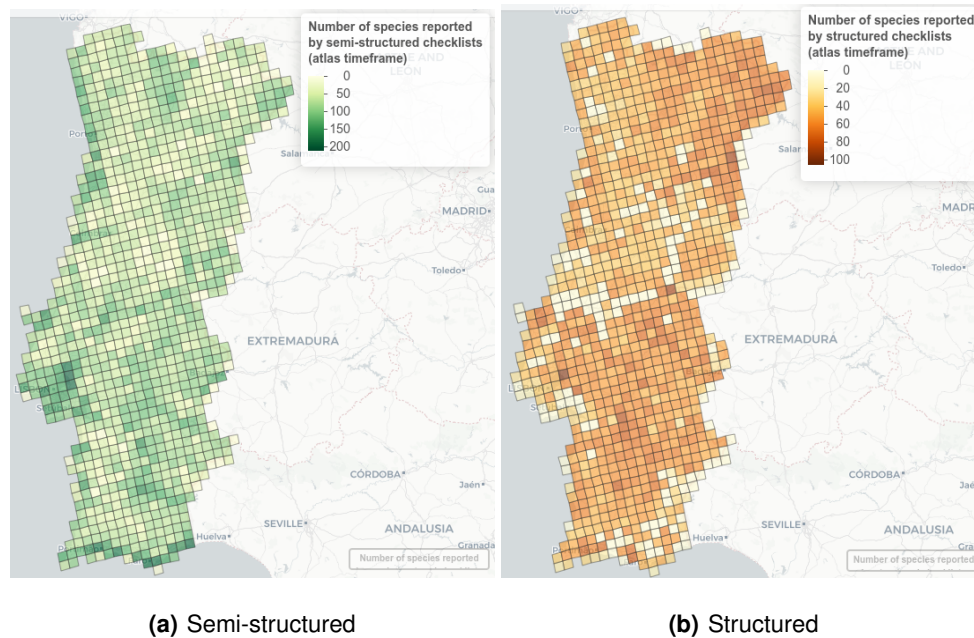
The standard deviations, present in the average plots for each of the categories, indicating how disperse the data is in relation to the corresponding average, helps to tell how much the values (in species reported) within each bin vary.

### 4.3.3 Mapping representation

Besides plotting graphs, several maps were created to provide a global image of a different set of metrics. The grid used for mapping these follow the same grid used by the atlas in its methodology, described in Subsection 2.4.1. It is through the mapped grids listed below that the user can interact with a specific cell in order to display additional information about it.

#### Total number of different species reported

Indicates per grid cell the number of different species reported. The image below depicts the number of species during the atlas timeframe:

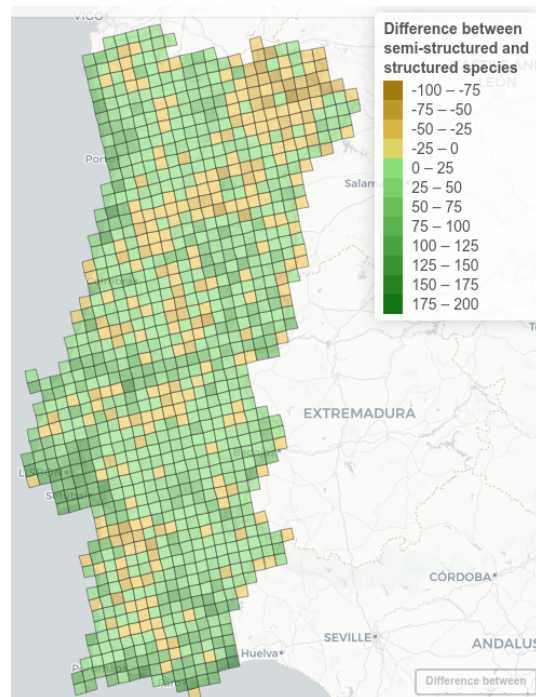


**Figure 4.10:** Maps depicting the total number of species reported by semi-structured and structured checklists, during the atlas timeframe.

The white cells in the Figure (b) above indicate tell that there have been some grid cells which have not been surveyed by the structured protocol, which will be confirmed by further metrics below. It is also worth noting the considerably smaller maximum of species reported on the structured approach's legend, in contrast to the semi-structured's. Below, the difference of species is also more noticeable with the difference between the two.

### Difference between the number of species reported

The following map indicates difference between the number of species reported by eBird's semi-structured observations in opposition to the number reported by the *Breeding Bird Atlas*, naturally, within the time-frame of the atlas. Hence, considering the Figure below, the positive values coloured green denote higher number of reported species from semi-structured checklists, and the other way around for the negative values in yellow.

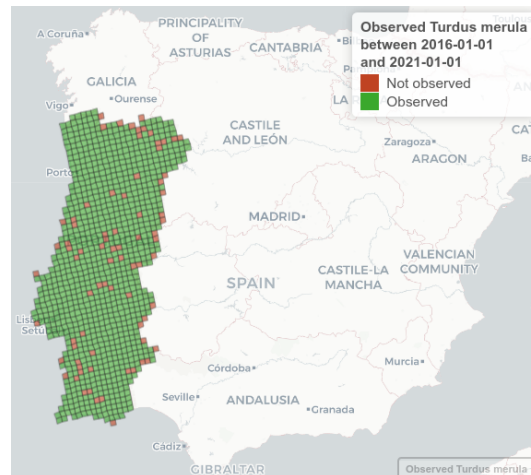


**Figure 4.11:** Map showing the difference between the number of species reported by semi-structured and structured checklists, during the atlas timeframe.

It is right away noticeable how more populated areas, specially throughout the coast, semi-structured lists depict a higher amount of species reported, due to the disparity in the number of checklists between the two approaches. However, the structured approach here demonstrates a clear superiority in the number of birds reported in some areas in inland Portugal, most notably in the *Bragança* district, in the northeast.

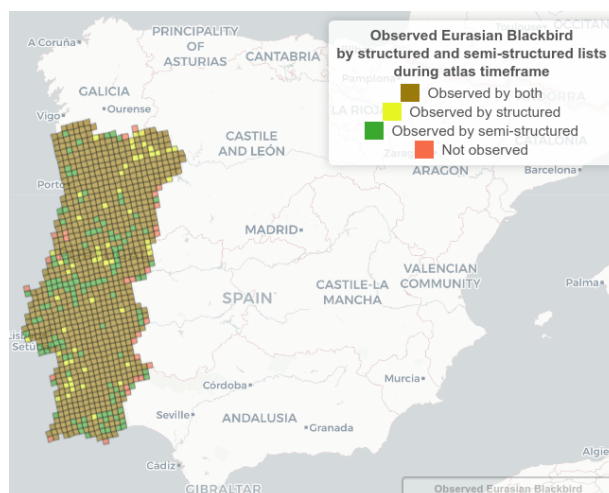
## Presence and absence of species

These maps are to be created if the user opts to simply generate a map of one species without comparing to structured data, by simply indicating where a given species has been reported, per grid cell.



**Figure 4.12:** Map depicting where species was observed between an arbitrary date.

If, on the contrary, the user selects the option for comparing with the structured data, a map like the one below is created. Similarly with the previous map, this one also distinguishes between those cells where the species have been observed and not observed by semi-structured reporting, but with the additional information on the observation by structured lists. These are mapped only during the atlas timeframe, for a fair comparison among all the grid cells.



**Figure 4.13:** Map depicting where Eurasian Blackbird have been observed during the atlas timeframe.

In the example above, with this species being one of the most common ones, it can be noted that, just like in Figure 4.10, most of the cells colored green also correspond to the those with 0 species reported by the structured approach.

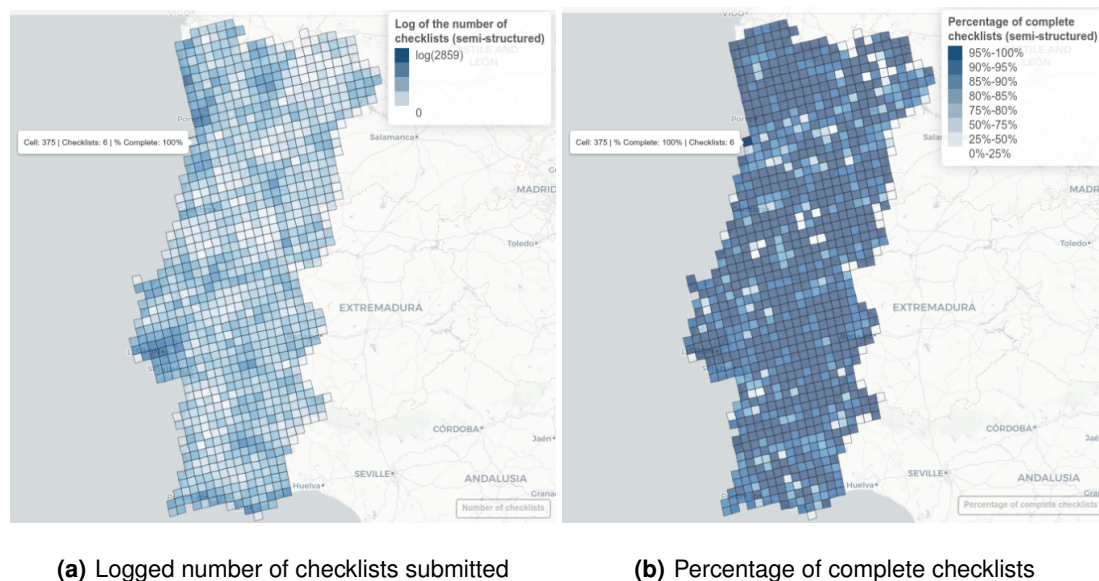
### Number of checklists submitted and completeness percentage

This map was created to introduce data regarding the nature of the checklists themselves. Due to the high disparity found in the number of checklists submitted among the cells, the *log* of its number was calculated, in order to properly view the differences.

Here, just like the mapping of the number of different species reported by semi-structured data, it is noticeable how urbanized coastal areas, along with renowned hotspots for birdwatching (such as and along the southern coast and in Tagus Estuary Natural Reserve) have submitted many more checklists. This is also consistent with the fact that coastal Portugal is substantially more populated than those in the interior. Conversely, less populated areas in the interior aren't as crowded, as expected.

Additionally, the light-blue color of the northeast are also consistent with the result obtained in Figure 4.11, which has seemingly resulted in a smaller count of different species in the area.

With regards to the percentage of complete checklists, a straightforward take from the graph 4.14 (b) is the unanimously high percentage of lists that are complete, even taking into account that these follow the *Travelling* and *Stationary* semi-structured protocols. In relation to the structured checklists, mapping these data wasn't the priority as its number of lists reported per each cell varies little, and are mostly complete by definition.



**Figure 4.14:** Maps depicting the logged total of semi-structured checklists in (a), and the its percentage of completeness in map (b).

### 4.3.4 Cohen's Kappa

Cohen's Kappa is a measurement of agreement between observations introduced by Jacob Cohen in 1960 [53] that quantitatively describes the reliability of two raters that are rating the same thing. In our particular case, we want to assess the agreement between the number of different species reported by eBird checklists from semi-structured observations, submitted by any user on the platform; and those that follow the structured Breeding Bird Atlas protocol, as explained in Section 2.4.1. In this way, this metric is expected to quantify the agreement between structured and semi-structured observations for each grid cell throughout the territory with regards to the different number of species reported.

		Second Rater		Total
		Positive	Negative	
First Rater	Positive	$a = f_{11}$	$b = f_{12}$	$(a + b) = r_1$
	Negative	$c = f_{21}$	$d = f_{22}$	$(c + d) = r_2$
Total		$(a + c) = c_1$	$(b + d) = c_2$	$N$

**Table 4.1:** Table depicting confusion matrices used for the calculation of the value of Kappa, depicting the theoretical matrix, with the variables relations. The variables  $a, b, c, d$  are the names given in this work's code, and are somewhat easier to visualize.

So, there are two raters, one corresponding to the semi-structured approach, the other to the structured approach. Let  $N$  be the number of cases, *i.e.*, the total number of different species ever recorded, and  $k$  be the number of categories in which a case can be rated, *i.e.*, either observation or no observation of a species ( $k = 2$ ). To calculate Cohen, a  $k$  by  $k$  confusion matrix is defined (Table 4.1), in which an element  $f_{ij}$  defines the number of cases that the first rater assigned a particular case to category  $i$  and the second to  $j$ . So,  $f_{jj}$  is the number of agreements for category  $j$ . Then:

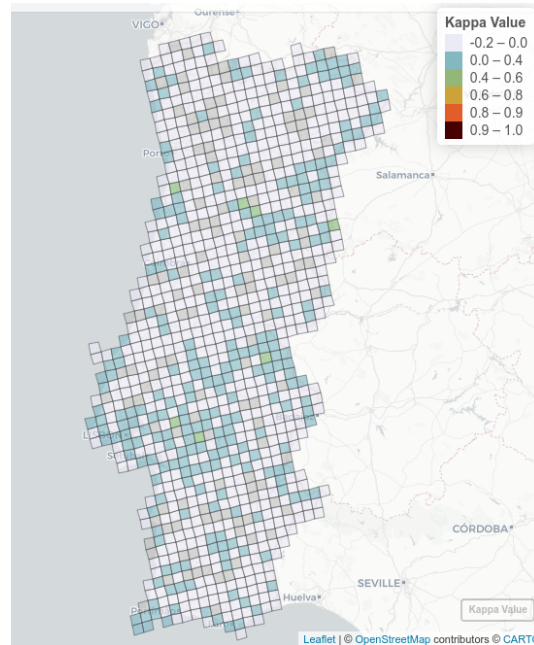
$$P_o = \frac{1}{N} \sum_{j=1}^k f_{jj}, \quad (4.1)$$

$$r_i = \sum_{j=1}^k f_{ij}, \forall i, \text{ and } c_j = \sum_{i=1}^k f_{ij}, \forall j, \quad (4.2)$$

$$P_e = \frac{1}{N^2} \sum_{i=1}^k r_i c_i, \quad (4.3)$$

where  $P_o$  the observed proportional agreement,  $r_i$  and  $c_j$  the row and column totals for category  $i$  and  $j$ , and  $P_e$  the expected proportion of agreement. The final measure of agreement  $\kappa$ , which is applied to each grid cell on the map, is given by:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (4.4)$$



**Figure 4.15:** Kappa coefficient for all species reported between 15th of March and 15th of July of the years 2015 to 2021.

Value of $\kappa$	Strength of Agreement
$<0.20$	Poor
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Good
$>0.80$	Very Good

**Table 4.2:** Interpretation of the Cohen's Kappa coefficient.

The values of Cohen's Kappa can be interpreted according to Table 4.3, with the maximum value being  $\kappa = 1$  corresponding to total agreement, and with  $\kappa = 0$  corresponding to agreement as expected by chance. Negative values may also show up.

This metric has been calculated for each grid cell in Figure 4.15. To get to the values of  $\kappa$ , a value  $N$  was needed for each grid cell, *i.e.*, the number of species ever recorded. Three different attempts to obtain this value were made. Initially,  $N$  was statically defined as all the species recorded across the whole Portuguese territory, ever - which, although it was the easiest value to get, does not represent the real species present at that timeframe, skewing the results. This was followed by considering all species ever recorded for each cell. Finally, we considered all species ever recorded for each cell, but within the atlas timeframe. Although the order between the values has generally been maintained, decreasing  $N$  resulted in decreasing the the number of species not observed either by the semi-structured or structured approaches (the value  $d$  in Table 4.1), which drastically reduced the values of  $\kappa$ . This ended up suggesting that the agreement is not as strong as initially thought.



		Structured approach		Total
		Present	Absent	
Semi-structured approach	Present	52	10	62
	Absent	9	19	28
Total		61	29	90

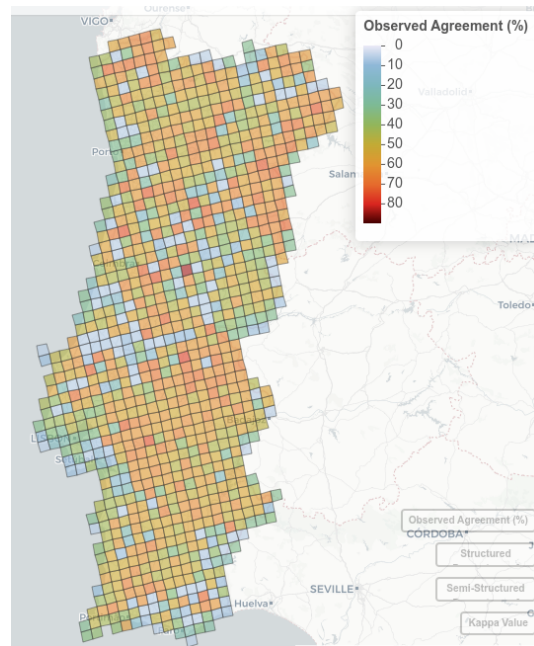
**Table 4.3:** Table depicting confusion matrix with the values measured for the grid cell with ID 266 throughout the atlas timeframe, which has presented the highest value of  $\kappa$  (0.512).

Since this metric did not allow us to fully portray the degree of agreement between semi-structured and unstructured observations, we chose to also map additional information regarding the agreement on species observation between the two approaches, below.

#### 4.3.5 Observable Agreement

Advantage was taken of Cohen's Kappa calculations to also map the observable agreement as a standalone metric. This indicates the percentage of bird species whose reporting has been the same in both parties, *i.e.*, the proportion of species that were observed and not observed by both semi-structured and structured approaches, relatively to a total of bird species ever reported in a given grid cell, referred to above as  $N$ .

This proved useful to clarify the agreement of both parties, while making up for some drawbacks revealed by Kappa.

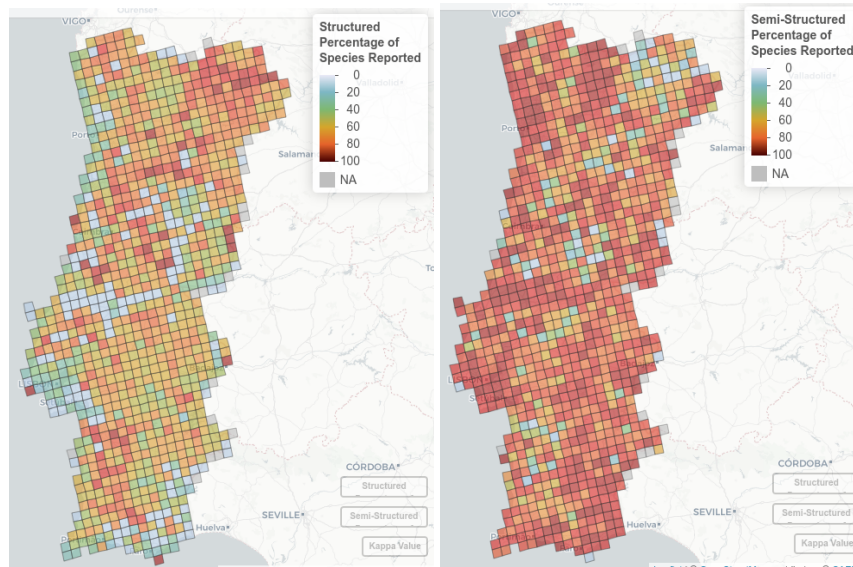


**Figure 4.16:** Observable agreement between structured and semi-structured observations, expressed as a percentage.



### 4.3.6 Percentage of Species Reported

The percentage of species reported gives information about the fraction of species that were reported by structured and semi-structured compared to the set of species ever recorded, using the  $N$  defined above. The following 4.17 exemplifies this metric:



(a) Percentage of species reported over the atlas timeframe by structured checklists.

(b) Percentage of species reported over the atlas timeframe by semi-structured checklists.

**Figure 4.17:** Comparison between structured and semi-structured percentages of species found over the atlas timeframe.

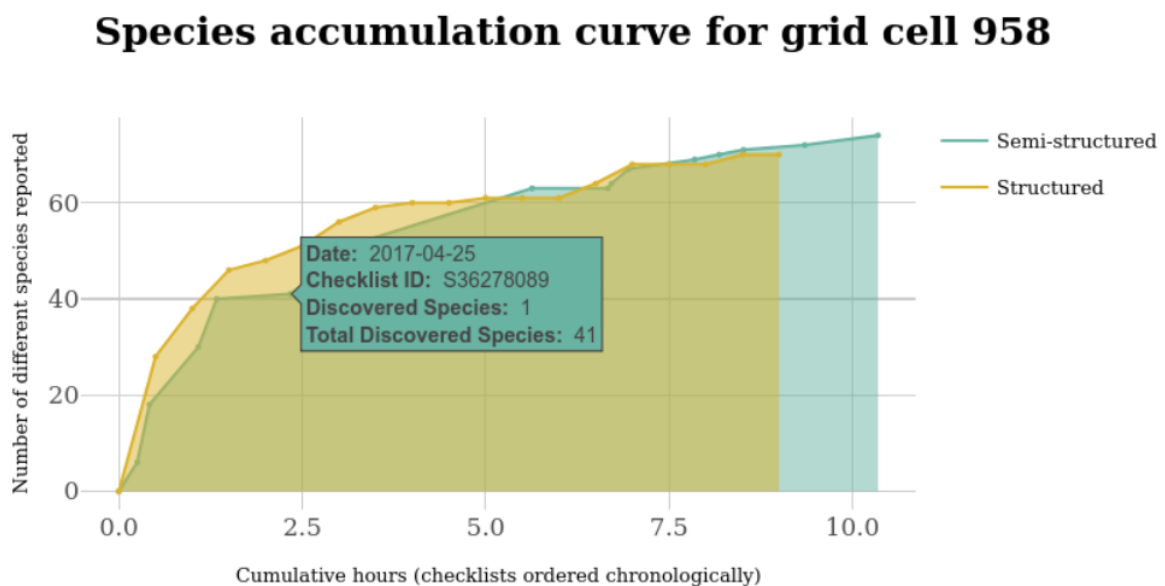
It is worth underlining the different interpretation derived with these values from that of Cohen's Kappa defined above. While the latter denotes the level of agreement in the number of species reported - which takes into consideration those that were not reported - between structured and semi-structured observations, here, that number is being compared for each procedure to the number of species ever reported throughout the atlas' time period. Meaning that one grid cell with a high agreement does not necessarily translate to having a high percentage on the number of different species reported on either approaches, but rather that the species reported and not reported are more alike.

### 4.3.7 Species Accumulation Curves

Species accumulation curves show the number of observed species or distinct classes of species as a function of sampling effort over a period of time. This usually provides a way of estimating the number of new species that can be discovered if additional effort is carried. It depicts a curve that will necessarily be increasing, and most commonly negatively accelerated, since, as the time goes by, the less likely it is to report new species.

These accumulation curves have been creatively used before in the field of citizen science, for instance, by Kelling et.al (2015) [42] to try to measure observer's skills. However, for our case, we would like to analyse, for a specific grid cell, how these evolve for every checklist present.

In order to create this plot, the checklists comprising the chosen interval are therefore ordered chronologically. We also made the choice of not only sorting the checklists by its submission date, but having the X axis corresponds to the cumulative sum of the checklists' distances, rather than simply plotting by the checklist ID. Thus, information about the effort taken between lists is also conveyed. Moreover, additional information is displayed on the tooltip of the graph, namely the date, ID and the number of new species found at a given point.



**Figure 4.18:** SAC plot comparing structured (in orange) and semi-structured checklists (in green).

The resulting plot in the figure above also features more information on the tooltip about the checklists, including the date at which it was submitted, its ID, and how many new species were added to the curve.

Due to the lesser number of the structured checklists, as compared to semi-structured ones, the visualization of the overlaid curves on the same graph can be often hindered due to the considerable different scales. As it will be mentioned in Chapter 5, the interactivity of these graphs allow the user to zoom in on the graph, which facilitating its viewing.

#### 4.3.8 Species Richness

Shannon's diversity index, also known as Shannon–Wiener index, was originally proposed by Claude Shannon in 1948 to quantify the entropy in strings of text [55]. This index can also be applied as a diversity index if we have available the number of individuals of each species reported in a given location. The calculation of this metric was done using the **vegan** [61] package, popular for providing standard tools of descriptive community analysis. The formula is defined as follows:

$$H = - \sum_{i=1}^M P_i \ln P_i \quad (4.5)$$

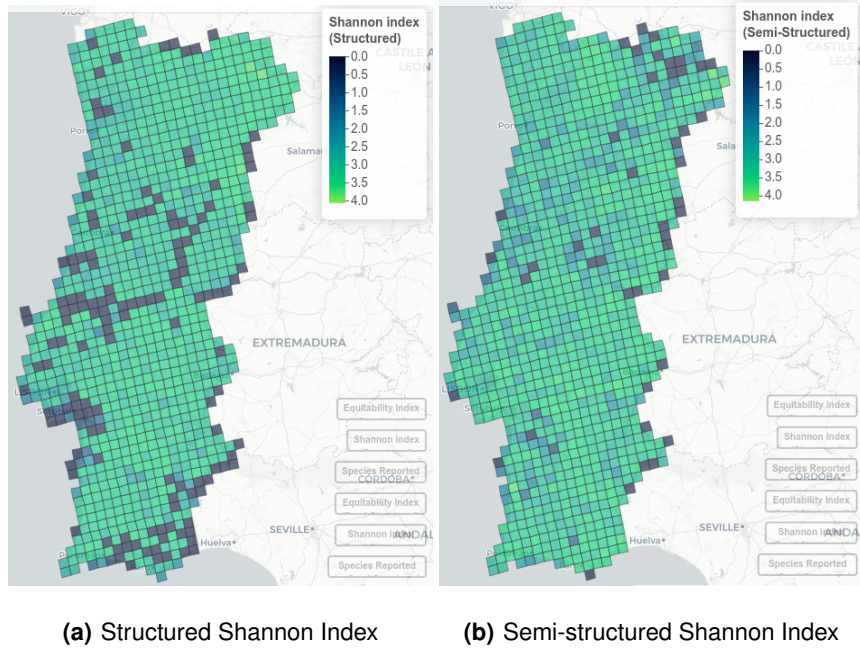
where:

$M$  the number of groups (*i.e.*, number of different species reported)

$P_i$  corresponds to the proportion of the entire community made up of species  $i$

The higher the value of  $H$ , the higher the diversity of species in a particular community. Conversely, a value of  $H = 0$  indicates that the community either has one, no species present. Since checklists with individuals count marked as "X" count as 0 towards the quantity of individuals reported, there is a possibility of having species that don't count towards this metric - which is the case for grid with ID 980, in the *Bragança* district, featuring 40 species reported, where none have counted species leading to a Shannon Index of 0.

Compared to simply mapping the number of different species, this metric goes a step further by taking into account how common those species that were reported are.



**Figure 4.19:** Comparison structured and semi-structured *Shannon Indexes*, over the atlas timeframe.

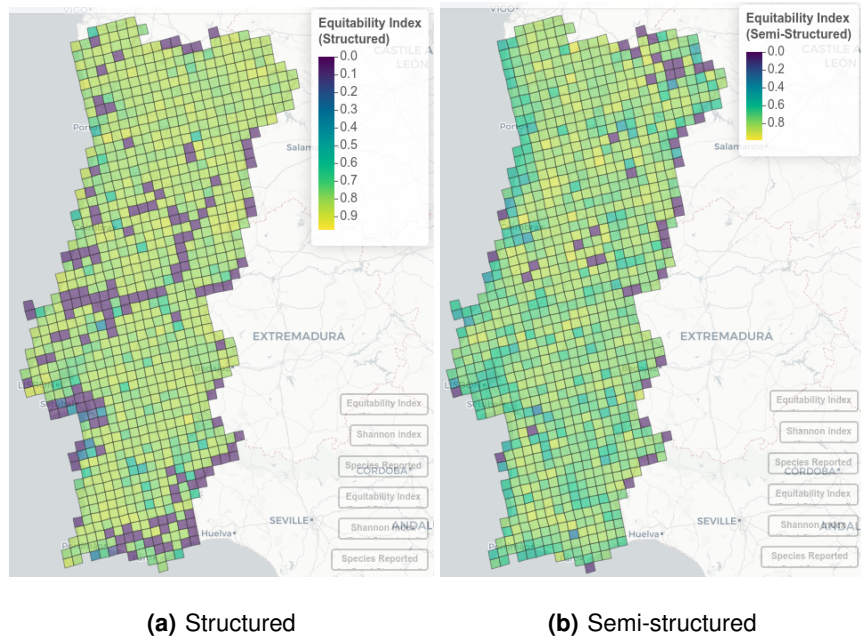
Again, the cells that haven't been surveyed by the structured approach are even more noticeable with this color palette. Also, coming back to cell with the *Bragança* district - it is interesting to see that the structured approach throughout the whole area features a higher diversity index than in the semi-structured one, with many misrepresentation occurring in that area in the the semi-structured map.

Besides this index, may also be useful to consider the **Shannon Equitability Index**, making use of the previously calculated Shannon Index, to get an idea of evenness, using the following formula:

$$E_H = \frac{H}{\ln M} \quad (4.6)$$

$H$  the calculated Shannon Index value  
 $M$  again, the number of groups (*i.e.*, number of different species reported)

This index is able to depict the degree of "evenness" across the different species reported, quantifying how similar the abundances of different species are in each community.

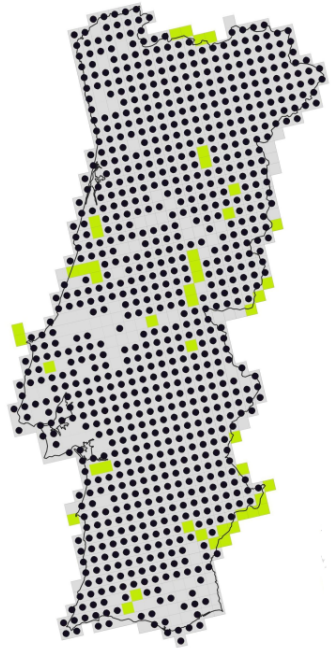


**Figure 4.20:** Comparison structured and semi-structured *Shannon Equitability*, over the atlas timeframe.

One interesting aspect of this map can be seen in the semi-structured one, where the coast of Portugal has definitely lower level of "evenness", meaning that the number of individuals vary at a greater scale, which could be result of the higher number of checklists. This is not observed in the structured approach, where the counts are more balanced.

Another popular ecologic metric used to measure species richness that was tested was the *Simpson's Diversity Index*, although the results were very similar to Shannon's.

The structured approach above, clearly depicts, as mentioned in other maps, that it has some grids where no observation has been made. We ended up being able to confirm this with the map below:



**Figure 4.21:** Breeding Bird Atlas' coverage and global search effort, following the same grid layout. Image kindly provided by Pedro Cardia.

The map's grid cells with a dot indicate that the cell has been surveyed according to the *Breeding Bird Atlas* protocol; the lime green cells indicate that the sum of the effort put on by semi-structured checklists, during the atlas time period, is lower than the minimum required by the structured approach to be considered surveyed - meaning that it can't be considered as visited by the atlas; and those cells painted grey denote the opposite. In any case, this serves as a confirmation that those subsquares that have not been surveyed by the structured checklists (with no dots present) are indeed correctly displayed in our analysis - showing that the structured checklists don't cover the whole territory, as well as verifying that the data from the semi-structured checklists on eBird are also taken into account in the atlas program.

# 5

## The Tool Proposed

### Contents

---

<b>5.1 R's Shiny Framework . . . . .</b>	<b>53</b>
<b>5.2 Development Workflow . . . . .</b>	<b>54</b>
5.2.1 Acquiring the Data and Analysis . . . . .	54
5.2.2 Pre-processing of the data . . . . .	54
5.2.3 Grid and Coordinate Reference System Used . . . . .	55
5.2.4 Database System . . . . .	56
5.2.5 Application's User Interface . . . . .	57
5.2.6 Deployment of the Tool . . . . .	58

---





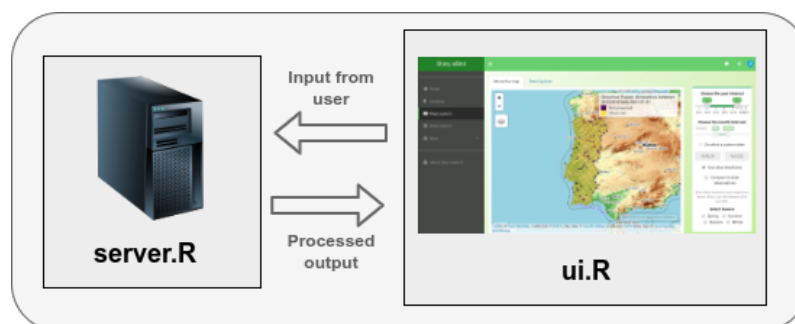
In order to develop an interactive visual tool capable of handling the freely available eBird data, the first go to was to explore different tools provided by the programming language R, our language of choice from the start, and commonly used in big data. It quickly became apparent that the best framework for this use case would be the framework **Shiny**. The name chosen for the tool was **Shiny eBird**, with the intentional typo being after the framework's name. In this chapter, we will go through the steps taken before, during development, and the outcome of said tool.

## 5.1 R's Shiny Framework

Shiny is a web application framework for R that simplifies the creation of reactive and responsive web applications containing data visualizations, making it possible to create web applications with virtually zero knowledge on HTML, CSS or Javascript languages, three pillars of web development. The logic behind a Shiny project consists on two main parts which can be implemented wither in the same or in the following separate R files:

- **ui.R file** , standing for User Interface, that solely focuses on the **frontend** of the application, *i.e.*, on the appearance of the tool, whose code is translated by the framework into HTML and CSS to be readily displayed on any browser;
- **server.R file** specifies all the processing and logic behind the scenes - from database access of the data to the creation of the graphs and maps in accordance with the user's input - which composes the **backend** of the application.

Both parties, naturally, are interconnected, resulting in the interactive outcome expected for the user, as illustrated from Figure 5.1. This rather simple and clean layout weighted as one of the factors in favor of using Shiny, apart from being an R package.



**Figure 5.1:** Simplified shiny app layout.

Notice that in the Figure above, the Server doesn't correspond to a real server, but rather to the server.R file, containing the instructions to the server it will be deployed.

## 5.2 Development Workflow

The diagram depicts the main steps for the development of the visual tool. It's important to lay emphasis on the fact that often exploring and perfecting of each stage happened concurrently, and thus it would be more proper to regard the workflow with a logical stance as well.

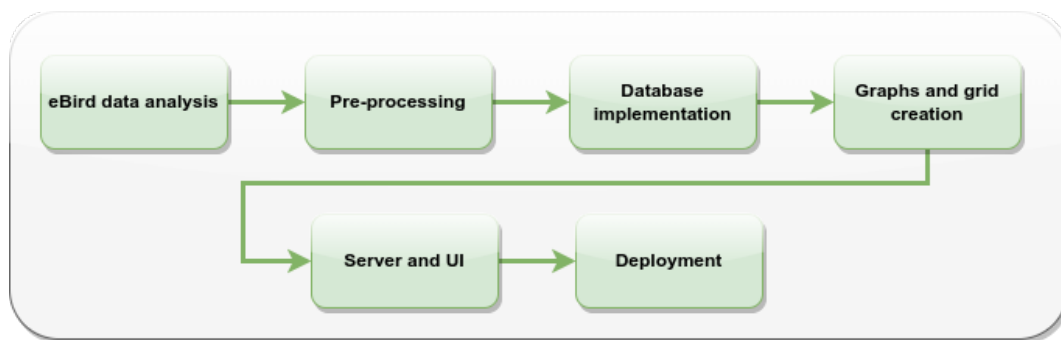


Figure 5.2: Diagram depicting the development workflow.

### 5.2.1 Acquiring the Data and Analysis

First and foremost, the initial step was to get access and download the data to be used in raw format. As explained briefly in Subsection 2.1.3, eBird has made this process considerably straightforward, by only requesting what data the user wants along with a concise description of its end goal. Our requested data was, then, downloaded via link received through email, in the a .txt file following a .csv table format, composed of all of eBird's observations in Portugal stored up until December 2021. A preliminary analysis of the data structure was carried to plan how to better handle it, leading to the next step.

### 5.2.2 Pre-processing of the data

The what was called the *pre-processing* stage relates to all the data handling of eBird data from the point it was acquired, until it was properly suitable to be used for the goals in mind. This key and often time consuming step would then allow to access eBird data locally more easily, and efficiently.

First and foremost, the program in which the code for all the work was created - Rstudio which is an integrated development environment (IDE) renowned for R programming dedicated to the R language.

A big part of initial data handling was done using the R package **auk** [57], mentioned in Subsection 2.1.3, with its biggest qualities being the filtration of data.

With the data imported onto R and filtered, many R packages of the collection **tidyverse** [66], namely **dplyr**, **tibble** and **purrr** to handle the data in dataframes (tabular structures where the raw data is

imported to in R), or **ggplot2** [62] for graphing.

One important aspect of this file was to have a static list of all the species ever encountered in the country, from which the dataset could then be organized. One could argue that the *auk* package could have been used for this purpose, but due to the sheer size of the dataset, even if divided into districts, could sometimes be too large for the memory available, specially on those with the PT-11 and PT-13 state codes, Lisbon and Porto's respectively. An easier way to get that list would be to use the eBird API with the function *ebirdregionspecies()*, via its R package **rebird** [70]. Sadly, this API would not be of much use besides this function, since the checklist data requests only goes up to 30 days.

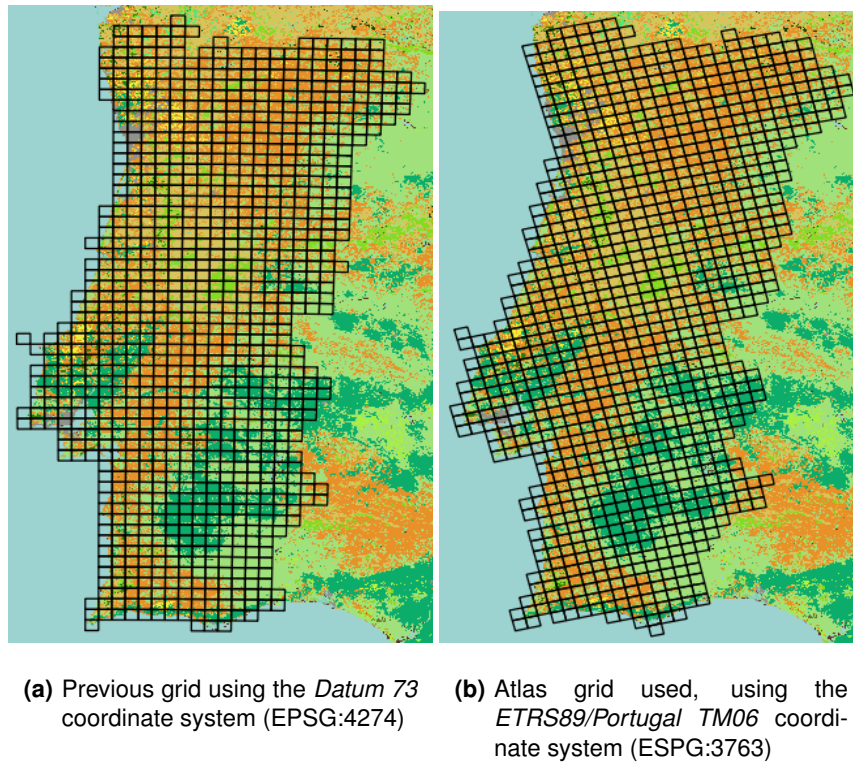
### 5.2.3 Grid and Coordinate Reference System Used

To systematize, keep track of surveyed areas and subsequently analyse the information via a systematic approach throughout the territory, a grid system is typically used. At a first experimental stage during this phase, the grid adopted to map the downloaded eBird data was the one used by the Atlas preceding the current *Breeding Bird Atlas* mentioned in 2.4.1 as we considered to include in our work with its data ranging from 1998 to 2005 [54].

As we came to discover, though, this followed a grid using a distinct coordinate reference system (CRS), which can be seen in Figure 5.3 (a). For this reason, only the grid used by the most recent atlas, present in eBird and containing richer information about the sampling effort, was used. This grid, whose files were generously shared by Pedro Cardia, follows the more recent *ETRS89/Portugal TM06* coordinate system bounded in mainland Portugal, with EPSG code 3763 (a standardized way of identifying, projecting, and performing transformations between them), and also features a grid composed of 10x10 kilometer cells which are then subdivided into 25 2x2 kilometer ones. Figure 5.3 (b) depicts the utilized grid, already seen in the maps present in Chapter 4.

In terms of programming, the transformation to the renowned World Geodetic System (also known as WGS84), coded EPSG:4326, is performed to allow its coordinates to be in accordance with those stored by eBird, and to further map it onto the interactive map. This was carried using the **sf** [60] R package.

Below, the figures depict the different grids considered.



**Figure 5.3:** Comparison between the grids considered. Images generated using QGIS.

## 5.2.4 Database System

Throughout the development, there were some abrupt changes in the approach taken regarding the way the information is structured. One of the biggest ones being the change from a file-based database, composed of thousand of files, to a database format, in a rather late stage of the work. Even though the usage of a this seems like an obvious choice, it wasn't at the time as the decision to proceed with a database meant that *auk*, very useful until then, would become unusable for data filtering during processing. On a positive note, it has allowed for a much quicker data access, where intensive operations that would have previously taken more than one hour to process, could be reduced to a minute, mostly thanks to not having to handle hundreds of files splitting the dataset. For this database, we used the relational database **SQLite** [67], which is *serverless* - meaning that it won't need a server to be connected to, making the process as straightforward as creating a database connection each time the Shiny application starts. The package **DBI** [68] was key for interacting with the database, including to create it and importing data, and subsequently for fetching the information from it.

The database layout followed the same one used - where each table corresponds to each folder in the old format - which divided the dataset into many multiple parts, by species, or by grid cell. However, this arrangement could be more efficient. There are a few cases where a single table would have been

preferred, such as when the output of the queries is of a size that can be handled at once. This would avoid to iterate through 1000 queries, which take more time compared to one query that groups the output by grid cell, in one query. This would reduce the processing time from around 20 to just a few seconds.

### 5.2.5 Application's User Interface

For the layout of the UI itself, the package **shinydashboard** [69] provided functions able to easily provide the dashboard look to the interface of the tool, with the different pages on a sidebar menu on the top left corner.

The tool is then composed of three pages - the first one being a homepage providing brief introduction on the objectives of the tool, some context on what it is representing. It also has described a handful of points regarding some of the technical details and concepts already alluded to in this document, namely the usage of the atlas, the grid utilized, *etc.*; the second one being where all the data is made available - with the options for the user to select the species, the timeframe, and whether or not to compare with the structured data; finally, the last and most brief page about the author.

The map where the metrics will be represented take a big part of the screen these was possible using the packages **leaflet** [64] together with the package **mapview** [59], which wraps the former and makes it even easier to map any type of data stored in dataframes while running R. These are the packages that managed to create the different maps listed in 4.3.3, and which are responsive to the user inside the application, by letting choose the grid cells and in some cases hover for additional information.

More features can be found in the application, besides the regular functioning of the tool. Among them are:

- **Interactive graphs** - As previously mentioned, two different approaches were taken to plot the graphs applied to each cell. One of them being the plotting below the map, where it would be possible to generate the **ggplot2** created graph using the **plotly** [65] R package. This package converts the ggplot2 figures into interactive ones, allowing to, for instance, display more information while hovering the plotted points, or zooming in on it.
- **Geolocalization** - If given the permission, the application can provide the user its approximate location, with the placement of a marker.
- **Map Tile Provider** - allows to change the representation of the map, such as with satellite imagery, offering the user's slight customization besides allowing for easier visualization as some map color palettes may be hard to view in the default mode.
- **Plot Download** - plots which are plotted using **plotly** can be downloadable, if that option is made available.

## 5.2.6 Deployment of the Tool

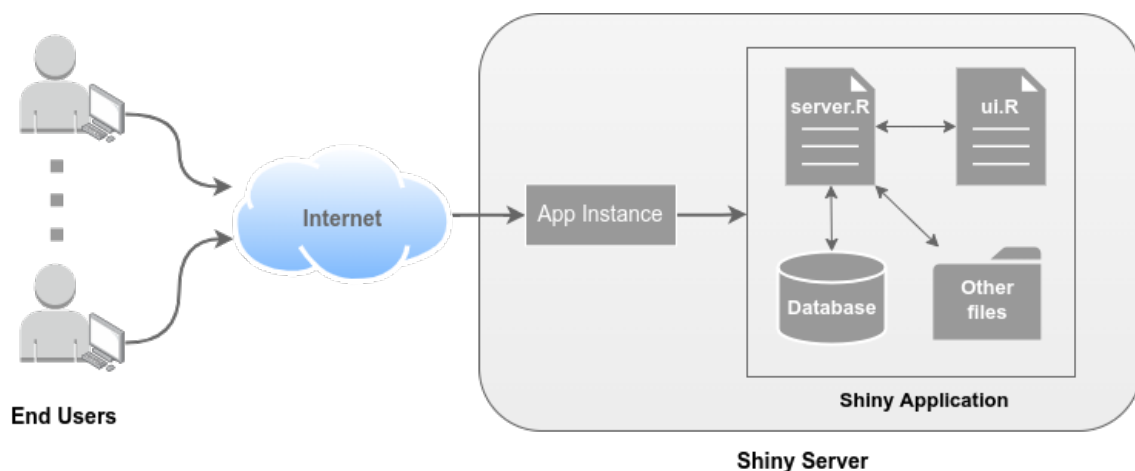
Last but not least, comes the deployment of the application. This was made using the online tool **Shinyapps.io**<sup>1</sup>, a self-service platform that hosts shiny applications on the web.

The drawback of this tool is the rather low monthly server uptime limit (*i.e.*, the amount of time in which one's server is online) set to the free tier users which narrows it to only 25 hours per month. The existing alternatives also require paid allowance for improving its availability.

In order to deploy the application via RStudio, one must use the **rsconnect** [71] package in a rather simple one time process of configuring the local Shiny application and binding it with the *shinyapps.io* account. Once this setup is complete, the application is one command away of being uploaded and updated with the function `deployApp()`, which will then will make it available at the following address:

[https://ebirdbias.shinyapps.io/ebird\\_tese\\_workspace/](https://ebirdbias.shinyapps.io/ebird_tese_workspace/)

With the tool deployed, the architecture's overview of the application can be depicted as in the the diagram below:



**Figure 5.4:** User-Shiny server architecture.

Where an instance is run to serve requests to a Shiny application from the end users - and more may be created if the traffic justifies it. The *Other files* here depicted may include any files used by the application, such as media (contained in a folder named */www*), or simply other R files with helper functions to facilitate coding in the main files. In our case, our project contains the database, the grid files for mapping, among other R files containing auxiliary functions used to process the data.

---

<sup>1</sup><https://www.shinyapps.io/>

# 6

## Conclusion

### Contents

---

6.1 Concluding Remarks . . . . .	61
6.2 Limitations and Future Work . . . . .	61

---





## 6.1 Concluding Remarks

One of the main objectives of this work was to be able to interactively display information about how the different regions are being reported, and with that characterize the structured and semi-structured approaches that have been considered, and highlight its irregularities. The mapping and graphing of these approaches using the presented set of distinct metrics, displayed in different formats, have allowed to reveal interesting results.

In a first phase, the structured data was expected to be regarded as the "correct" data source, due to its rigorous nature, to then be used to evaluate semi-structured data performance. However, in many fronts, the grid cells that were observed by semi-structured observations, revealed to have richer information about bird species - often due to its greater amount of data gathered and larger geographical coverage, specially in urban centers and in the most notorious hotspots for birdwatcher. Still, in the cells where the eBird volunteer participation is lower, such as in northeastern Portugal and in some inland areas, the structured observations managed to keep up and even surpass the outcome of semi-structured's both in the number and richness of the species observed. Besides the number of different lists reported, metrics that were used to describe the agreement between the approaches - in particular Observable Agreement - have also shown that the semi-structured checklists lead the way in terms of the different number species reported.

Collectively, semi-structured observations often outperformed Atlas' data in areas where eBird's community is more active - even though it's more prone to the different types of bias - whereas in some more remote areas the structured approach has the upper hand. As Galván et al. (2021) [56] put it: no bird database is perfect.

## 6.2 Limitations and Future Work

With regards to the outcome of this work, we believe that there was a slight shift in the course of our work that lead us to a more focus in the comparison of structured and semi-structured data, whereas initially it was more aimed at complex statistical models capable of modeling distributions. Throughout the work, our curiosity in using the structured approach grew stronger, specially given that the data structure was the same. Also, with the *Breeding Bird Atlas'* methodological approach having a rather low number of checklists in various locations, the comparison between the two via a distribution model such as the STEM that were mentioned, didn't seem to be balanced in terms of spatial coverage and number of checklists, steering the work towards a more direct juxtaposition of the two types of data.

Also, the outcome of the tool turned out to become slightly further directed to a more into a niche target audience, with its best use for those already familiarized with the metrics that ended up being used (and perhaps also for those familiarized with the grid system used). At first, the tool was conceptualized

to grow into a more eBird user friendly tool to be used in any scenario among the eBird community. This would be due to some unawareness regarding Shiny's limitations mixed with a slight misjudgement on our part on how much data actually had to be processed for each metric - specially in those cases that require the processing of data for the whole country's grid at once - and the prioritization of metrics that would allow the analysis carried. However, even with some limitations, many obstacles were overcome.

Regarding the future work, there are other possible approaches to the problem at hand, such as:

- The usage of eBird data to create a statistical model capable of modelling distribution and abundance of species.
- Assess the rarity of the species reported by either approaches, and perform a deeper analysis on grid cells that unexpectedly reported a considerable smaller amount of species in the structured checklists.
- Include a structured comparison with Incidental and Historical data as well, regarding it as a less semi-structured approach compared to this work's semi-structured data.

Considering the endless number of possibilities of handling eBird's information-rich data, the following points consist of some ideas that were at some point during the development considered to be implemented but ended up being abandoned due to time constraints. However, these could be explored further since they seemed achievable in the context of Shiny and R:

- Mapping on a selected cell on the grid the 2x2 kilometer grid with pinpointed information about the checklists locations within that cell, allowing for the user to view exactly where the data were surveyed. It could be accompanied with visual information about the number of checklists present at each spot, such as by using differently sized points on the map, for an easier way of identifying more popular areas.
- Being able create a downloadable report of the graphs and maps that were presented to the user.
- Displaying more information about a species in the case of a single one being selected. Information such as an image or brief description of the species would enrich the user experience, and seems feasible via the R package Wikipedia API wrapper called WikipediR<sup>1</sup>.
- Implementing a "Data Explorer" allowing the user to examine in a table format all the checklists' information present in the selected grid cell.
- Optimization of the database usage, by minimizing the somewhat unnecessary number of queries executed each time a user generates a new map. Also, other ways of structuring its data could also improve the processing rate.

---

<sup>1</sup> <https://github.com/Ironholds/WikipediR>

# Bibliography

- [1] Heigl, F., Kieslinger, B., Paul, K., Uhlik, J., Dörler, D. Opinion: Toward an international definition of citizen science. *Proc Natl Acad Sci U S A*. 2019 Apr 23;116(17):8089-8092. doi: 10.1073/pnas.1903393116. PMID: 31015357; PMCID: PMC6486707.
- [2] Miller-Rushing, A., Primack, R. and Bonney, R. (2012), The history of public participation in ecological research. *Frontiers in Ecology and the Environment*, 10: 285-290. <https://doi.org/10.1890/110278>
- [3] Amano, T., Lamming, J., Sutherland, W., Spatial Gaps in Global Biodiversity Information and the Role of Citizen Science, *BioScience*, Volume 66, Issue 5, 01 May 2016, Pages 393–400, <https://doi.org/10.1093/biosci/biw022>
- [4] Callaghan, C., Martin, J., Major, R., Kingsford, R. (2018). Avian monitoring - comparing structured and unstructured citizen science. *Wildlife Research*. 45. 176-184. 10.1071/WR17141.
- [5] Scistarter. (2021). Project Finding Tool. <https://Scistsarter.org/finder>
- [6] Sullivan, B., Aycrigg, J., Barry, J., Bonney, R., Bruns, N., Cooper, C., Damoulas, T., Dhondt, A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W., Iliff, M., Lagoze, C., La Sorte, F., Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*. 169. 31-40. <https://doi.org/10.1016/j.biocon.2013.11.003>.
- [7] Dickinson, J.L., Shirk, J., Bonter, D., Bonney, R., Crain, R.L., Martin, J., Phillips, T. and 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. <https://doi.org/10.1890/110236>
- [8] Altwegg, R., Nichols, J.D. Occupancy models for citizen-science data. *Methods Ecol Evol*. 2019; 10: 8– 21. <https://doi.org/10.1111/2041-210X.13090>

- [9] Sullivan, B.L., C.L. Wood, M.J. Iliff, R.E. Bonney, D. Fink, and S. Kelling. 2009. eBird: a citizen-based bird observation network in the biological sciences. *Biological Conservation* 142: 2282-2292.
- [10] La Sorte, F., Lepczyk, C, Burnett, J., Hurlbert, A., Tingley, M., Zuckerberg, B. (2018). Opportunities and challenges for big data ornithology. *The Condor*. 120. 414-426. 10.1650/CONDOR-17-206.1.
- [11] GBIF.org (2021), GBIF Home Page. Available from: <https://www.gbif.org>
- [12] Kelling, S., Yu, J., Gerbracht, J., Wong, W., "Emergent Filters: Automated Data Verification in a Large-Scale Citizen Science Project," 2011 IEEE Seventh International Conference on e-Science Workshops, 2011, pp. 20-27, doi: 10.1109/eScienceW.2011.13.
- [13] Kelling, S., Gerbracht, J., Fink, D., Lagoze, C., Wong, W.-K., Yu, J., Damoulas, T., Gomes, C. (2012). A Human/Computer Learning Network to Improve Biodiversity Conservation and Research. *AI Magazine*, 34(1), 10. <https://doi.org/10.1609/aimag.v34i1.2431>
- [14] eBird Basic Dataset. Version: EBD\_relMar-2021. Cornell Lab of Ornithology, Ithaca, New York. Mar 2021.
- [15] Fink, D., T. Auer, A. Johnston, M. Strimas-Mackey, O. Robinson, S. Ligocki, W. Hochachka, C. Wood, I. Davies, M. Iliff, L. Seitz. 2020. eBird Status and Trends, Data Version: 2019; Released: 2020. Cornell Lab of Ornithology, Ithaca, New York. <https://doi.org/10.2173/ebirdst.2019>
- [16] Strimas-Mackey, Miller, E., Hochachka, W. (2018). auk: eBird Data Extraction and Processing with AWK. R package version 0.3.0. <https://cornelllabofornithology.github.io/auk/>
- [17] Roche, J., Bell, L., Galvão, C., Golumbic, Y., Kloetzer, L., Knoben, N., Laakso, M., Lorke, J., Mannion, G., Massetti, L., Mauchline, A., Pata, K., Ruck, A., Taraba, P., Winter, S. (2020). Citizen Science, Education, and Learning: Challenges and Opportunities. *Frontiers in Sociology*. 5. 10.3389/fsoc.2020.613814.
- [18] LaDeau, S.L., Han, B.A., Rosi-Marshall, E.J. et al. The Next Decade of Big Data in Ecosystem Science. *Ecosystems* 20, 274–283 (2017). <https://doi.org/10.1007/s10021-016-0075-y>
- [19] Elith, J., Leathwick, J.R., 2009, Species Distribution Models: Ecological Explanation and Prediction Across Space and Time: *Annual Review of Ecology, Evolution, and Systematics*, v. 40, iss. 1, 677–697 p.
- [20] Fink, D., Hochachka, W.M., Zuckerberg, B., Winkler, D.W., Shaby, B., Munson, M.A., Hooker, G., Riedewald, M., Sheldon, D., Kelling, S. Spatiotemporal exploratory models for broad-scale survey data. *Ecol Appl*. 2010 Dec;20(8):2131-47. doi: 10.1890/09-1340.1. PMID: 21265447

- [21] La Sorte, F.A., Jetz, W. Avian distributions under climate change: towards improved projections. *J Exp Biol* 15 March 2010; 213 (6): 862–869. doi: <https://doi.org/10.1242/jeb.038356>
- [22] La Sorte, F.A., Fink, D., Blancher, P.J., Rodewald, A.D., Ruiz-Gutierrez, V., Rosenberg, K.V., Hochachka, W.M., Verburg, P.H., Kelling, S. Global change and the distributional dynamics of migratory bird populations wintering in Central America. *Glob Chang Biol.* 2017 Dec;23(12):5284-5296. doi: 10.1111/gcb.13794. Epub 2017 Jul 24. PMID: 28736872.
- [23] Callaghan, C.T., Gawlik, D.E. (2015), Efficacy of eBird data as an aid in conservation planning and monitoring. *J. Field Ornithol.*, 86: 298-304. <https://doi.org/10.1111/jofo.12121>
- [24] Ruiz-Gutierrez, V., Bjerre, E.R., Otto, M.C., et al. A pathway for citizen science data to inform policy: A case study using eBird data for defining low-risk collision areas for wind energy development. *J Appl Ecol.* 2021; 00: 1-8. <https://doi.org/10.1111/1365-2664.13870>
- [25] Johnston, A., Hochachka, W., Strimas-Mackey, M., Ruiz-Gutierrez, V., Robinson, O., Miller, E., Auer, T., Kelling, S., Fink, D. (2019). Best practices for making reliable inferences from citizen science data: case study using eBird to estimate species distributions. 10.1101/574392.
- [26] Kelling, S., Johnston, A., Bonn, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Fernandez, M., Hochachka, W.M., Julliard, R., Kraemer, R., Guralnick, R. Using Semistructured Surveys to Improve Citizen Science Data for Monitoring Biodiversity, *BioScience*, Volume 69, Issue 3, March 2019, Pages 170–179, <https://doi.org/10.1093/biosci/biz010>
- [27] Burgess, H., DeBey, L., Froehlich, H., Schmidt, N., Lambers, J., Tewksbury, J., Parrish, J. (2016). The science of citizen science: Exploring barriers to use as a primary research tool. *Biological Conservation.* 208. 10.1016/j.biocon.2016.05.014.
- [28] MacPhail, V., Colla, S. (2020). Power of the people: A review of citizen science programs for conservation. *Biological Conservation.* 249. 108739. 10.1016/j.biocon.2020.108739.
- [29] Tiago, P., Ceia-Hasse, A., Marques, T.A. et al. Spatial distribution of citizen science casuistic observations for different taxonomic groups. *Sci Rep* 7, 12832 (2017). <https://doi.org/10.1038/s41598-017-13130-8>
- [30] Geldmann, J., Heilmann-Clausen, J., Holm, T.E., Levinsky, I., Markussen, B., Olsen, K., Rahbek, C. and Tøttrup, A.P. (2016), What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity Distrib.*, 22: 1139-1149. <https://doi.org/10.1111/ddi.12477>
- [31] Strimas-Mackey, M., W.M. Hochachka, V. Ruiz-Gutierrez, O.J. Robinson, E.T. Miller, T. Auer, S. Kelling, D. Fink, A. Johnston. 2020. Best Practices for Using eBird Data. Version 1.0.

- <https://cornelllabofornithology.github.io/ebird-best-practices/>. Cornell Lab of Ornithology, Ithaca, New York. <https://doi.org/10.5281/zenodo.3620739>
- [32] Courter, J., Johnson, R., Stuyck, C., Lang, B., Kaiser, E. (2012). Weekend bias in Citizen Science data reporting: Implications for phenology studies. *International journal of biometeorology*. 57. 10.1007/s00484-012-0598-7.
  - [33] Laney, J.A., Hallman, T.A., Curtis, J.R., Robinson, W.D. The influence of rare birds on observer effort and subsequent rarity discovery in the American birdwatching community. *PeerJ*. 2021 Jan 21;9:e10713. doi: 10.7717/peerj.10713. PMID: 33552730; PMCID: PMC7827972.
  - [34] Troudet, J., Grandcolas, P., Blin, A. et al. Taxonomic bias in biodiversity data and societal preferences. *Sci Rep* 7, 9132 (2017). <https://doi.org/10.1038/s41598-017-09084-6>
  - [35] Chen, D. and Gomes, C. (2019). Bias Reduction via End-to-End Shift Learning: Application to Citizen Science. *Proceedings of the AAAI Conference on Artificial Intelligence*. 33. 493-500. 10.1609/aaai.v33i01.3301493.
  - [36] Ponti, M., Hillman, T., Kullenberg, C., Kasperowski, D. (2018). Getting it Right or Being Top Rank: Games in Citizen Science. *Citizen Science: Theory and Practice*. 3. 10.5334/cstp.101.
  - [37] Xue, Y., Davies, I., Fink, D., Wood, C., Gomes, C. 2016. Avicaching: A Two Stage Game for Bias Reduction in Citizen Science. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 776–785.
  - [38] Zizka, A., Antonelli, A., Silvestro, D. (2021), sampbias, a method for quantifying geographic sampling biases in species distribution data. *Ecography*, 44: 25-32. <https://doi.org/10.1111/ecog.05102>
  - [39] Pearson, R. (2010). Species' Distribution Modeling for Conservation Educators and Practitioners. *Lessons in Conservation*. 3.
  - [40] Jun, Y, Wong, W., Hutchinson, R. (2010). Modeling Experts and Novices in Citizen Science Data for Species Distribution Modeling. *Proceedings - IEEE International Conference on Data Mining, ICDM*. 10.1109/ICDM.2010.103.
  - [41] Matutini, F., Baudry, J., Pain, G., Sineau, M., Pithon, J. How citizen science could improve species distribution models and their independent assessment. *Ecol Evol*. 2021; 11: 3028– 3039. <https://doi.org/10.1002/ece3.7210>
  - [42] Kelling, S., Fink, D., La Sorte, F. A., Johnston, A., Bruns, N. E., and Hochachka, W. M. (2015). Taking a 'Big Data' approach to data quality in a citizen science project. *Ambio* 44, 601–611. doi: 10.1007/s13280-015-0710-4

- [43] Fink, D., Damoulas, T., Bruns, N. E., La Sorte, F. A., Hochachka, W. M., Gomes, C. P., Kelling, S. (2014). Crowdsourcing Meets Ecology: Hemisphere-Wide Spatiotemporal Species Distribution Models. *AI Magazine*, 35(2), 19-30. <https://doi.org/10.1609/aimag.v35i2.2533>
- [44] Daniel, D., Damoulas, T., Dave, J. (2013). Adaptive Spatio-Temporal Exploratory Models: Hemisphere-wide species distributions from massively crowdsourced ebird data. *Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI 2013*. 1284-1290.
- [45] Johnston, A., Fink, D., Reynolds, M.D., Hochachka, W.M., Sullivan, B.L., Bruns, N.E., Hallstein, E., Merrifield, M.S., Matsumoto, S. and Kelling, S. (2015), Abundance models improve spatial and temporal prioritization of conservation resources. *Ecological Applications*, 25: 1749-1756. <https://doi.org/10.1890/14-1826.1>
- [46] Fink, D., Auer, T., Johnston, A., Ruiz-Gutierrez, V., Hochachka, W. M., and Kelling, S.. 2020. Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications* 30( 3):02056. 10.1002/eap.2056
- [47] Palacio, R.D., Negret, P., Velásquez-Tibata, J., Jacobson, A. (2020). A data-driven geospatial workflow to improve mapping species distributions and assessing extinction risk under the IUCN Red List. 10.1101/2020.04.27.064477.
- [48] Hefley, T.J., Hooten, M.B. Hierarchical Species Distribution Models. *Curr Landscape Ecol Rep* 1, 87–97 (2016). <https://doi.org/10.1007/s40823-016-0008-7>
- [49] Roth, R. (2013). Interactive Maps: What we know and what we need to know. *Journal of Spatial Information Science*. 6. 59-115. 10.5311/JOSIS.2013.6.105.
- [50] Midway, S.R. (2020). Principles of Effective Data Visualization. *Patterns*, Volume 1, Issue 9. <https://doi.org/10.1016/j.patter.2020.100141>.
- [51] Adlard, E., A. I. Fagundes. (2020). Iberian Network for Seabirds and Marine Mammals - Portugal mainland counts during 2019. *Sociedade Portuguesa para o Estudo das Aves*, Lisboa (non published report)
- [52] Alonso, H., Coelho, R., Gouveia, C., Rethoré, G., Leitão, D., Teodósio, J. (2021). Relatório do Censo de Aves Comuns 2004-2020. *Sociedade Portuguesa para o Estudo das Aves*, Lisboa (relatório não publicado).
- [53] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46.

- [54] Instituto de Conservação de Natureza e da Biodiversidade (Lisboa). (2008). Atlas das aves nidificantes em Portugal:(1999-2005). Assírio e Alvim.
- [55] Shannon, C.E. (1948), A Mathematical Theory of Communication. Bell System Technical Journal, 27: 379-423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- [56] Galván, S., Barrientos, R., Varela, S. (2021). No Bird Database is Perfect: Citizen Science and Professional Datasets Contain Different and Complementary Biodiversity Information. *Ardeola*, 69(1), 97-114.
- [57] Strimas-Mackey, M., Miller, E., Hochachka, W. (2018). auk: eBird Data Extraction and Processing with AWK. R package version 0.3.0. <https://cornelllabofornithology.github.io/auk/>
- [58] Strimas-Mackey, M., Ligocki, S., Auer, T., Fink, D. (2021). ebirdst: Tools for loading, plotting, mapping and analysis of eBird Status and Trends data products. R package version 1.0.0. <https://cornelllabofornithology.github.io/ebirdst/>
- [59] Appelhans, T., Detsch, F., Reudenbach, C. and Woellauer, S. (2021). mapview: Interactive Viewing of Spatial Data in R. R package version 2.10.0. <https://CRAN.R-project.org/package=mapview>
- [60] Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* 10 (1), 439-446, <https://doi.org/10.32614/RJ-2018-009>
- [61] Oksanen, J., Simpson, G.L., Blanchet, F. G., et.al (2022). vegan: Community Ecology Package. R package version 2.6-2. <https://CRAN.R-project.org/package=vegan>
- [62] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- [63] Chang, W., Cheng, J., Allaire, JJ, Sievert, C., Schloerke, B, Xie, Y., Allen, J., McPherson, J., Dipert, A. and Borges, B. (2021). shiny: Web Application Framework for R. R package version 1.7.1. <https://CRAN.R-project.org/package=shiny>
- [64] Cheng, J., Karambelkar, B. and Xie, Y. (2022). leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.1.1. <https://CRAN.R-project.org/package=leaflet>
- [65] C. Sievert. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC Florida, 2020.
- [66] Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686, <https://doi.org/10.21105/joss.01686>
- [67] Owens, M. (2006). The definitive guide to SQLite. Apress.



- [68] Hadley Wickham and Kirill Müller (2021) R Special Interest Group on Databases (R-SIG-DB) DBI: R Database Interface. R package version 1.1.2. <https://CRAN.R-project.org/package=DBI>
- [69] Chang, W., and Borges, B. (2021). shinydashboard: Create Dashboards with 'Shiny'. R package version 0.7.2. <https://CRAN.R-project.org/package=shinydashboard>
- [70] Maia, R. et. al. (2021). rebird: R Client for the eBird Database of Bird Observations. R package version 1.3.0. <https://CRAN.R-project.org/package=rebird>
- [71] Atkins, A., McPherson, J., and Allaire, JJ (2021). rsconnect: Deployment Interface for R Markdown Documents and Shiny Applications. R package version 0.8.25. <https://CRAN.R-project.org/package=rsconnect>



