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

Rodrigo Freitas
rodrigo.d.freitas@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

June 2022

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. One of the most notable platforms created for this purpose, eBird, is our case study for 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. 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

1. Introduction

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].

This 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[1]. 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.

[1] https://www.birds.cornell.edu/home/

1.1. Goals

This work seeks 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.

2. Background

2.1. 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 1, 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[2].

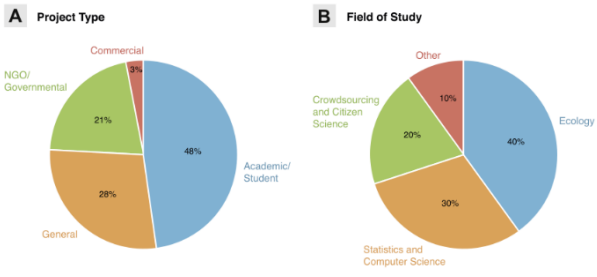


Figure 1: Circle charts illustrating the end-uses of 1100 eBird data requests. Image obtained from Sullivan et al. (2014) [6].

eBird acts as a reliable case study for several scientific education initiatives and raises awareness among the general public [17]. eBird's collected data has been making its contribution to studies and innovations across a wide span of fields [6], allowing for more accurate estimates across wide spatial and temporal scales.

[2]https://ebird.org/data/request

2.2. 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.

2.2.1 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], its 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].

2.3. eBird's Survey Structure

Two characteristics named in the Section **??** 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.

2.4. 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 **??**, 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*[3], 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*[4]. 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 [5].

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 subsquares, 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.

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[6] that shows through an interactive timeline bird migration movements of four bird species in continental Europe.

Besides eBird's visualization tools, another rather

[3] https://spea.pt/censos/censo-aves-comuns/

[4] https://www.spea.pt/wp-content/uploads/2020/12/Metodologia-campo_v6_20201209.pdf

[5] https://spea.pt/censos/iii-atlas-aves-nidificantes/

[6] https://birdmigration.climate.copernicus.eu/the-progression-of-bird-migration

valuable source of inspiration was the R Shiny gallery[7], which is comprised of contributions from the community showcasing examples of interactive tools built using the Shiny framework[63] 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 said framework.

3. Related Workd

3.1. Data Visualization

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[8] that shows through an interactive timeline 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[9], which is comprised of contributions from the community showcasing examples of interactive tools built using the Shiny framework[63] 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 said framework.

4. Approach

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**), were not taken into account for our solution, since the former, as the name suggests, 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.

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 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.

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

[7] https://shiny.rstudio.com/gallery/

[8] https://birdmigration.climate.copernicus.eu/the-progression-of-bird-migration

[9] https://shiny.rstudio.com/gallery/

Blackbird (*Turdus merula*) was chosen for these graphs due to being a common species in checklists throughout the territory.

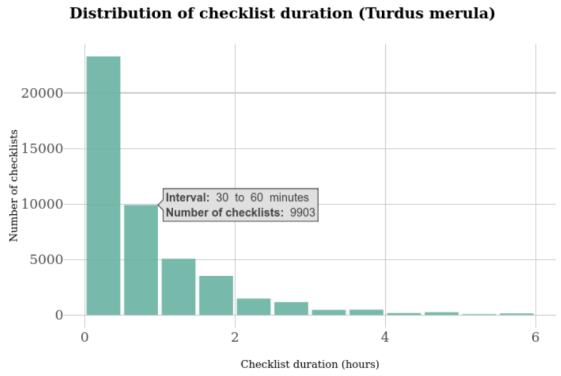


Figure 2: Distribution of the semi-structured checklists' duration where the species *Turdus merula* was reported.

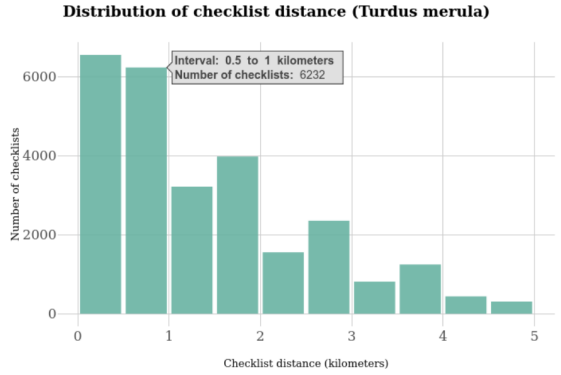


Figure 3: Distribution of the semi-structured checklists' distance where the species *Turdus merula* was reported.

4.3.2 For Each Grid Cell

Graphs were also created to be plotted with the data of a specific grid cell, which 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. Plots of checklists per grid cell feature: a plot that 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; the average number of checklists reported per weekday for the grid cell number 10 (here with a notable spike on Saturday), containing with info on each days average checklists duration and species reported; the average number of different species reported per duration and distance bins, pictured below for the distance, with additional information on the number of checklists and standard deviation.

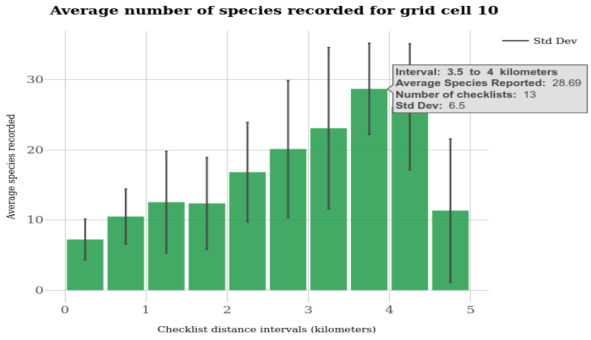
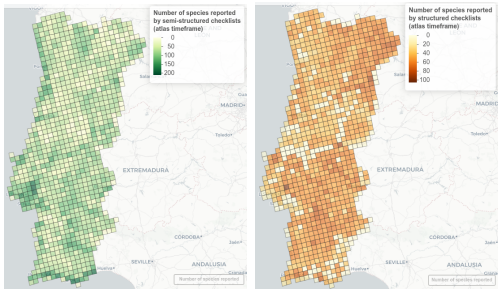


Figure 4: Average number of species reported per **distance** bin.

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. 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**, which indicates per grid cell the number of different species reported. It tells us that for the structured approach that there have been some grid cells which have not been surveyed by the structured protocol, and highlight the difference in the quantity of reported species.



(a) Semi-structured (b) Structured

Figure 5: Maps depicting the total number of species reported by semi-structured and structured checklists, during the atlas timeframe.

- **Difference between the number of species reported**, indicating the 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.

- **Presence and absence of species** indicating where a given species has been reported, per grid cell, for two scenarios: observations only for the semi-structured approach with a simple grid on whether or not the selected species was observed, and that together with the mapping of the presence of the structured approach, during the atlas timeframe, if the user opts for the comparison.
- **Number of checklists submitted and completeness percentage**, introducing 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. It made it 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.

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.

To calculate Cohen, a k by k confusion matrix is defined, 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:

$$P_o = \frac{1}{N}\sum_{j=1}^{k} f_{jj}, \tag{1}$$

$$r_i = \sum_{j=1}^{k} f_{ij}, \forall i, \text{ and } c_j = \sum_{i=1}^{k} f_{ij}, \forall j, \tag{2}$$

$$P_e = \frac{1}{N^2}\sum_{i=1}^{k} r_i c_i, \tag{3}$$

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 κ, which is applied to each grid cell on the map, is given by:

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{4}$$

The values of Cohen's Kappa can be interpret according a 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.

To get to the values of κ, a value N was needed for each grid cell, *i.e.*, the number of species ever recorded.

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, which drastically reduced the values of κ. This ended up suggesting that the agreement is not as strong as initially thought.

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.

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 and semi-structured compared to the set of species ever recorded, using the N defined above.

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

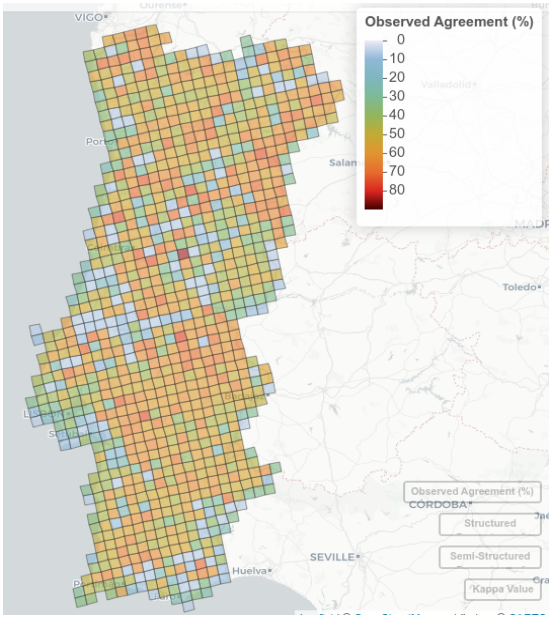


Figure 6: Observable agreement between structured and semi-structured observations, expressed as a percentage.

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 shows 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.

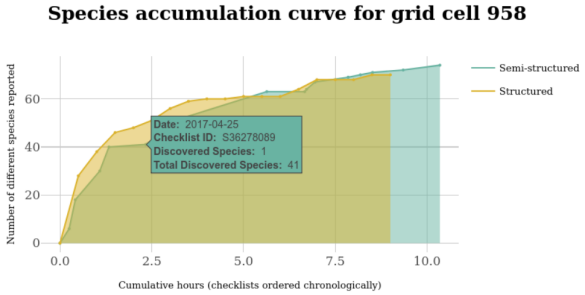


Figure 7: SAC plot comparing structured (in orange) and semi-structured checklists (in green).

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:

$$H = -\sum_{i=1}^{M} P_i \ln P_i \tag{5}$$

where:

- M is the number of groups (*i.e.*, number of different species reported)
- P_i is 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 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.

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:

$$E_H = \frac{H}{\ln M} \tag{6}$$

· 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.

5. The Tool Proposed

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 section, 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 with beautiful 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:

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.

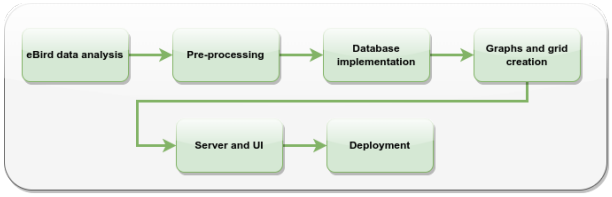


Figure 8: Diagram depicting the development workflow.

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. 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) renowed for R programming dedicated to the R language.

A big part of initial data handling was done using the R package **auk**[57], 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 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). 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 ESPG 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.

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.

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 the 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 connection each time the Shiny application starts. The package **DBI**[68] was key for interacting with the database, after being created from importing all data to one, to fetching information form it.

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] toghether 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.

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**[10], 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/

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

[10] https://www.shinyapps.io/

the the diagram below:

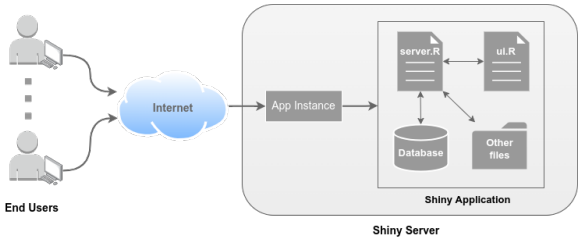


Figure 9: User-Shiny server architecture.

Where an instance is run to serve requests to a Shiny application from the end users. 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.

6. Future Work and Concluding Remarks

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, 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, regarding it as a less semi-structured approach.

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.
- 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[11].
- Implementing a "Data Explorer" allowing the user to examine in a table format all the checklists' information present in the selected 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.

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 it's 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.

[11] https://github.com/Ironholds/WikipediR

References

[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