Politecnico di Milano
Department of Civil and Environmental Engineering
Doctoral Program in Environmental and Infrastructure Engineering

POLITECNICO
MILANO 1863

Integration of Machine Learning And Citizen Science To Address The Challenges of Public Engagement and Data Validation

Doctoral Dissertation
Maryam Lotfian
Matr. 901559

Supervisor: **Prof. Maria Antonia Brovelli**
Co-Supervisor: **Prof. Jens Ingensand** (University of Applied Sciences and Arts Westerns Switzerland, HEIG-VD)
Tutor: **Prof. Giovanna Venuti**
Coordinator: **Prof. Riccardo Barzaghi**

XXXIII Cycle (2017 - 2021)

This page intentionally left blank

Abstract

The number of citizen science (CS) projects has grown significantly in recent years, owing to technological advancements. One important aspect of ensuring the success of a CS project is to consider and address the challenges in this field. Two of the main challenges in CS projects are sustaining participation and improving the quality of contributed data. Despite the studies that have been conducted to address these two challenges, there is still a need for new approaches, one of which is the use of artificial intelligence (AI) and machine learning (ML) in CS projects.

Therefore, the objective of this thesis was to investigate the integration of ML and CS, as well as the role of this integration in addressing CS challenges. A comprehensive review conducted in this study of motivational factors in CS projects indicated that *interest in learning about science* and *receiving feedback* were strong motivations among participants in the majority of CS projects. Typically, experts verify the data and provide feedback to participants. However, due to large amounts of data, this manual data verification can be time-consuming. Thus, in this research, it was investigated how the integration of ML and CS can, on the one hand, automate and speed up the data validation process, and on the other hand, increase participant engagement and sustain participation by providing real-time informative feedback.

To that end, a biodiversity CS project was implemented with the goal of collecting and automatically validating observations as well as providing participants with real-time feedback. ML algorithms were trained to model species distribution using environmental variables (e.g., land cover) and species data from an existing CS project, and then to validate a new contributed observation based on the likelihood of observing a species in a specific location. Furthermore, volunteers were given real-time feedback on the likelihood of observing a species in a particular location, as well as species habitat characteristics. Moreover, a user experiment was conducted, and the results indicated that participants with a higher number of contributions found the real-time feedback to be more useful in learning about biodiversity and stated that it increased their motivation to contribute to the project. Besides that, as a result of automatic data validation, only 10% of observations were flagged for expert verification, resulting in a faster validation process and improved data quality by combining human and machine power. Finally, based on the findings of the experiments and the discussions that followed, we made some recommendations for CS practitioners to consider before designing a new project or improving an existing one.

The future objective of this research is to focus more on the challenges of ML and CS integration, and to investigate how this integration can be applied in other CS fields besides biodiversity.

This page intentionally left blank

Acknowledgements

I would like to express my sincere appreciation to my supervisor, **Prof. Maria Brovelli**, for her guidance throughout my thesis and for assisting me in seeing the clear path from beginning to end. Being that she is an internationally recognized person in the geospatial world, she assisted me in connecting with many experts all over the world.

I would like to express my heartfelt gratitude to my co-supervisor, **Prof. Jens Ingensand**, for welcoming me into his lab and for his unwavering support of my PhD studies, as well as for assisting me in finding solutions to my problems with his wise suggestions. We met weekly throughout my thesis, and even when his schedule was packed and I suggested we skip this week's meeting, his exact words were "No, we should find a time to meet, your thesis is very important". Furthermore, in addition to scientifically guiding me, whenever I was panicking or overstressed, he was helping me to break down the problems and calm down with his very supportive and warm words.

I would like to express my gratitude to **Dr. Sven Schade** from the Joint Research Centre and **Prof. Ali Mansourian** from university of Lund for taking the time to review the thesis and provide very valuable and constructive feedback.

I would like to thank **Prof. Andres Perez-Uribe** from the TIC department at HEIG-VD for his kind assistance when I had machine learning questions, and **Prof. Olivier Ertz** from the COMEM department at HEIG-VD for his very helpful feedback, particularly on UX design, during my PhD.

I would like to thank my two colleagues, Thibaud Chassin (my coffee buddy) and Nicolas Blanc, with whom I shared the best and worst moments of my PhD, and who were always willing to help and brainstorm with me when I got stuck. Moreover, I would like to thank my current and former colleagues at HEIG-VD and PoliMi: Romain Sandoz, Dr. Daniele Oxoli, Dr. Carlo Biraghi, Gorica Brati, Katarina Spasenovic, Juan Fernando Toro Herrera, Simon Oulevay, Sarah Composto, Dr. Timothée Produit, Dr. Eylül Kilsedar, and everyone else who helped me in any way.

I would like to express my gratitude to my amazing friends Yasaman, Hemad, Simon, Leona, Kanaha, Faiza, Maria Elena, Shokouh, Mozhdeh, and all of my other friends for their help and moral support.

Last but not least, I would like to thank my family, particularly my parents, who have always been very supportive, and my sister, Fereshteh, who has always been warm and reassuring. I would like to give special thanks to my husband, Andres Felipe, for everything, from brainstorming ideas that could help me implement my application to understanding my ups and downs due to stress and supporting me with his unconditional love.

"I like crossing the imaginary boundaries people set up between different fields—it's very refreshing. There are lots of tools, and you don't know which one would work. It's about being optimistic and trying to connect things."

Maryam Mirzakhani

Contents

Abstract **i**

Acknowledgements **iii**

List of Tables **ix**

List of Figures **xi**

1 Introduction **1**

1.1 Motivation 1

1.2 Thesis focus and objectives 4

1.3 Thesis outline 5

2 Citizen Science and Related Work **7**

2.1 Introduction 7

2.2 Definition of the terms 8

2.3 Citizen Science typology 11

2.4 Selected Challenges in Citizen Science 16

2.4.1 Challenge 1: Public engagement 16

2.4.2 Challenge 2: Data quality 20

2.5 Summary 24

3 Integrating Machine Learning In Citizen Science Projects 25

3.1 Introduction 25

3.2 Types of Machine Learning and Applications 28

3.3 The Influence of Machine Learning on Citizen Science Project's Phases 31

3.3.1 Machine Learning for Engaging the Public and Sustaining Participation . 33

3.3.2 Machine Learning for Data Collection 34

3.3.3 Machine Learning for Data Validation 35

3.4 Selected Case Studies of Combination of Citizen Science and Machine Learning . 36

3.5 Benefits and Challenges of Integration of Machine Learning in Citizen Science . 42

3.5.1 Benefits and Challenges for Participant' Engagement 42

3.5.2 Benefits and Challenges for Data Quality 44

3.5.3 Ethics 45

3.6 Summary 46

4 Hypotheses 48

5 Public Engagement and Establishment of a Conceptual Framework For Motivational Factors 53

5.1 Introduction 53

5.2 Motivational factors 54

5.3 Citizen Science Case Studies of Analyzing Participants' Motivations 59

5.4 Motivational framework 66

5.5 Validation of the proposed framework using a case study 72

5.6 Discussion and evaluation of hypotheses 1 to 3 77

5.7 Summary 81

6 Data validation using machine learning 84

6.1 Introduction 84

6.2 Species Distribution Modeling 86

6.2.1 Species Distribution Modeling (SDM) generation 88

6.2.2 Presence-only versus Presence-absence 89

6.2.3 Common Species Distribution Modeling (SDM) algorithms 91

6.3 Case study: BioSenCS application 106

6.4 Image and date validation 114

6.5 Location validation 117

6.5.1 Data preparation for species distribution modeling 117

6.5.2 Spatial cross validation 124

6.5.3 Comparison of algorithms to generate species distribution models for our case study 129

6.5.4 Implementation of the Application Programming Interface (API) for location validation 141

6.6 User experiment 146

6.7 Discussion and evaluation of hypotheses 4 to 6 153

6.8 Summary 157

7 Discussion of Hypotheses and Recommendations for Citizen Science Practitioners 160

7.1 Introduction 160

7.2 Discussion of Hypotheses 161

7.3 Recommendations for Citizen Science Practitioners 165

8 Conclusions and Future Work **169**

8.1 Summary and Main Findings 169

8.2 Future Work 173

A Using Flickr Data to Perform Species Distribution Modeling: An Early Proposal **176**

B Other case studies of combining CS and ML **183**

B.1 AId Lakes: AI for water quality monitoring 183

B.2 MoDoS 190

List of Acronyms **192**

Bibliography **197**

List of Tables

3.1 Example of case studies of combination of machine learning and citizen science . 41

5.1 Some of the motivational factors to contribute to OpenStreetMap 58

5.2 Case studies reviewed in this article for each Citizen Science (CS) typology . . . 60

5.3 A framework to classify volunteers' motivation in citizen science projects 68

6.4 Environmental variables used in this study to generate SDM 123

6.5 Algorithms trained in this study to generate Species Distribution Modeling (SDM) 129

6.6 Average Rankings of the algorithms 134

B.1 Indicators of R-CNN model performance 189

List of Figures

1.1 Examples of Citizen Science Applications Displayed in Google Play Search . . . 3

2.1 An example of user-generated spatial content in Wikimapia project, displaying points with shop/store label in the city of Lausanne. *Source: https://wikimapia.org/* 8

2.2 Intersection of Citizen Science (CS) and Volunteered Geographic Information (VGI): Geographic Citizen Science (Haklay, 2013) 11

2.3 Levels of participation in Citizen Science (CS), adapted from (Haklay, 2013) . . 14

2.4 Typology of Citizen Science (CS) projects 16

2.5 Reader to leader framework, adapted from (Preece & Shneiderman, 2009) 18

2.6 Participation inequality in OpenStreetMap adapted from Wood's analysis(Wood, 2014) for the users that contributed at least one edit. Few registered users contribute the majority of data (blue side); the majority of users contribute less than 100 edits (orange side). 18

2.7 Various steps to encourage users to contribute to OpenStreetMap. Adapted from (Wood, 2014) 19

2.8 The four aspects of data evaluation. Adapted from (Balázs et al., 2021) 22

3.1 Relationship between Artificial Intelligence, Machine Learning, and Deep Learning. *Source: Adapted from ("Deep learning", 2021)* 30

3.2 A taxonomy showing the integration of Machine Learning (ML) and Citizen Science (CS) based on the three Citizen Science (CS) phases of public engagement, data collection, and data quality (own work). 32

3.3 Screenshot from Braindr application where citizen scientists are required to label the MRI images by selecting pass or fail. *Source: https://braindr.us/* 40

3.4 Benefits and challenges of combining Citizen Science (CS) and Machine Learning (ML) (own work) 43

5.1 Eyewire interface. *Source: https://eyewire.org/explore* 66

5.2 The interactive map in BioPocket application. Participants can check the type and location of actions, as well as the biodiversity points of interests around them 73

5.3 Distribution of BioPocket survey respondents within Switzerland. The regions with no responses are not labelled on the map 75

5.4 Motivation ranking to participate to BioPocket project. Motivational factors were given scores from 1 to 8, and the ranking was based on the average score give to each. 76

5.5 Correlation between motivation types based on the scores given by respondents. All the correlations are statistically significant with p-values < 0.001. 77

5.6 Boxplots of respondent's residence types versus motivation of spending time in the nature (on the left), and gaining recognition among others (on the right). The residence types are aggregated to two types, i.e., Apartment and Villa. . . . 78

5.7 Violin plots comparing the scores given to the motivation "spending time in the nature" among respondents living in apartments and villas 80

5.8 Violin plots depicting the distribution of scores (from 1 to 5, y-axis) assigned to location authorization and creating a new account or logging in with an existing account, versus education levels on the x-axis. 81

6.1 Steps to perform species distribution modeling 89

6.2 A statistical explanation of Maximum-entropy (MaxEnt) for ecologists. Adapted from (Elith et al., 2011), *Source: ("Maxent", n.d.)* 94

6.3 Architecture of an artificial neural network with two hidden layers. *Source: https://cs231n.github.io/neural-networks-1/* 95

6.4 Linear threshold unit. *source: https://www.oreilly.com/library/view/neural-networks-and/9781492037354/ch01.html* 96

6.5 The graph of Sigmoid function (blue line), and Tangent Hyperbolic (TanH) function (red line). x axis: weighted some of input neurons $z = W^t X$, y axis: the value of $\sigma(z)$ or $\tanh(z)$ 97

6.6 Rectified Linear Unit (ReLU) activation function, $ReLU(z) = max(0, z)$ 98

6.7 A very basic visual example of a decision tree model for land cover classification. *Source:* (Horning et al., 2010) 101

6.8 A basic visualization of random forest algorithm for classification/regression, *source: https://ai-pool.com/a/s/random-forests-understanding* 102

6.9 Support Vector Machine. *source: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm* 103

6.10 A summary of Bayes theorem, *Source: Adapted from (Geller, n.d.)* 104

6.11 Screenshot of the Augmented Reality (AR) mode (left) and 2D map (right) of BioSentiers application 107

6.12 Information page of the species 108

6.13 The BioSentiers experiment with pupils 109

6.14 Django's Model-View-Template pattern("Django's Structure – A Heretic's Eye View - Python Django", n.d.) 110

6.15 The high-level architecture of BioSenCS application 111

6.16 Home page (a), and various buttons (b) 112

6.17 Clusters of all observations (c), and Adding/Editing observation point (d) 112

6.18 Querying information of species observation (e), and the participants' leaderboard (f) . . . 113

6.19 The automatic data validation procedure applied in BioSenCS . . . 114

6.20 Automatic date (a) and image feedback (b) in BioSenCS application . . . 115

6.21 An example of Clarifai predicted tags and their probabilities for an observation contributed to BioSenCS . . . 117

6.22 Presences (blue points) and Pseudo-absences (red points) for Carrion Crow in Switzerland . . . 120

6.23 Example of six environmental variables computed over Switzerland with resolution of $2km^2$ . . . 122

6.24 Leave-one-out cross-validation. *source: ("Cross Validation and Model Selection", n.d.)* . . . 125

6.25 k-fold cross-validation for k=5. *source: ("Cross Validation and Model Selection", n.d.)* . . . 125

6.26 Block cross-validation: the three methods propose by Roberts et al. (2017) to arrange spatial folds, a)Unique, b)Systematic, and c)Random . . . 126

6.27 The result of blockCV package (Valavi et al., 2018) for getting the block size for spatial block cross validation taking into account the spatial auto-correlation range among the environmental variables used in our study . . . 127

6.28 An example of defining spatial blocks for our study area using two different block sizes and two different methods of assigning folds: a) Unique folds for each block b) 5 folds randomly assigned to the spatial blocks . . . 129

6.29 The architecture of the Deep Neural Network (DNN) we trained to generate Species Distribution Modeling (SDM) . . . 131

6.30 Confusion matrix to evaluate accuracy of a binary classification problem . . . 132

6.31 The Receiver Operating Characteristics (ROC) curve. The gray part shows the area under the curve. *Source: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5* 133

6.32 The box plots comparing the AUC among the four trained algorithms to generate SDM 134

6.33 Comparison of the algorithms for the species where Naive Bayesian (NB), Random Forest (RF), and Deep Neural Network (DNN) have Area Under (ROC) Curve (AUC) below 70%, Balanced-RF performs better for such species 135

6.34 Variable importance derived from Balanced-RF for two species of a) Golden eagle, and b) Tufted duck 137

6.35 Average environmental variable importance derived from Balanced-RF for all species 138

6.36 Maps of binary classification (a), and probability of occurrence (b) of Golden eagle within Switzerland 139

6.37 Maps of binary classification (a), and probability of occurrence (b) of Common kingfisher within Switzerland 140

6.38 The architecture of BioLocation Application Programming Interface (API) . . . 142

6.39 The process of automatic location validation and real-time feedback generation . 143

6.40 Extraction of environmental variables in a neighbourhood of 1km around the location of observation added by the participant 143

6.41 Location feedback if probability of occurrence of species is higher (a) and lower (b) than 50% 144

6.42 Adding bird species name in BioSenCS application: User does not know the species name (a), participant ask for suggestions to add the species name (b), and participant knows the species name and type the name and an auto-complete of bird names is offered (c) 145

6.43 Number of users who visited BioSenCS application through the three weeks experiment . . . 146

6.44 The contributions of BioSenCS' participants during three weeks of user testing . 147

6.45 Map of all observations collected in Switzerland . . . 148

6.46 Average scores given to the three questions regarding user interface, application language, and species common names proposed in English . . . 150

6.47 Average scores given to two questions regarding receiving automatic feedback. minimum score 1, and maximum score 5 . . . 151

6.48 Top 20 bird species observed during the three-week user experiment . . . 152

6.49 Average scores given to each motivation, 1 indicates very weak motivation, and 5 indicates very strong motivation . . . 153

6.50 Examples of images of bird observations that were flagged. Meaning that the probability of the presence of bird tag in the image was less that 85%. . . . 156

A.1 Location of Flickr images with bird tags in Switzerland . . . 177

A.2 An example of an image with bird tag, which was filtered out using Clarifai . . . 178

A.3 The workflow of evaluating the correlation between distribution of Flickr bird images and CORINE land cover classes . . . 179

A.4 The most frequent tags from the downloaded Flickr images (x-axis: Flickr tags, y-axis: frequency of tags) . . . 180

A.5 Density map of the Flickr images with bird tags . . . 181

A.6 Species distribution maps and the models' performances generated using Maxent for Common Kingfisher using the datasets of eBird (top), and Flickr (down) . . . 182

A.7 Statistical comparison of the species distribution maps generated using eBird and Flickr datasets . . . 182

B.1 Sample of dataset images after data augmentation . . . 185

B.2 Architecture of the implemented Convolutional Neural Network (CNN) model . 185

B.3 CNN performance. x axis: number of epochs, y axis Top: Accuracy; Bottom: Loss. 186

B.4 Steps of detection of water quality phenomena using faster Region-based Convolutional Neural Network (R-CNN) algorithm 187

B.5 Intersect over Union, visual explanation 188

B.6 True positives for foams and algae (correctly detected and located) 189

B.7 True and false positives (clouds, right) for foams 189

B.8 True positives for algae and false positive for foams (reflection of light, left and clouds, right) 190

B.9 False negatives for algae (not detected) 190

B.10 Summary diagram of the entire methodology 191

Chapter 1

Introduction

1.1 Motivation

The participation of general public in scientific projects, known as Citizen Science (CS), has been around for centuries, however the term CS was coined in the 1990s and has grown in popularity since then (Vohland et al., 2021). The Christmas Bird Count project[1], which began in 1900 with the goal of inviting people to collect observations of local birds in North America, is among the early CS projects (Butcher et al., 1990). CS has a variety of goals, including but not limited to public knowledge production, increasing scientific literacy (Aristeidou & Herodotou, 2020), connecting the general public to science and promoting "scientific citizenship" (Haklay et al., 2021), improving communities, and moving toward an inclusive society (Vohland et al., 2021). Furthermore, because a large number of CS projects involve public data collection for scientific projects, CS aims to assist the members of academic institutes in obtaining data and information from citizen scientists that would otherwise be difficult to obtain (Cohn, 2008).

The collection of meteorological observations was one of the early focuses of public participation in science (Strasser & Haklay, 2018); however, over time, the focus of CS expanded to various fields such as biodiversity, with citizen scientists contributing the vast majority of biodiversity data (Strasser & Haklay, 2018), and also other areas of science such as astronomy (Lintott

[1] https://www.audubon.org/conservation/science/christmas-bird-count

et al., 2008), medicine (Swan et al., 2010), biology (Cooper et al., 2010), and so on. With recent advances in technology, particularly mobile technology, new smartphone functionalities, and new programming language frameworks, there are now many web or mobile applications. As a result, over the last decade, a large number of CS mobile/web applications in various domains have been developed (Schade & Tsinaraki, 2016). Searching for CS in Google Play[2] or Apple Store[3] yields hundreds of different applications in a variety of fields, demonstrating how CS has grown (See Figure 1.1). This increase in the number of CS projects has resulted in the collection of large amounts of data (Dalby et al., 2021; Sagiroglu & Sinanc, 2013) in many fields, most notably biodiversity (Kullenberg & Kasperowski, 2016). Regardless of the increase in CS projects, knowing how to keep a CS project running efficiently is one of the most important aspects of CS (Conroy, n.d.). The success of a CS project is dependent on a variety of factors, including a clear definition of objectives, a well-designed project, time management, effective communication tools, and so on (Westreicher et al., 2021). However, given the primary goals of CS, involving citizens in science and assisting scientific projects, two major questions must be addressed: How to motivate citizens to contribute to CS (public engagement)? and Is the data collected useful for scientific projects (data quality)? Several studies have been conducted to answer these two questions (Curtis, 2015a; Kosmala et al., 2016a; Leocadio et al., 2021), yielding interesting outcomes and frameworks for others to consider before designing their CS project. However, the aforementioned questions continue to be a source of concern in CS projects, with researchers looking for new approaches to finding answers. Machine Learning (ML), which presents new opportunities in CS, is a recent focus for answering these questions.

ML is a subset of Artificial Intelligence (AI) and is the study of computer algorithms that can learn patterns based on data and improve themselves over time as more data is available (Popenici & Kerr, 2017). As a result, the vast majority of ML algorithms are data hungry, which means that the more data fed to a ML algorithm, the better the machine performs in making predictions (Halevy et al., 2009; van der Ploeg et al., 2014). As a result, one of the major challenges in ML is a lack of sufficient labeled data[4] to train ML algorithms (Keshavan

[2]https://play.google.com/store/apps

[3]https://apps.apple.com/us/app/apple-store

[4]A sample dataset that has been tagged with one or more labels, allowing certain types of ML algorithms to learn from it and then make predictions on previously unseen data.

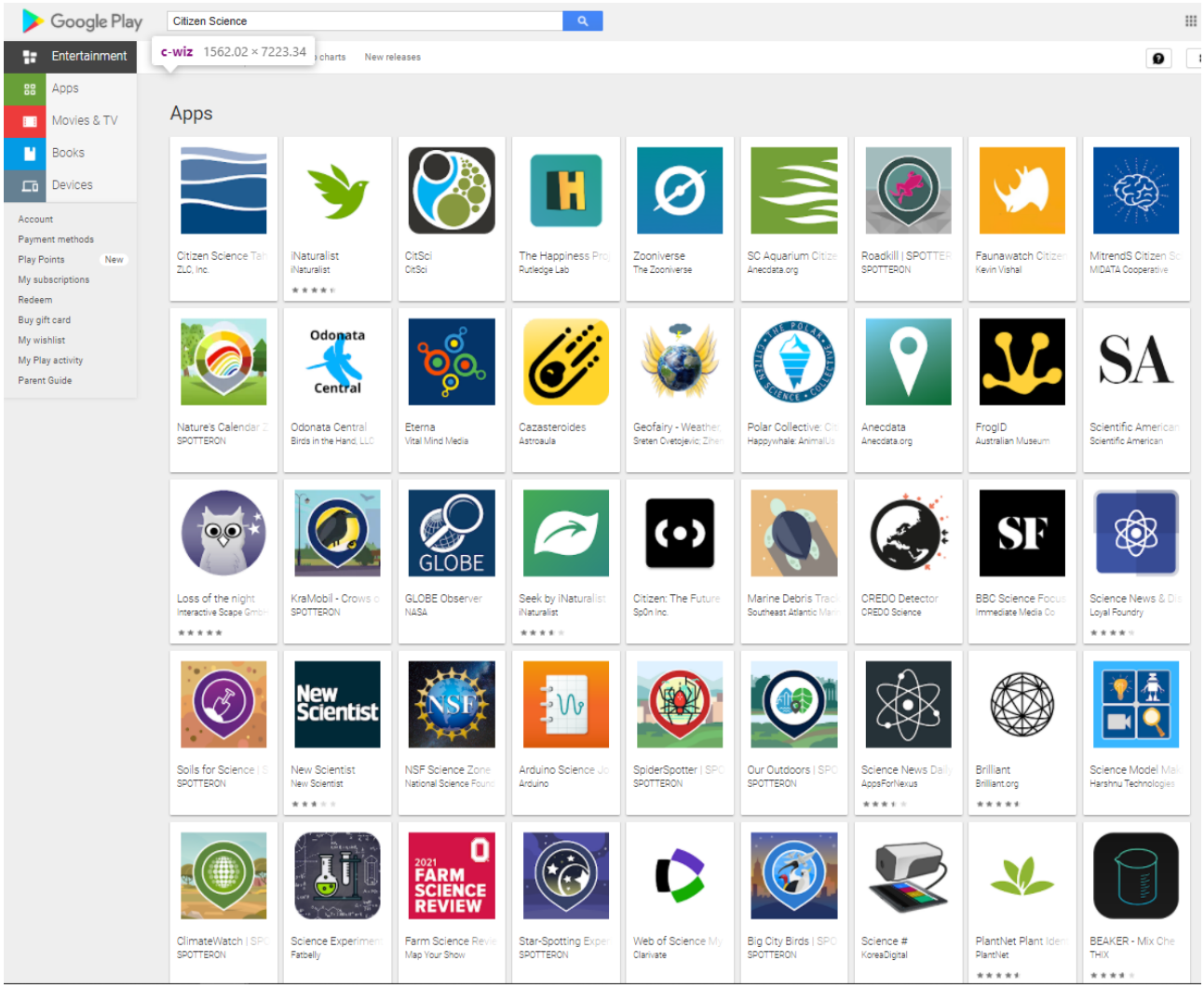


Figure 1.1: Examples of Citizen Science Applications Displayed in Google Play Search

et al., 2019). As previously stated, the growth of CS leads to big data collection, which can be a way of addressing the lack of sufficient data for ML algorithms, thereby forming a partnership between CS and ML. Nevertheless, aside from CS's assistance in providing input data for ML algorithms, what are the benefits of this partnership for CS? The majority of studies in which CS and ML are combined, have focused on using citizens' contributions to collect labelled data for ML algorithms, but to the best of our knowledge, only a few studies have focused on using ML algorithms to address challenges in CS projects.

Keeping the foregoing in mind, the primary goal of this thesis is to investigate how the combination of CS and ML can benefit in addressing some of the CS challenges, as well as the challenges that arise from this combination, with a particular emphasis on public engagement and data quality. While the integration of ML and CS discussed in this thesis has been primarily focused

on biodiversity projects, we provide a general framework and guidelines that can be used when integrating ML and CS in various domains.

1.2 Thesis focus and objectives

As previously stated, the focus of this thesis is to delve deeply into the two challenges of public engagement and data quality assurance in CS, as well as to identify the applied methods in the existing literature on these challenges. Due to the lack of a comprehensive study on the factors that motivate citizens to contribute to CS, and because the majority of studies are case specific, one of the goals of this research is to conduct an exhaustive literature review on available studies and present and categorize participants' motivations based on the designs and typologies of CS projects.

Furthermore, another focus of this thesis is to investigate the existing data validation approaches and to discuss the limitations of the most common data validation approach, manual data verification by experts. The two main limitations of expert data verification are that it is time consuming to verify large amounts of data (e.g., some global projects like eBird[5] have millions of new records each year), and that volunteers may not receive any feedback regarding their contribution or that the time gap between contribution and feedback from expert is very long, which can be demotivating for the participants. As a result, in this study, we focus on the role of ML in (semi)-automating data validation as well as providing informative feedback to volunteers in order to motivate them to retain their participation.

To achieve the aforementioned research focuses, we define a set of objectives that we aimed to accomplish in this thesis. The objectives are as follows:

- Establish a conceptual framework to classify participants' motivations based on the CS project typology.

[5]https://ebird.org/home

- Develop a taxonomy of the various ways in which ML and CS can be combined to address CS challenges.
- Assess the benefits and challenges of combining CS and ML, with a focus on public engagement and data quality.
- Analyze the effect of real-time machine-generated feedback on motivating participants to continue contributing to CS projects.
- Evaluate the role of integrating ML in CS towards simplifying data quality assurance.

Aside from the extensive literature review, in this thesis, a mobile and a web application are developed, which are being tested with the general public in order to achieve the aforementioned objectives. As previously stated, in addition to the general focus of this research on the role of ML in addressing CS challenges, this research focuses particularly on biodiversity CS projects and presents two case studies called BioPocket and BioSenCS. The BioPocket case study focuses on validating our conceptual framework on participants' motivations to contribute to CS projects. The BioSenCS case study aims to collect and validate biodiversity observations and automatically filtering the data using species distribution models generated by ML algorithms. Furthermore, in BioSenCS, real-time feedback is generated for participants as a result of machine predictions, such as the likelihood of observing a species in a specific location and species habitat characteristics. Finally, a user experiment is conducted and discussed in this thesis to evaluate the developed approach in BioSenCS.

1.3 Thesis outline

The remaining chapters of this thesis are structured as follows:

Chapter 2: In this chapter, we define some of the common terms used for public contribution with the focus on CS, as well as different types of CS and the two main challenges of public engagement and data quality.

Chapter 3: In this chapter we present the integration of CS and ML as well as a taxonomy of this integration. Furthermore, we present several use cases in which CS and ML are combined, and we categorize these use cases based on the field of science and the ML algorithms used. Finally, we discuss the benefits and challenges of combining CS and ML, with a focus on public engagement, data quality, and ethics.

Chapter 4: We defined a set of hypotheses taking into consideration our objectives in the introduction chapter, as well as the state of the art and the arguments presented in chapters 2 and 3. Thus, in this chapter, we present our hypotheses and their associated research questions, which we aimed to evaluate in chapters 5 and 6 on public engagement and data quality, respectively.

Chapter 5: In this chapter, we present a motivational framework that we developed as a result of a review of CS literature on volunteers' motivations. Finally, we evaluate the framework using the results of a survey we conducted on our own biodiversity case study, and we discuss the results as well as the hypotheses 1 to 3 that were presented in Chapter 4.

Chapter 6: In this chapter, which was the core of our work in putting the combination of CS and ML into practice, we present our biodiversity CS application, where we used ML algorithms to validate the location of biodiversity observations, and to generate real-time feedback for the participants. We present the methodologies used to implement the application and train our ML models, followed by a user experiment to evaluate our approach. Finally, we used the results of our user experiment to discuss the hypotheses 4 to 6 presented in Chapter 4.

Chapter 7: In this chapter, we provide an overview of the hypotheses discussed in chapters 5 and 6, as well as a set of recommendations for CS practitioners based upon the discussion. The goal of this chapter is to refocus the reader's attention on the thesis objectives while also preparing the reader for the conclusion/final chapter.

Chapter 8: In this final chapter, we present our main findings related to the use of ML and CS to address public engagement and data quality. We also discuss the thesis's limitations and future direction, as well as some suggestions for potential continuation of this research.

Chapter 2

Citizen Science and Related Work

2.1 Introduction

Web 2.0; a term coined by Darcy DiNucci in 1999 (DiNucci, 1999) and later popularized by Tim O'Reilly in 2004 (O'reilly, 2005); has changed the way people interact with the internet over the last two decades (Antoniou et al., 2010). As a result, Web 2.0 enabled internet users to share content with one another. This new technology is the foundation of many well-known web-based platforms such as Wikipedia[1], Twitter[2], and Facebook[3]. Advances in technology and new functionalities in mobile devices such as cameras, Global Navigation Satellite System (GNSS), and sensors, enabled users to contribute various types of content such as spatial data, images, and videos, as well as record and share other types of data such as temperature or noise. Depending on the purpose of these contributions, such as whether it is a fun activity like uploading photos to Instagram[4] or a contribution to a research project like recording biodiversity data, there are various terms and definitions to differentiate diverse types of data collection projects. In the following section, we define some of the most commonly used terms followed by discussing the challenges of such types of public content generation, with a focus on Citizen Science projects.

[1]https://en.wikipedia.org/
[2]https://twitter.com/
[3]https://facebook.com/
[4]https://www.instagram.com/

2.2 Definition of the terms

Various terms of *User Generated Content* (Krumm et al., 2008), *crowdsourcing* (Howe et al., 2006), *Volunteered Geographic Information* (Goodchild, 2007), and *Citizen Science* (Bonney et al., 2009a) are often used interchangeably (See et al., 2016), which although related, each of these terms have their own definition that distinguishes them from one another.

User Generated Content (UGC): "User Generated Content (UGC), is any form of content, such as images, videos, text, and audio, that has been posted by users on online platforms such as social media and wikis" [5]. The rise of UGC happened at the same time Web 2.0 got popularized by O'Reilly (O'reilly, 2005). The social media platforms, Internet forums, blogs, and wikis are good examples of UGC. Typically, users create a community, and content is generated through a number of communities. One example of UGC communities in the geo-spatial world is the mapping communities in projects such as Wikimapia[6] (See figure 2.1). Accordingly, user-generated spatial content (Antoniou, 2011) is a specific type of UGC that should be considered.

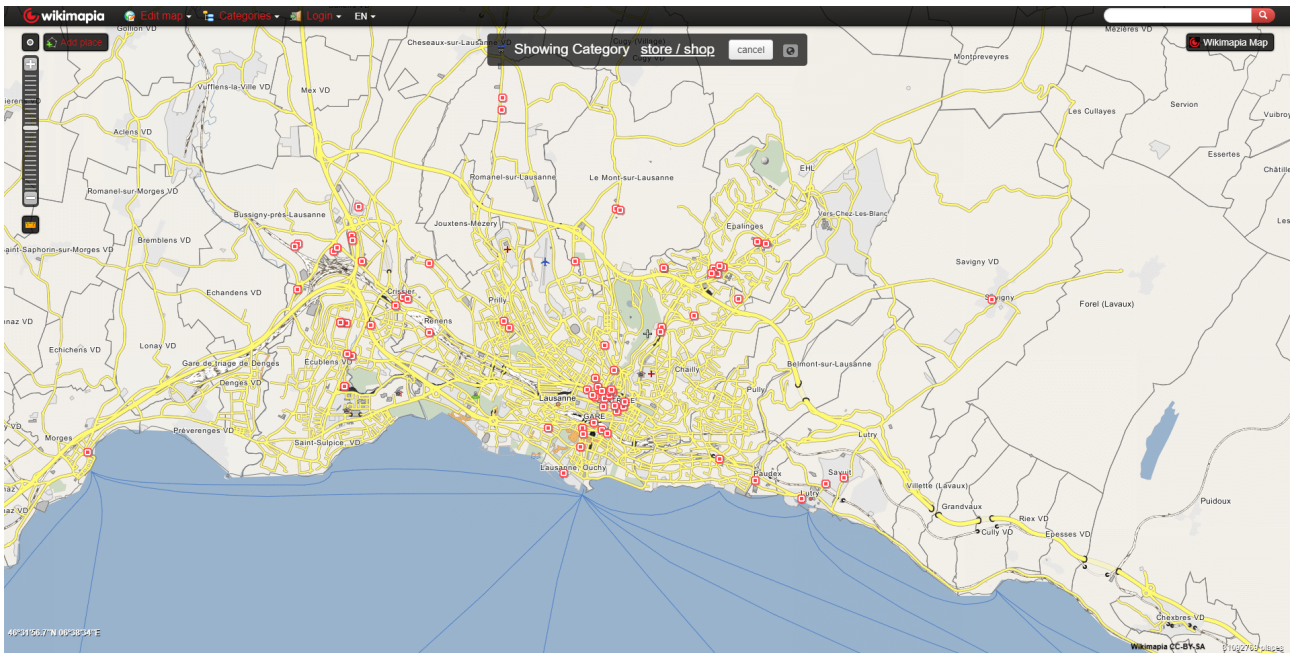


Figure 2.1: An example of user-generated spatial content in Wikimapia project, displaying points with shop/store label in the city of Lausanne. *Source: https://wikimapia.org/*

[5] https://en.wikipedia.org/wiki/User-generated_content
[6] https://wikimapia.org/

Crowdsourcing: The term crowdsourcing initially was coined by Howe (Howe et al., 2006). Despite a number of existing definitions of crowdsourcing (Estellés-Arolas & González-Ladrón-de Guevara, 2012), Lexico dictionary[7] (powered by Oxford) defines crowdsourcing as "the practice of obtaining information or input into a task or project by enlisting the services of a large number of people, either paid or unpaid, typically via the internet". Social media platforms such as Twitter, Facebook, and Instagram are among the most known examples of crowdsourcing where large amounts of data are shared by people. Crowdsourced data may or may not contain spatial information. Furthermore, although crowdsourced data may not be directly collected for a scientific project, it can greatly assist scientific research. Twitter data, for example, has been used in a number of studies, the majority of which have been related to natural disaster management (Carley et al., 2016; Chatfield & Brajawidagda, 2013; Chen et al., 2016).

Volunteered Geographic Information (VGI): Volunteered Geographic Information (VGI) is a term introduced by Goodchild (Goodchild, 2007) and it refers to the voluntarily collection or creation of geographic data. As mentioned by Haklay (Haklay, 2013), VGI activities can range from "fun" activities like georeferencing photos of trips or outdoor activities (Sun et al., 2013) to more focused and serious activities like mapping to help disaster management, such as digitizing roads and buildings after an earthquake in humanitarian OpenStreetMap[8] (Bonafilia et al., 2019; Moradi, 2020). Furthermore, depending on the objectives of data collection, VGI contributions can be categorized to two types of passive and active VGI (See et al., 2016). Passive VGI occurs when users willingly provide geotagged information but the collection of geospatial data is not their primary goal. Instead, their data may be used for other purposes such as using Twitter data for studies on behavior analyses (Hara, 2015) or disaster detection (Farnaghi et al., 2020), or using geotagged images from Flickr[9] to analyze distribution of some phenomena such as biodiversity (Lotfian & Ingensand, 2021). Furthermore, active VGI is when volunteers knowingly collect data for a particular project with a specific purpose; for example mapping features in the OpenStreetMap project. There are benefits and drawbacks to both

[7] https://www.lexico.com/definition/crowdsourcing
[8] www.openstreetmap.org
[9] https://www.flickr.com/

active and passive VGI (See et al., 2017): in active VGI, data are collected within a defined protocol and specified standards for a specific project, whereas data in passive VGI require extensive pre-processing to be usable for analysis; on the other hand, in passive VGI, there is a large amount of data ("big data"), whereas in active VGI, additional efforts are required to keep participants engaged and motivate them to contribute data on a regular basis.

Citizen Science (CS): CS is defined as the voluntarily contribution of non-professional scientists in scientific projects (Cohn, 2008; Silvertown, 2009). CS involves volunteers to contribute to science in various ways, from contributing in data collection to forming the scientific project, problem definition, formulating the hypothesis and interpretation of the results (See section 2.3 for more detail on various levels of participation) (Aristeidou et al., 2017; Haklay, 2013). Public participation in scientific projects benefits both researchers and volunteers. On the one hand, researchers will be able to obtain large amounts of data that would be difficult or expensive to obtain without the contributions of volunteers, and on the other hand, volunteers will have the opportunity to learn about science and scientific procedures by collaborating with people working at research institutions, as well as being part of a scientific community (Cohn, 2008; Land-Zandstra et al., 2016). Initially CS projects were concentrated mainly on environmental projects, while over time CS projects have been expanded to include many areas such as astronomy (Lintott et al., 2008), medicine (Petersen et al., 2019), biology (Khatib et al., 2011), etc. However, the majority of CS projects today continue to focus on biodiversity research (Schade & Tsinaraki, 2016).

Although not all CS projects collect geospatial data, the majority of them do, particularly in the ecological field. When a citizen's contribution to a CS project involves the collection of geographic data, this is where CS and VGI intersect, and is referred to as *Geographic Citizen Science (GCS)*, a term coined by Haklay (Haklay, 2013) (See figure 2.2). The Christmas Bird Count project[10] (See also (Butcher et al., 1990)), which began in 1900 with the goal of conducting an accurate census of bird species for conservation biology purposes, is one of the longest running examples of GCS projects.

[10]https://www.audubon.org/birds

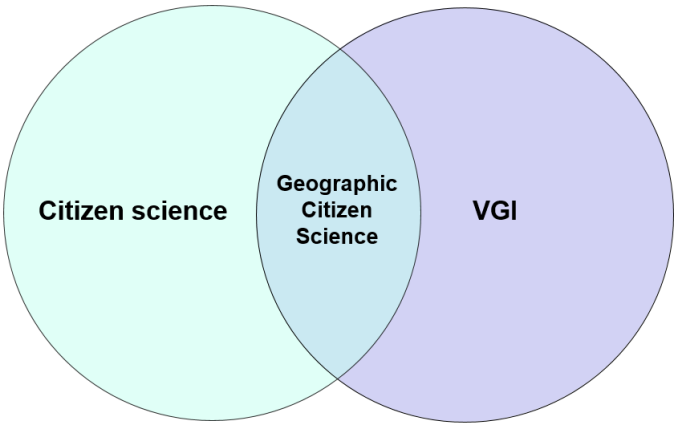


Figure 2.2: Intersection of CS and VGI: Geographic Citizen Science (Haklay, 2013)

The definitions of these various terms help to clarify them when we use them later in this thesis. Although the focus of this research is on CS projects, many of the aspects discussed further, particularly with regard to participant motivations and data quality, can also be applied in other projects, especially VGI projects. Section 2.3 delves deeper into the details of CS projects, such as the various phases of a project, how the projects are classified, and some of the notable CS projects.

2.3 Citizen Science typology

Haklay et al. (2021) identify several aspects to consider when describing a CS project, which are as follows:

- Types of participants: CS participants can range from a wide group, such as school pupils, tourists, nature enthusiast, museum visitors, activists, and so on.

- Scientific field: CS projects can take place in a wide range of scientific disciplines. Furthermore, the outcomes of these projects can range from knowledge outcomes, such as journal articles or production of information that can be used by participants to address local concerns, to practical policy outcomes, and tangible outcomes, such as open data repositories.

- Open science dimensions: Given the growing importance of open science and the integration of CS as part of the European Open Science, it is necessary to consider and apply open science practices in CS, such as the use of open data, the release of code as open source, open access publications, and so on.

- Technology use and accessibility: The tools and technologies required, such as desktop computers and scientific instruments, as well as access to such technologies and the necessary skills to use them, are important factors to consider in CS projects.

- The temporal dimension: The timeline of CS projects can vary; they can occur only once, be short-term (a few days or weeks), occur infrequently (once a year or once a month), or be long-term (every day and/or over a long period of time, as many biodiversity CS projects are).

CS projects include five key phases, with participants engaging in all or some of the phases depending on the project type. The following are the primary phases for each CS project ("Basic Steps for Your Project Planning | CitizenScience.gov", n.d.; Bonney et al., 2009b):

- Defining the problem: Exploring the problem that needs to be solved by answering questions, such as why this issue is important, who the stakeholders are, and what will be achieved.

- Designing the project: Identifying the objectives, allocating the necessary resources (funding, team members, equipment, etc.), and defining the project planning.

- Building a community: Encouraging the general public to participate in the project and sustaining their engagement by establishing a trusting relationship with the volunteers.

- Data collection, quality assurance, and analysis: Designing data collection tools, training volunteers, determining how to store data, filtering and cleaning collected data, analyzing data to detect trends, and sharing data with participants or other practitioners.

- Sustain and improve the project: Maintaining project funding by searching for different sources of funding, and sustaining participation by communicating with volunteers and receiving/giving feedback from/to them.

Several classifications for CS projects are defined according to the types of volunteer contribution (Bonney et al., 2009a; Haklay, 2013; Roy et al., 2012). Taking into account the aforementioned project phases, Bonney et al. (2009a) described three types of CS projects: contributory projects, in which scientists design the project and members of the public contribute primarily to data collection; collaborative projects, in which scientists design the project and members of the public contribute not only to data collection but also to data analysis and/or interpretation of the findings; and finally, co-created projects are those in which the project is designed in collaboration with scientists and members of the public, and some members of the public are involved in most, if not all, of the project steps. Shirk et al. (2012) supplemented the above classification with two new categories: contractual projects and collegial contributions. Contractual projects are those in which communities ask professional researchers to conduct a specific scientific investigation and report on the findings. Collegial contributions are those made by non-credentialed individuals who conduct independent research with varying degrees of expected recognition by institutionalized science and/or professionals.

In addition to Bonney's classification of CS projects, Haklay has established a ladder that categorizes levels of participation in such projects (Haklay, 2013) (See figure 2.3). Participants can be involved in problem definition, data collection, and data analysis, as briefly mentioned in the previous section, and Haklay's framework is based on the levels of cognitive engagement of volunteers in a project. Level 1, also known as crowdsourcing, is when participation is primarily limited to data collection with little cognitive engagement; as we progress through the levels, cognitive engagement increases and volunteers become more involved in various stages of the project; level 4, known as Extreme CS, is when volunteers are involved in problem definition, data collection, and data analysis.

Aside from the levels of participation in CS, there can be various typologies of CS projects depending on the designs, objectives, and types of tasks to be performed. Haklay (Haklay,

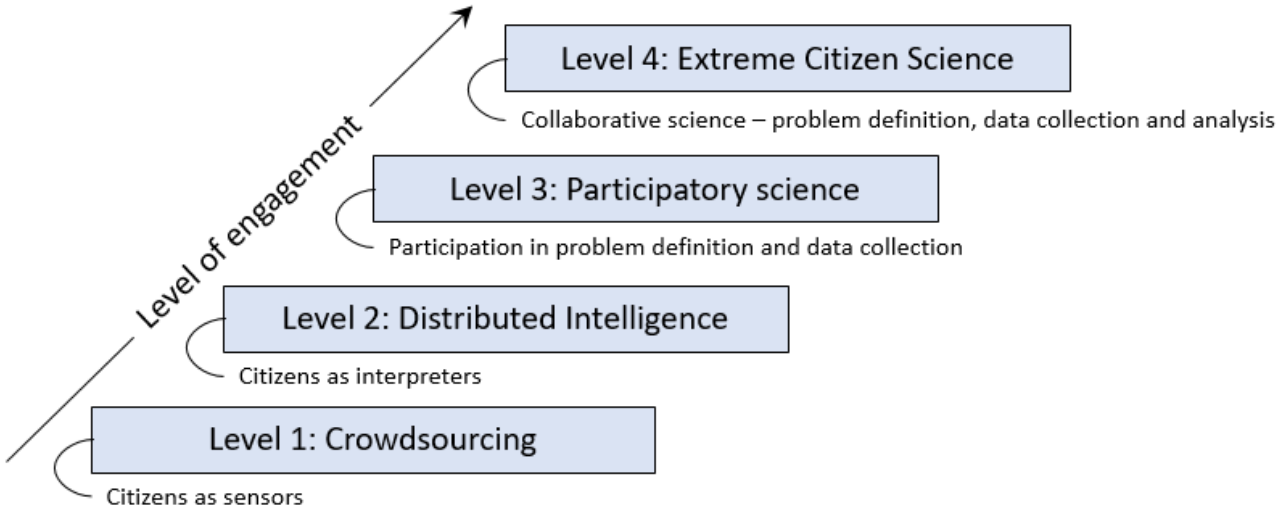


Figure 2.3: Levels of participation in CS, adapted from (Haklay, 2013)

2013) categorises CS projects into three types: classic CS, environmental management, and Citizen Cyber-Science (CCS).

Classic CS: Classic CS, mainly refers to the collection of observations from a broad and dispersed community of observers. Most of the cases in this category concentrate on the collection of environmental data such as biodiversity, water quality or meteorological observations, while also including other fields such as archaeology (Haklay, 2013). The evolution of mobile technology has made communication and observation collection easier in classic CS. Moreover, almost all of the collected observations in classic CS are geolocated, thanks to the integrated GPS chips in mobile devices, and therefore, are part of GCS projects (See section 2.2 for more detail) (Haklay, 2013).

Public Governance: The second type refers to the role of citizen participation in policy formulation and decision making. Alan Irwin's focus is very much in line with this typology, where his goal was having a more democratic science and technology policy in which citizens' voices must be taken seriously in decision-making (Irwin, 1995). A recent article by Gobel et al. (2019) defines the relationship between CS and governance using empirical findings from the DITOs[11] (Doing-It-Together Science) project. As a result, they define four governance modes: CS as a source of information for policy-making, CS as an object of research policy, CS as a policy instrument, and CS as socio-technical governance. Furthermore, because this

[11]http://www.togetherscience.eu/

typology has largely covered the environmental domain, Haklay refers to it as environmental management (Haklay, 2013). Despite increasing attention in this typology, there is yet inadequate evidence as to whether or not it will lead to the creation of environmental policies and regulations (European Commission and Directorate-General for Environment, 2018).

Citizen Cyber-Science (CCS): The third category, CCS (also known as Virtual CS (Lintott et al., 2013; Reed et al., 2013; Wiggins & Crowston, 2011)), is defined by Francois Grey (Yadav et al., 2018) as the use of web-based computer interactions to allow volunteers to contribute in scientific research through collaboration with scientists. Zooniverse[12] is a CS web portal with hundreds of CCS projects (Lintott et al., 2013). Haklay (Haklay, 2013) proposed a categorization of CCS with three subcategories: Volunteered Computing (VC), Volunteered Thinking (VT), and participatory sensing.

In VC projects, volunteers devote their computing resources to provide processing power to support computer-intensive tasks in a scientific project (Yadav et al., 2018). SETI@home[13] and Folding@home[14] are two of the known examples of VC projects. However, in VT projects, volunteers are using their cognitive ability to perform tasks (Yadav et al., 2018), such as identifying objects in an image (Lintott et al., 2008; Westphal et al., 2006). Some VT projects are part of VGI when the tasks involve geodata analysis, for instance georeferencing historical images in the sMapShot project (Produit & Ingensand, 2018). Lastly, participatory sensing is the latest form of CCS projects that uses mobile phone capabilities such as cameras, GPS receivers, different transceivers (mobile network, Wi-Fi, Bluetooth), and microphones to sense the environment; emotional maps (MacKerron & Mourato, 2013), air-quality monitoring (Dutta et al., 2009), water quality monitoring (Brovelli et al., 2019), and noise level monitoring (Maisonneuve et al., 2010) are examples of this category.

Curtis (Curtis, 2015b) proposed "CS games" (Holliman & Curtis, 2015) as the third category of CCS. She suggested that it is a more appropriate category than Haklay's notion of "participatory sensing". Participatory sensing mainly regards data collection rather than data analysis,

[12]https://www.zooniverse.org/
[13]https://setiathome.berkeley.edu/
[14]https://foldingathome.org/

and therefore, it fits better in classic CS typology. Figure 2.4 shows the CS typology discussed above, which illustrates the categorisation defined by Haklay, except for participatory sensing; the third CCS category; which is replaced by CS games (as suggested by Curtis).

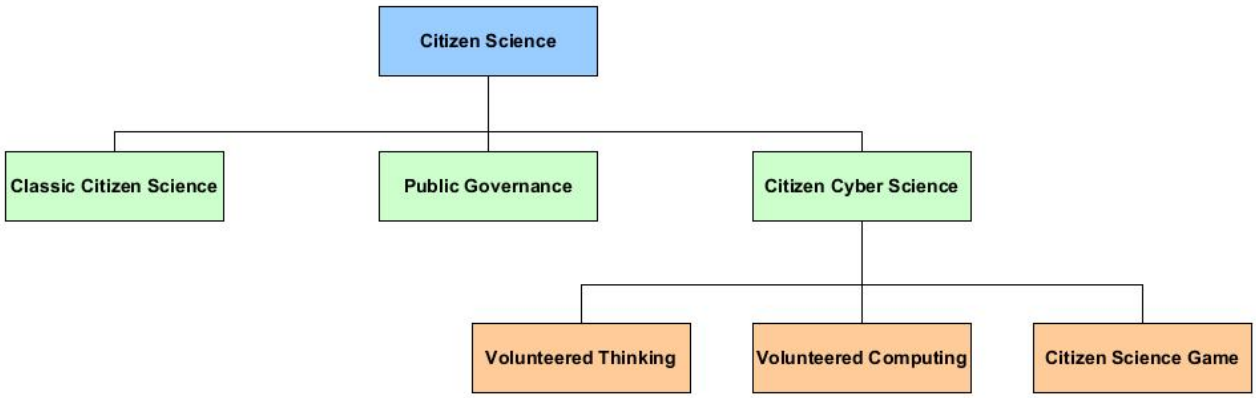


Figure 2.4: Typology of CS projects

2.4 Selected Challenges in Citizen Science

Although the rise of CS projects has been a huge benefit to science and researchers, simplifying traditional tasks such as data collection and analysis, there are several challenges to initiating and maintaining a successful CS project. For example, technological challenges related to project design and implementation, ethical challenges such as confidentiality of participants and contributed data, financial requirements such as for buying and maintaining servers or hiring experts, and so on (PARTHENOS, n.d.). Among the existing challenges, public engagement and sustaining participation (De Moor et al., 2019; MacLeod & Scott, 2021), as well as data quality assurance (Balázs et al., 2021; Hecker et al., 2018), are two that are critical in any CS project.

2.4.1 Challenge 1: Public engagement

Participants are at the heart of CS projects, and their active participation is what keeps a CS project flowing efficiently (Fritz et al., 2017). However, it is often hard to engage people to join a

project, to begin contributing to a project, and, finally, to sustain the volunteers' participation. Furthermore, it is critical to understand how to encourage non-active participants to contribute more and how to keep active participants motivated and engaged. Understanding participant behavior is an essential initial step toward addressing the challenge of motivating participation and increasing public engagement.

2.4.1.1 Patterns of participation

"*All large-scale, multi-user communities and online social networks that rely on users to contribute content or build services share one property: most users don't participate very much. Often, they simply lurk in the background.*" (Nielsen, n.d.).

According to Nielsen (Nielsen, n.d.), the concept of participation inequality was first introduced by Will Hill (Hill et al., 1992); this term refers to the fact that the majority of content in online forums is generated by a minority of volunteers, while the majority of people are simply using the generated information. User participation usually follows 90-9-1 rule (Arthur, n.d.; Nielsen, n.d.), implying that 1 per cent of the users participate very frequently, 9 per cent of the users participate "from time to time", and 90 per cent of the users are "lurkers", indicating that they only read or observe information without contributing. Although some projects have hundreds of registered participants, only a few of them actively make contributions (Curtis, 2015b). Preece and Shneiderman (Preece & Shneiderman, 2009) analysed the patterns of contributions to online communities and social networks such as YouTube or Wikipedia and, as a result, developed a framework called "Reader-to-Leader", showing the evolution of user participation in online social communities from readers, to contributors, collaborators, and ultimately leaders (See figure 2.5).

In addition, we can mention OpenStreetMap, as a well-known VGI project where participation inequality among volunteers has been discussed several times (Budhathoki, 2010; Mooney & Corcoran, 2012). Wood's analysis (Wood, 2014) illustrated that only a small number of participants among the registered users (about 150,000 out of 2 million registered users at the time of analysis) contribute more than 100 points of data (Figure 2.6). This pattern has persisted

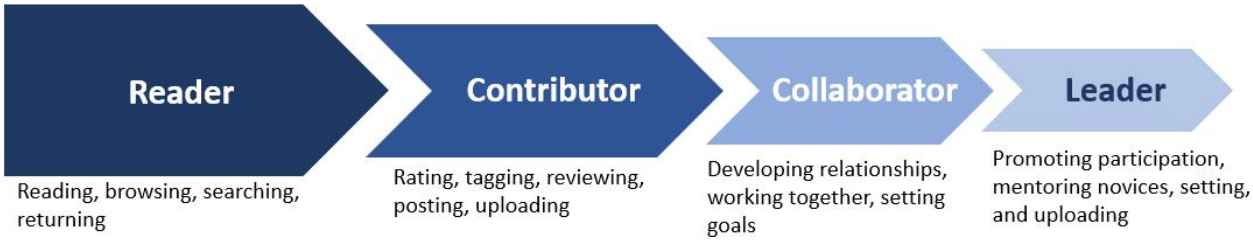


Figure 2.5: Reader to leader framework, adapted from (Preece & Shneiderman, 2009)

until now, since OpenStreetMap statistics show that only 1% of all users contribute regularly[15]. Wood also noted that it is essential to understand how to **recruit new participants**, and to **encourage registered participants to contribute** (make edits) and **active participants to continue with their contributions** (Figure 2.7).

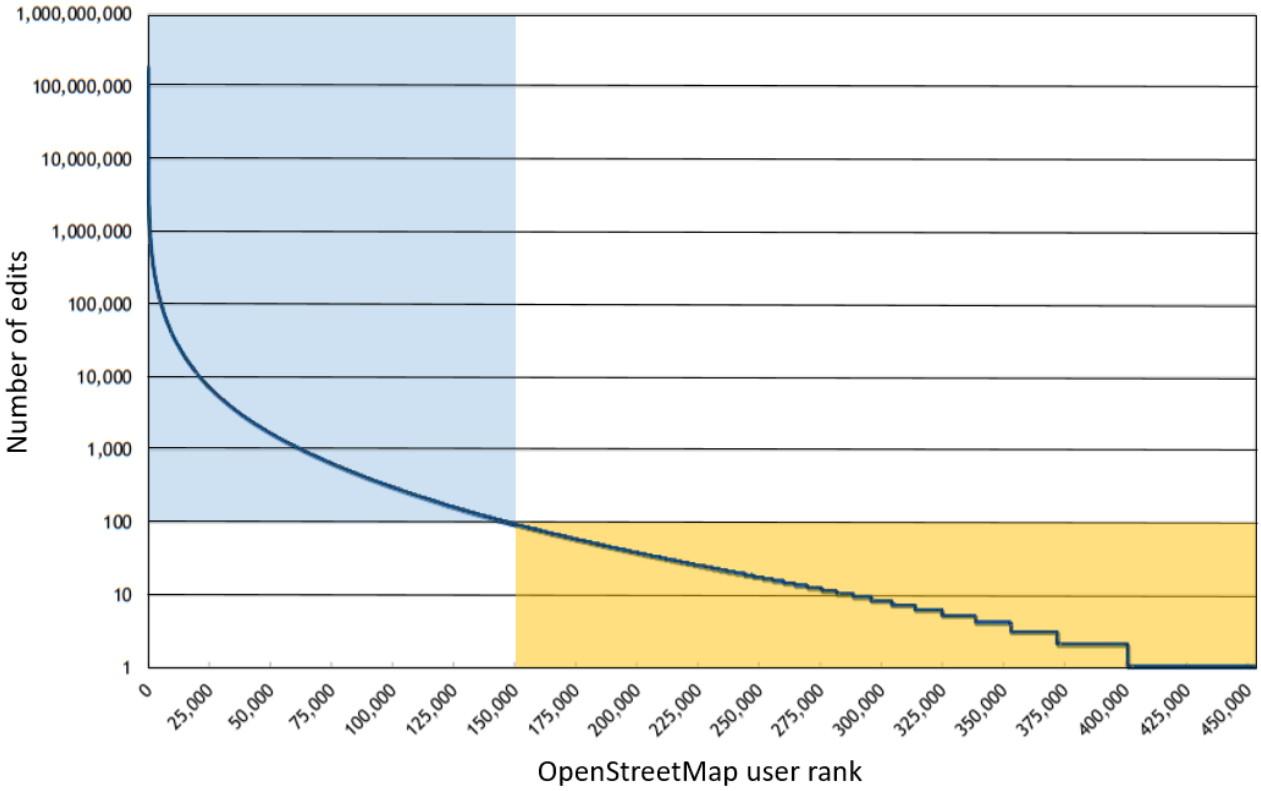


Figure 2.6: Participation inequality in OpenStreetMap adapted from Wood's analysis(Wood, 2014) for the users that contributed at least one edit. Few registered users contribute the majority of data (blue side); the majority of users contribute less than 100 edits (orange side).

In addition, Budhathoki and Haythornthwaite (Budhathoki & Haythornthwaite, 2013) classified OpenStreetMap participants into serious and casual mappers, taking into consideration the

[15]https://wiki.openstreetmap.org/wiki/Stats

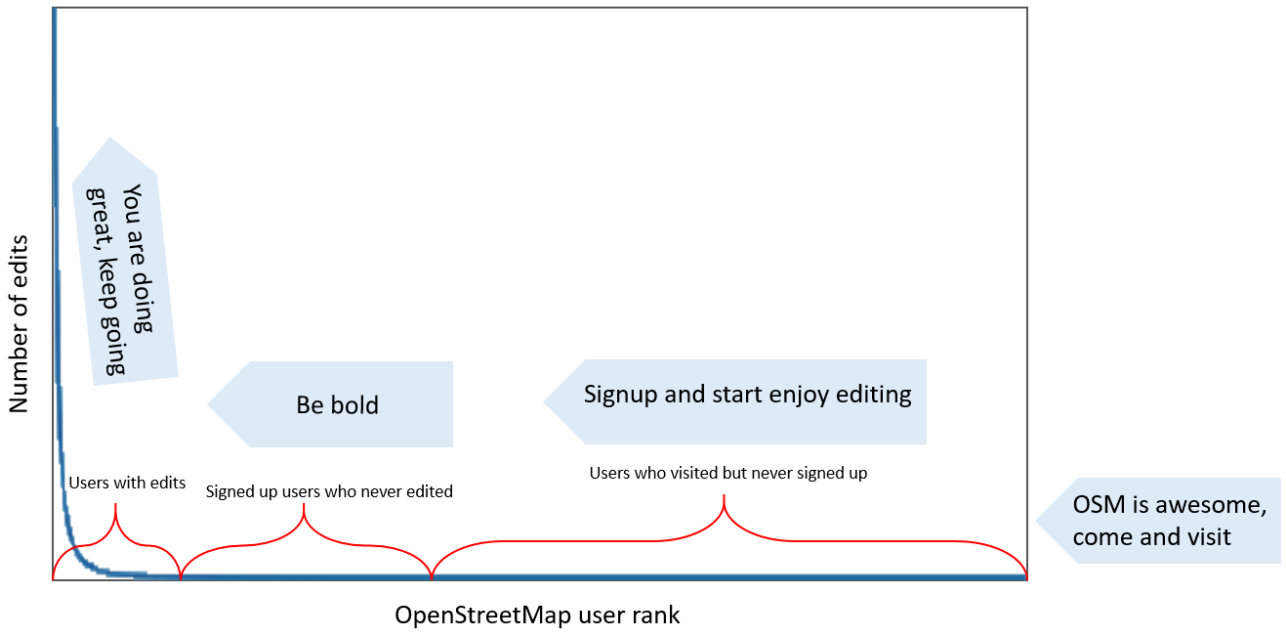


Figure 2.7: Various steps to encourage users to contribute to OpenStreetMap. Adapted from (Wood, 2014)

amount of their contribution (number of contributed nodes), the duration of contribution, and the frequency of contribution. The survey results of the motivations of these two classes of OpenStreetMap participants illustrated that serious mappers were more driven by a desire to be part of a community, and by learning and career-based factors, whilst for casual mappers, the thought of creating free and accessible maps for others was a stronger motivational factor (Budhathoki & Haythornthwaite, 2013).

As stated by Haklay (Haklay, 2016), participation inequality does not only appear in online projects, but can also be observed in other types of projects such as environmental volunteering. Boakes et al. (2016) studied patterns of contribution in three CS biodiversity projects, i.e., Greenspace Information for Greater London Community Interest Company (GiGL[16]), iSpot[17] and iRecord[18], and they considered Greater London as their study area. In order to better understand the patterns of contribution of observers (biodiversity recorders), as well as to search for spatial and taxonomic biases in the observations, they analysed the data set from the three aforementioned projects. They used a methodology to cluster volunteers based on

[16] www.gigl.org.uk
[17] www.ispotnature.org
[18] www.brc.ac.uk/irecord/

three engagement metrics: Activity Ratio (number of days a volunteer was active divided by the total days they are linked to the project), Relative Activity Duration (number of days a volunteer was active divided by the overall study observation period in days), and Variation in Periodicity (average number of days elapsed between two sequential active days of an individual divided by the total average of days elapsed between active days of all individuals). As a result of this clustering, they identified three classes of volunteers: "dabbler", "steady", and "enthusiast". The dabbler category has the largest activity ratio, the lowest number of observations, the highest variation in periodicity and the least activity duration relative to the other two categories. Enthusiasts, on the other hand, are predominantly long-term observers and also the category that spends most time on the project and collection of observations. The category of enthusiasts consists of the smallest number of participants. Finally, the steady group has a profile type between the two categories of dabbler and enthusiast. The authors noted that dabblers were the clusters with the largest number of participants in all of the three projects evaluated, and that the number of participants gradually decreased from steady to enthusiast, a similar behaviour to that described in the Reader-to-Leader framework (see Figure 2.5). Moreover, similar to studies of online social communities and OpenStreetMap, the authors found that a few volunteers contribute with many observations and many volunteers contribute with a small number of observations. They stated that in the GiGL project, despite of having 1 million recorders, only four participants contributed over 10 percent of the records, while about 35 percent of participants contributed with only one record.

Considering this participation pattern and participant behavior, understanding how to recruit more volunteers, increase casual volunteer engagement, and sustain active volunteer participation, is dependent on identifying the motivations of each group to contribute to CS/VGI projects.

2.4.2 Challenge 2: Data quality

Data quality is an index of measuring how well a data set fits a particular purpose. Data quality metrics are based on characteristics of data quality such as accuracy, completeness, consistency,

reliability, uniqueness, and timeliness (Wang et al., 1995).

Despite some scientists' skepticism about the ability of unpaid volunteers to produce high-quality data, there is a large body of literature that shows that CS data can be as accurate as, if not more accurate than, data collected by professionals (Kosmala et al., 2016b).However, because volunteers in CS and VGI projects come from diverse backgrounds (e.g., different education levels, age groups, expertise, etc.), the data they contribute must be validated before it can be used in scientific analysis. The case of OpenStreetMap is a well-known example of data quality assurance, with several studies focusing on evaluating the quality of added features (e.g. roads and buildings) (Abdolmajidi et al., 2015; Antunes et al., 2015; Brovelli & Zamboni, 2018; Mondzech & Sester, 2011).

As mentioned, when it comes to data quality assurance, several factors must be considered, such as accuracy, timeliness, completeness, accessibility and so on. The literature on data quality in CS is mostly project-specific, and a framework or general guidance on dealing with data quality is lacking, even in projects in similar domains (Balázs et al., 2021). There are various biodiversity CS projects, for example, but there is no general framework for data validation in such projects. Balazs et al. (2021) argue that in order to reuse data from CS projects, a protocol for ensuring a minimum standard of data quality across different CS projects is required. The authors define four aspects to evaluate CS data: data quality, data contextualization, data reuse, and data interoperability. Data quality refers to ensuring the validity and reliability of the data, data contextualization refers to communicating how a specific data set is created, for example by providing metadata, data reuse refers to clarification on data ownership and future accessibility, as well as using open data and open standards, and finally data interoperability refers to the development of a standard system to simplify data reuse across various projects and systems (Balázs et al., 2021).

The first aspect of data evaluation in CS projects illustrated in figure 2.8 is data quality: ensuring data validity and reliability. This task is known as data validation in CS projects and is typically addressed by experts in the field. However, depending on the task's complexity and the volunteer's experience, there are a variety of approaches for enhancing data quality (Kosmala

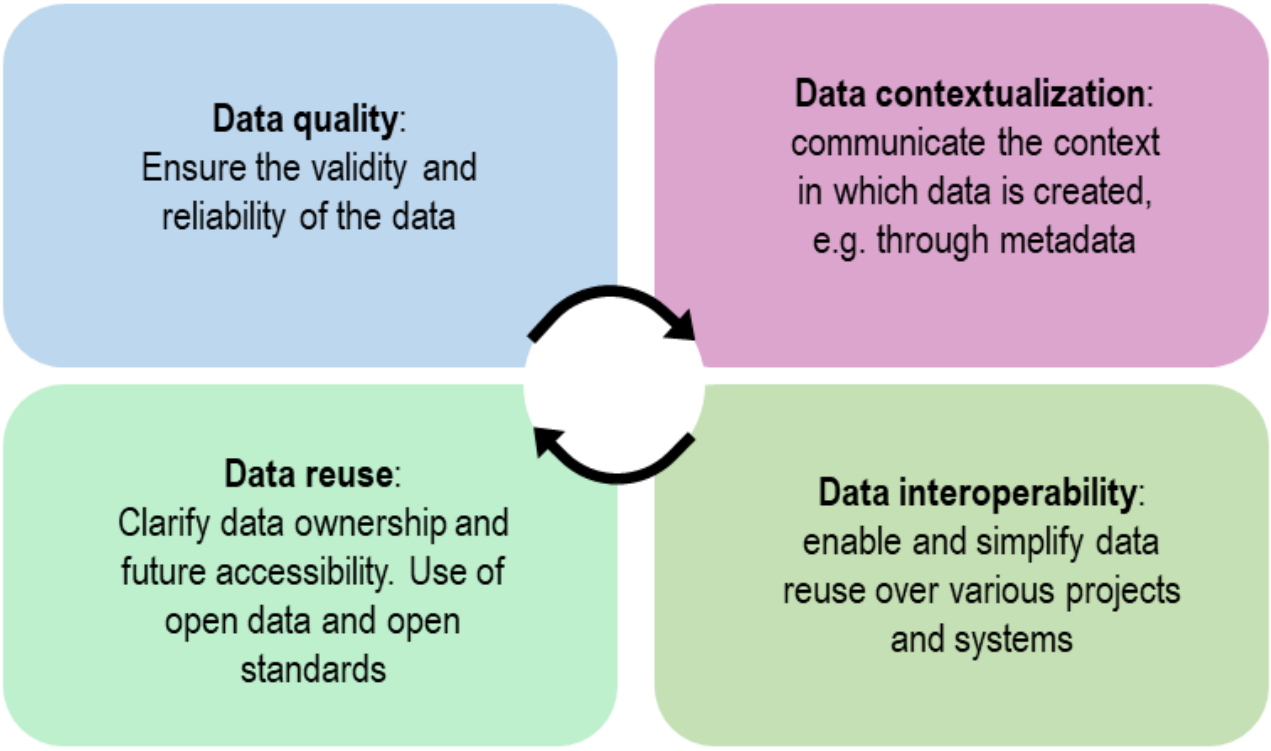


Figure 2.8: The four aspects of data evaluation. Adapted from (Balázs et al., 2021)

et al., 2016b). Wiggins et al. (2011) conducted a survey on the mechanisms used to validate the quality of collected data in several CS projects. As a result of this survey, 63 complete responses were obtained. They identified 18 mechanisms used in projects for data validation, with expert review being the most commonly used method (77 percent of responses), followed by photo submission in biodiversity studies as another method of validation. Furthermore, they mentioned that several projects reported on other methods of data validation combined with expert review. Balazs et al. (2021) classified the validation approaches in four categories: peer verification, expert verification, automatic quality assessment, and model-based quality assessment.

- Peer verification: Peer verification refers to the involvement of more experienced participants in validating data collected by other new or inexperienced participants, or in assisting them in data collection, in a manner similar to Wikipedia's peer review approach (De la Calzada & Dekhtyar, 2010).

- Expert verification: Identifying stakeholders or some of the specific volunteers as experts to validate the data contributed by other participants is what expert verification entails

(Balázs et al., 2021). In most CS projects, but especially in biological studies, expert verification is the most commonly used approach (Adriaens et al., 2021; Baker et al., 2021).

- Automatic quality assessment: The use of software-based systems to automatically assess the quality of collected data in a CS project is referred to as automatic quality assessment. This can be accomplished, for example, by using data mining and statistical analysis to detect outliers or by employing AI to perform automatic data quality assessment (Wessels et al., 2019; Wiggins et al., 2011).

- Model-base quality assessment: Finally, model-based quality assessment goes beyond automatic data validation by performing validation as a result of understanding how a specific phenomenon behaves, such as by modeling the phenomenon of interest in space and time. For example, modeling the level of expertise of volunteers and using the model to validate contributed data is one of the model-based-validation approaches, a method used to estimate the observation skills of eBird observers (Kelling et al., 2015).

A recent article on solutions for CS data validation on invasive alien species (Adriaens et al., 2021), mentions a new approach that can be used to collect higher quality data. This approach includes the use of chatbots and conversational agents that can mimic human language and communicate with participants in order to provide assistance in the form of automatic feedback for guiding participants and simplifying alien species identification.

Besides the validation approaches, there are various phases in a project when data validation can be applied. Balazs et al. (2021) stated four stages where data validation can be performed in a CS project including *at the project planing and design stage*, *during the project*, *after the project (before data publication)*, and *after the project (after data publication)*. Depending on the stage, a different validation mechanism (or a combination of two or more validation mechanisms) can be used. For example, in the initial phase, expert validation is primarily used, but during the project, one or a combination of all of the above approaches can be used, and after the project and data publication, expert and peer verification are typically used.

2.5 Summary

The rise of CS and VGI projects has been a great point for scientists and for assisting scientific projects in achieving data beyond traditional data collection approaches and benefiting from the cognitive engagement of citizens to assist in addressing scientific problems.

Despite numerous studies on how to motivate participants in CS projects and how to evaluate the quality of collected data, these two issues remain major challenges in such projects, and new approaches to addressing them are required. As mentioned in the data quality section, the majority of projects rely on expert validation as their primary method. While the use of expert knowledge is critical in a CS project, relying solely on expert review has its own drawbacks, such as the validation task being time-consuming, lack of sufficient volunteer experts, and a large time gap between the moment volunteers make a contribution and receiving feedback (if they receive any) that can demotivate volunteers. Although there are new ways to automate data validation, this is still in its early stages, and more research should be conducted to determine how to optimize data validation automation or which factors are more important in data validation automation. Researchers are increasingly focusing on the use of AI to address a variety of scientific challenges, and this is also becoming a recent focus in CS projects. Thus, in the following chapter, we discuss how integrating ML into CS can benefit in addressing the two aforementioned challenges, as well as the existing and potential ways that ML can be integrated into CS projects.

Chapter 3

Integrating Machine Learning In Citizen Science Projects

3.1 Introduction

This chapter is based on a published literature review we conducted on integration of machine learning and citizen science (Lotfian et al., 2021).

The simulation of human intelligence in machines, known as AI, is widely applied in various domains, and the number of scientific publications in this area is significantly increasing (Perrault et al., 2019). AI is a term used when machines can perform tasks which simulate the human mind such as learning, reasoning, and solving problems (Shinde & Shah, 2018). Thus, ML is a sub-field of AI, defined as the study of developing computer algorithms, which use data to learn patterns, make predictions, and improve their performance over time by more data (Popenici & Kerr, 2017). The majority of ML algorithms require large amounts of labeled data, which has resulted in a close collaboration between ML and CS projects (Rzanny et al., 2017; Terry et al., 2020). The reason for this is the rise of CS, that has led to large data set collections in a variety of scientific domains (Wright et al., 2019), which can be a valuable input source for ML algorithms.

Although the combination of ML and CS is not new (Ceccaroni et al., 2019), until recently, these two fields have mostly been implemented separately (McClure et al., 2020). The integration of ML and CS can result in producing a new learning paradigm for citizen scientists through human–computer interactions (Franzen et al., 2021). Moreover, it can result in increasing interdisciplinary collaborations among researchers as well as members of the public in various fields such as computer science, ecology, astronomy, and medicine, to name a few (McClure et al., 2020). This integration has been focused primarily on object detection in images and videos with the main focus on automatic species identification in biodiversity projects (Norouzzadeh et al., 2018; Willi et al., 2019). A well-known example is the iNaturalist project[1], which has included automated species identification suggestions since 2017 using images obtained from observers. The automatic identification has improved over the years as more images are used to train the model, and the latest model release was in March 2020 by the time of writing this thesis (Ueda, 2020). The automatic species identification in iNaturalist has provided citizen scientists the opportunity to learn about species and to minimize the occurrence of erroneous observations (Van Horn et al., 2018).

The objective of combining CS and ML is not limited to providing data for ML algorithms and automating identification tasks. The goal is also to combine human and machine intelligence to help enhance CS tasks such as automated data collection, processing, and validation, as well as to increase public engagement. There are potential challenges and opportunities in integrating ML and CS that must be discussed. In this thesis, we aim to address the following research questions:

- What are some examples of successful CS projects where ML is integrated?
- What ML techniques have been used in these projects?
- What CS tasks have been affected by ML in such projects?
- What are the benefits and challenges of integrating ML in CS for practitioners and citizen scientists?

[1] https://www.inaturalist.org/

To answer these research questions, we explore use cases where ML and CS can be combined. We have reviewed successful CS projects, highlighting the typologies of techniques used in such projects and categorizing them in light of the effect of ML on CS tasks. Although the opportunities and challenges of integrating ML and CS have been addressed in a few recent articles (Ceccaroni et al., 2019; Franzen et al., 2021; McClure et al., 2020), the main emphasis has been on the transparency of using ML in CS in terms of how the ML algorithms use CS data (Franzen et al., 2021), the effects of AI on human behavior and improving insights in CS (Ceccaroni et al., 2019), and the effects of this combination in ecological monitoring in terms of having cheaper or more efficient ways for data collection and data processing (McClure et al., 2020). While these are key issues to explore, to the best of our knowledge, the integration of ML and CS has received less attention in terms of how this integration can affect the usual processes in a CS project, from volunteer involvement to evaluating the quality of their contributions (See chapter 2 for more details). Our primary objective is to explore how some CS tasks can be automated using ML and whether this automation is beneficial or detrimental for the project and its participants. Rather than being overly broad, we broke down the forms of ML combinations in various CS phases and discussed the benefits and challenges of this integration in each step. We outline how ML can be integrated in each step, including what has already been applied, what can be applied in the future, and what the current and potential challenges and benefits of this integration are for each step. The focus of our research here is primarily on the use of ML in CS rather than AI in general for two reasons: first, the majority of recent CS projects involving AI are more focused on using ML and deep learning rather than AI in general; and second, the extent of AI applications is vast, and investigating its combination with CS would be beyond the scope of this research, so we focused on ML as a subset of AI.
In the following section, we will discuss the ML paradigm, and then we will evaluate the impacts of ML on CS tasks by reviewing some notable use cases. Finally we discuss the potential benefits and challenges of integrating ML in CS projects.

3.2 Types of Machine Learning and Applications

ML is a subset of AI, which was first introduced in 1955 by Arthur Samuel when he applied learning to his droughts (checkers) algorithm (Shinde & Shah, 2018). Samuel defined ML as a "field of study that gives computers the ability to learn without being explicitly programmed" (Géron, 2019). ML algorithms build models which learn using the input data (known as training data) and are able to make predictions based on the learnt experience. There are three main ML types, known as supervised learning, unsupervised learning, and reinforcement learning (Géron, 2019).

- Supervised learning: In supervised learning, the training data are labeled, and the task is to map the input (independent variables) to the output (dependent variables). The two typical types of supervised learning are classification, where the output variable is categorized, and regression, where the output variable is continuous (Géron, 2019). The most widely known algorithms of supervised learning are k-Nearest Neighbors (KNN), linear regression, logistic regression, Support Vector Machines (SVM), decision trees, Random Forest (RF), and Neural Network (NN).

- Unsupervised learning: In unsupervised learning, the training data are not labeled, and the goal is to identify structures and patterns in the data (Géron, 2019). The typical types of unsupervised learning include clustering (grouping similar input data), dimension reduction (extracting meaningful features from the data), and association (exploring the data to discover relationships between attributes) (Géron, 2019). Some of the most known algorithms of unsupervised learning are k-means, one-class SVM, Hierarchical Cluster Analysis (HCA), and Principal Component Analysis (PCA).

- Reinforcement learning: In reinforcement learning, the learning algorithm, also called the agent, observes the environment and learns through a system of rewards and punishments. Reinforcement learning is commonly used in robotics, such as walking robots and self-driving vehicles, as well as in real-time decision making and game AI (Géron, 2019).

Deep learning, a sub-field of ML (See Figure 3.1 for the relationship between AI, ML, and deep learning), is concerned with algorithms known as Artificial Neural Network (ANN) (will be discussed in details in Chapter 6) that attempt to simulate the structure and functions of a biological brain (Paeglis et al., 2018). Since there is a significant body of literature on AI and ML algorithms, we will not go into the details and will briefly discuss some of the common AI, ML, and deep learning techniques applied largely in scientific projects:

- Computer Vision (CV): CV is an interdisciplinary scientific field which aims at developing techniques so that computers can identify and understand the contents in digital images and videos. In other words, CV aims at enabling computers to identify elements in images in the same way as humans would do. The advances in ANN and deep learning have had great impact on CV, which in some cases outperforms the human power to identify objects (Borji & Itti, 2014). Some popular applications of CV include self-driving cars, face recognition, etc. (Géron, 2019). Moreover, starting in the year 2020 and with the COVID-19 pandemic[2], CV has been applied in monitoring and detecting social distancing among people (Saponara et al., 2021). CV has also been commonly used in species identification, with Plant@net[3] and iNaturalist being two well-known CS examples. A class of deep learning which is commonly used in CV is the Convolutional Neural Network (CNN).

- Natural Language Processing (NLP): NLP is a subfield of linguistics, computer science and AI that deals with human–computer interactions through the use of natural language, which means that NLP aims to enable computers to read and understand human language (Chowdhury, 2003). The mechanism involves the machine capturing the human's words (text or audio), processing the words and preparing a response, and returning the produced response (in the form of audio or text) to the human. Language translation applications such as Google Translate or DeepL[4], as well as personal assistant applications (e.g., Siri or Alexa), are common uses of NLP in people's daily lives.

[2]https://en.wikipedia.org/wiki/COVID-19_pandemic
[3]https://identify.plantnet.org/
[4]https://www.deepl.com/translator

- Acoustic identification: Acoustic identification is a technique based on pattern recognition and signal analysis, where the acoustic data are processed and features are extracted and classified. Main applications of acoustic identification are in species detection (Stowell et al., 2019). For example, BirdNet[5] is an application to identify bird species based on the bird song.

- Automated reasoning: Automated reasoning is a branch of AI that seeks to train machines to solve problems using logical reasoning (Robinson & Voronkov, 2001). In other words, in automated reasoning, the computer is given knowledge and can generate new knowledge from it, which it then uses to make rational decisions. Automated reasoning is mainly used to assess if something is true or false or whether an event will occur or not.

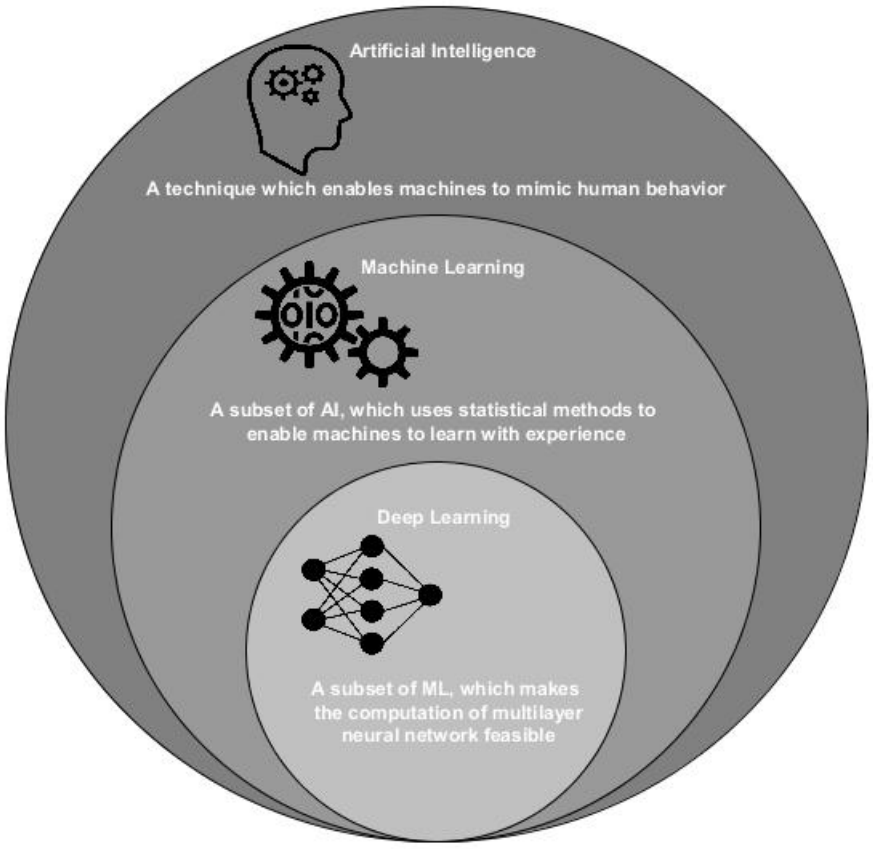


Figure 3.1: Relationship between Artificial Intelligence, Machine Learning, and Deep Learning. *Source: Adapted from ("Deep learning", 2021)*

[5]https://birdnet.cornell.edu/

3.3 The Influence of Machine Learning on Citizen Science Project's Phases

When it comes to the combination of ML and CS, the role of CS as a possible solution to the problem of a lack of training data in ML algorithms (Keshavan et al., 2019; Wright et al., 2019) has been discussed more intensively than the role of ML in addressing challenges in CS projects. Ceccaroni et al. (2019) explored the AI technologies used in CS projects and the opportunities and risks that are expected to be encountered due to the increase in the use of AI in CS. The authors define three categories for the use of AI in CS including "assisting or replacing humans for completing tasks", "influencing human behavior", and "improving insights". The first category describes the role of AI in fully or partially automating tasks that were previously performed by humans: for example, tasks related to automatically detecting and classifying data, such as classifying species based on images or sounds (Joppa, 2017; Mac Aodha et al., 2018; Parham et al., 2018). The second category discusses the aim of AI to influence human behavior (Deng et al., 2012) and to extend the educational and social benefits of CS to the general public (Joshi et al., 2018). The third category discusses the impact of AI on identifying patterns in CS data for informing research and policies or on facilitating the understanding of CS concepts using ontologies. Another study by McClure et al. (2020) discusses the integration of AI and CS in ecological monitoring. Rather than delving into the details of how AI and CS can be combined, the authors addressed the challenges and opportunities of performing ecological monitoring using only CS, only AI, or a combination of the two. The opportunities and challenges are discussed in the context of six categories, including efficiency, accuracy, discovery, engagement, resources, and ethics. Efficiency refers to the benefits that CS and ML can provide for scientific projects, such as facilitating data collection and automating laborious tasks, as well as the ability to perform extensive data processing when human and machine power are combined. Accuracy refers to the possibility of integrating human and machine intelligence to produce high-quality data or the challenge of providing incorrect and misleading information. Discovery explores the advantages of complex species identification and serendipitous discoveries

made through the partnership of citizen scientists and deep learning. Engagement explores the impact of CS and AI on multidisciplinary engagement. Resources highlights the role of citizen scientists and machines in saving human and financial resources by, for example, freely contributing data and automating complex tasks, but it also covers the challenges of training citizen scientists, large data requirements, and the need for ML experts. Ethics highlights the challenges of potential information misuse when integrating AI. Another recent study by Franzen et al. (2021) also discusses the opportunities and challenges of human–computer interaction in CS with a focus on the concept of transparency when integrating ML in CS projects, which means that information about data use, ML algorithms, and data processing must be transparent and communicated to participants.

Thus, our goal is to expand the existing literature on the integration of CS and ML by focusing not only on the scientific outcomes of CS projects, but also on the participants, who are at the heart of the projects. We therefore address the integration of ML into various components of a CS project (See chapter 2 for details about various phases in CS), and focus on the impacts of ML on three categories: engaging people and sustaining their participation, data collection, and data validation (Figure 3.2).

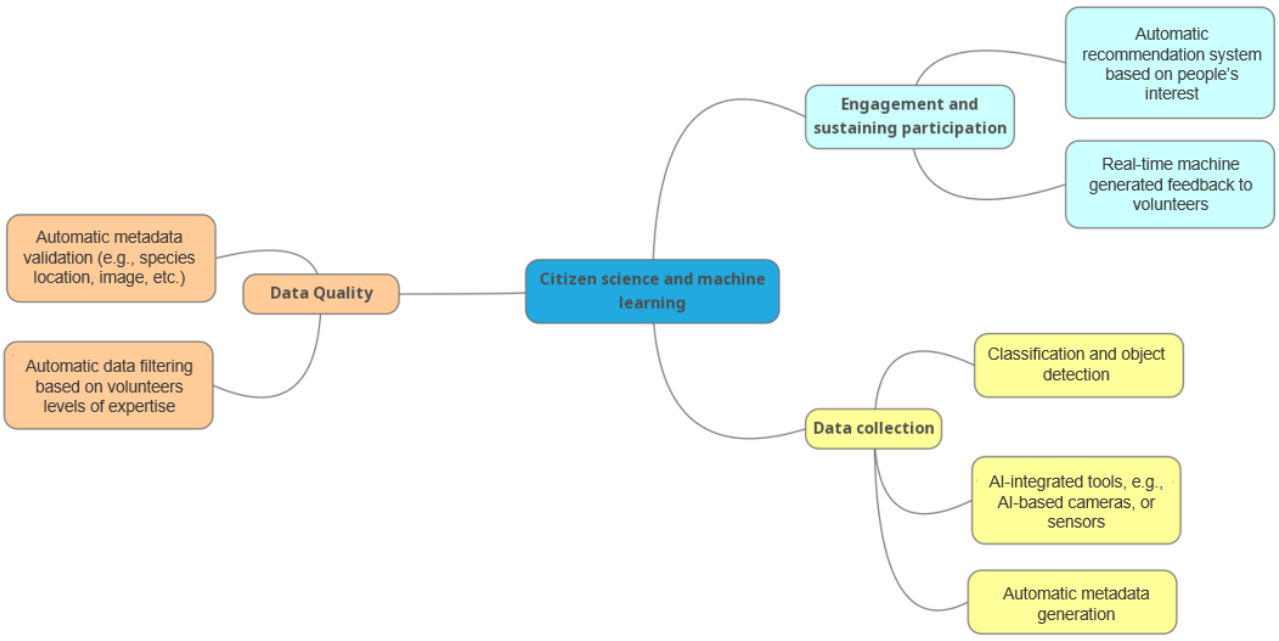


Figure 3.2: A taxonomy showing the integration of ML and CS based on the three CS phases of public engagement, data collection, and data quality (own work).

3.3.1 Machine Learning for Engaging the Public and Sustaining Participation

As mentioned in Chapter 2, a key aspect in a successful CS project is to understand how to motivate the public to participate in a project and how to sustain their participation (Rotman et al., 2014). Depending on the objectives and designs of the CS project, various approaches have been used to engage people (Lotfian et al., 2020). We discuss two potential approaches in using ML towards engaging participants and sustaining participation:

- Automatic community search: The traditional approaches such as word-of-mouth, social media posts, direct emails, workshops, etc., while beneficial for building a community, can be time consuming or require financial resources (for instance, for organization of workshops or ads in newsletters). Antoniou et al. (2016) have proposed a guidance tool to provide information to volunteers so that they can find the VGI project of their choice based on their motivations and interests. To automate what they have proposed, ML algorithms can be used to find and classify the potential target participants based on their interests and to introduce a project to them accordingly. Several studies have been conducted to apply ML algorithms to extract relevant information from social media (e.g., Twitter or Instagram) posts, such as where the images were taken, what type of content is contained in the image, or what topic is mostly discussed in the textual posts (Devaraj et al., 2020; Park et al., 2019). As a result, similar approaches can be adapted to CS projects by employing ML techniques such as CV and NLP to identify people's interests from social media posts and linking them to the relevant CS project. Furthermore, to the best of our knowledge, the use of ML in user profiling to create a recommendation system (Barnard, 2012; Kanoje et al., 2016) where CS projects are recommended to people based on their socio-demographic details is not used as a way to engage people to contribute to CS projects. Moreover, the use of chat bots in CS projects can be a potential approach in engaging and sustaining participation, which has been applied in few studies (Adriaens et al., 2021; Schade et al., 2020; Tinati et al., 2017b). Chat bots may also help as a

real-time guide for participants.

- Automatic feedback to participants: As discussed in Chapter 2, participants may become discouraged if they do not receive feedback on their contributions (Ingensand et al., 2015; Kelling et al., 2011). Moreover, due to massive amounts of data, it is time-consuming to provide feedback to all participants, or often, feedback from experts is provided after a long time has passed (Bonter & Cooper, 2012; Kelling et al., 2011). In order to inform participants regarding the quality of their contributions and to update them regarding the project advancements, automatic informative and user-centered feedback can be generated using ML algorithms (Lotfian et al., 2019). The participants can be informed about the quality of their contribution and how they can enhance it and can learn from the feedback provided (Adriaens et al., 2021) (e.g., learning about biodiversity through feedback regarding species habitat characteristics). Thus, human–computer interaction through machine-generated feedback can be a strategy for increasing and sustaining participation in CS projects.

3.3.2 Machine Learning for Data Collection

As mentioned in Chapter 2, depending on the typology of CS, volunteers contribute differently. For example in classic CS, contribution often requires little cognition engagement, such as collecting biodiversity data (e.g., photographs of species), recording noise (Guillaume et al., 2016) or air pollution levels. However, in other typologies such as volunteered thinking, human cognition is more employed to collect information, such as in labeling and identifying objects in images; or in more complicated projects, training prior to data contribution is required to complete tasks, such as identifying protein structures in the Foldit project (Cooper et al., 2010) or georeferencing historical images in the sMapShot project (Produit & Ingensand, 2018). Thus, by incorporating ML techniques into CS, the data collection task can be partially or fully automated. As a result, considering the two key types of data collection, we define two possible approaches in which ML can be integrated in this step:

- Machines as sensors (adapted from citizens as sensors): The integration of ML in the projects where limited human cognition is required for data collection, can be performed using AI-based tools, such as AI-integrated cameras. A well-known example in ecological studies is the use of camera traps to automatically capture images of species (Wiggers, n.d.). Moreover, sensors integrated with ML techniques can automatically record measurements such as noise recording (Monti et al., 2020) or air pollution (Van Le & Tham, 2017).

- Collective intelligence: For other projects where cognition is involved for data collection, ML algorithms can learn to automate certain tasks, such as object detection in images/videos, which is the most common technique, or more complex tasks, such as automated prediction of protein structures using deep learning (Panou & Reczko, 2020).

3.3.3 Machine Learning for Data Validation

As mentioned earlier (See Chapter 2), due to large amounts of data being contributed to CS projects, manual expert validation can be very time intensive. Thus, automatic or semiautomatic data validation can be applied by filtering potential erroneous data, considering both the contributed information and the ability and experience of participants in contributing data. Two types of potential automatic validation approaches can be the following:

- Automatic data quality assurance: The static comparison of the contributed data with reference data sets has been used in biodiversity CS projects to perform automated filtering of unusual observations (Kelling et al., 2011). However, rather than comparing the submitted data with the historical records, the ML algorithms could be used to perform model-based validation and real-time verification of the newly contributed data. For example, species distribution models can be used to validate the spatial accuracy of biodiversity observations, or a CNN algorithm can be used to validate images labeled by the participants.

- Classification of participant's level of expertise: The level of expertise and experience

in contribution varies among participants in CS projects. For example, in biodiversity monitoring projects such as eBird[6] or iNaturalist[7], some participants contribute observations casually, while others are very involved and experienced and may even be considered as expert volunteers not only to contribute data but also to verify others' observations (Kelling et al., 2015). Thus, the contributors' previous records can be used in ML algorithms to classify the participants (e.g., by assigning them scores based on their level of expertise), and the newly contributed data can be validated based on the classification of the participants' levels of expertise.

Figure 3.2 illustrates a taxonomy of possible combinations of ML and CS, which is classified according to the CS phases, including the three discussed categories of public engagement, data collection, and data validation. Some of these ML integrations have already been applied in current CS projects, such as the automatic species identification or the classification of observers' levels of expertise in eBird, which will be explored in greater detail in the next section. Nevertheless, there are some other potential integration that, to the best of our knowledge, have not been applied in current projects, notably in terms of the role of ML in engaging participants through user profiles and recommendation systems. The following section presents and categorizes the case studies, taking into account the taxonomy discussed in the current section.

3.4 Selected Case Studies of Combination of Citizen Science and Machine Learning

In this section, we present some of the case studies in which ML and CS are combined, with the goal of developing a typology of such projects based on the AI and ML applications outlined in section 3.2 and the various integration of ML on CS tasks outlined in section 3.3. We begin by categorizing the case studies based on the field of science and then present the most commonly

[6]https://ebird.org/home
[7]https://www.inaturalist.org/

used approaches in each category. The categorization of the use cases is shown in Table 3.1.

Environmental science: The most common approach in environmental studies is training ML algorithms using the images/videos labeled by citizen scientists to automate species identification and/or classification. Some of the common applied methods are as follows:

- Camera traps for species identification: when it comes to the combination of ML and CS in biodiversity research, one of the most common approaches is the use of camera traps, where cameras are installed in nature to take photos of species, and the photos are then labeled by citizen scientists to feed and train ML algorithms (Green et al., 2020; Willi et al., 2019). Citizen scientists may, depending on the project, be involved in only one or all the activities of camera placement, submission of images, and labeling and classification of images/videos from camera traps (Green et al., 2020). MammalWeb (Hsing et al., 2018), eMammal (McShea et al., 2016), and WildBook (Berger-Wolf et al., 2017) are three examples of projects focused on camera traps data, and depending on the projects' goals, they invite volunteers to either collect or classify images (Table 3.1). The use of contributed images to train CNN algorithms for automatic wildlife identification can result in the implementation of software packages such as the R package MLWIC (Machine Learning for Wildlife Image Classification) (Tabak et al., 2019), which can be useful for environmental studies, particularly for ecologists. Another approach of integrating human and machine intelligence in camera trap projects is to invite volunteers to observe species images and confirm machine predicted labels in each image (Willi et al., 2019). This approach helps to balance the time required for labeling images while maintaining high quality classification, and human intelligence is used for verification and identifying more challenging species that are difficult for machines to classify.

- Species identification based on images and metadata: the majority of species identification projects use only images to train ML algorithms (Weinstein, 2018). However, the identification of some species only with images and in the absence of other metadata is very complex both for humans and machines, and only human experts are able to distinguish among various images. Including metadata such as the spatial and temporal

distribution or the ability of observers to identify species can increase ML predictive performance and provide more confidence in species identification. One example in this case is a study performed by Terry et al. (2020) to identify ladybirds using both images and metadata such as location, date, and observer's expertise (Table 3.1). Another example is the eBird project (Yu et al., 2010), where a probabilistic model has been developed to classify observers as experts and novices, taking into account their experience in making contributions (Table 3.1). Another project, BeeWatch, invites citizen scientists to identify bumblebee species in images (Van der Wal et al., 2016), and it employs Natural Language Generation (NLG) to provide volunteers with real-time feedback (Table 3.1). Experiments conducted by the BeeWatch researchers with project participants, revealed that the automatically generated feedback improved the participants' learning and increased their engagement (Van der Wal et al., 2016).

- Marine life identification: unlike other species, marine life identification by combining ML and CS has rarely been discussed (Langenkämper et al., 2019). In an article by Langenkämper et al. (2019), the authors focused on combining ML and CS in annotation of marine life images. Citizen scientists are requested to annotate the images (digitize a bounding box around the species in the image); however, there is a possibility that volunteers may miss identifying the species (false negative), annotate a species which is not present in the image (false positive), or place the bounding box incorrectly. Despite all of the possible annotation errors, the authors conclude that merging CS with ML in marine life studies has considerable promise, providing that citizen scientists receive sufficient training prior to image annotation (Table 3.1).

- Automatic wildlife counts from aerial images: estimating wildlife abundance is an important aspect of biodiversity conservation studies. One approach is to count the species in aerial images. However, if done entirely manually, this is an extremely time consuming and labor-intensive process. A study focused on the counts of wildebeests in aerial images (Torney et al., 2019) has illustrated promising results in obtaining accurate counts by combining CS and deep learning (Table 3.1). In this study, the counting is done by both citizen scientists and machines (a trained CNN algorithm), and while the results

indicate that the machine performance is faster and more accurate than the human, the authors state that the citizen scientists' contributions are essential in providing training data to feed the algorithm.

Neuroscience: similar to environmental studies and species identification tasks, citizen scientists' input can be very valuable in amplifying the gold standard data generated by neuroscience experts. In (Keshavan et al., 2019), an approach is proposed to amplify expert-labeled MRI (Magnetic Resonance Imaging) images using CS and deep learning. This approach involves three main steps. First, the experts label a collection of MRI images. Second, to amplify the labels, a web application called Braindr[8] is implemented that presents a 2D brain slice to citizen scientists, and they are required to pass or fail the image taking into account its quality (check Figure 3.3). Finally, in the third step, a deep learning algorithm is used to verify the quality of the CS labels compared to the expert-labeled MRI images. Once the high-quality data are available, they are used to train a CNN algorithm to automate labeling the MRI images.

Astronomy: the involvement of the general public in online astronomy projects started in 2008 with the first release of the Galaxy Zoo project (Lintott et al., 2008). Traditionally, the classification of galaxy images in Galaxy Zoo was done by citizen scientists, but with advances in ML, the classification task was automated using amateurs and expert labels as input training data (Jiménez et al., 2020). The Milky Way project is another well-known project in this field, with the goal of involving volunteers in identifying bubbles in images collected from space telescopes (Kendrew et al., 2012), and to automate the identification, the volunteers' labels were then used to train a RF algorithm called Brut (Beaumont et al., 2014). The authors mentioned that the combination of ML and CS in astrophysical image classification has opened a new path towards obtaining large scale classified data sets, which would have been more complex to achieve if each of these fields (CS and ML) were applied separately.

Table 3.1 illustrates that the majority of projects that combine CS and ML are in environmental science, which is also true for CS projects in general, where the number of biodiversity CS projects far outnumbers projects in other domains (Pettibone et al., 2017). Furthermore, the

[8]https://braindr.us/

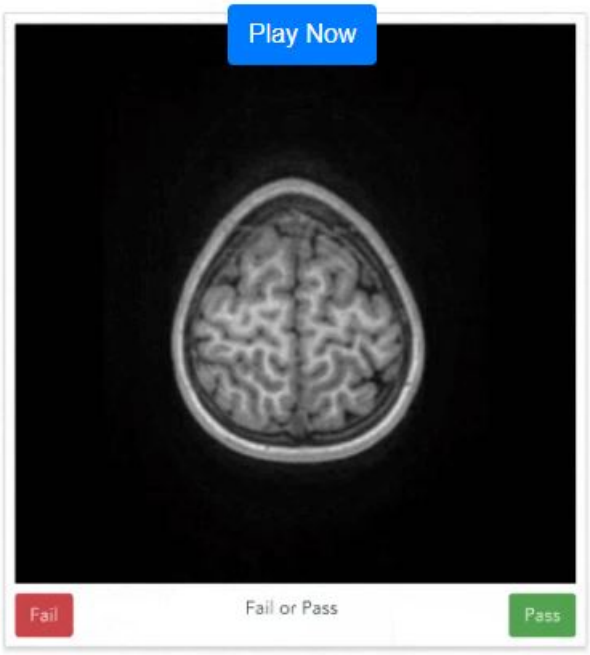


Figure 3.3: Screenshot from Braindr application where citizen scientists are required to label the MRI images by selecting pass or fail. *Source: https://braindr.us/*

table shows that, regardless of the area of science, the integration of CS and ML comprises primarily the use of labeled data from citizen scientists to feed ML algorithms. Typically, trained models are used to automate data collection (mostly labeling and object detection tasks in online CS projects) and data validation (automatic filtering and flagging the erroneous contributions). In contrast, the use of ML in CS to increase and sustain participation has received far less attention, with the BeeWatch project being the only one (among the studied use cases) that has directly evaluated the effects of automatic feedback on engagement.

Furthermore, while in most projects, once the model is trained, the identification or the labeling tasks can be completely automated, the majority of authors argue that the role of citizen scientists does not fade away and that human cognition can be used to perform more challenging tasks, such as verifying machine predictions or identifying rare species. Given these current projects and the prospect of further possible ML and CS integrations, the next section discusses the benefits and challenges that may arise as a result of this combination.

Table 3.1: Example of case studies of combination of machine learning and citizen science

Science Field	Case Studies	Impact on Citizen Science Task	Machine Learning Technique	Brief Objective
Environmental science	Wildlife species identification using camera traps (Willi et al., 2019) (Green et al., 2020) (Hsing et al., 2018; McShea et al., 2016) (Berger-Wolf et al., 2017)	Automate data collection: - Automatic identification of species in images - Automatic photo capturing using AI-integrated cameras Automate data validation: - Real-time validation of new labeled images	Supervised learning: computer vision and use of CNN	- The labeled images by citizen scientists are used to feed and train CNN algorithms to automate wildlife identification in images. - Volunteers can contribute other types of information besides images, such as species habitat, or they can focus on more challenging tasks, such as rare species identification.
	Ladybird identification based on images and metadata (Terry et al., 2020)	Automate data validation: - Auto-filtering of new observations - Auto-identification of species based on images and metadata	Supervised learning:NN for metadata only, CNN for image only, and a combined model for metadata and images	Train ML algorithms to automatically identify ladybird species using images along with the structured metadata (date, location, and citizen scientists' experiences)
	eBird, use observers' expertise to verify the contribution (Johnston et al., 2018; Yu et al., 2010)	Automating data validation: - Screening of new observations based on observer's ability	Using probabilistic models and automated reasoning based on observers' previous contributions	Classification of citizen scientists to experts and novices to improve identification of new species, and pass the rare species detection task to the expert observers
	BeeWatch, identification of bumblebees (Van der Wal et al., 2016)	Generate automatic feedback to: - Improve participants' learning to identify bumblebees - Increase participants' engagement	Natural Language Generation (NLG)	Automatically generate feedback with the aim of improving participants' ability to identify bumblebees and increasing their engagement
	Marine life identification (Langenkämper et al., 2019)	Automate data collection and validation: - Identification of marine life in images - Auto-detection of location (ROI) of marine species in the images	Supervised learning: computer vision and CNN (use of transfer learning)	Improving marine species identification by combining the power of citizen scientists and deep learning
	Automatic species count from aerial images (Torney et al., 2019)	Automatic data validation: - Auto-filtering erroneous contributions caused by volunteer miscount - Automatic validation of species reports based on the expected density	Supervised learning: computer vision and use of CNN	Combining the power of citizen scientists and deep learning to improve wildlife counting in aerial images for conservation purposes
Neuroscience	Braindr (Keshavan et al., 2019)	Automatic data collection: - Automatic labeling of MRI images Automatic data validation: - Validation of new added labels by citizen scientists	Supervised learning: computer vision and use of CNN	Amplification of expert-labeled MRI images with the help of citizen scientists, followed by the use of the amplified labels to train an algorithm to automatically replicate the labeling task of experts
Astronomy	Galaxy Zoo (Jiménez et al., 2020)	Automatic data collection: - Automatic classification of galaxy images	Supervised learning: computer vision and use of CNN	Classify galaxy images by training an ML algorithm based on citizen scientists' input
	Milky Way (Beaumont et al., 2014)	Automatic data collection: - Auto-detection of bubbles in space telescope images Automatic data validation: - Auto-filtering of amateurs' contributions	Supervised learning: Random Forest algorithm	- Detect bubbles in space telescope images by feeding an ML algorithm using labels provided by citizen scientists and experts - Let the participants spend time on labeling the more challenging images

3.5 Benefits and Challenges of Integration of Machine Learning in Citizen Science

Although it is discussed that the combination of ML and CS offers more benefits than when they are implemented in isolation (McClure et al., 2020), there are several points that need to be considered prior to the integration of ML and CS. When integrating ML in CS, it is critical to maintain transparency. In other words, participants must be given sufficient information about how their contributions are being used, as well as specifics about data processing and ML algorithms. Furthermore, it is critical to consider the potential biases in model predictions that may be caused by algorithm mis-training (See (Mehrabi et al., 2021) for examples of discriminatory systems caused by the biased data in training the algorithms).

In this section, we discuss the benefits of combining CS and ML, as well as the potential challenges that can arise if ML is not used cautiously in CS projects. The benefits and challenges of ML and CS integration are discussed in the scope of *engagement*, *data quality assurance*, and *ethics* (check Figure 3.4). Data collection is not listed as a separate category in the section of benefits and challenges since the impacts of ML on this step are integrated into the categories of engagement and data quality.

3.5.1 Benefits and Challenges for Participant' Engagement

- Benefits: As mentioned earlier, one of the benefits of AI for community building in CS projects is to encourage engagement by targeting the potential volunteers through social media. Another important factor in CS is the impact of the interaction with and feedback to the participants on the basis of their contributions (Tang et al., 2019; Zhou et al., 2020). Thus, the use of ML in CS in providing automated feedback to the participants might promote engagement through human–computer interaction and result in sustaining participation. Furthermore, the intelligently generated feedback can provide participants with useful knowledge about the research subject, allowing them to learn while contribut-

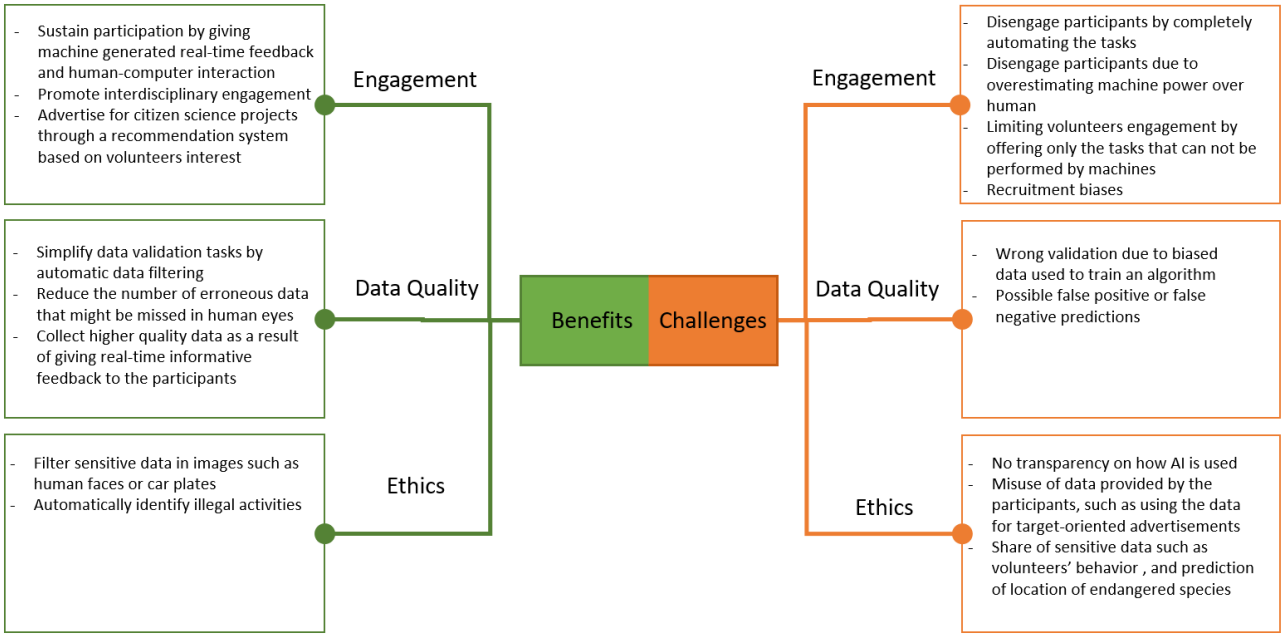


Figure 3.4: Benefits and challenges of combining CS and ML (own work)

ing, which can be another factor in increasing participation (e.g., BeeWatch project (Van der Wal et al., 2016), or the use of chatbots in identification of invasive alien species (Adriaens et al., 2021)). Another potential benefit of combining ML and CS is to encourage interdisciplinary engagement among volunteers and researchers, which can lead to collaborations from several scientific fields (McClure et al., 2020). Finally, automating certain simple tasks allows volunteers to concentrate on more complicated ones, such as identifying common species from camera trap images using CNN and leaving the identification of the unusual species to volunteers. However, there is another side to the task automation, which is discussed in the challenges section.

- Challenges: The use of ML in CS could result in the automation of most of the tasks, which may demotivate participants because they are fully or partially being replaced by machines. As previously mentioned in the use cases, in most projects, CS data is used to train ML algorithms, and then the tasks can be performed entirely by machines, effectively replacing humans. While it has been mentioned that in the case of task automation, citizen scientists would then concentrate on more challenging tasks, some participants would like to contribute to CS projects to fill their spare time with activities that make them feel good, such as helping science or spending time in nature (see (Lotfian et al.,

2020)), which are not inherently challenging. For example, in the sMapShot project (Produit & Ingensand, 2018) (a CS project for georeferencing historical images), there is strong competition among participants of higher age groups, and the incentive system plays an important role in motivating them; therefore, if the computer performs the task more efficiently, motivation is expected to drop, and thus participation will decline. One solution is that, considering all activity levels among participants, participants are allowed to contribute with their task of interest even if the task can be fully automated by machines, and thus the contributions can be helpful in retraining the algorithms to have a better performance. Another recommendation is to incorporate new forms of contributions to fill in the gap caused by automated tasks. Furthermore, another potential risk is the overestimation of AI power in CS projects, such as trusting model predictions over expert volunteers, which could result in disengaging the participants (Ceccaroni et al., 2019).

3.5.2 Benefits and Challenges for Data Quality

- Benefits: The use of ML in CS will speed up the process of big data validation, reducing the workload of manual data quality assurance for experts (Bonter & Cooper, 2012; Lotfian et al., 2019). Prescreening and filtering data (for example, removing empty images or low-quality images in camera trap projects), flagging erroneous observations, and submitting only flagged observations for expert verification will save a lot of time and allow the experts to concentrate on the scientific aspects of the project rather than the manual filtering of all data. Furthermore, the generation of real-time informative and user-centered feedback for participants with information about their contributions will improve the participants' knowledge on the subject, their proficiency, and, as a result, the quality of data they contribute over time. Another finding from the BeeWatch project (Van der Wal et al., 2016) concerning the impact of feedback on volunteers was that NLG feedback resulted in increased learning, and the identification accuracy was higher for those who received informative feedback than for those who only received confirmation

of correct identification.

- Challenges: Although the benefits of automatic filtering and validation have been discussed, the efficiency and reliability of automated validation and feedback are highly dependent on the data used to train the ML algorithms. For example, if the training data are biased in some way, such as spatially or temporally, the automated data validation based on the trained model is also biased and could provide participants and experts with misleading information (McClure et al., 2020). In addition to bias in the data, it is critical that the data used to train the model are of a gold standard and validated by experts, since the trained model will be used to verify new data, and if the input data are uncertain, the model will predict false detections (McClure et al., 2020), such as failing to identify a species, in the case of a false negative, or incorrectly detecting an abnormal shape in an MRI image, in the case of a false positive. It is important to keep in mind that machine intelligence should not be overestimated in comparison to human intelligence. In other words, when participants receive machine-generated feedback on their contribution, the decision to either modify or retain the contribution should be made by the participants, and human experts will make the final confirmation in such cases. It is also necessary to note that when a model is trained on data from a specific region, it cannot necessarily be applicable in other areas, and doing so can result in misevaluation and the generation of misleading information. Furthermore, training algorithms for small data sets (such as rare species, see (Norouzzadeh et al., 2018)) or multitype data sets (such as a mix of images and metadata, see (Terry et al., 2020)) and learning how to tune the parameters of the algorithms to achieve the desired performance are hard challenges that must be considered prior to performing automated data validation in CS projects.

3.5.3 Ethics

- Benefits: The use of ML can be advantageous in filtering sensitive information from CS data, such as human faces or license plates in images. Furthermore, ML can be used to detect illegal actions, such as illegal animal trades, by sentiment analysis using information

posted on social media platforms such as Twitter (Di Minin et al., 2019).

- Challenges: One major concern of integrating ML in CS is the use of data collected from participants for other commercial reasons, which may go against the participants wishes and result in their disengagement from the project. Thus, it is critical to be transparent and communicate effectively with participants on how their inputs are being used in the algorithms, rather than simply creating a black box project in which the participants function is limited to producing data and feeding the algorithms (Ceccaroni et al., 2019; Franzen et al., 2021; McClure et al., 2020). As discussed in (Ceccaroni et al., 2019), technology giants like Google and Facebook offer target-oriented advertisement services by selling personal information, which can be a danger for the future of AI-based services used in CS projects, as it may lead to a lack of confidence on the part of participants to freely share their contributions and personal information. Another ethical issue that may emerge from ML-based CS projects is the sharing of sensitive data that may be deceptive or result in geoprivacy violations, such as predicting the position of endangered species or predicting participant activity based on the history of their contributions.

3.6 Summary

The combination of ML and CS is mainly discussed in terms of providing data to train algorithms, rather than how this combination can impact and benefit CS projects and the participants. In this chapter, we discussed some potential ways that ML can be integrated in CS and address the challenges of public engagement and data quality assurance, as well as reviewing some use cases and finally discussing the benefits and challenges of this integration. Despite existing projects and articles on the integration of ML and CS, this topic is still in its early stages and requires additional research discussing other benefits and challenges, as well as proposing alternative use cases to those that have been applied.

The integration of ML and CS indicate that this combination demonstrates considerable potential for both fields. However, there are some consequences to this as well, as advancements

in AI and the superior power of computers, in some cases better than humans, raise the possibility of completely replacing humans in CS projects. Nevertheless, there are certain tasks that cannot be performed without human input, such as activities that involve *imagination*, *critical thinking*, and *communication skills*. Furthermore, when combining ML and CS, it is critical that the primary goal of CS, engaging the general public in scientific projects and knowledge sharing with the public, does not fade away as a result of giving machines too much control. In addition, it is critical to apply transparency to the project and effectively communicate with volunteers about how ML is being integrated and how the ML algorithms are using participants' input. Finally, prior to integrating ML in CS, the possible challenges and benefits must be thoroughly investigated to determine which one has more weight, as well as to understand how to mitigate risks and maximize benefits from ML integration in the project at all levels, from public engagement to data quality assurance. An argument for future expansion of this chapter is to evaluate the risk of losing participants due to the use of ML or any type of automation in CS projects, and to look for ways to mitigate this risk, such as by involving participants' perspectives on the use of ML, or by creating the possibility of adaptability so that automation is used only if the participants agree with it.

Chapter 4

Hypotheses

The main focus of this thesis, as mentioned earlier in the introdcution chapter, is on the two challenges of public engagement and data quality in CS projects. To that end, the objectives are first to identify and categorize the motivational factors that encourage volunteers to contribute to CS projects, and then to investigate how the integration of ML techniques in CS projects can help in improving data quality and increasing/sustaining participation in such projects. To accomplish our goals, we developed a set of hypotheses to evaluate, and for each hypothesis, we asked one or more research questions, which will be discussed in the following two chapters. We categorize the hypotheses to two categories of *public engagement* and *data quality*. We evaluate and discuss the hypotheses in Chapter 5 (motivational framework for CS projects) and Chapter 6 (use of ML techniques in performing real-time data validation and providing real-time feedback to participants). We used a multi-approach to evaluate the hypotheses, using statistical tests to find correlations between variables, or a more general approach that used directly the data from the surveys with the participants in cases where there were insufficient respondents to strongly support the significance of statistical tests. Hypotheses 1-3 were discussed in Chapter 5 and primarily evaluated using statistical tests, whereas hypotheses 4-6 were discussed in Chapter 6 and evaluated using both statistical tests and direct survey results or indirectly using application log files. However, in cases where there was insufficient data to strongly support the hypotheses with statistical tests, we plan, as a future work, to conduct

the experiment using other methods of engaging participants and gathering their feedback to see if we can have a larger community of participants to obtain significant statistical tests. Our hypotheses and the research questions we aim to discuss are as follows:

User engagement

H1: The design, objective, and type of a CS project are correlated with the participant's motivations to contribute to that project.

- H1RQ1: What are the motivational factors to encourage the public to contribute to CS projects?
- H1RQ2: Are the motivations of CS participants correlated with the designs, objectives and thus the typologies of CS projects?

By posing these three research questions, we want to illustrate that there are a variety of reasons why people contribute to CS projects, and that these can be related to the typology of CS projects. To evaluate and discuss these research questions, we must first identify the motivations of volunteers who have contributed to various CS projects, then determine whether these motivations follow a pattern based on CS typology, or, in other words, whether we can classify these motivations based on CS project typology.

H2: Volunteers' motivation to contribute to a CS project is correlated to where they live.

- H2RQ1: Is there a correlation between the area of participants' residences and their motivations to contribute to a CS project?

This hypothesis and its corresponding research question can be better discussed in the context of CS projects in which the contributors participate in outdoor activities. By posing this research question, we want to demonstrate whether people living in different areas, such as densely urbanized areas or suburbs, would approach contributing to a CS project in different

ways. Furthermore, if that is the case, we want to look into possible ways to encourage various groups living in different zones to be more engaged.

H3: The socio-demographic information of participants is correlated with their motivations and perspectives on contributing to a CS project.

- H3RQ1: Are demographic variables such as age or education level correlated with volunteers' motivations to contribute to a CS project?
- H3RQ2: Does volunteers' familiarity with new technologies, particularly mobile technologies, impact their behaviour to contribute to a CS project?

To engage volunteers from various socio-demographic groups, it is essential to first assess whether these differences impact their motivations to contribute to CS projects, and if so, to investigate what strategies to use to encourage engagement for each specific socio-demographic group, and to design the project accordingly. Therefore, the goals of this hypothesis and its corresponding research questions are to evaluate different groups of participants' perceptions toward contributing to a CS project, as well as to investigate the relationship between socio-demographic variables and the use of a CS mobile/web application.

H4: Giving real-time feedback to the participants can result in increasing engagement.

- H4RQ1: Does giving feedback to volunteers' can increase their motivations to contribute to a CS project?
- H4RQ2: Does giving feedback to volunteers' can result in sustaining their participation?

We discussed in Chapter 3 that one approach of integrating ML to CS is to provide real-time feedback to participants. Therefore, the goal of this hypothesis and its corresponding research questions is to determine whether or not providing real-time feedback tailored to each participant's contribution can result in participants continuing to contribute to the project. In

other words, by asking these research questions, we want to evaluate the influence of informative machine-generated feedback on volunteer engagement.

Data quality

H5: Automatic data filtering can simplify and accelerate data validation task, and improve data quality.

- H5RQ1: Would using automatic data filtering result in reducing the number of erroneous contributions?
- H5RQ2: Would using automatic data validation help in improving data quality in CS projects?

As we discussed in Chapter 3, ML can be integrated into CS to perform automatic data validation. The goal of this hypothesis and its associated research questions is to determine whether automatic data validation can, on the one hand, speed up the data validation process by automatically filtering the data and only sending the flagged ones for expert review, and on the other hand, improve data quality by combining human and machine power.

H6: Giving real-time feedback to volunteers increases their knowledge about scientific domain of the project (e.g., biodiversity) and results in collection of higher quality data.

- RQ8: How giving real-time feedback to participants can improve the quality of their contributions?
- RQ9: Would informative machine generated feedback to the participants improve their learning curve, and encourage them to learn about science (e.g., biodiversity)?

Aside from discussing the role of real-time feedback in increasing engagement, it is also important to investigate its role in educating volunteers by including useful information in real-time feedback. The goal of this hypothesis and its associated research questions is to investigate

whether providing volunteers with machine-generated informative feedback can help them learn about science. Another goal is to see if providing feedback that includes specific information about participants' contributions can help them learn more and, as a result, contribute higher quality data.

Chapter 5

Public Engagement and Establishment of a Conceptual Framework For Motivational Factors

5.1 Introduction

This chapter is based on our published article on establishment of a framework for motivations of citizen science participants, and verifying this framework with our own case study (Lotfian et al., 2020).

As mentioned in Chapter 2, understanding the motivations of CS and VGI participants is critical to the success of a project. Furthermore, as previously stated, the motivations for contributing to a project vary depending on the design and objective of the project. Despite the availability of several studies on motivation in CS and VGI, a comprehensive framework that categorizes motivations while taking project typology and objective into account is lacking (Raddick et al., 2013). Therefore, in this chapter, we reviewed several studies addressing the motivations of participants from various CS projects with different typologies (classic CS, volunteered thinking, volunteered computing, and gamified CS, See Chapter 2 for more detail), and as a result, we

defined a framework that categorizes the motivations according to the project's typology, as well as the strength of the motivation in recruiting volunteers and sustaining their participation. Finally, as a preliminary discussion of the framework, we present the results of an online survey that we conducted to identify the motivations of potential participants to contribute to our biodiversity CS project, BioPocket.

The motivational factors commonly identified for volunteer activities and VGI projects are presented in the following section. We then go over the CS case studies and extract the motivations from these studies, followed by a discussion of the motivational framework we defined taking into account various CS typologies. We then present our BioPocket project and the results of the survey we conducted to determine the motivations of BioPocket participants. Finally, we discuss and evaluate our hypotheses 1,2 and 3 (refer to Chapter 4 for all the hypotheses) taking into account the motivational framework and the results of our BioPocket survey.

5.2 Motivational factors

To better understand why people contribute to CS projects, it is essential to understand why people take part in voluntarily activities in the first place. Clary et al. (1998) provides six potential functions (Volunteer Functions Inventory (VFI)) that explain why people are motivated to do voluntarily activities. The suggested functions are as follows:

1) Values: volunteering provides the opportunity for individuals to express values towards altruistic and humanitarian concerns for others.

2) Understanding: volunteering provides individuals the opportunity for learning and gaining new skills and knowledge

3) Social: volunteering provides the opportunity for interactions with others, or engaging in an activity, which is viewed positively by others.

4) Career: individuals might obtain career-related benefits from volunteering activities (e.g.

open source software developers, or contributors to Stack Overflow[1])

5) Protective: volunteering may reduce guilt over being more fortunate than others. In a study of Red Cross volunteers (Frisch & Gerrard, 1981), some of the volunteers indicated that they volunteer to avoid negative inner feelings.

6) Enhancement: volunteering provides the opportunity for some people to obtain satisfaction related to ego enhancement and self-esteem.

Batson et al. (2002) studied "community involvement" and concluded that motivation can be categorised into four types based on the ultimate goal: egoism, altruism, collectivism, and principlism. Egoism refers to the increase of one's own welfare, altruism refers to the increase of other people's welfare, collectivism refers to the improvement of the welfare of a group, and finally, principlism is the maintenance of one or more moral principles, such as justice. They noted that each of these types of motivation has its own strengths and weaknesses; thus, approaches should be identified in order to increase community involvement, in such a way that the strength of one type of motivation will outweigh the weakness of another (Batson et al., 2002).

In addition to motivational studies in general volunteering, Ryan and Deci (Ryan & Deci, 2000) studied motivation in the area of formal education. Depending on the reasons and objectives which give rise to an action, they distinguished two types of motivation: intrinsic and extrinsic (Ryan & Deci, 2000). Intrinsic motivation refers to performing an activity because it is satisfying or enjoyable. Extrinsic motivation, however, refers to doing something because of its outcome, such as receiving monetary rewards, feeling forced, or facing punishment. Some studies have discussed that while intrinsic motivation promotes creativity, extrinsic motivation prevents it (Hossain, 2012). It has been stated in other papers that a balance between intrinsic and extrinsic motivation, especially in gamified contributory platforms, can involve more participants and maintain their participation (Curtis, 2015a; Tinati et al., 2017a). From a survey conducted by Lakhani and Wolf (Wolf, 2005) on participants in Free and Open Source Software (FOSS) development, they concluded that the most common form of motivation among par-

[1]https://stackoverflow.com/

ticipants was enjoyment-based intrinsic motivation, such as how one feels creative from being part of a project. In addition, a survey on the contributors of open-source projects showed that both internal (e.g., altruism) and external motivation (e.g., career-related motivation, direct compensation) played significant role in increasing participation (Hars & Ou, n.d.).

These motivational studies are considered as the basis for understanding the motivation of volunteers in CS and VGI projects. Taking into account all of the above studies, Budhathoki (Budhathoki, 2010) studied the motivation of OpenStreetMap participants and, as a result, set out a detailed list of possible forms of motivation for encouraging individuals to contribute to VGI projects. Below is the summarised list of intrinsic and extrinsic motivations defined by Budhatohoki (Budhathoki, 2010):

Intrinsic Motivations:

- Unique ethos
- Learning
- Personal enrichment
- Self-actualisation
- Self-image
- Fun
- Recreation
- Instrumentality
- Self-efficacy
- Meeting own needs
- Freedom of expression
- Altruism

Extrinsic Motivations:

- Career
- Strengthen social relations
- Project goal
- Community
- Identity
- Reputation
- Monetary return
- Reciprocity
- System trust
- Networking
- Social political

The types of motivation for individuals to contribute to OpenStreetMap vary, depending on how people are engaged in making contributions. There are a variety of ways that people engage to contribute to OpenStreetMap, such as by joining a local user group or a local chapter[2]. In addition, a very popular way of encouraging people to contribute is by arranging mapping parties or Mapathons (Coetzee et al., 2018). Most Mapathons are organised for humanitarian reasons, such as mapping areas in developing countries where there are large gaps in OpenStreetMap data, or crisis Mapathons, where individuals are encouraged to start mapping shortly after a natural disaster occurs (Quill, 2018). In addition, Kamptner and Kessler analysed the effect of a small-scale disaster (such as a building fire) on user contributions, and their study found a spike in OpenStreetMap data (ways, nodes and attributes) following incidents (Kamptner & Kessler, 2019).

Although Mapathons have a major effect on the increase in OpenStreetMap contributions, a study by Hristova et al. (2013) indicated that these events are not enough to sustain participation, and that newcomers typically stop participating after the event is over, although experienced users continue to contribute. Khanal et al. (2019) therefore analysed the Digital

[2]https://wiki.openstreetmap.org/wiki/How_to_contribute

Table 5.1: Some of the motivational factors to contribute to OpenStreetMap

	Motivational factors	Detailed motivations to contribute to OpenStreetMap	Example studies
Intrinsic	Altruism	Helping to map rural areas in developing countries Mapping to help after natural disasters	(Kamptner & Kessler, 2019; Khanal et al., 2019; Quill, 2018)
	Personal enrichment	Learning digital leadership skills Learning about information technology Understanding about geography and mapping	(Coetzee et al., 2018; Green et al., 2019; Khanal et al., 2019)
	Fun	Enjoying the organized activities in mapping parties Improve the gameplay experience in applications which use OSM	("Gamification - OpenStreetMap Wiki", n.d.; Johnson, 2019)
Extrinsic	Strengthen social relations	Being part of a mapping community Social networking and making friends	(Coetzee et al., 2018; Fritz et al., 2017)
	Career	Including learned (mapping) skills in professional CV Receiving mapping certificates for future career	(Coetzee et al., 2018; Fritz et al., 2017; Green et al., 2019; Khanal et al., 2019)
	Monetary rewards	Receiving school credits Receiving an award (e.g. a mug or a USB key)	(Fritz et al., 2017; Green et al., 2019)

Internship and Leadership (DIL) strategy to address OpenStreetMap data disparities in rural Nepal by engaging and retaining the participation of youth mappers through virtual internships.

In addition, people also contribute to OpenStreetMap because they want to collect data for other applications (mostly in the form of games) that use OpenStreetMap. For example, an analysis of the impact of Pokémon GO[3] (an augmented reality game) on OpenStreetMap contributions in South Korea showed a significant rise in the number of daily contributors and daily edits after the game was released, in order to enhance the in-game map (Johnson, 2019). They found that "park" and "water body" features were created or edited more by contributors who were motivated to contribute because of Pokémon GO than by other frequent OpenStreetMap contributors. They also stated that the increase in the number of edits/contributors fell to regular levels around one and a half months after the release of the game.

Although it is difficult to derive all forms of motivation among OpenStreetMap participants given the broad possibilities of contributing to this project, table 5.1 shows some of the key forms of motivation that we have taken from the literature for the above mentioned contribution possibilities.

[3]https://pokemongolive.com/

5.3 Citizen Science Case Studies of Analyzing Participants' Motivations

In this chapter, we reviewed the motivational studies for different types of CS projects (See Chapter 2, Section 2.3 for the typology of CS) and extracted the motivation types, taking into account the conceptual motivational studies listed in section 5.2. To do so, we automatically downloaded 100 articles from Google Scholar with the keywords of "CS", "VGI", "motivation", "participation" and "engagement", from which we extracted the ones directly related to motivational studies in CS and VGI projects. A Python[4] script was used to retrieve the articles automatically, providing the titles, authors, publication years and number of citations for each article. We found use cases for classic and CCS (VT, VC, and CS game) projects, but did not find studies analysing participant motivation in environmental management projects. Table 5.2 illustrates the studied use cases in each CS typology. The use cases were selected on the basis of three key criteria: the results of each use case are published in peer-reviewed journals, the number of citations, and the sample size of the participants whose motivation was analysed in each use case (minimum 100 respondents). All use cases conducted online surveys or interviews to determine the motivation types of their participants, along with other information, such as socio-demographic background. Various methods to identify motivations were used:

- Volunteers are presented with a list of possible types of motivations and are asked to rank them according to their priorities.
- Volunteers are presented with a list of possible types of motivations and are asked to give a score to each.
- Volunteers are asked open questions about their motivation, and hence, motivational factors are derived by coding free text responses. The approaches used to categorise the text responses are different and the theories used in such use cases are stated in this section.

[4]https://www.python.org/

Table 5.2: Case studies reviewed in this article for each CS typology

Citizen Science Typology		**Case Study**
Classic CS		Dutch biodiversity monitoring (Ganzevoort et al., 2017) Great Pollinator (Domroese & Johnson, 2017) Water quality monitoring (Alender, 2016)
Citizen Cyber Science	Volunteered Thinking	Galaxy Zoo (Lintott et al., 2008) Stardust@home (Westphal et al., 2006) Planet Hunters[5]
	Volunteered Computing	SETI@home (Anderson et al., 2002) Folding@home (Beberg et al., 2009a)
	CS game	Foldit (Cooper et al., 2010) Eyewire (Tinati et al., 2017a)

We studied three use cases for classic CS: Dutch biodiversity recorders (Ganzevoort et al., 2017), Great Pollinator Project[6] (Domroese & Johnson, 2017), and water quality monitoring (Alender, 2016). In the use case of Dutch national biodiversity recorders, an online survey was designed to address the following questions: What are the characteristics of citizen scientists with respect to their activities and socio-demographic backgrounds? What are their motivation to contribute to biodiversity monitoring? And what are their views on data ownership and data sharing? (Ganzevoort et al., 2017). The most remarkable conclusion drawn from the authors' analysis of responses to the second question illustrated that the biodiversity recorders were mainly motivated by intrinsic motivation such as: learning about nature, spending time in the nature, and the desire to help by contributing to science and nature conservation (Ganzevoort et al., 2017). In contrast, social motivation, such as interacting with others or building friendships and self-development, were given low importance (Ganzevoort et al., 2017).

Volunteers in the Great Pollinator project are encouraged to collect bee observations with the aim of understanding more about bees and other pollinators in New York City. Surveys were carried out to explore who the active volunteers are, their motivation for joining the project, the benefits they received by contributing to the project, and the difficulties they encountered when gathering and submitting data (Domroese & Johnson, 2017). The results of the survey highlighted that the major motivation for most respondents was to learn about bees, accompanied by contributing to a scientific project and spending time in nature (Domroese & Johnson, 2017). In contrast, social factors, such as being motivated by participating with family and

[6]http://greatpollinatorproject.org/

friends, or participating in a team, were not strong forms of motivation for the participants, and just five percent of respondents listed them as their motivation.

Unlike the first two, the third classic CS use case is not about capturing biodiversity; rather, its objective is about monitoring water quality (Alender, 2016). The top three forms of motivation among participants contributing to several water quality monitoring projects are helping the environment, helping a community, and being connected with nature (Alender, 2016). Moreover, like other projects, building social networks and being known among a community are given little importance; however, working with like-minded people was an important motivation. They also compared motivation types among various age groups; some variations among younger and older participants were observed for social and career-related motivation. Even though career-based motivation and recognition among others were given low scores for all age groups combined, they were reported as stronger reasons to participate among younger participants.

For the first CCS typology, VT projects, Galaxy Zoo (Lintott et al., 2008), Stardust@home (Westphal et al., 2006), and Planet Hunters were reviewed as the most well-known use cases in this category. Galaxy Zoo (the first and most well-known virtual CS project in Zooniverse) aims to encourage citizen scientists to classify the shape of galaxies in images (Lintott et al., 2008). As a result of open-ended interviews with Galaxy Zoo volunteers, 12 motivational factors were identified from the responses using the "grounded theory" (Raddick et al., 2013). The result of the interviews formed the basis for a second survey to analyse once more the motivation of Galaxy Zoo participants (Raddick et al., 2013). In the second survey, three approaches were used to determine the motivation type: first, rating the 12 motivational factors using a Likert scale; second, obtaining additional motivation as free text responses beyond these 12 predefined ones; and finally, asking respondents to indicate which motivation is the most important for them. The framework used in Galaxy Zoo to explain motivation is based on a theory called "expectancy-value", which states that an individual's motivation to perform an action depends both on the outcome they expect to occur (expectancy) and on the value they put on achieving that outcome (Atkinson, 1957). The results illustrated that the main motivations were compassion for project objectives and interest in scientific content, while

learning about science and participating in a social community appeared to be less motivating. Moreover, the interesting conclusion was that even though contribute was not the highest rated motivation in the Likert scale, it was selected as the primary motivation by a large group of respondents (nearly 40%), regardless of their age, gender or level of education (Raddick et al., 2013). Furthermore, the authors mentioned that respondents who rated contribute as their primary motivation rated science higher, meaning that contributing to a scientific project seemed to be a strong motivator. Another interesting outcome was the negative relationship between the two motivation types, help and discovery, indicating that people whose motivation was to help science by classifying galaxies were less concerned about finding new and rare shapes in images of galaxies.

The next VT use case is Stardust@home, where volunteers (also known as "Dusters") classify images from the National Aeronautics and Space Administration (NASA) Stardust spacecraft by searching for the tracks left by interstellar dust particles (Westphal et al., 2006). Prior to studying the motivation types among Stardust@home volunteers, a research model was defined, which included six motivation categories that may result in increasing participation: collective motivation (valuing the project goals), norm-oriented motivation (expectations of the reaction of other important people such as family, friends or colleagues), identification (being identified as a member of a group), intrinsic motivation, reputation, and social interactions (Nov et al., 2011). The survey results illustrated that collective and intrinsic motivations were most widespread, while identification and norm-oriented motivations were of secondary importance. Similar to Galaxy Zoo, social interaction was not rated as a strong motivator.

The third VT use case is Planet Hunter, which is also one of the Zooniverse projects; its goal is to analyse data obtained by the NASA Kepler Space Observatory. The Kepler Space Observatory monitors the stars and tracks their brightness every half hour. When a change of the brightness is detected, it could be attributed to a planet passing in front of it. Participants are required first to answer a question about the shape of the light curve and then to state whether or not there is a change in brightness. An online survey was performed to explain four major aspects related to Planet Hunter participants (Curtis, 2015b): demographic characteristics, patterns of participation, motivation and reward, and interaction and contribution. Out of

160,000 registered volunteers at the time of the survey, only 118 individuals responded to it, and similar to other online CS projects, the majority of respondents were male. An open-ended question was asked to understand volunteers' motivation to participate in Planet Hunters, and the responses were coded to 13 different categories. Similar to other CS projects, interest in science and making a contribution to scientific research were the top two motivators. The possibility of making a new discovery was viewed as a strong motivation by approximately 22 percent of volunteers, and interestingly, similar to Galaxy Zoo, it was noted that it is important for volunteers to be recognised for their new discoveries by being included in the article acknowledgment or article co-authorship. The opportunity to learn was a motivator only for four per cent of respondents, although it should be noted that interest in learning about exoplanets comes into the category of interest in science. Moreover, giving importance to the project's goal and having fun were mentioned by a few people as their reasons for participating in Planet Hunters.

For the next CCS category, VC, we studied the survey results of the two use cases Folding@home[7] and SETI@home (Anderson et al., 2002). Launched in 1999, SETI@home is the first and remains a successful VC project. It uses radio data to search for intelligent life in outer space. In order to study the motivation of SETI@home participants (Nov et al., 2010), the authors developed a framework which groups motivation into four main categories, considering whether they are intrinsic or extrinsic, and whether they are self-oriented or project-oriented. The four categories are as follows:

1)Enjoyment (intrinsic and self-oriented): enjoyment of interacting with others
2)Value (intrinsic and project-oriented): valuing the project' objectives
3)Reputation (extrinsic and self-oriented): gaining reputation among others in the project
4)Enhancement (extrinsic and project-oriented): seeing project's results get published in a scientific journal

In addition to the four categories listed above, two social factors related to the participation of volunteers in the project were addressed. The first, team affiliation, refers to volunteers who

[7]https://foldingathome.org/

participate as a team, such as a university team or a national team; the authors hypothesised that team affiliation would result in increased contributions. The second social factor is the effect of project membership on participation. The authors hypothesised that lifetime membership in VC projects would be associated with a reduced level of contribution to the project. The analysis of their survey data illustrated that the two self-oriented motivations, i.e., enjoyment and reputation, received higher scores compared to the project-oriented motivations, value and enhancement. The results showed that enjoyment and reputation were not related to the level of contribution, and this, as they concluded, might have been due to the nature of VC projects where contribution is passive, and enjoyment from social interactions in such projects is different from that in other kinds of online community-based projects. While enhancement was given a low score, the results showed a positive statistically significant relationship between enhancement and contribution level, which could be because seeing the results of the project published in scientific journals and acknowledging the majority of active participants is unconsciously causing people to make greater contributions. In addition, as hypothesised by the authors, affiliation in a team is positively linked with level of contribution; this is an interesting finding that demonstrates that even in VC projects where social interactions are not required for contribution, being part of a social group still plays an important role. Finally, there was a negative association between being a long-term member of the project ("the tenure of membership") and the level of contribution; the authors found this to be in line with contribution to other social communities, where contributors lose interest in projects when the excitement of community involvement fades over time. However, if the participants are affiliated with a team, the chances of losing interest are less compared to when they are contributing individually.

The next reviewed VC use case is the Folding@home project, which has the objective of understanding how protein molecules fold into their final 3D structure (Beberg et al., 2009b). The project also aims to simulate this folding process millions of times more slowly than the natural process and to use this simulation to study the folding aspects in order to find cures for diseases such as cancer, Alzheimer's, and so on (Beberg et al., 2009b). Therefore, volunteers' computing power is being used to help Folding@home scientists perform the intensive computational simulation tasks. Curtis (Curtis, 2018) analysed participants of Folding@home to understand

their demographic characteristics, to see why people contribute to the project, and finally, to see whether or not volunteers contribute in the same way. She found that the majority of participants were male (98 percent) with a large number either professionals or students in Information Technology (IT) or related fields. She concluded that the key reason to contribute to Folding@home was to make a contribution to science or a worthy cause. Several participants were very interested in providing large-scale computing power, and sharing images of their advanced computer setups with other participants appeared to be a strong self-expression motivation to continue contributing to the project. The competition aspect and the desire to learn more about computer hardware were also strong motivators, especially among active participants. Similar to SETI@home, being part of a team was another important motivational factor.

Finally, for CS game, we reviewed the two well-known projects, Foldit[8] and Eyewire[9]. Foldit is one of the first online CS games with an established online community of participants; its objective is to deduce protein structures (Cooper et al., 2010). Based on an online survey and semistructured interviews (Curtis, 2015a), 14 motivations to contribute to Foldit were identified, in which making a contribution to science, background interest in science, and intellectual challenge were the top three. Moreover, it was illustrated that the opportunity to collaborate with others and the communication features in Foldit, such as interaction between participants, were more motivating than the traditional game elements, such as pointing system, badges, and a leader board (Curtis, 2015a). Of the large number of active participants, the desire to play computer games was an initial motivational factor for a minority. Nevertheless, the gamified features in Foldit are helpful in sustaining participation.

The second use case, Eyewire (Tinati et al., 2017a), is a web-based CS game asking people to identify connected areas in 3D transformed Functional Magnetic Resonance Imaging (fMRI) in order to generate a detailed map of the human brain (see Figure 5.1). In order to understand motivation among Eyewire participants, a thematic coding approach was developed to identify motivational elements from the free text responses to the question of "Why do you play Eyewire?" (Tinati et al., 2017a). A thematic coding was applied to the free text responses, and as

[8]https://fold.it/

[9]https://eyewire.org/

a result, 18 motivation factors were extracted, which were then clustered into four categories: contributing to science; learning and personal interests; community and communication; and gaming and entertainment. The key reasons for participation appeared to be related to the value of the project, such as helping a beneficial cause, learning, and advancing scientific knowledge (Tinati et al., 2017a). While gaming (the possibility to play a game) was not rated as a strong motivator, the effect of design features and game elements on participants' behaviour illustrated that these features, which facilitate communication among participants, resulted in increased contributions (similar to Foldit). Thus, the authors concluded that maintaining a balance between the game mechanics and the ultimate goal of learning about science and scientific tasks is the key to success in CS games.

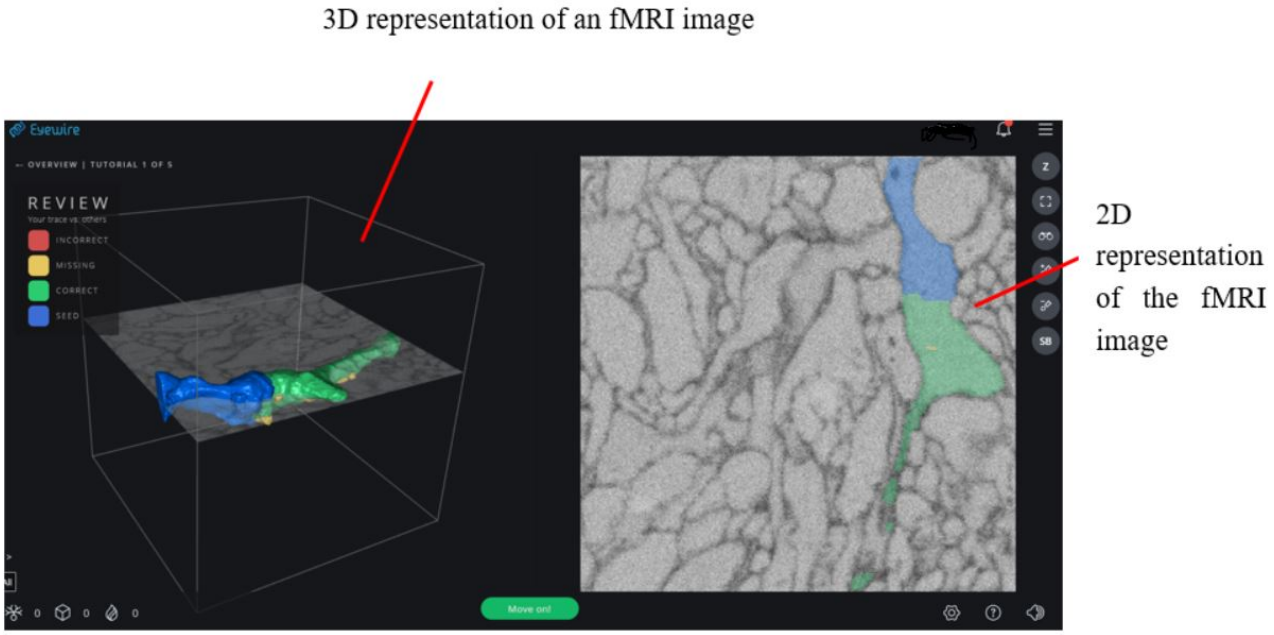


Figure 5.1: Eyewire interface. *Source: https://eyewire.org/explore*

5.4 Motivational framework

In view of the study of motivation in voluntary activities and the analysis of various use cases of CS typologies, we identified a framework that categorises the motivations by taking into account the types of CS projects. Since the number of use cases studying motivation on the basis of the level of participation are limited, and even those that have considered it are not complete

and require further analysis, this framework does not classify motivation on the basis of the participation level, but rather, classifies it based on the strength of motivation types as rated by volunteers in each use case. Therefore, in addition to the project typology, motivation types are classified as primary (highly rated) and secondary (low-rated). This framework is a generalised version of Curtis's classification of the motivation types of Foldit participants (Curtis, 2015a). Both frameworks classify motivation types on the basis of internal and external factors, and are therefore equivalent in their first two general levels. However, Curtis focused primarily on motivational factors in gamified CS projects (Curtis, 2015a), while our framework included all the other possible typologies (classic CS, gamified CS, VT, and VC). In addition, Curtis defined the framework based on the survey responses of Foldit participants, whereas we considered at least two use cases for each CS typology. In our research, we considered the importance given to each motivation type by participants and categorised the results accordingly; this was not done in Curtis' research. Therefore, we conclude that our framework is an improved version because of the addition of more use cases. Moreover, since we included more information on the strength of the motivation types, we suggest that our framework can help CS practitioners to obtain the information needed to recruit volunteers or to sustain participation in their projects. Table 5.3 presents the framework, which includes three main levels:

Level 1: The most general level, that classifies motivation types into two categories: intrinsic or extrinsic. As mentioned earlier, intrinsic motivations are when people perform an action because of the inner positive feeling they get; in comparison, extrinsic motivations are those when people perform an action because of what they receive in return, such as monetary rewards or reputation.

Level 2: At this level, intrinsic and extrinsic motivations are broken down into subcategories. Taking into account the motivation to contribute in VGI projects described by Budhathoki (Budhathoki, 2010), as well as motivation in the reviewed use cases, we identified the key motivational factors in CS projects. In addition to CS and VGI projects, to identify the categories in this level (in order to have a comprehensive overview of motivational factors), we considered the motivations to contribute to other forms of voluntary activities, such as FOSS projects (Wolf, 2005). These categories were defined in such a way that all the motivational

Table 5.3: A framework to classify volunteers' motivation in citizen science projects

Level 1	Level 2	Level 3							
		Classic Citizen Science		Citizen Cyber-Science					
				Citizen Science Game		Volunteered Thinking		Volunteered Computing	
		Primary	Secondary	Primary	Secondary	Primary	Secondary	Primary	Secondary
Intrinsic	Altruism	contribution to science nature conservation helping science and environment		contribution to science desire to support a worthy cause (such as finding a cure for a disease)		contribution to science helping the scientists	giving importance to goals of the project	desire to help scientific projects and contribution to something worthy offering personal computing power to scientific analysis	sharing acquired knowledge over years with others
	Fulfilment	interest in learning about nature and environment	understanding the scientific process	interest in science learning new knowledge and mastering a new topic		interest in science and desire to be connected to science learning about science (for example learning about galaxies)		enthusiast in fully utilizing the computing power	learn and acquire new skills
	Enjoyment	spending time in the nature outdoor recreation		satisfaction of solving complicated problems intellectual challenge	playing computer games	making new discoveries	having fun by doing the activities such as classifying or identifying objects in images	satisfying feeling of being useful	
Extrinsic	Community	working with like-minded people	interaction with others meeting new people group data collection	interaction with other participants being part of a diverse community with shared goals			being identified as a member of a group social interaction with others	being affiliated to a team	interaction with others and build relationship
	Ego enhancement		gain recognition among other participants	compete with other players gain recognition as top player in the leaderboard	improve self-status among other players	to be offered co-authorship in scientific articles to be acknowledged in scientific articles being named after a discovery	being among firsts to make scientific discoveries	seeing the results of the project published in a journal paper being acknowledged in an article as an active participant	competition self-expression towards having a higher computing power
	Future return		career-based benefits	gain knowledge from the game for personal reasons such as school grade or career	receiving badges and points monetary rewards		receiving extra awards such as t-shirt, certificates, or virtual badges	obtain credits for a course	

factors found in the reviewed use cases were allocated to a category, and no motivation type remained unclassified. Further to that, certain motivation types could be allocated to two categories; for example, in a CS game, earning a badge could be classified as reputation and also expecting a tangible reward. In addition, certain motivations were integrated into just one category, for example, the future return category included the expectation of monetary rewards, or the expectation of obtaining valuable information (e.g., to find/promote a career), or even the expectation of achieving school grades as a result of participation. Therefore, given the review of previous motivational studies and the CS use cases with different objectives and designs, we believe that new additional motivations from other use cases will fit into the defined categories at this level. However, this framework can be further expanded, given that new motivation types are identified that could not be assigned to any of the specified categories. Accordingly, we ended up with six categories of motivation types at level 2, three intrinsic and three extrinsic. The categories at level 2 are as follows:

The intrinsic motivation types are subcategorised as "altruism", "fulfilment", and "enjoyment":

- *Altruism*: Volunteers are motivated to contribute to CS projects to help scientists advance their research or contribute to a worthy cause.

- *Fulfilment*: Volunteers participate in CS projects (in order) to fulfil an interest in science or the desire to gain new knowledge.

- *Enjoyment*: Volunteers contribute to CS projects because they want to enjoy doing an activity other than their profession, such as having fun while playing a computer game or spending time in nature (e.g., bird watchers).

The extrinsic motivation types are subcategorised as "community", "ego enhancement", and "expected future returns":

- *Community*: Volunteers contribute to CS projects because of the social aspects of volunteering, such as meeting new people or being part of a team with a shared goal.

- *Ego enhancement*: Volunteers contribute because they want to improve their self-esteem or to have a good reputation among others. Examples include being mentioned in the website of a project as the most active participant, or being acknowledged in a paper.

- *Expected future returns*: Volunteers participate because they expect to achieve something in exchange for their contribution, such as monetary rewards, certificates, school grades, or to win points and badges in a gamified project.

Level 3: In this level, the motivation types were broken down according to details which were specific to each project. The categorisation of the first and second levels was more general and independent of project type; however, in this level, the motivation types were classified, taking into account the CS typology (check Table 5.3). In addition, a general pattern was identified in such a way that motivation types which had been given top priority in one use case were given high importance in other use cases of the same typology. The same pattern was observed for low-ranked motivation types. Taking into account this pattern and the importance given to motivation in each typology, motivation types were further categorised into two other sub-categories: primary and secondary motivation. Primary motivation types, as defined in this framework, were of considerable importance to the majority of volunteers in the studied project. For example, in all classic CS use cases, high scores were given to the motivation *helping the environment* by a majority of respondents, or in VC projects, *recognition among others (ego enhancement)* was a strong motivator to contribute among a broad number of respondents, even if passively reported (e.g., the effect of acknowledging participants on increasing their level of contribution). In addition, in most cases, the primary motivating factors were essential to sustain the participation of enthusiasts and serious contributors. On the other hand, secondary motivators were those given lower ranks by volunteers, and were selected as the central motivating factors by a smaller number of participants, relative to primary motivations. For instance, in gamified CS projects, the desire to play a computer game was the main motivation for a relatively small group of participants, or in VT projects, receiving monetary rewards was mentioned as a central motivation by few respondents in all the studied VT use cases. Secondary motivators are mostly important to encourage casual volunteers to contribute, or

to encourage potential volunteers to initiate participation. This categorisation does not imply dismissing people driven by secondary motivators, but it is a means of recognising the different levels of motivation when designing a project to achieve a balance among keeping committed volunteers engaged in the project, increasing the level of contribution of casual volunteers, and recruiting new participants.

The classification in Table 5.3 indicates that the motivation types among participants in one type of CS project are similar, and different than those of another category. Although motivating factors varied based on the project type, certain motivational factors were common in all types of CS projects. For instance, contribution to a scientific project and interest in science were reported as strong motivations in almost all use cases. Looking more into specific types, we found that in classic CS projects, intrinsic motivations play a more important role in engaging participants than extrinsic ones. Nonetheless, both intrinsic and extrinsic motivations are important in engaging people in CCS projects. For instance, ego enhancement, such as being offered a paper co-authorship, was reported as a strong motivational factor in VT projects. Moreover, contrary to what seems to be the case in CS games, traditional game elements such as pointing systems, leader boards, and badges are not the primary reasons for recruiting volunteers; in other words, volunteers do not start participating in CS games because of their enthusiasm for playing games, but because they are interested in contributing to science (Iacovides et al., 2013). Nevertheless, game elements are helpful in maintaining participation by, for example, creating a sense of competitiveness amongst the players or feeling like they are part of a community when playing as a team (Iacovides et al., 2013). Some of the participants in VT and VC projects, however, were strongly motivated by future returns and not necessarily by monetary rewards, but by achieving better school grades or earning certificates, for instance. Although participants from the same CS type tended to have similar motivations, the strength of motivation types among participants of one use case may be different from another use case. For instance, in VT projects, the motivation of valuing the project goals was mentioned in all VT use cases, but it was rated as a stronger motivator in Stardust@home than in Planet Hunters.

A comparison of the motivational framework in Table 5.3 with that of the OpenStreetMap in

Table 5.1 reveals that both in reviewed CS use cases and OpenStreetMap, learning and altruism are the two primary motivators among participants. However, altruism in OpenStreetMap primarily refers to humanitarian mapping purposes (e.g., mapping affected areas after natural hazards), while in CS, it is mostly focused on preserving nature or contributing to science, or helping scientists to accomplish the objectives of their projects. In CS projects, social interactions provide greater motivation for participating in online CS projects compared to classic CS, while in OpenStreetMap, social interaction is an important motivator for joining Mapathons and contributing to OpenStreetMap. As mentioned earlier, depending on the broad variety of ways of contributing to OpenStreetMap, the motivation types differ greatly, and an in-depth comparison of motivation types in OpenStreetMap and CS is a subject for further study.

5.5 Validation of the proposed framework using a case study

This section presents a discussion of the framework using a classic CS project in addition to what was previously reviewed. Therefore, we first present our biodiversity classic CS project, called BioPocket[10] (source code of the landing page is available on GitHub[11]), and then present the results of an online survey, which was conducted to understand the motivation types among potential BioPocket participants and their socio-demographic variables. The survey was designed in Google Forms[12], and the R[13] and Python programming languages were used to analyse the responses.

BioPocket is a CS biodiversity project implemented as a hybrid mobile application using free and open source framework called Apache Cordova[14], and the code is available on GitHub[15]. The objective of BioPocket is encouraging citizens to learn about biodiversity and take actions

[10]https://biopocket.ch/
[11]https://github.com/MediaComem/biopocket-landing-page
[12]Link of BioPocket survey: https://biopocket.ch/Questionnaire.html
[13]www.r-project.org
[14]https://cordova.apache.org/
[15]https://github.com/MediaComem/biopocket-mobile

in favour of it (Lotfian et al., 2018). A variety of activities that can be performed to promote biodiversity are specified in this project; these actions are categorised according to a number of criteria, such as theme, difficulty, importance, etc. Actions may range from simple ones, such as taking pictures of species, to more complicated ones, such as building a birdhouse or constructing a pond in a garden. Participants can learn how to undertake an activity by following the instructions given in the application. Details are given, such as what supplies are needed for each particular action, or how long each action takes. In addition, participants may monitor what action is being taken in their neighbourhood, using the interactive map in the application (Figure 5.2).

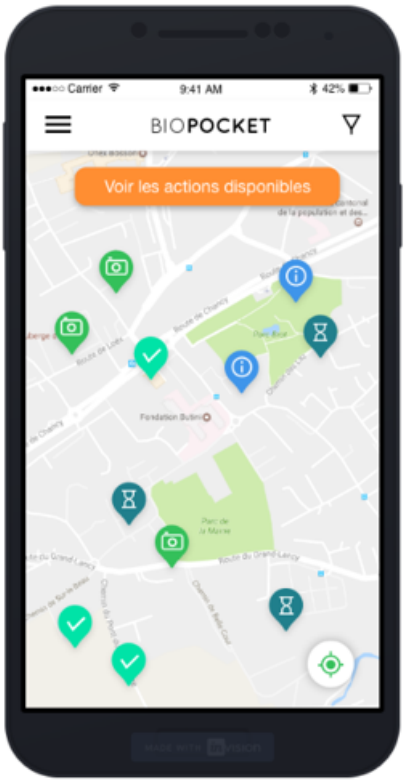


Figure 5.2: The interactive map in BioPocket application. Participants can check the type and location of actions, as well as the biodiversity points of interests around them

We conducted an online survey to understand the characteristics of potential participants and their motivations to participate in BioPocket. In this survey, we gathered information in four categories: socio-demographic information, backgrounds of participants on environmental activities, reasons to contribute to BioPocket, and their views on the usage of mobile apps (in terms of security aspects and experience with the use of mobile applications). In the section of

socio-demographic information, we obtained data on the age, level of education and occupation of the respondents. In addition, in order to conduct a spatial analysis, we collected the postal codes of the respondents and the type of housing in which they live (e.g., apartment, villa, etc.). In analysing the responses, it was our objective to understand:

- The correlation between motivation types
- The correlation between motivation types and other variables (age, education, residence type, etc.)

The survey was distributed via email and a Facebook campaign, resulting in 94 responses. The majority of respondents were between 15–24 and 25–34 years of age, i.e., 41% and 38% respectively, and the number of male respondents was almost 14% higher than that of female respondents. The majority of respondents were from the west part (French-speaking part) of Switzerland (Figure 5.3), with the cantons of Vaud and Geneva comprising 60.9% and 17%, respectively (as set out in the Facebook campaign for the distribution of ads to citizens from French-speaking cantons).

In addition to demographic information, a ranking was obtained on the basis of the average score given to each motivation type. The scores ranged from 1 to 8; Figure 5.4 presents the ranking of the motivating factors. Similar to the motivation types of classic CS in our motivational framework (Table 5.3), in this survey, the participants also gave a higher score to intrinsic, nature-related motivations, as the top three motivating factors were *helping nature*, *spending time in nature*, and *learning about biodiversity*. On the other hand, extrinsic motivation types, such as *social interactions, gaining recognition among others (ego enhancement)* and *expecting future returns*, such as *awards or certificates*, were given less priority and were ranked lowest. These findings are in accordance with the results from the use cases related to the classic CS typology in the framework (See section 5.4).

Correlations between motivation types were obtained using the Pearson correlation coefficient (Kirch, 2008) shown in Figure 5.5. There was a **statistically significant positive correlation** (p-values < 0.001) between intrinsic motivations of **helping nature, learning about**

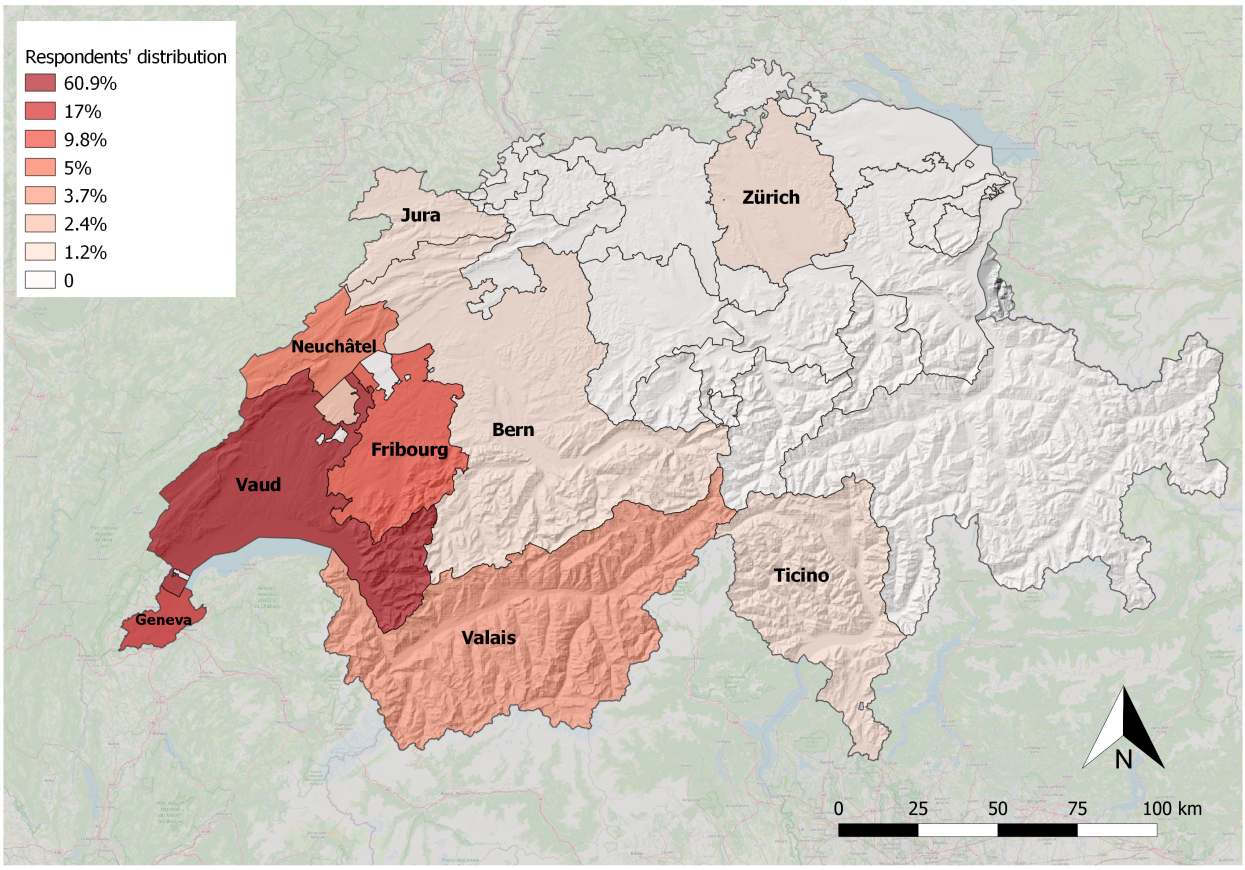


Figure 5.3: Distribution of BioPocket survey respondents within Switzerland. The regions with no responses are not labelled on the map

biodiversity and **spending time in nature**, as well as between the three extrinsic motivations, i.e., **social interactions, gaining recognition among others**, and **receiving awards or certificates**. On the other hand, these intrinsic and extrinsic motivation types were negatively correlated, meaning that respondents who gave higher scores to intrinsic nature-related motivations gave lower scores to extrinsic motivations. This corresponds to what was observed in the classic CS use cases in the framework (See section 5.4).

Correlations between the motivation types and other socio-demographic variables (e.g., age and education) were obtained using the Polyserial correlation coefficient (Olsson et al., 1982); the results illustrated weak correlations. On the other hand, to evaluate the correlation between numerical and nominal variables, we used a statistical measure called the Freeman Theta coefficient (Buck & Finner, 1985), which ranges from 0 to 1, where 0 implies no association between variables and 1 indicates a perfect association. The results of this test on motivation and res-

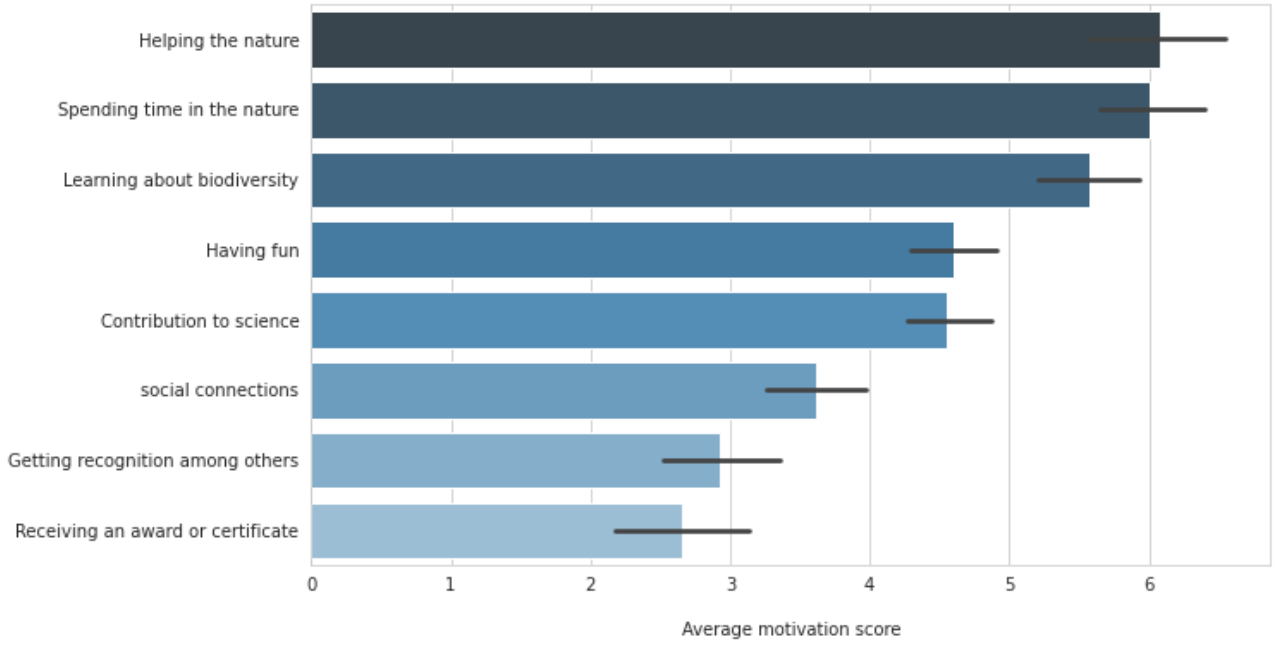


Figure 5.4: Motivation ranking to participate to BioPocket project. Motivational factors were given scores from 1 to 8, and the ranking was based on the average score give to each.

idence type illustrated that there was a moderate association between residence type and the two motivations of spending time in the nature and gaining recognition among others, with theta coefficients of 0.32 and 0.34, respectively. To further illustrate the results, we generated boxplots for the residence type versus the two aforementioned motivating factors (in Figure 5.6, residence types are aggregated into two types: apartment and villa). From the box plots, it can be seen that people living in apartments (mainly located in densely urbanised areas) placed more value (about 1 score on average) in spending time in nature relative to people living in villas (mainly located in open urban areas surrounded by more green areas compared to densely built areas). In contrast, gaining recognition among others was given higher scores by people living in villas than those living in apartments. Although these results require further investigation, a tentative conclusion would be that people living in densely urbanised zones are motivated to contribute to biodiversity CS projects because they want to escape city life and to spend more time in nature, while people living in less urbanized areas appear to be motivated not only by intrinsic factors related to biodiversity, but also by social interactions and being recognised among others. Therefore, the area where people live can have an important influence on volunteer participation. CS practitioners should consider this factor while recruiting and sustaining participation (e.g., designing an interactive app to teach volunteers

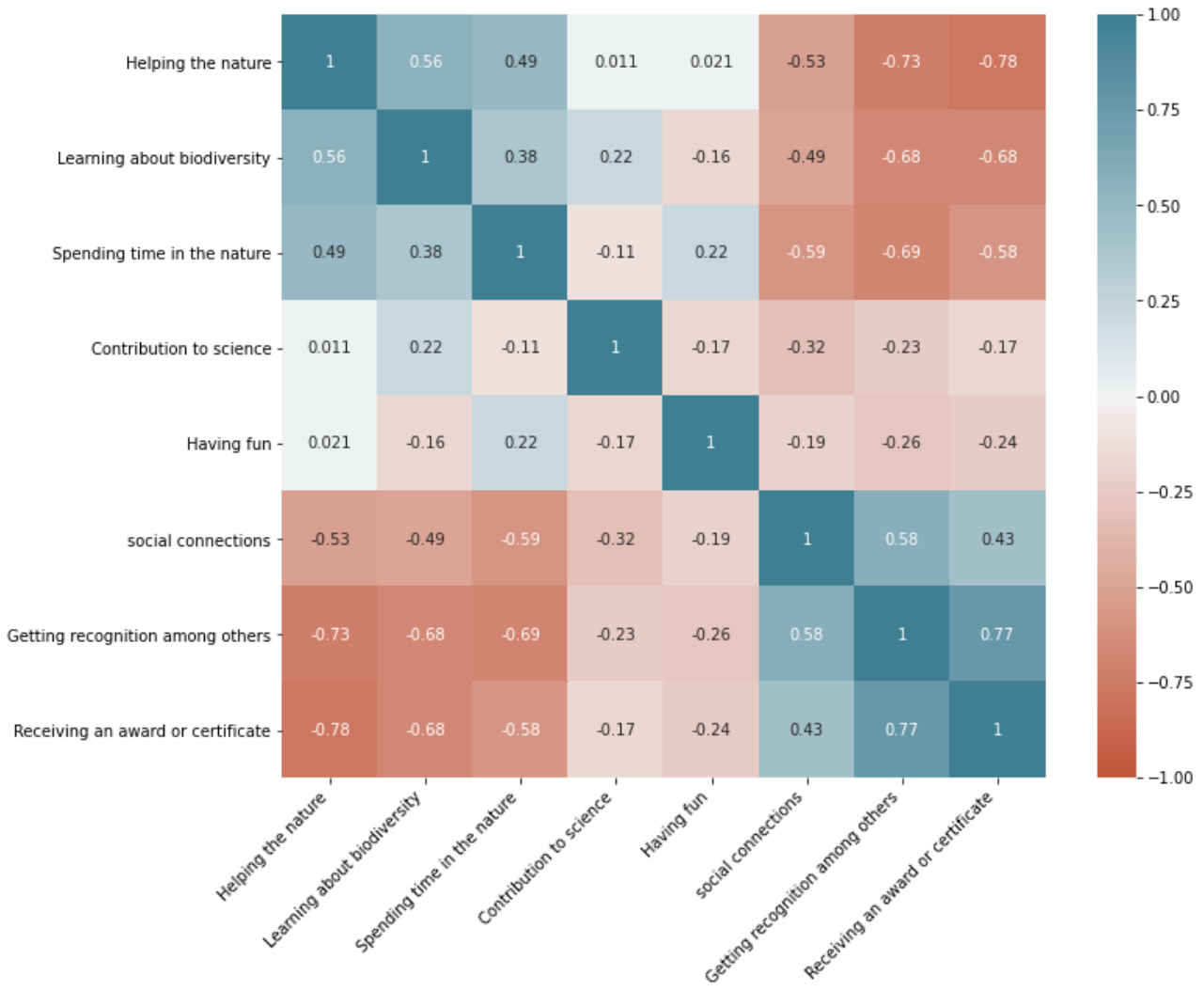


Figure 5.5: Correlation between motivation types based on the scores given by respondents. All the correlations are statistically significant with p-values < 0.001.

about biodiversity, acknowledging volunteers, and organising social events to perform group data collection).

5.6 Discussion and evaluation of hypotheses 1 to 3

Following the presentation of the motivational framework and the participants' motivations in the BioPocket use case, this section of the chapter discusses three of the hypotheses (H1 to H3) presented in Chapter 4.

H1: The design, objective, and type of a CS project are correlated with the participant's moti-

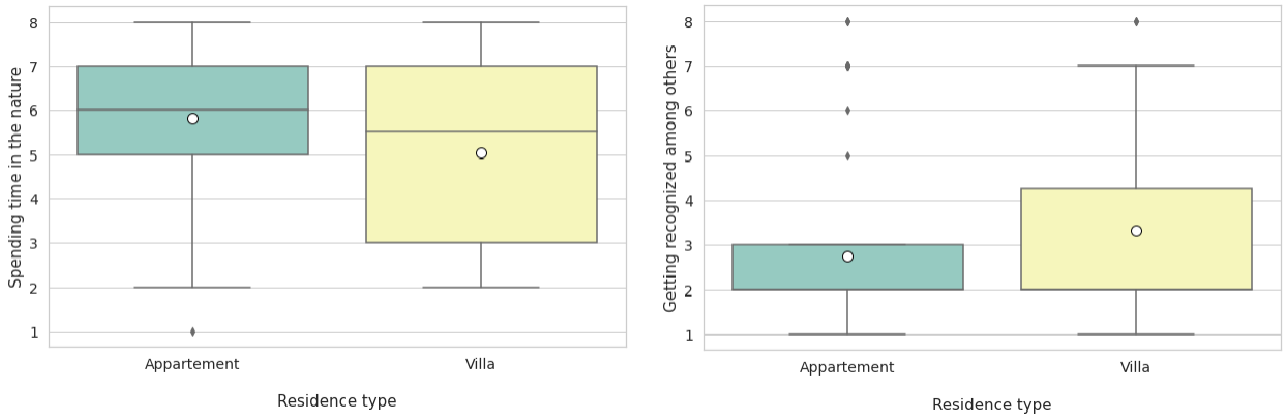


Figure 5.6: Boxplots of respondent's residence types versus motivation of spending time in the nature (on the left), and gaining recognition among others (on the right). The residence types are aggregated to two types, i.e., Apartment and Villa.

vations to contribute to that project.

Based on the conducted literature review in this chapter on participants' motivations to contribute to various typologies of CS projects, it was illustrated that participants from projects with the same typology have similar motivations. Establishing a conceptual framework to categorize motivations based on CS typology is one way to demonstrate that the motivations of the participants are associated to the project design and objective. Verifying this hypothesis with statistical measures necessitates a large sample of motivational studies in each CS typology, but the majority of studies are focused on the classic CS type, and there are not many case studies, particularly for online CS projects. Another way to verify this hypothesis was to compare the motivations of participants in the BioPocket use case to the motivations in classic CS typology in the framework. We indicated that the strong motivations expressed in the BioPocket survey were similar to those observed in other classic CS use cases. For example, in BiPocket, motivations such as *spending time in nature* and *learning about the environment* were among the primary motivations, while motivations such as *ego enhancement* and *expectation of future return* were among the low ranked motivations. Furthermore, the BioPocket survey reported a statistically significant negative correlation between intrinsic and extrinsic motivations, which corresponds to the motivations in the established framework for classic CS typology.

H2: Volunteers' motivation to contribute to a CS project is related to where they live.

This hypothesis is most applicable to classic CS in which participants engage in outdoor activities. To the best of our knowledge, the existing CS literature does not take into account the relationship between participants' motivations and the zone in which they live. However, as previously stated, in the BioPocket survey, we obtained the respondents' zip code as well as the type of residence they live in. As previously mentioned in the box plots in figure 5.6, a difference in the average score given to the two motivations *spending time in nature* and *gaining recognition among others* was observed among people living in different residence types. We used a t-test ("Student's t-test", 2021) to determine whether these differences were statistically significant. The result indicated that the average score for the motivation *spending time in nature* is statistically higher among respondents living in apartments than among those living in villas, with a p-value of 0.04. However, for the motivation *gaining recognition among others*, while the average score is lower for people living in apartments compared to those living in villas (negative t-test value), the difference is not statistically significant (p-value of 0.22). Furthermore, figure 5.7 depicts the violin plots for the motivation *spending time in nature*, and it shows that respondents living in apartments gave this motivation a higher score. Although further research with more use cases is required to validate this hypothesis, preliminary evidence from the BioPocket use case suggests that where people live is an indicator of the factors that motivate people to contribute to classic or, in particular, biodiversity CS projects.

H3: The socio-demographic information of participants is correlated with their motivations and perspectives on contributing to a CS project.

As previously stated, Polyserial correlation was used to determine the correlation between a continuous and a categorical variable with at least two categories. The correlation scale runs from -1 (perfectly negative correlation) to 1 (perfectly positive correlation), with 0 indicating no correlation. In the BioPocket use case, the correlations between motivations and age groups, gender, and education were either very weak or non-existent. The only finding was a weak but statistically significant correlation (cor = 0.20, P-value < 0.01) between *contributing to a scientific project* and education level, implying that people with higher education were more motivated to contribute to science. However, more use cases should be investigated because studies have shown a correlation between participants' demographics and their motivations to

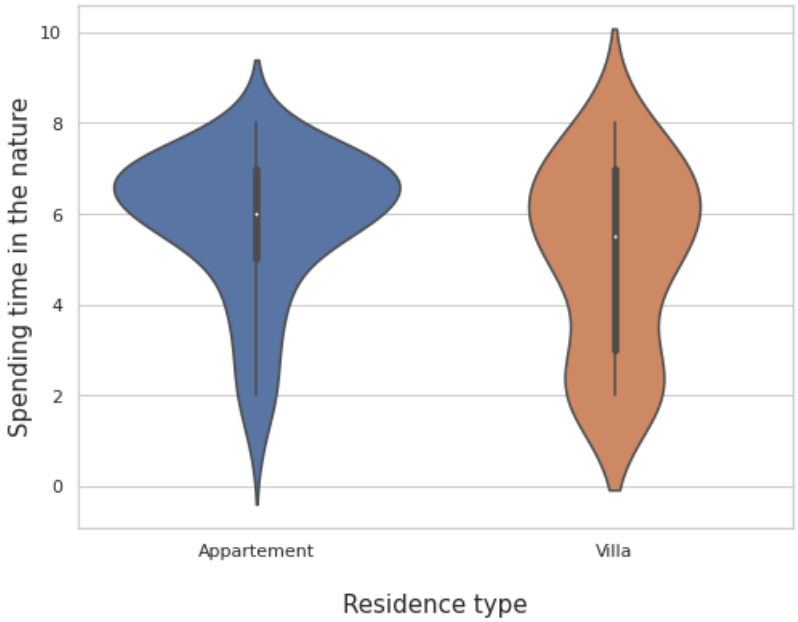


Figure 5.7: Violin plots comparing the scores given to the motivation "spending time in the nature" among respondents living in apartments and villas

contribute to CS projects (West et al., 2021).

Furthermore, we asked respondents three questions to see if there is an association between age and education level and views on contributing to a CS through a mobile or web application. First, respondents were asked to rate their familiarity with technology, such as using smartphones or new applications. Second, would participants grant access to their location when contributing to a CS application? Third, whether they prefer to create a new account in an app or login with their existing accounts (e.g., Facebook or Google account). Although a weak correlation was found between age groups and level of familiarity with technology, there was a positive correlation (cor = 0.32, p-value < 0.01) between level of education and level of familiarity with technology. Similarly, the correlation between age and willingness to authorize location access was weak, but there was a negative correlation (cor = -0.3) between level of education and willingness to authorize location access when contributing to a CS application, indicating that people with a higher education level gave a lower score to willingness to authorize location access when contributing to a CS application (See figure 5.8).

Finally, the correlations between age, level of education, and login with existing accounts were both negative, with values of -0.28 and -0.38, indicating that participants of higher age groups and education levels prefer to create new accounts rather than using their existing accounts

(See figure 5.8). All of the tests had p-values less than 0.01, indicating that the correlations, while not very strong, were statistically significant. As a result, we can supports the hypothesis that socio-demographic factors such as age or education level are related to sentiments toward using a CS application, but we did not find strong evidence to support the impact of socio-demographic factors on motivations, which requires further investigation.

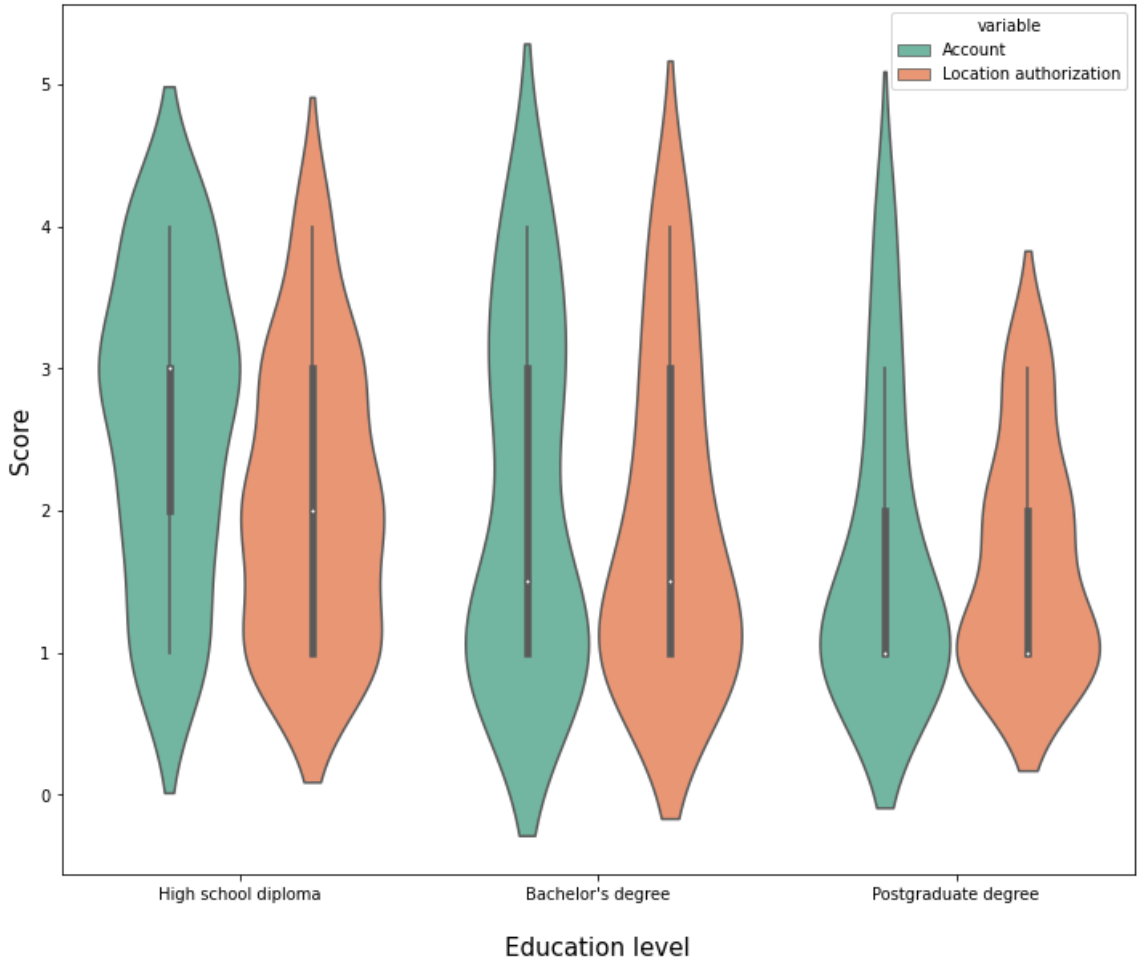


Figure 5.8: Violin plots depicting the distribution of scores (from 1 to 5, y-axis) assigned to location authorization and creating a new account or logging in with an existing account, versus education levels on the x-axis.

5.7 Summary

Understanding how to recruit and retain volunteers is a long-standing concern in CS and VGI projects, and subsequently, several studies have been conducted to investigate volunteer motivations. In this chapter, we reviewed several use cases and developed a framework to classify

the motivations of CS participants while taking project typology into account. The framework demonstrated that the design of a project is related to the motivations of its participants, as projects of the same typology reported similar motivations for their participants. However, in almost all of the studied use cases, the motivations of contributing to science and assisting scientists were reported as strong motivations to initiate contribution. The most important aspect in all use cases is to maintain a balance between attaining the objectives of the project and engaging volunteers. It is essential that project designers create the project in ways that meet/suit the expectations of a wide number of participants. For instance, in CS game projects, if the only emphasis is on game elements and the opportunity to learn about science is absent, a substantial number of participants may stop participating soon after their initial participation. As another example, in VT and VC projects, it is important to share the findings and new publications of the projects with the volunteers in order to let them know that their participation is appreciated and recognised.

Our proposed motivational framework in this thesis can be useful for researchers working on VGI and CS projects to understand the potential factors which can motivate individuals to join and contribute to their specific project. Finally, some suggestions for CS practitioners and project coordinators are offered below:

- *Interest in learning something new*: include creative learning opportunities in the project using new technology such as augmented reality (Ingensand et al., 2018), or add more information such as text and videos to a web page or project web/mobile application. In addition, project coordinators can provide the opportunity for volunteers to meet with experts by organising workshops and seminars (or activities such as mapping parties).

- *Contribution to a scientific project*: provide updates to participants on their contribution and let them know that their efforts have been taken into account, and communicate the results to them. Besides the general feedback on final results and publications, it is important to give detailed feedback to volunteers on their individual contributions, either through expert assistance or through automatic feedback generated using ML techniques, and to provide participants with informative information on the project topics (Lotfian

et al., 2019).

- *Interest in social interactions*: provide an opportunity for volunteers to interact with others through social forums, chatting functionality in the project application, and organising events such as field activities for group data collection.

- *Ego enhancement*: inform volunteers that they will be acknowledged in publications and on social media. Moreover, some volunteers would like to receive written cards, while some prefer to gain recognition by receiving certificates. An additional possibility would be to acknowledge top ranked contributors in scientific seminars and conferences (Fritz et al., 2017).

Given the motivational factors identified in this chapter, the goal of this thesis is to show how to integrate ML techniques into a CS project (discussed in Chapter 3) to increase motivation and sustain participation. One of the chapter's conclusions was to keep volunteers informed about the progress of a project, for example, by providing feedback, and in Chapter 3 we discussed the integration of ML in CS to provide automatic feedback to participants. As a result, in the following chapter, we implement a methodology for performing real-time data validation as well as providing real-time feedback to participants, and finally testing this approach with the public and evaluating the impact of machine generated feedback on their engagement.

Chapter 6

Data validation using machine learning

6.1 Introduction

We discussed the two main challenges of public engagement and data quality assurance in CS projects in Chapter 2. We also mentioned that *expert review* is one of the most commonly used data validation approach, according to surveys conducted on CS projects (Adriaens et al., 2021; Wiggins et al., 2011). Moreover, we discussed how, as the number of CS projects grows, large amounts of data are being collected, necessitating the use of automation to validate CS data. As discussed in Chapter 3, an integration of ML and CS can be used to automate data quality assurance. According to our literature review of the use cases on the integration of ML and CS, the majority of CS projects, such as camera trap projects in biodiversity CS, use AI to automatically identify objects in images. Few studies used other metadata (aside from images) to automatically validate data, and when they did, they relied heavily on historical databases to identify outliers (e.g., if the count of species in a neighbourhood exceeded a threshold). In addition, in Chapter 3, we discussed how the integration of ML and CS can be beneficial in encouraging public engagement (See the taxonomy we defined in section 3.3). One method for increasing motivation is to provide participants with informative and user-centered feedback (rather than just general repeated text regardless of the user contribution), which can be generated using machines. Furthermore, based on our motivational framework in Chapter

5, we categorized various motivations based on CS typology, and it is useful to see which of the motivations in the framework can be fulfilled using ML algorithms when integrated in CS projects.

Thus, in this chapter, we investigate how we can put our previously discussed elements into practice in our biodiversity CS project. Thereby, the goal of this chapter is to validate biodiversity observations (with a particular emphasis on birds species) by generating species distribution models using existing data from other CS projects (e.g., eBird). In other words, the goal of this chapter is to validate biodiversity observations by taking into account species habitat characteristics. Furthermore, in order to investigate the impact of automated feedback to participants on increasing their motivations, we implemented our application to provide informative feedback to participants based on the user's location and environmental variables in the neighborhood surrounding that location. We conducted an experiment with our participants to assess the impact of automatic data validation and machine-generated feedback on data quality and public engagement. This project was our main focus on the combination of CS and ML, but we also worked on other use cases where CS and ML are combined outside the area of biodiversity, which two of these use cases are briefly explained in Appendix B.

The chapter is structured as follows: We will first go over the specifics of species distribution modeling, explaining what it is, how it can be generated, and what types of data are required. This is followed by the introduction of BioSenCS, our biodiversity CS case study. We discuss the project's objectives, how it was implemented, and how we used automatic data validation with a focus on species distribution modeling, as well as the various data sets and algorithms we used, and the evaluation and comparison of the algorithms. Finally, we conducted a user experiment to evaluate our approach and to explore how automatic data validation and real-time feedback affected data quality and public engagement.

Despite the fact that this chapter is longer than the others, we decided to leave it as is rather than split it into two separate chapters. The main reason was to keep the thesis's current structure, which follows the two challenges of public engagement and data quality, but separating the section that presents species distribution modeling from the section that presents our case

study would have disrupted the thesis's flow. Thus, to avoid disrupting the flow of the thesis, we did not separate section 6.2 from our case study.

6.2 Species Distribution Modeling

Several factors, including but not limited to habitat destruction, disease, competition, predation, and displacement of indigenous species due to the presence of invasive species, can pose a threat to species and biodiversity (Van Dyke & Lamb, 2020). Aside from the reasons stated above, one of the primary concerns of ecologists is the impact of climate change on biodiversity, and as a result, researchers seek to investigate how to address and mitigate these impacts on the environment and biodiversity (Van Dyke & Lamb, 2020). A multidisciplinary science that focuses on the issue of biodiversity loss is called conservation biology (Soulé, 1985).

Conservation biology is a field of study concerned with biodiversity and species protection, with the goal of addressing the issue of biodiversity loss while taking species and ecological process knowledge into account (Pullin, 2002). Three key criteria must be considered when performing biodiversity conservation including *spatial*: where should conservation be done? *temporal*: when should conservation be done? , as well as *methodology and strategy*: how should conservation be carried out? (Redford et al., 2003). To answer these questions, financial resources are required; however, due to potential social and economic crises, such investments for biological conservation are difficult to obtain (Pi, 2016). Given these challenges, there is a vital need for other strategies to address conservation biology issues and assist in decision making. Among the available strategies, species distribution modelings are regarded as powerful potential tools for addressing conservation issues.

Species Distribution Modeling (SDM) is a class of numerical models that explain how the presence or absence of a species at a given location is related to environmental (e.g. temperature, precipitation, etc.) and landscape characteristics (e.g. landcover, elevation, slope, etc.) (Elith & Leathwick, 2009). They are used to gain ecological and evolutionary insights as well as to predict distributions across landscapes, which requires spatial and/or temporal extrapolation.

SDM can be used to understand how a species' distribution is correlated with its location, as well as to predict the locations of species occurrence where no data is available. "There are various other terms for SDM including bioclimatic models, ecological niche models, habitat models, resource selection functions, range maps, correlative models or spatial models." (Elith & Leathwick, 2009). Joseph Grinnell (Grinnell, 1917, 1924) defines the term "niche" in ecological niche modeling as the conditions under which a species can survive and reproduce; thus, understanding environmental characteristics as well as the availability of other species play important roles in modeling species habitat (Pi, 2016). Later, George Evelyn Hutchinson defined "niche theory" (Leibold, 1995) in which he introduced two terms: fundamental niche, defined as the environmental conditions in which a species could survive and reproduce in the absence of other species, and realized niche, defined as the environmental conditions in which a species actually lives while taking other species into account. A more recent theory is then defined by Hubbel (Hubbell, 2005), in which stochastic population events such as death, birth, or immigration determine community composition at the local scale rather than species niche differences. These theories were the first steps toward ecological research. Having this brief introduction about SDM, the next step is to see how SDM generation initiated and changed over time.

SDM was initially based on linear regressions, but as modeling advances and new algorithms have been introduced, it has advanced to use new modeling techniques and algorithms (Wintle et al., 2005). The same is true for advancements in the data used to generate SDM. Initially, ecologists had access to limited geospatial data such as latitude, longitude, or elevation; however, advances in Geographic Information System (GIS) and the availability of new tools and software such as widely accessible satellite images, the possibility of obtaining 3D terrain models, and so on have made it easier to obtain a broader range of data to use for SDM (Elith & Leathwick, 2009). SDM generation is more challenging for organisms that move in space (mobile species) than for sessile organisms (fixed location) (Elith & Leathwick, 2009). The reason is that sessile species are attached to their surroundings (e.g., soil, rocks, etc.) and move mainly by external forces (e.g., water currents) ("Sessility (motility)", 2021) and thus the environment around them does not change frequently; therefore, it is easier to characterize

their environment and model their distribution. While it is more difficult to model the habitat of mobile species, the likelihood of their occurrence over a region of interest can be predicted. In the following subsections, we will go over the data needed and algorithms that are frequently used to generate SDM in greater detail.

6.2.1 SDM generation

SDM algorithms are typically fitted to a collection of species data set without regard for other species. Joint Species Distribution Models (JSDMs), on the other hand, take species dependence into account when modeling species distribution (Taylor-Rodríguez et al., 2017). In this thesis, the species dependence is not taken into consideration, and the models are generated independently for each species. To generate SDM two sets of data are required:

- Species occurrence data: the locations (often point-based) where the species has been observed, which are collected in a variety of ways, including natural museum records, field observations by biologists, and crowdsourcing and CS projects.
- Environmental variables: environmental variables include both climate data such as temperature and precipitation as well as landscape characteristics such as elevation, slope, soil type, land cover, and so on. The importance of various environmental variables in modeling species distribution varies according to species and organisms (Syphard & Franklin, 2009).

The two types of data mentioned above will be used in an algorithm to explore the relationship between environmental variables and the location of species occurrences, and as a result, to predict the suitable habitat characteristics required for the species to survive. The model then allows us to predict the distribution of species in unknown locations by estimating the likelihood of a species being observed in a specific location given a set of environmental variables ("Introduction to Species Distribution Models", n.d.). The steps to generate SDM are illustrated in Figure 6.1 which include:

1) Data preparation

2) Choose an algorithm

3) Feed and train the algorithm using the input data

4) Evaluate the performance of the algorithm

5) Predict species distribution over the whole study area

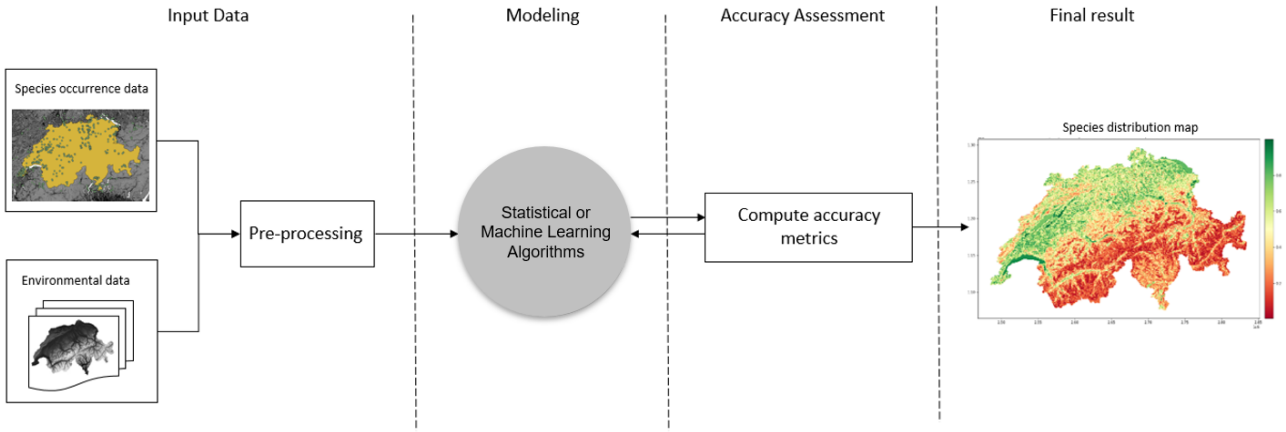


Figure 6.1: Steps to perform species distribution modeling

The species data set contains both presences (where the species is observed) and sometimes absences (where the species is not observed). Some algorithms only require presence data, whereas others require both presence and absence data. In the next section we explain presence-only, presence-absence, and presence-background (or presence pseudo-absence) data sets.

6.2.2 Presence-only versus Presence-absence

As previously stated, some algorithms require only *presence* data and are known as presence-only algorithms (such as Maxent (Phillips et al., 2006), which will be discussed further in this section), while others require both *presence* and *absence* data. True absences are difficult to obtain, particularly for mobile species that move throughout space, so obtaining presence data is much easier than absence data. As a result, one solution is to create artificial absences known as pseudo-absences or background data (these two terms are not necessarily always the

same, but they are frequently used interchangeably). There are several methods for generating pseudo-absence data, but the three most common approaches are as follows (Senay et al., 2013):

- Random pseudo-absence selection: In this method, absence data are derived at random from the environment, except in locations where presence points exist (this is mainly known as background data).

- Pseudo-absence selection based on geographical extent: The pseudo-absence data are generated within a certain distance of the presence points in this method. Some approaches choose the radius (distance to presence points) based on the model's performance, so several radius are chosen, and the radius where the model achieves the best performance is considered the extent. Other methods choose the radius at random or based on the knowledge of experts in the distribution of species in that area.

- Pseudo-absence selection based on environmental variables: The pseudo-absences in this approach are chosen from background data that is environmentally dissimilar to the presence data. In this method, suitable environmental areas are removed from the background, and then pseudo-absence points are chosen at random from the remaining zones.

Some other methods combine the approaches mentioned above, such as selecting pseudo-absence points by defining both a buffer around the presence points and taking into account environmental variables that are not similar to the presence locations (Senay et al., 2013). However apart from the mentioned methods, there are some other important factors to consider before generating pseudo-absences, such as *the ratio of presence-absence data, the optimum number of pseudo-absence points*, and *the impact of the pseudo-absence generation method and number of pseudo-absence points on model performance* (Barbet-Massin et al., 2012). Barbet-Massin et al. (2012) have compared various algorithms with different number of pseudo-absence data, and compared the model performances. Thus, one can run several models with varying numbers of background points and compare and select the model that performs best.

6.2.3 Common SDM algorithms

While there are various algorithms for generating SDM, researchers continue to look for the models that can have higher performances (Araujo & Guisan, 2006) (See section 6.5.3.2 for more details on algorithms performance). The three main goals of ecological modeling are to achieve reality, generality, and precision (Li & Wang, 2013). Typically, two of these three goals will be reached, with the third being sacrificed. Despite the availability of several models to estimate the correlation between the environment and the locations of species, different models can produce different estimates, so it is critical to determine which model is appropriate for which data types (Li & Wang, 2013). One of the known ways to compare SDM models is using a package in R called *biomod2* which offers the possibility of training and comparing up to 10 algorithms and then recommends the one with the best performance for the given input data (Thuiller et al., 2009).

SDM algorithms are classified into four types: profile models, statistical regression models, machine learning models, and geographical models. Each of these categories contains one or more algorithms for investigating the species-environment relationship. Statistical models and machine learning models have received the most attention in the literature of these four categories. Thus, we will look at some of the algorithms in these two categories.

6.2.3.1 Statistical Models:

Generalized Linear Models (GLMs): This algorithm is an extension of classical Linear Regression (LR), with the exception that some of the assumptions in LR are not respected in GLMs; for example, unlike LR, there is no assumption that the response variable (species occurrence) follows a normal distribution in GLMs (Kienast et al., 2012). The three main components of a GLMs are probability distribution of the response variable, the linear combination of predictor variables (Linear predictor), and the function that links the mean of the response variable to the linear prediction (link function). So although the relationship between the response variables and the predictor variables is not linear, the link function transforms the

response variables in such a way that the transformed response and the predictors are linearly related. In SDM, the predictors are the environmental variables and the response is the score of the habitat suitability for the species. In addition, in LR the parameters of the predictors are usually estimated using Least Squares (LS) method, but in GLMs the parameters are estimated using Maximum Likelihood Estimation (MLE), meaning that the parameters are estimated in such as way that they maximise the likelihood of the model predicting the actual observed data.

Generalized Additive Models (GAMs): GAMs are extension of GLMs, and it is used when the relationship between variables is very complex and cannot be simply modeled with standard linear or non-linear models (Guisan et al., 2002). GAMs have also the three main components as GLMs with the difference that the coefficients of predictor variables in linear predictor are replaced with a smoothing function. Moreover, similar to GLMs, MLE method is used for the estimation of the coefficients.

Multivariate Adaptive Regression Splines (MARS): MARS is designed for multivariate non-linear regressions, and uses an aggregation of linear regression models to build the final predictor (Friedman, 1991). To aggregate various pairs of the regression models, a function is used which is called *hinge function* and the merge point (or split point) is called the *Knot* of the function. The hinge function compares the environmental value at knot with the environmental values on the left (left function: $Max(0, EnvVariable - knot)$) and right (right function: $Max(0, knot - EnvVariable)$) sides of the knot. The left and right functions are called *basis* functions. MARS generates many basis functions, and then fits a least-square model to the result of these functions as the final predictor. This process is done iteratively until the best fit is found as the final predictor.

6.2.3.2 Machine Learning Models:

Maximum-entropy (MaxEnt): MaxEnt is a ML algorithms which is used in several domains such as NLP (Berger et al., 1996). As previously stated, there are algorithms that can estimate the species-environment relationship solely based on presence data; which are known

as presence-only models. MaxEnt is one of the most well-known presence-only algorithms in ecological modeling (Phillips et al., 2006). MaxEnt is composed of two main components ("Maxent", n.d.):

- Entropy: The model is calibrated to find the most uniform distribution over the whole region of study.
- Constraints: There are rules that constrain the predicted distribution, which these rules are based on the values of environmental variables at locations where the species have been observed.

MaxEnt compares the locations of species occurrences to the entire environment of the study area. To accomplish this, it samples a large number of points in the background environment, known as background points. It is important to note that here background points differ from pseudo-absence points, since the background points are used to define the environment of the region under study rather than the location of absences.

MaxEnt begins by computing the probability densities of the values of environmental variables. Thus, the probability densities are computed both for locations where species are observed (f1(z)), and also over the whole study area for the background points (f(z)) (Elith et al., 2011). Therefore, in order to estimate feature importance and relative habitat suitability, MaxEnt computes the ratio of f1(z)/f(z) which is known as MaxEnt's "raw" output. Figure 6.2) illustrates the probability densities for two environmental variables of temperature and precipitation for the presence and background points.

MaxEnt selects the distribution that maximizes the raw output, or in other words, the distribution that maximizes the similarity between the environmental characteristics of the background and those of the presence points. To avoid overfitting, MaxEnt employs two generalization approaches. First, to avoid fitting the model to the exact values of features, MaxEnt uses a method of relaxing the constraints by fitting the models to the confidence intervals of the constraints rather than their exact means, variances, and so on. Second, it reduces features by excluding those that do not significantly improve the model.

Even thought MaxEnt is one of the most used algorithm in SDM, like any other algorithm there are some limitations in using it. One limitation is that MaxEnt predicts the environmental suitability rather than probability of occurrence, which makes it difficult to compare MaxEnt with other algorithms. Another limitation is that MaxEnt makes predictions based on the assumption that the species can occur anywhere in the study region with a probability threshold of 0.5, and thus the final estimation is based on this assumption which can not always be appropriate especially for rare species.

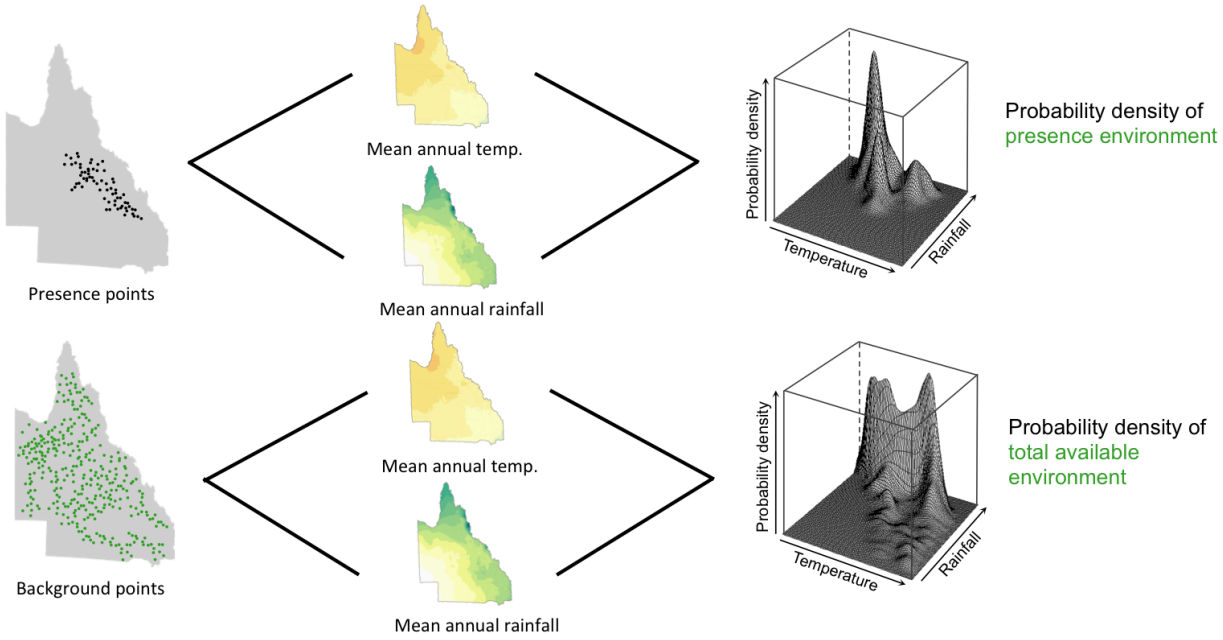


Figure 6.2: A statistical explanation of MaxEnt for ecologists. Adapted from (Elith et al., 2011), *Source: ("Maxent", n.d.)*

Artificial Neural Network (ANN): Human brain is a collection of more than 10 billion interconnected neurons, which are responsible to receive and process information (Abraham, 2005). ANN refers to the algorithms which are inspired by the interconnected networks of neurons in biological brain (Abraham, 2005). The processing elements of neural networks known as nodes or artificial neurons, receive input signals and using the connection weights the output is generated (Abraham, 2005). The network improves learning each time by adjusting the weights. The architecture of a multi layer neural network includes an input layer, one or more hidden layers, and an output layer (Figure 6.3).

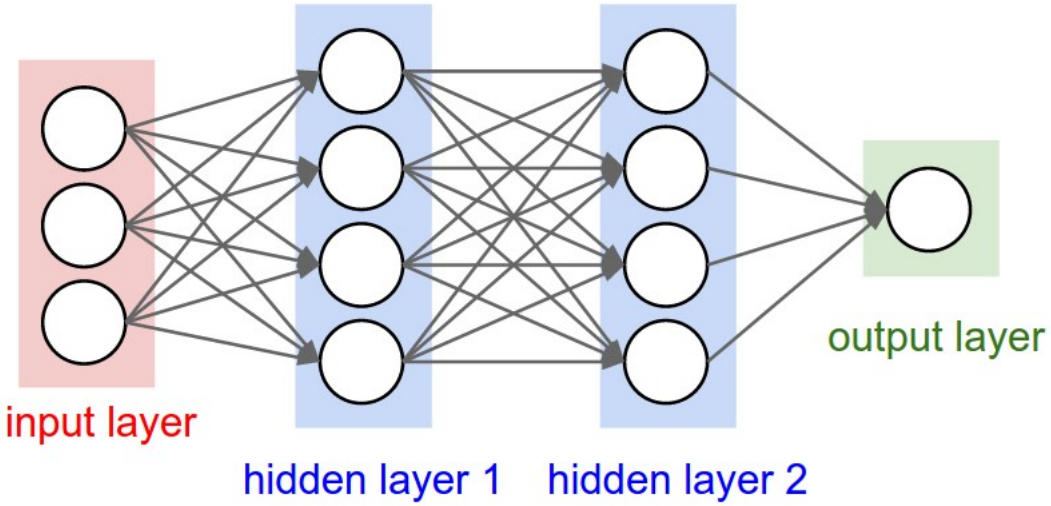


Figure 6.3: Architecture of an artificial neural network with two hidden layers. *Source: https://cs231n.github.io/neural-networks-1/*

One of the simplest ANN architectures is the Perceptron, invented in 1957 by Frank Rosenbalt (Rosenblatt, 1958). Perceptron is based on an artificial neuron called Linear Threshold Unit (LTU) (Géron, 2019) (Figure 6.4). The input connection is associated with a weight, and LTU computes a weighted sum of the input (Equation 6.1), and finally a step function is applied to the weighted sum to compute the output (Equation 6.2).

$$z = w_1 x_1 + w_2 x_2 + ... + w_n x_n = W^T X \tag{6.1}$$

$$h_w(X) = step(z) = step(W^T X) \tag{6.2}$$

The Perceptron learning algorithms includes four main steps:

1) Randomly initializing the weights

2) Using the initialized weights, computing the output y given the input x

3) Update the weights: $wj(t+1) = wj + \alpha(d-y)x$ *[d is the desired output]*

4) Repeat the steps 2 and 3 until we reach an error smaller than a threshold, or based on a given number of steps

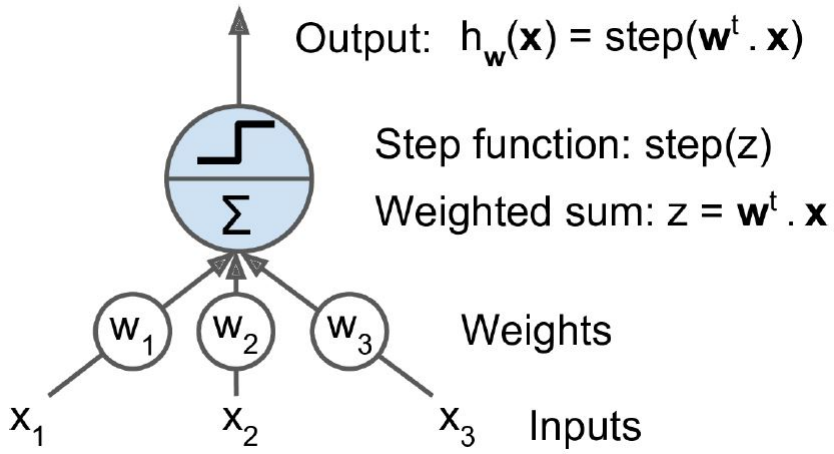


Figure 6.4: Linear threshold unit. *source: https://www.oreilly.com/library/view/neural-networks-and/9781492037354/ch01.html*

A Perceptron is simply composed of a single LTU layer and it outputs linear binary classification. The most used step function in Perceptron is the Heaviside step function (Géron, 2019) (Equation 6.3).

$$\begin{cases} 0 & \text{if } z \leq 0 \\ 1 & \text{if } z > 0 \end{cases} \tag{6.3}$$

The problem in using Healviside step function specially in Deep Neural Network (DNN) is that a small change in weights can result in a complete change in the network and as a result the output. In other words, a small change in weights can change the value of z (See equation 6.3) from negative to positive and vice versa, and this can change the output from 0 to 1. One solution to this is the use of other activation functions rather than Heaviside step function. Activation functions in NN help the network to learn complex patterns from the data. The activation functions are applied on the output from the previous neuron to create the input for the next neuron. Some of the most known activation functions are as follows:

1) Sigmoid: Sigmoid function is used for binary classification, and is defined in Equation 6.4 (Han & Moraga, 1995). It generates values between 0 to 1 (Figure 6.5). Due to being computationally expensive and the vanishing gradient problem (Hochreiter, 1998)(small values of gradient in backpropagation and thus updated weights are almost close to the old weights

and not possible to improve the learning), Sigmoid is no more being used especially in recent networks.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{6.4}$$

2) TanH: Tangent Hyperbolic (TanH) is very similar to Sigmoid function, or in other words it is a scaled Sigmoid function (Nielsen, 2015), generating values between -1 and 1 (Figure 6.5).

$$tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \tag{6.5}$$

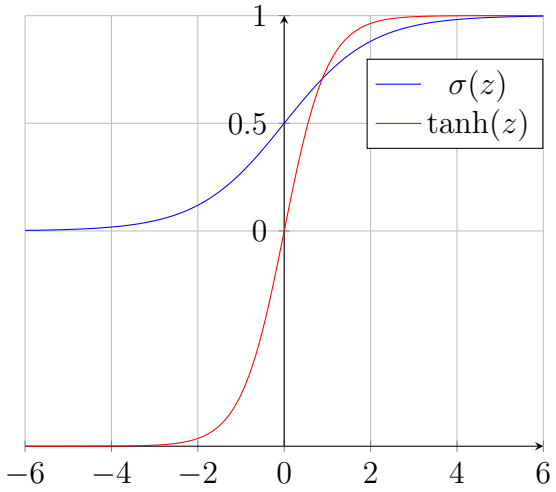


Figure 6.5: The graph of Sigmoid function (blue line), and TanH function (red line). x axis: weighted some of input neurons $z = W^t X$, y axis: the value of $\sigma(z)$ or $\tanh(z)$

3) Softmax: Similar to Sigmoid, Softmax also produces outputs values between 0 and 1 (Gao & Pavel, 2017), but contrary to Sigmoid, Softmax is used for multi-class classification. It is applied in the last layer of the network, and it generates the probability of predicting each class.

4) Rectified Linear Unit (ReLU): ReLU is the most commonly used activation function. It is computationally easy, and it addresses the issue of vanishing gradient. ReLU returns the same value if the value of z is positive and returns 0 otherwise (Agarap, 2018) (See Equation 6.6 and Figure 6.6).

$$\begin{cases} 0 & \text{if } z \leq 0 \\ z & otherwise \end{cases} \tag{6.6}$$

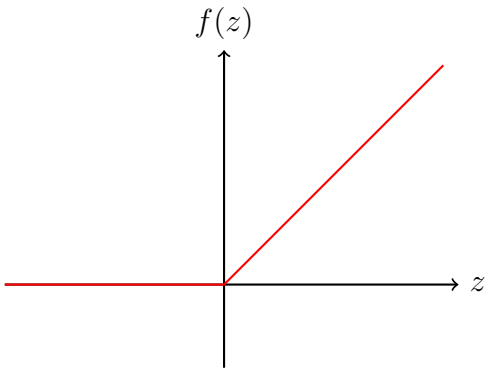


Figure 6.6: ReLU activation function, $ReLU(z) = max(0, z)$

There are other activation functions such as Leaky ReLU (Maas et al., 2013) and ReLU6 ("ReLU6 — PyTorch 1.10.0 documentation", n.d.) to address some limitations of ReLU. Moreover, there are the non-linear activation functions such as swish (Ramachandran et al., 2017) or h-swish (Howard et al., 2019), which we do not go into their details in this thesis.

As previously stated, the network learns by adjusting its weights in each loop, as demonstrated by the steps of how Percetron learns. To determine how well the NN performs, the cost function is computed in each iteration. The cost function calculates the difference between the model prediction and the expected output value (Nielsen, 2015). Depending on the choice of algorithm there are various ways to compute the cost. For instance one of the most known ways to compute cost for regression problems is the Mean Squared Error (MSE) (Nielsen, 2015) computing as $\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2$ where $f_\theta(x_i)$ is the prediction output for the training data x_i, and y_i indicates the true value. Thus, the idea is to find a set of θ parameters that minimize the cost. The terms loss and cost functions are frequently used interchangeably; however, the loss function computes the error over a single training example, whereas the cost function computes the error over the entire training data set. Cost function can be computed as shown in Equation 6.7, where m is the total number of training set, y is the vector of expected outputs, and $\hat{y}$ is the vector of predictions.

$$J = \frac{1}{m}\sum_{i=1}^{m}(y - \hat{y})^2 \tag{6.7}$$

An algorithm known as backpropagation is used in NN to compute a set of parameters that minimize the cost. Backpropagation (Rumelhart et al., 1986) computes the partial derivatives of the cost function and updates the weights accordingly. The output is then computed again in the next loop using the new parameters, and depending on the value of the cost (if it is within an acceptable error threshold), the parameters will be updated or not. This weight-updating procedure will be repeated until the model achieves the desired accuracy or the maximum number of runs (Nielsen, 2015). So the new weights will be updated using the following equation:

$$\theta_{new} = \theta_{old} - \alpha(\frac{\partial Cost}{\partial \theta}) \tag{6.8}$$

Where in equation 6.8, θ is the vector of network parameters (weights and biases) and α is the learning rate. Learning rate is a hyper-parameter (a parameter used to control the learning process) that is used in each backpropagation iteration to indicate the speed of the steps towards the local minima (Nielsen, 2015). If α is too small, the steps will be too short and thus slow gradient descent. On the contrary, if α is too large, the steps of the gradient descent can be very large and it can overshoot the local minima, and might fail to converge.

To train a DNN to generate SDM, both presence and absence data are required. The use of DNN, and particularly CNN, in modeling species distribution was not widespread until recently (Botella et al., 2018; Deneu et al., 2021; Vos et al., 2019).

Decision trees: The decision tree algorithm, proposed by (Breiman et al., 1984), is a non-parametric ML algorithm that can be used to solve classification or regression problems. Decision tree, as the name implies, is constructing a tree-like model with nodes and leaves. The node is where the tree splits into branches based on the conditional question in the node, and the leaf is where the tree cannot split any further, or in other words, it is the decision

of a branch. Classification trees are used to classify categorical data sets, such as land cover classification (Friedl & Brodley, 1997), whereas regression trees are used to predict continuous values, such as house price (Fan et al., 2006). It is essential to understand how to split a node and when to stop splitting when creating a decision tree. Splitting involves taking a subset of input training data and breaking it down to smaller and smaller subsets until further subsets cannot be formed (Song & Ying, 2015). Stopping, on the other hand, means establishing a set of rules to prevent the tree from becoming overly complex and, as a result, suffering from lack of generalization. Some of the stopping rules include first defining a minimum number of training data to be used in each node to avoid complexity and building a model for each small subset of the data, and second defining the depth of the tree (number of steps from the root node) (Song & Ying, 2015). Figure 6.7 illustrates a very basic example of a decision tree for land cover classification.

To generate SDM using the decision tree algorithm, both presence and absence data are required. A number of studies have used the decision tree algorithm to model species distribution (Debeljak et al., 2007; Džeroski & Drumm, 2003; Kampichler et al., 2000; Kobler & Adamic, 1999; Ogris & Jurc, 2007), but due to some of its limitations, such as being unstable (significant change in the tree as a result of small change in the data) or having low prediction accuracy, other powerful algorithms (built upon decision trees such as RF) are preferred ("Classification Tree", 2021).

Random Forest (RF): RF (Breiman, 2001), as the name suggests, is an ensemble model of several decision trees that, like decision trees, can be used for classification and regression problems. The bagging algorithm is used by RF to generate the final prediction. Bagging is the process of selecting random subsets of training data with replacement so that each subset contains zero, one, or more than one copy of the training data examples. After which, for each sample, a weak algorithm (e.g., decision tree) is used to model the sample, and the final prediction is the ensemble of all the weak models (Kuhn, Johnson, et al., 2013). RF fits many decision trees to subsets of training data. Then, each tree creates a classification, which is referred to as the tree votes for that class, and the class with the most votes is chosen as the final prediction from among all the trees in the forest (Horning et al., 2010). In classification

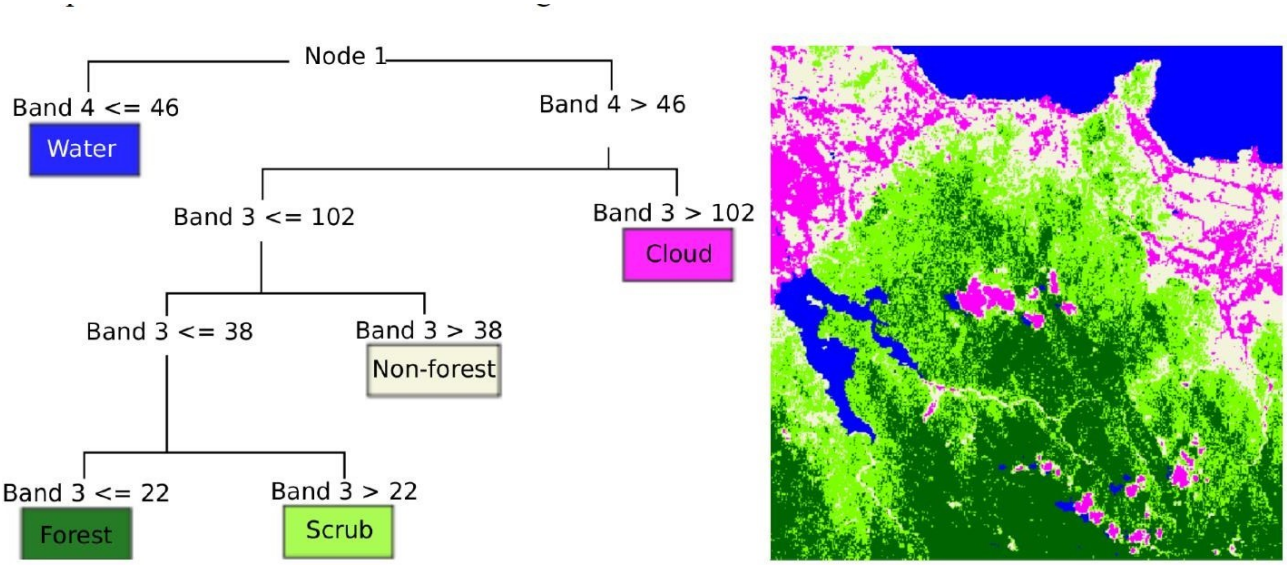


Figure 6.7: A very basic visual example of a decision tree model for land cover classification. *Source:* (Horning et al., 2010)

problems, the class with the most votes is chosen as the model prediction, and in regression problems, the average of the terminal node values is calculated. The success of the RF algorithm is dependent on how it generates each decision tree that makes up the forest. A subset of training data (approximately 2/3) is randomly selected with replacement to build and train each decision tree, and the remaining 1/3 of data is used to test the model's accuracy; this testing sample is known as the Out-Of-Bag (OOB) sample (Horning et al., 2010). The model performance obtained from this left out sample is usually called the OOB accuracy estimate (Horning et al., 2010). As a result, since the classification error is estimated internally in the run (the OOB error), there is no need for cross-validation or a separate testing data set in RF. However, when working with spatial data, it is always advisable to evaluate model performance using a different testing data set due to spatial auto-correlation between training and testing sets (LOTFIAN, 2016).

Balanced Random Forest: In SDM, background data (pseudo-absences) are typically sampled as a large number of points to represent the environment of the region under study, resulting in significantly imbalanced data sets, particularly for species with few presence samples. This class imbalance can cause the model to be biased towards the majority class (He & Garcia, 2009; Kaur et al., 2019), meaning that the model can have higher accuracy on the majority class and perform poorly on the minority class (Kaur et al., 2019). There are various

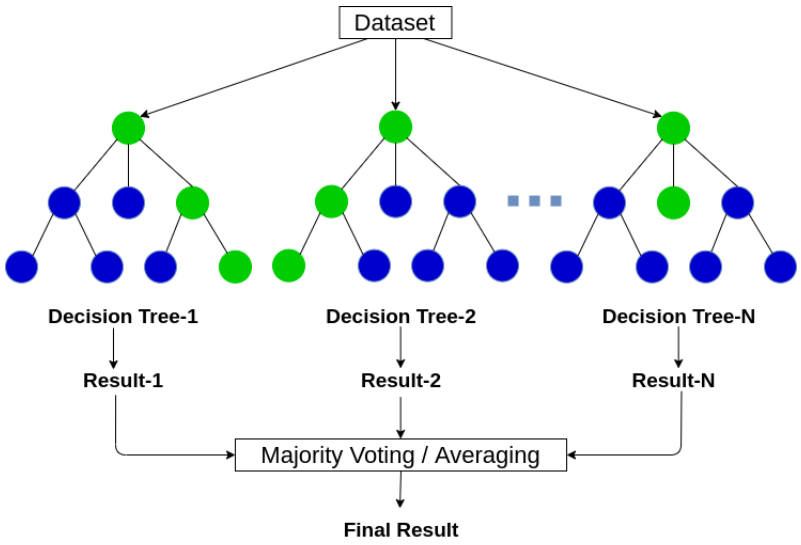


Figure 6.8: A basic visualization of random forest algorithm for classification/regression, *source: https://ai-pool.com/a/s/random-forests-understanding*

ways to deal with imbalanced data in ML algorithms (See (Kaur et al., 2019) for a systematic review of the solutions for the imbalanced data challenges in ML). One of the proposed solutions in the literature for classification of imbalanced data is the resampling method among which oversampling and undersampling are the two basic and most known data resampling approaches (Mohammed et al., 2020; Yap et al., 2014). Oversampling increases the number of samples in minority class by either creating duplicates of the samples or producing new samples from the minority class (Kaur et al., 2019; Mohammed et al., 2020). Two of the most known oversampling approaches are Random Oversampling which randomly creates duplicates of the minority class, and Synthetic Minority Oversampling Technique (SMOTE) which synthesizes new samples from the minority class (Chawla et al., 2002). Undersampling however, either removes or selects a subset of samples from the majority class (Mohammed et al., 2020). One of the simplest and widely applied undersampling approaches is the Random Undersampling method which randomly deletes examples from the majority class.

So, returning to the RF algorithm, as previously stated, the default RF produces bootstrap samples by randomly sampling the training data without regard for class labels. As a result, some bootstrap samples may contain very few or no examples of the minority class. One solution is to use one of the above data sampling methods on the bootstrap samples. In this thesis, we used a Python package called *Imbalanced-learn* (Lemaître et al., 2017) to train a Balanced

RF classifier (*BalancedRandomForest* in the mentioned package) which uses an undersampling method to select the bootstrap samples.

Support Vector Machine (SVM): SVM is a supervised ML algorithm that, like the previous algorithms, can be used for classification and regression problems. In SVM, the data are plotted in n-dimensional space, where n is the number of features. A hyper-plane is then drawn to divide the plotted data into two classes; however, the question is how to define the right hyper-plane that best classifies the two classes. When selecting the hyper-plane, there are several rules of thumb to consider, but the two main ones are selecting the hyper-plane that best separates the two classes and selecting the hyper-plane that maximizes the distance between the data of each class. The maximum distance between the data points of the two classes is called the maximum margin, and the training instances that are at the border (or on the edge) of the positive and negative hyperplanes are called support vectors. Figure 6.9 illustrates the hyperplane that separate the two classes, maximum margin, and support vectors (Géron, 2019). Although SVM has been used in a number of studies to model species distributions, it is not one of the most widely used approaches in ecological modeling (Drake et al., 2006; Raes et al., 2018).

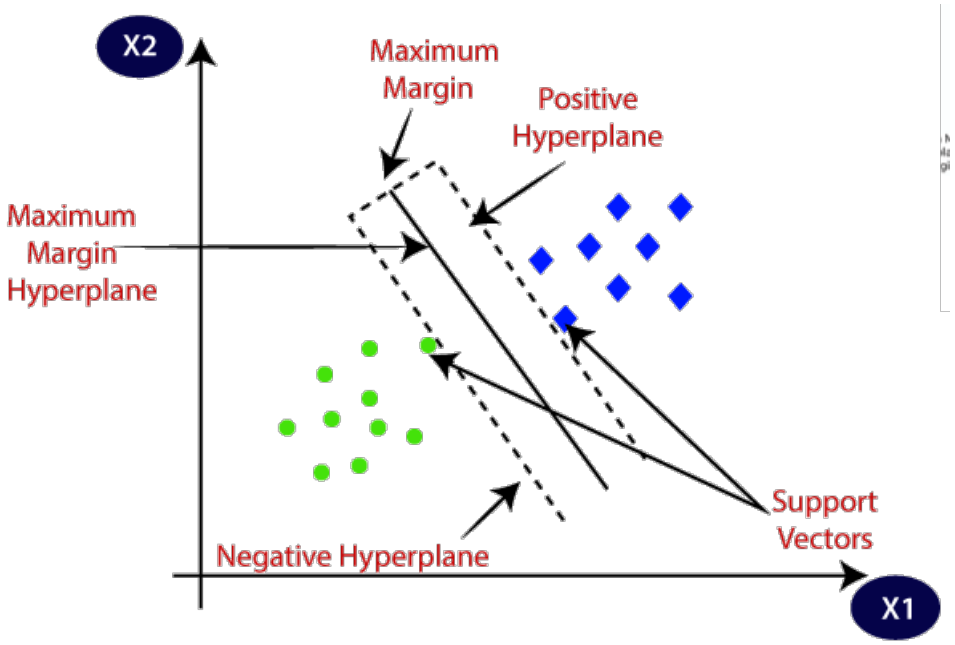


Figure 6.9: Support Vector Machine. *source: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm*

Naive Bayesian (NB): NB is a supervised algorithm based on the Bayes theorem, which is

named after the philosopher Thomas Bayes ("Bayes' theorem", 2021). Furthermore, the term *Naive* is based on a naive assumption that the effects of attributes on a given class are independent of one another, which is not the case in reality (Leung, 2007). Bayes' theorem describes the probability of an event occurring based on prior knowledge of the conditions associated with that event ("Bayes' theorem", 2021). Equation 6.9 illustrates the Bayes Theorem.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{6.9}$$

Where $P(A|B)$ known as *posterior probability* is the probability of hypothesis A to be true given the data B. $P(B|A)$ is the probability of data B given that hypothesis A is true. P(A), known as *prior probability of A*, is the probability that hypothesis A is true given any data. Finally, the P(B) is the probability of data B given any hypothesis. In equation 6.9, B is usually considered as evidence, and A as hypothesis, so $P(A|B)$ is usually written as $P(H|e)$, which is the likelihood of hypothesis H being true given the evidence e. All of the preceding explanations have been summarized in the figure 6.10.

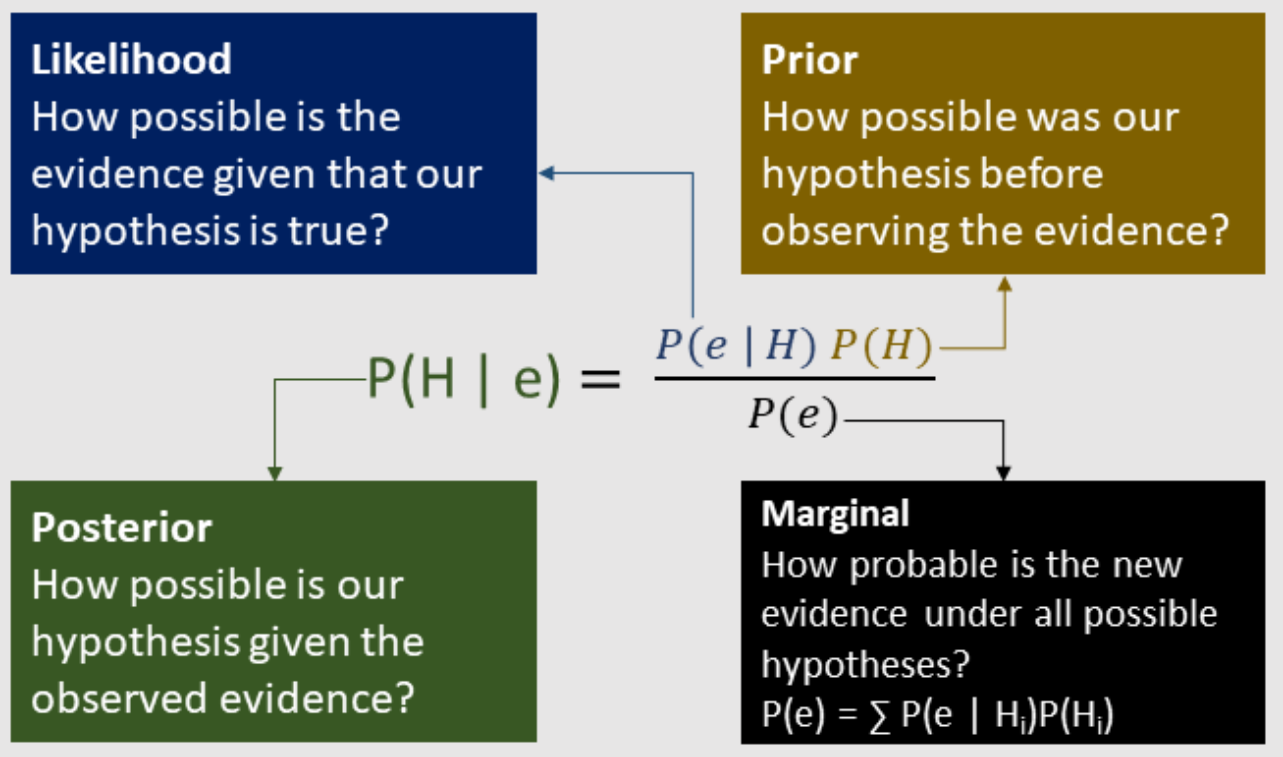


Figure 6.10: A summary of Bayes theorem, *Source: Adapted from (Geller, n.d.)*

Taking into account the Bayes theorem, here is how NB classifier works:

1) Let T be a set of training samples with k classes, $C = \{C_1, C_2, ..., C_k\}$, and each sample includes an m-dimensional vector, where m is the number of features. Given X to be one sample of the whole training data set T, then for each X there are m attributes $X = \{X_1, X_2, ..., X_m\}$.

2) The sample X will be predicted in class C_i if and only if the posterior probability of sample X being in class C_i is bigger than any other class:

$$P(C_i|X) > P(C_j|X) \tag{6.10}$$

Where j is any value between 1 and k (the total number of classes), and $j \neq i$.

Referring to the Bayes equation (equation 6.9), as P(X) is equal for all classes, the class C which maximizes $P(X|C_i)P(C_i)$, will be the predicted class for sample X.

3) As we mentioned, sample X includes a set of attributes, so the $P(X|C_i)$ should be computed for each attribute separately ($\{P(x_1|C_i), P(x_2|C_i), ..., P(x_m|C_i)\}$) and then multiply all together.

4) Calculating the probability differs depending on whether the features are categorical or continuous:

- Categorical features: If the features are categorical, $P(x_m|C_i)$ is computed as the number of x_m samples in class C_i, divided by the total number of samples in class C_i.

- Continuous features: For continuous features, the assumption is that the values have a Gaussian distribution with mean μ and variance σ, and the Gaussian is computed as:

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} exp^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.11}$$

  Thus, $P(x_m|C_i) = g(x_m, \mu, \sigma)$

In species distribution modeling, the features are the environmental variables in a neighbourhood around the location where the observation is made, and C is the class of species occurrence (presence or absence). Therefore $P(C_{present}|X)$ is the probability of observing a species given a set of environmental variables (X). Although NB is not among the widely used SDM algorithms (Raes et al., 2018), its performance has shown promising results (Altartouri & Jolma, 2013).

Now that we presented the required data set to generate SDM as well as the the various algorithms, we can apply some of these algorithms to our own use case and explain how we used SDM to validate the location of biodiversity observations in our project. However, we must first introduce our case study, how it is implemented, and how we intend to perform automatic data validation and real-time feedback generation in our application taking into account ecological modeling knowledge.

6.3 Case study: BioSenCS application

BioSenCS is a biodiversity CS project that we developed with the following objectives:

- Simplify data validation by automatically validating bird observations using ML algorithms
- Improve data quality through automatic filtering
- Give the participants real-time machine-generated feedback
- Encourage public engagement as a result of automatic feedback
- Increase participants' knowledge about biodiversity using machine-generated feedback
- Improve data quality as a result of automatic feedback

BioSenCS project initially created with the goal of being a complementary work to another project that we had, called BioSentiers. BioSentiers (Ingensand et al., 2018) is an Augmented

Reality (AR) mobile application, which aims at increasing pupils' knowledge about biodiversity and connect them with the nature through a gamified (Pokemon-like) application. In this project, pupils observe natural space through the camera of a portable device (tablet or smartphone), with the virtual elements representing biodiversity Points of Interest (POI) overlaying on the screen (See Figure 6.11).

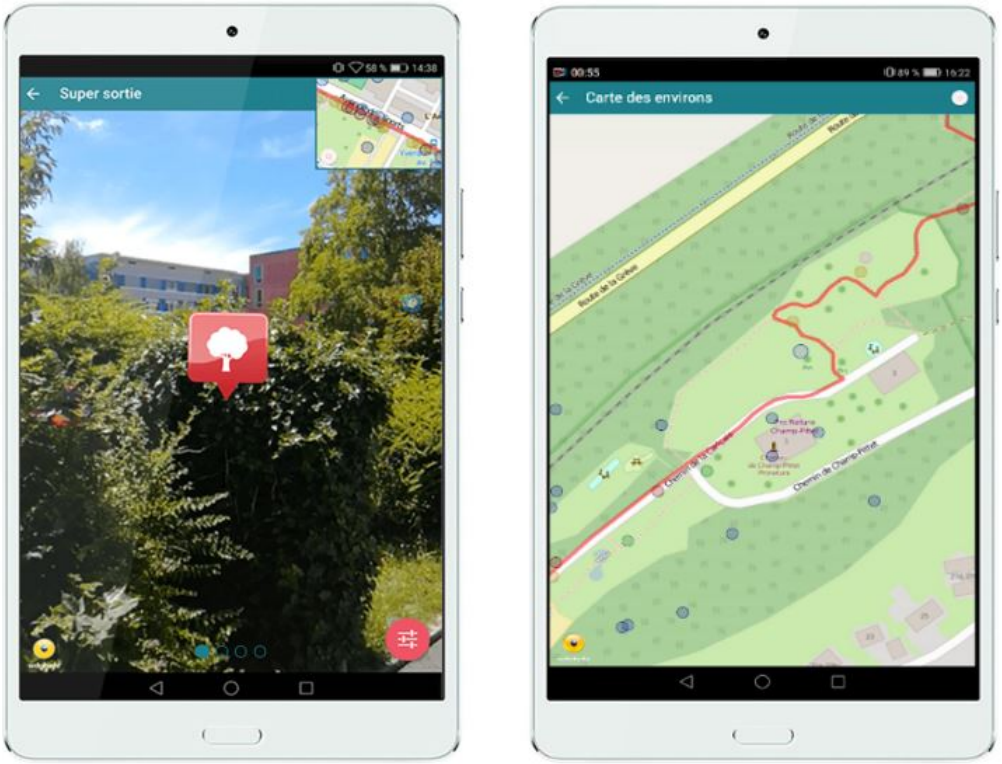


Figure 6.11: Screenshot of the AR mode (left) and 2D map (right) of BioSentiers application

By tapping the screen on each POI, pupils are guided about their distance (in meters) from the species in the nature. Once they get close enough to the object, they can tap the screen on the virtual object, and a modal opens containing an image and information regarding the species (See Figure 6.12). Therefore, pupils can compare the image with the real species, and enhance their knowledge about biodiversity. If a pupil sees the virtual object in the real world he/she needs to confirm having identified the object. In this way, the application can be entertaining, while at the same time keeping its fundamental goal. We ran an experiment in November 2017 with 15 pupils aged around 9-12 years old (See Figure 6.13) to evaluate the behaviour of the pupils while using the app, mainly checking whether they follow the path through the AR view or 2D map, or whether they use this app only with a motivation to have fun by finding the

species and go find another one, or whether they actually check the species information that pops up on the screen after finding the species. The results were published in proceedings of AGILE conference 2018 in Lund (Ingensand et al., 2018). One task for the continuation of this work was to figure out how to update the POI and feed the BioSentiers database with more new species. This guided us through the implementation of a CS project to collect new observations and enrich the BioSentiers database. The goal of this CS project shifted from simply enriching the BioSentiers database to implementing a biodiversity CS project to collect and automatically validate biodiversity observations.

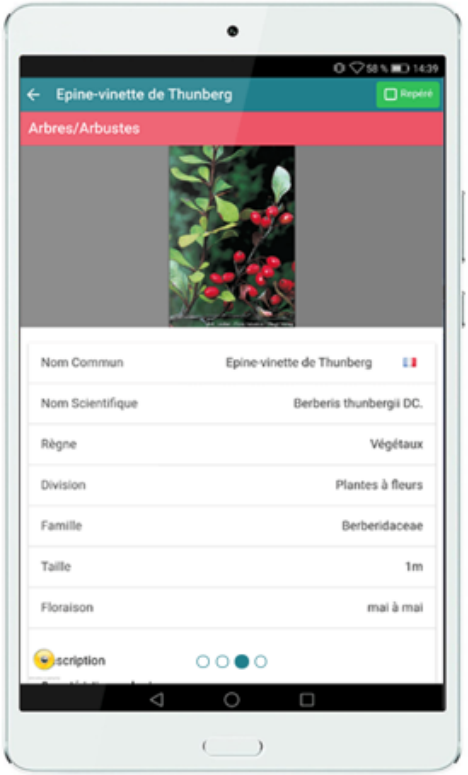


Figure 6.12: Information page of the species

BioSenCS[1] is thus an extension of the BioSentiers project, with the goal of collecting biodiversity observations and, more importantly, applying automatic data validation to the biodiversity observations while volunteers contribute (Lotfian et al., 2019). BioSenCS is built as a Progressive Web Application (PWA), which means it is designed to be used on smartphones and tablets, as participants primarily collect data while being out in the field. Although many CS applications are designed as native mobile applications, we believe there are several advantages to

[1]https://biosentiers-cs.heig-vd.ch/

Figure 6.13: The BioSentiers experiment with pupils

starting a CS project as a PWA. One advantage is that PWAs are independent of the operating system, making them easier to implement and update. Another advantage is that it is often difficult to convince people to install an application on their phone when starting a project and advertising for the project (See Chapter 5, Section 5.6), but if it is a PWA user can easily open the application by clicking on its URL (Uniform Resource Locator). So, we believe that using a PWA to build and launch a project and test a prototype with the initial community of participants is an efficient way to get started, and then once the community is established and user needs are identified, it can be a good step for transition to native applications and benefit from the additional functionalities that native applications provide for users. However, we should note that having a PWA has its own set of drawbacks, such as compatibility with different browsers and the lack of support for several functionalities, but for an experimental prototype, these were not an issue for us throughout this thesis.

BioSenCS is implemented in a Django framework[2], which is a Python-based free and open-source web framework, and it simplifies the developments of a web applications with its Model-View-Template (MVT) architectural pattern (Figure 6.14).

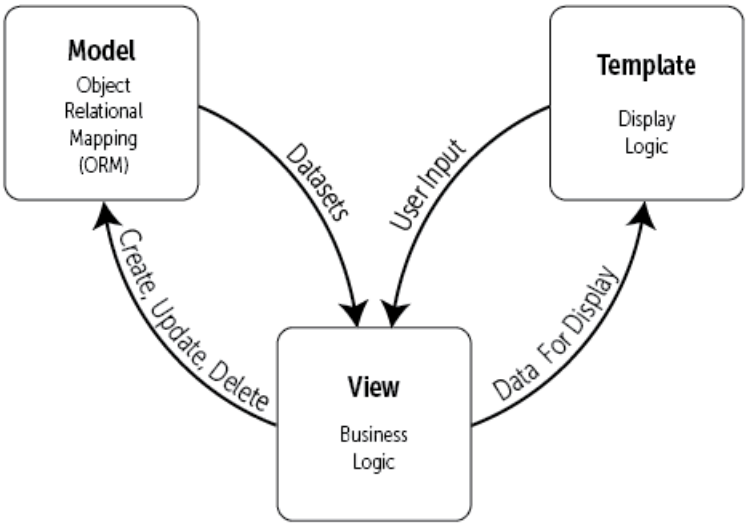


Figure 6.14: Django's Model-View-Template pattern("Django's Structure – A Heretic's Eye View - Python Django", n.d.)

We used PostgreSQL[3]/PostGIS[4] database for constructing our data models and preserving the collected observations. The high-level architecture of BioSenCS application is illustrated in figure 6.15, and the source code is available on GitHub[5].

In this project, a participant requires first to register as a user to the application, and then to start adding observations. An observation include location, organism type (flower, tree, bird, butterfly), species name, date, image, and a description of the observation. For creating the geometry features (point data), we used Django-Leaflet[6], which is a Python package allowing one to use Leaflet[7] (an open source JavaScript library used to build web mapping applications) in a Django project. The Django-Leaflet map features allow users to add, modify, and delete an observation point on the map. As the basemaps, we added OpenStreetMap[8] and ESRI World

[2]https://www.djangoproject.com/
[3]https://www.postgresql.org/
[4]https://postgis.net/
[5]https://github.com/mlotfian/Biosentiers-CS-functionality
[6]https://pypi.org/project/django-leaflet/
[7]https://leafletjs.com/
[8]https://www.openstreetmap.org/

satellite imagery tiles[9]. The user can see his/her observations (My observations) or the anonymous observations added by other users (All observations). The observations can be retrieved from our database as GeoJSON format[10] (an open standard geospatial data format based on JavaScript Object Notation (JSON)), and displayed on the map. Some of the screenshots of BioSenCS application are illustrated in Figures 6.16, 6.17, and 6.18.

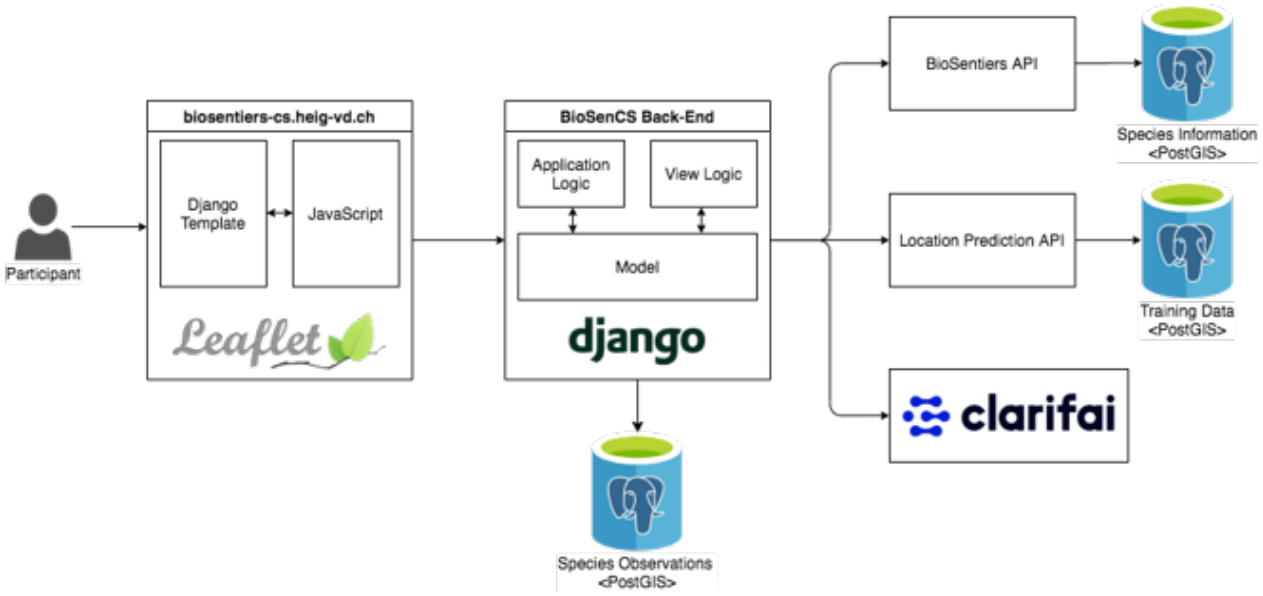
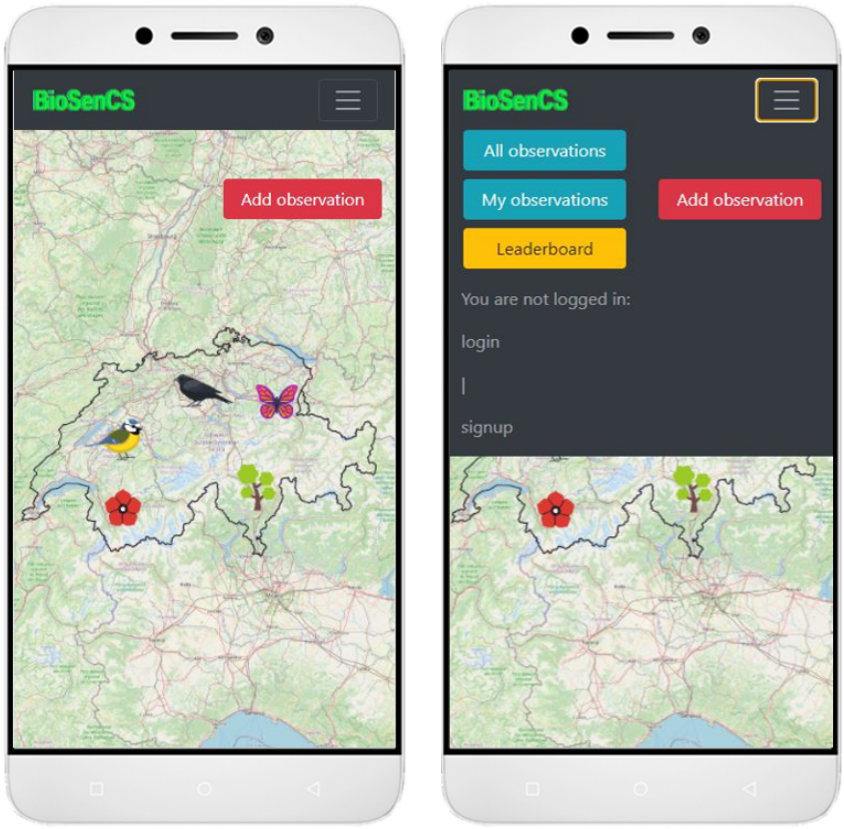


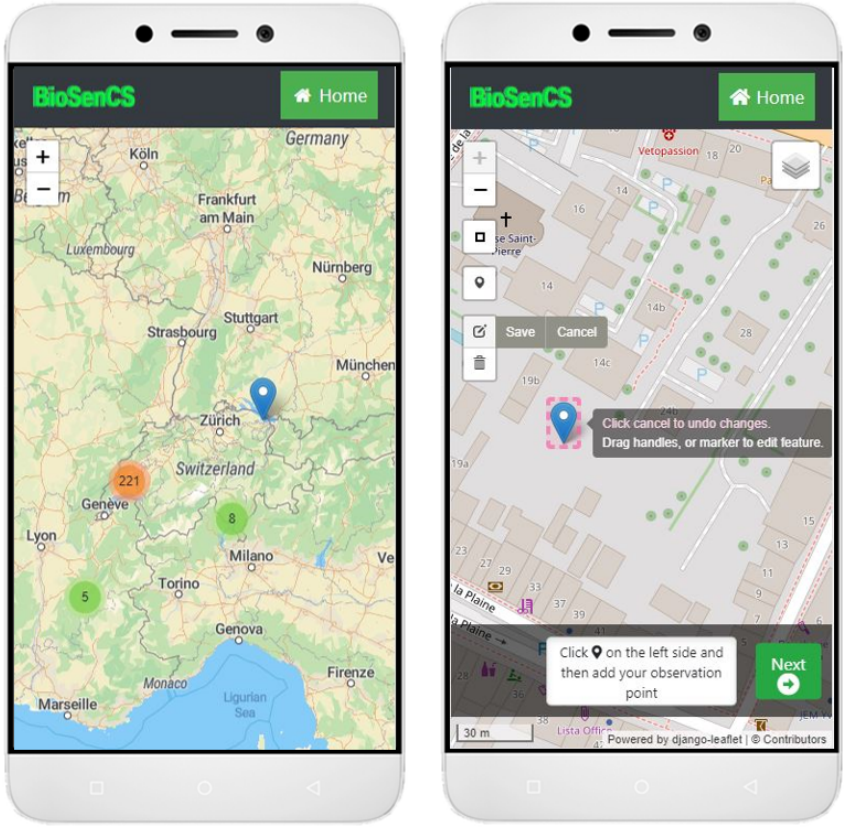
Figure 6.15: The high-level architecture of BioSenCS application

[9]http://services.arcgisonline.com/ArcGIS/rest/services/World_Imagery/MapServer
[10]https://www.ogc.org/standards/eo-geojson



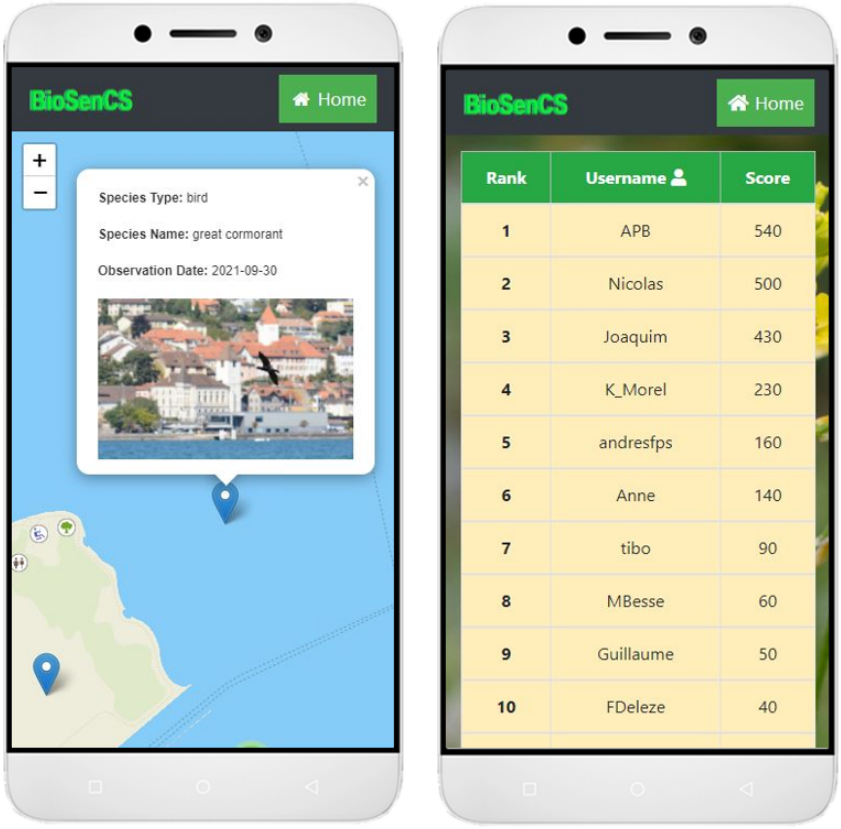
(a) (b)

Figure 6.16: Home page (a), and various buttons (b)



(c) (d)

Figure 6.17: Clusters of all observations (c), and Adding/Editing observation point (d)



(e) (f)

Figure 6.18: Querying information of species observation (e), and the participants' leaderboard (f)

One of the main goals of this project, as previously stated, was to apply an automatic validation or filtering of the observations. The validation process is illustrated in figure 6.19 and it works as follows: When a user submits an observation to the application, the observation goes through the automatic filtering process, and if the observation fails the automated filter criterion, it is flagged as unusual. In this case, the user will receive feedback (first feedback) explaining why the species was flagged as an outlier, and there are two possible outcomes: first, the user can modify the observation and resubmit it using the information in the machine-generated feedback, or second, the user can keep the observation as is and confirm the submission. In the second scenario, the observation will be forwarded to the final expert validation, and if more information is needed, the expert will send the user additional feedback (second feedback). Therefore, our two objectives here are to reduce the number of observations that must be controlled by the expert and to simplify the data validation task, and on the other hand to

give real-time feedback to the participants regarding their observation towards keeping them motivated and sustaining their participation to the project. We used three different types of automatic verification and feedback generation: date, image, and location validation. Two of these validations, image and location, use ML algorithms to perform auto-validation, while date filtering is done by comparing observation dates to an existing database provided by ecologists. Location filtering is only applied to bird species, but image and date validation is performed on all species types. The following subsections go over the specifics of these auto-validations.

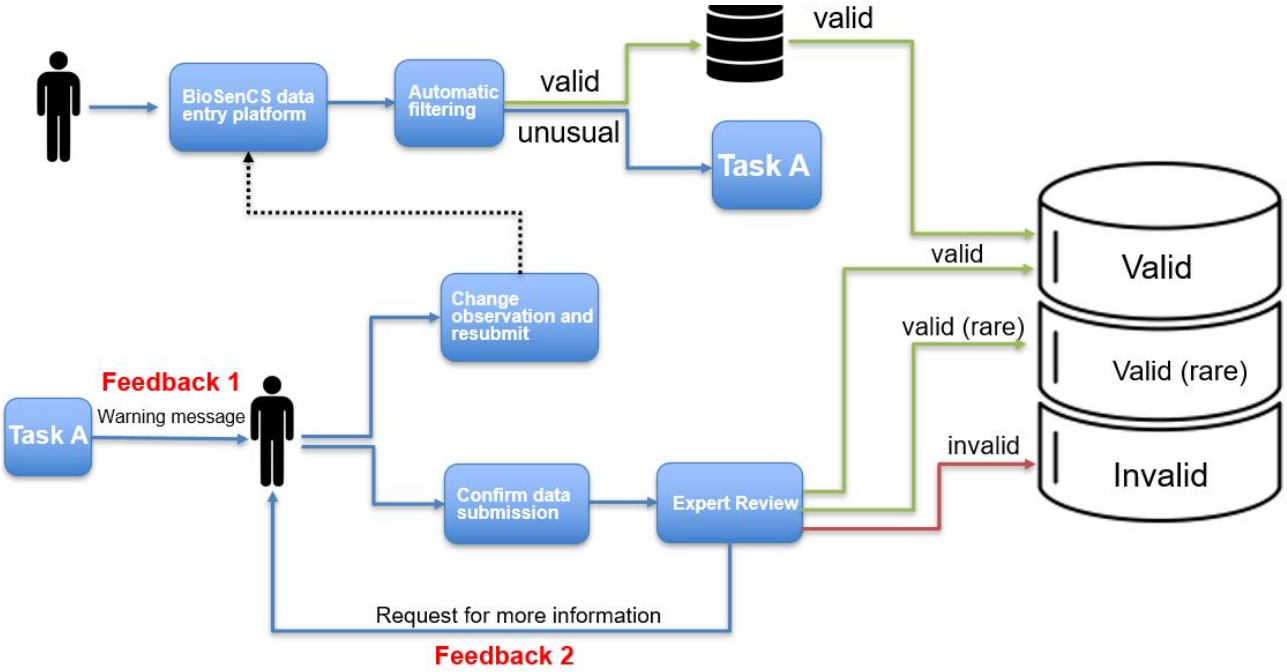


Figure 6.19: The automatic data validation procedure applied in BioSenCS

6.4 Image and date validation

Date validation: In the BioSentiers project (Ingensand et al., 2018), the initial information added to create biodiversity POIs was collected by ecologists which are saved in a PostgreSQl/-PostGIS database, and are available through BioSentiers Application Programming Interface (API)[11]. One of these information was the visibility period of the species (the attributes `periodStart` and `periodEnd`), which we used this information to verify whether or not, given a species name, the observation date falls within the species visibility period. If the obser-

[11]https://biosentiers.heig-vd.ch/api/species

vation date is outside the species visibility period, the observation is considered as an outlier and the participant receives a feedback with information about the months the species can be observed usually, and asking the participant to verify the added observation (e.g. species name or the date). The final decision is however given to the participant, and the participant is not forced to modify the observation. The observation is flagged in the database in a boolean attribute `flagDate` to be verified by experts later on. Figure 6.20 (a) illustrates the real-time date feedback to the participant.

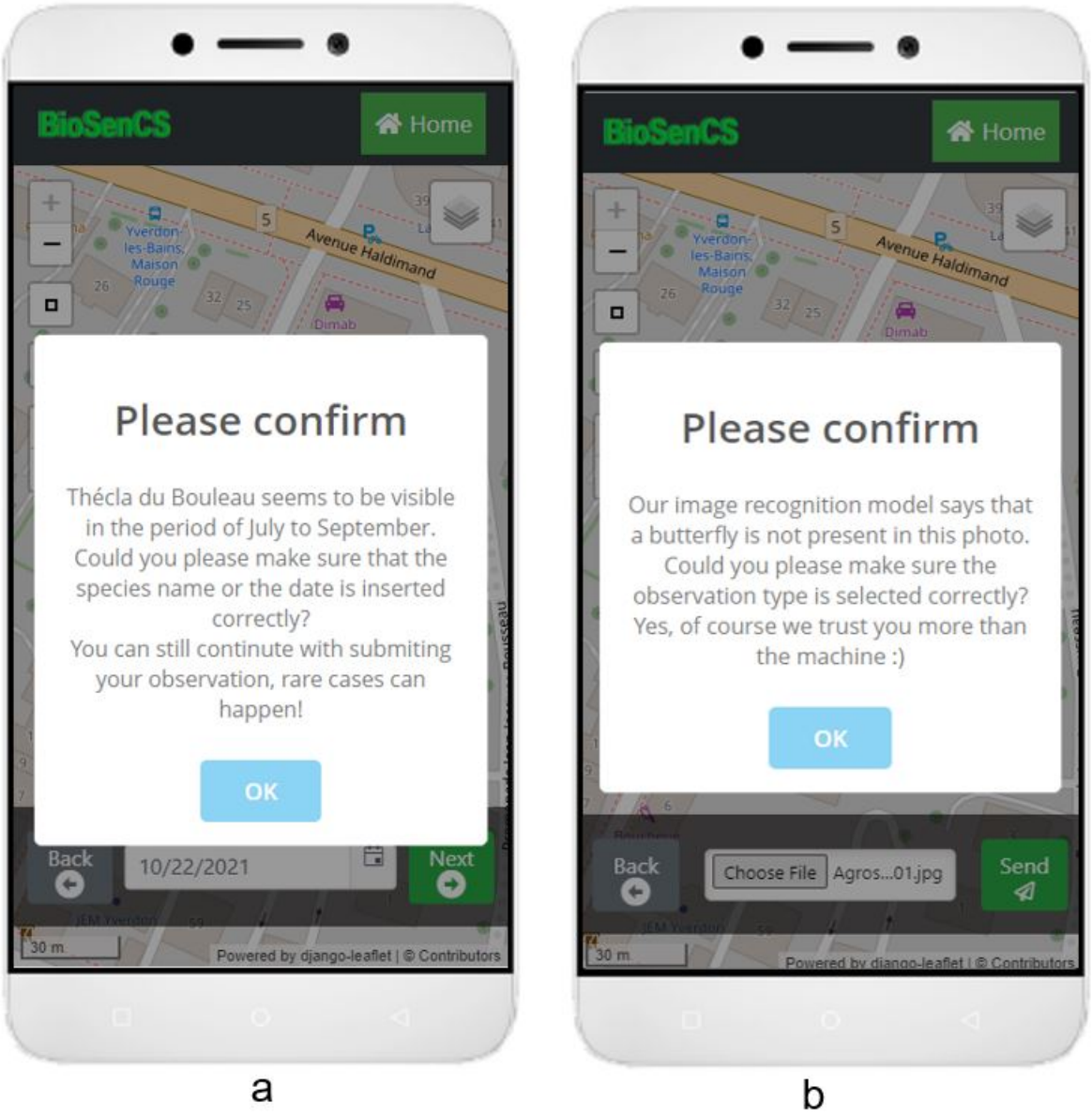


Figure 6.20: Automatic date (a) and image feedback (b) in BioSenCS application

Image filtering: The image filtering screens submitted images that do not include the reported

organism (bird, flower, tree, or butterfly), but it does not verify the image at the species level at the moment, which means that the image filtering does not check images of specific species. To perform image filtering, we used an AI platform called Clarifai[12]. Clarifai is an AI company that specializes in computer vision, NLP, and automatic speech recognition. It provides pre-trained models[13] as well as the option of training a model with a custom dataset. Clarifai offers services via its API (1000 free API calls per month), which has a fast response time and can be integrated into AI-powered mobile or web applications. The reasons for using Clarifai rather than training our own model for image validation were twofold: first, we wanted to save time by using available services to quickly create the first version of the application so that we could test it with users and get their input on automatic validation and feedback before moving on to location validation, and second, Clarifai was suitable for general image screening (image contains bird or not, but not checking the image of the species). We used its general model[14] to determine, for example, whether an image with bird tag really contains a bird or not. Once an image is sent to Clarifai's API, the model generates a set of possible tags that are present in the image along with their probability scores (See figure 6.21). We flagged an observation and sent a feedback to the participant, if the probability of having the organism in the uploaded image was less than 85 percent. Figure 6.20 (b) illustrates the real-time image feedback to the participants.

We used image and date filtering in the first version of the application to see how it could help us filter the observations and how participants would react to such feedback. However, the main focus of our work was to implement real-time validation of observation location, which we included in the second version of the application. The subsection that follows goes into detail about the approach and dataset we used for location validation.

[12]https://www.clarifai.com/
[13]https://www.clarifai.com/developers/pre-trained-models
[14]https://www.clarifai.com/models/image-recognition-ai

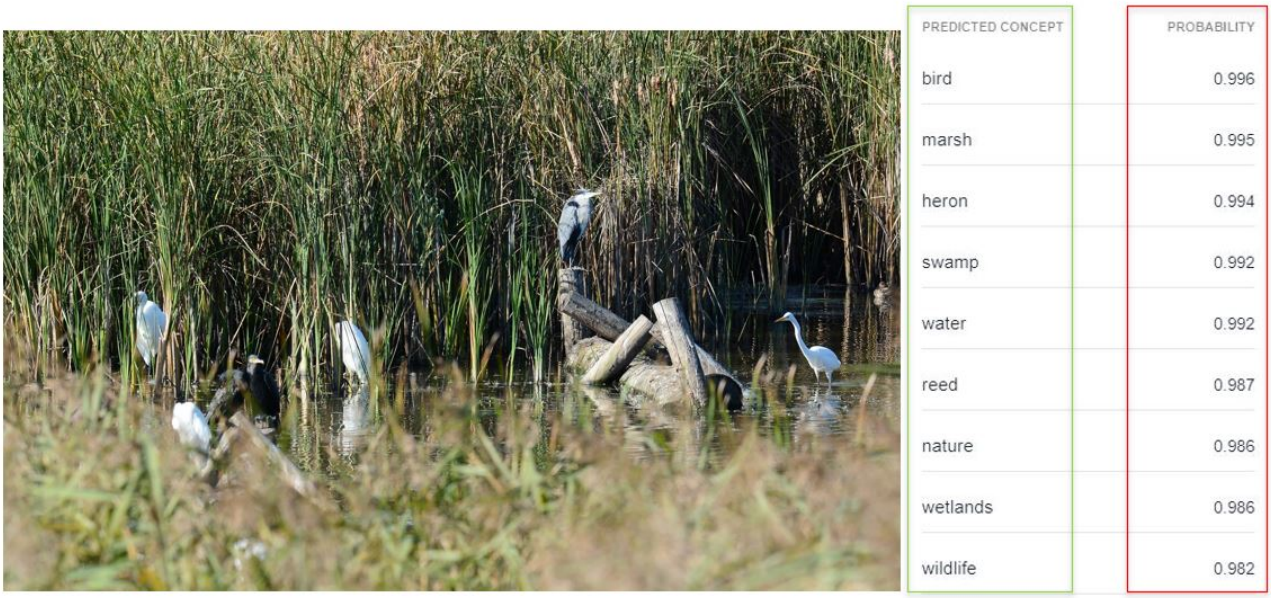


Figure 6.21: An example of Clarifai predicted tags and their probabilities for an observation contributed to BioSenCS

6.5 Location validation

To perform location validation, we determined how the environmental variables surrounding the observation location corresponded to the species habitat characteristics. To accomplish this, we generated the distribution of the species in relation to the environmental variables in our study area. Thus, we used the previously mentioned information about SDM (Section 6.2), particularly with regards to the required dataset and the ML algorithms to generate SDM. Thereby, we present the dataset we used with the steps on data preparation, the algorithms we trained to generate SDM, the evaluation and results of the algorithms, and finally, we discussed how we used the generated SDMs to validate the location of a new observation and to provide real-time feedback to the participants.

6.5.1 Data preparation for species distribution modeling

The three main steps of preparing data for SDM in this study are *obtaining raw data*, *generating pseudo-absences*, and *adding environmental variables to our data set*. The steps are discussed as follows:

1) *Obtaining and filtering eBird data*: With nearly 600 million bird observations from all over the world as of January 2019 (Strimas-Mackey et al., 2020a), eBird (Sullivan et al., 2014) is one of the largest CS projects (if not the largest) for collecting bird observations. eBird is highly valued by researchers worldwide due to its open access and wide-ranging spatial and temporal coverage (Strimas-Mackey et al., 2020a). What distinguishes eBird from other biodiversity CS projects is the method by which data are collected in order to account for robust observations, which means that not only the location and time of observations are collected, but also information about the effort made to collect an observation (Kelling et al., 2018). As a result of this data collection method, eBird is regarded as a semi-structured project (Strimas-Mackey et al., 2020a). A checklist structure is used for data collection in eBird. Each checklist represents a collection of observations made during a single event, such as taking a short walk around your house or watching birds in your yard. Each checklist contains information such as the time and location of observation, the number of individual species observed, media files (video, image, or audio), the effort used to obtain observations such as the number of people involved in collecting observations, the distance travelled to collect observations, the survey protocol (traveling, stationary, and incidental), and the time an observer spent collecting observations. A *complete checklist* is one that contains all of the bird observations that a participant was able to detect and identify. Accordingly, eBird contains two main data sets: the eBird Basic Dataset (EBD), which contains all of the information about species observations, and the Sampling Event Data (SED), which contains only the checklist data (SED contains no information about species names).

The eBird data can be requested from the eBird platform (https://ebird.org/data/download), indicating the date frame and region of interest. We obtained the observations for Switzerland from January 2016 to July 2020. Since the EBD data set is massive and cannot be opened with Excel, one solution would be to extract a subset of data using an R package called *auk* (Strimas-Mackey et al., 2021), which is designed for extracting and processing of eBird data. However, we used PostgreSQL/PostGIS database to open the eBird data and perform filtering and data pre-processing. The initial filter we applied was to make sure that all the records are validated using the Boolean `isValidated` attribute in the data set. Another filter we applied was choosing the

taxonomic category at species level[15] using the attribute `category`. Moreover, we selected only the complete checklists by filtering the Boolean attribute `all_species_reported`. Finally, we limited the species to those with at least one hundred distinct observation points. We obtained 322778 records (out of 400450) with 101 species as a result of data filtering.

It is worth noting that, prior to using the eBird data set, we investigated whether or not we could generate SDM using Flickr data, but our findings indicated that Flickr data can only be used as a complementary source to other (semi)-structured data sets and cannot be used exclusively in ecological modeling (See Appendix A for details of the Flickr data analysis).

2) Generating pseudo-absence data: Initially, we used the complete checklist to prepare a data set of presence-absence, in which absence is considered when a species is not reported in a complete checklist. However, absences are dependent on the observer's ability to detect a species as well as other detectability factors such as weather, time of day, and so on (Strimas-Mackey et al., 2020a). As a result, it is possible that the absences are not the true absence data in many cases, which is why we decided to use only the presence data from eBird and use the presence points to generate pseudo-absences for each species.

There are several ways to generate pseudo-absences, as discussed earlier in this chapter. In this thesis, we used a method of randomly sampling absences by taking into account the spatial extent around each presence point. We generated pseudo-absences with radius of 5 kilometers around each presence point, and we sampled pseudo-absences outside of these limits. For each species, we randomly sampled 5000 pseudo-absence points. Figure 6.22 illustrates an example of presence/pseudo-absences for an species called Carrion crow[16] with distance of 5 kilometers from the presence points.

3) Adding environmental variables: After generating the presence/pseudo-absence data sets for all 101 species, we added the environmental variables to each of the species data set. For each record (presence or absence) we created a neighbourhood of size $2km^2$ around the point, and we computed and extracted the environmental variables within this zone. In this thesis, the

[15]eBird taxonomic system: https://ebird.org/science/use-ebird-data/the-ebird-taxonomy
[16]Carrion crow: https://www.vogelwarte.ch/en/birds/birds-of-switzerland/carrion-crow

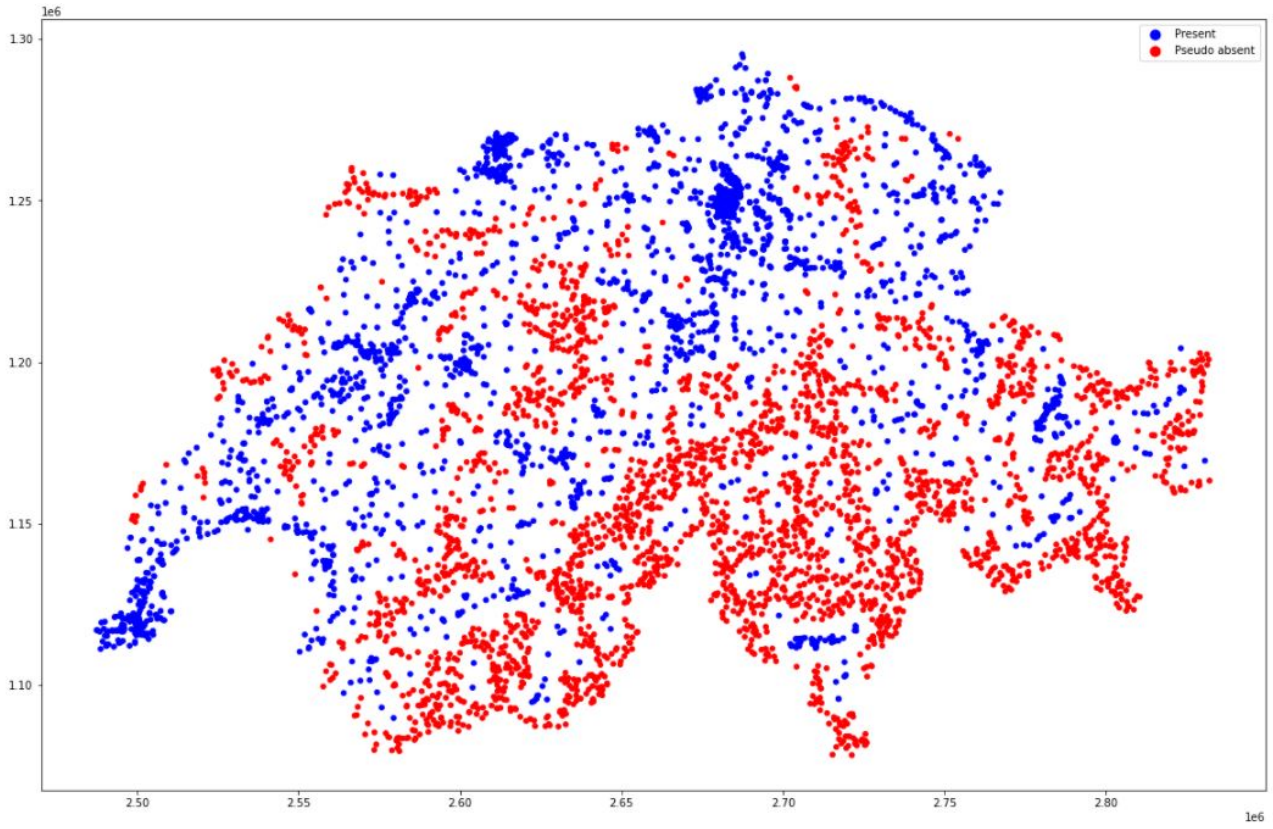


Figure 6.22: Presences (blue points) and Pseudo-absences (red points) for Carrion Crow in Switzerland

following environmental variables were used to generate SDM:

Average elevation: We used Digital Elevation Model (DEM) obtained from Swisstopo[17] with resolution of 200m. This is an acceptable resolution for our study as the zones in which we computed the environmental variables are with size of $2km^2$. Therefore, the average elevation was computed within each zone.

Average slope: Slope was computed from the DEM, and like elevation, the average slope was computed within each zone.

Land cover: The CORINE (Co-ordination of Information on the Environment) Programme was implemented by the European Commission between 1985 to 1990. CORINE Land Cover (CLC)[18] was designed with the aim of standardization of land cover data in Europe towards supporting environmental policy developments (Büttner, 2014). Standard CLC nomenclature

[17]https://www.swisstopo.admin.ch/en/geodata/height/dhm25200.html
[18]https://www.eea.europa.eu/data-and-maps/data/copernicus-land-monitoring-service-corine

includes three hierarchical levels (the level of details increase as we go from level 1 to level 3) in five major groups at level 1 including (1) artificial surfaces; (2) agricultural areas; (3) forests and semi-natural areas; (4) wetlands; and (5) water bodies, and with a total of 44 classes in level 3 (Büttner, 2014).

In this study, we used 1-level CLC classes for some variables, such as artificial surfaces, wetlands, and water bodies, but for others, such as vegetation surfaces, we needed more detailed information (3-level CLC). The ratio of each class occupied in the zone of $2km^2$ was then computed. For each class, the ratio ranges from 0 to 1, with 0 indicating that the class is absent in the zone and 1 indicating that the class completely occupies the zone.

The reasons for using CORINE land cover rather than Swiss data are due to two factors: first, Swiss data layers are not always homogeneous, and data varies from canton to canton; and second, using CORINE land cover makes reproducing this analysis on other European countries easier.

Average annual NDVI: The NDVI (Normalized Difference Vegetation Index) was obtained from the Swiss Data cube portal[19]. The mean annual and mean seasonal data sets of NDVI for Switzerland can be downloaded free of charge from this portal. We obtained the annual mean NDVI from 2016 to 2018 (the most recent ones available by the time of doing our analyses), and we generated the average map of these years. The generated NDVI map was then used to calculate the average values within each zone.

A data set was created for each species which included all of the presence/pseudo-absence points as well as the environmental variables for each point. To do so, we implemented a function to first create a neighbourhood of $2km^2$ around each point and then to extract the environmental variables mentioned above within each zone. Data set generation for all 101 species took around 8 hours. Table 6.4 contains a detailed list of the environmental variables used to generate SDM in this study.

In addition to computing environmental variables around observation points for each species, we

[19]https://www.swissdatacube.org/

generated environmental variables across all of Switzerland to create a prediction surface, which is later used to generate species distribution maps within Switzerland (See section 6.5.3.2). The prediction surface's resolution is also $2km^2$, which is the same as the resolution used to prepare the species data set. Figure 6.23 represents six examples (out of the total 19) of environmental variables computed over Switzerland.

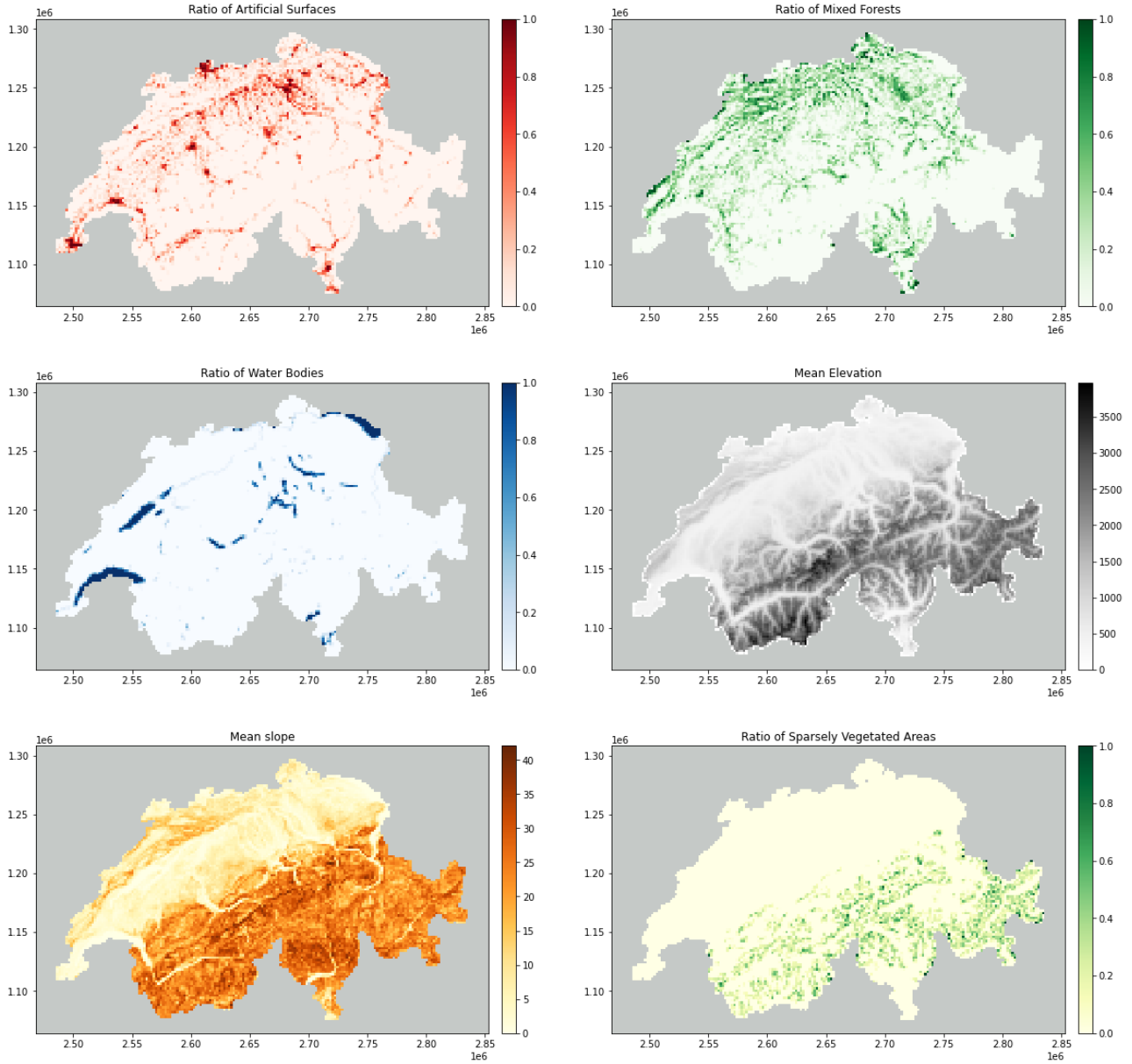


Figure 6.23: Example of six environmental variables computed over Switzerland with resolution of $2km^2$

Table 6.4: Environmental variables used in this study to generate SDM

Environmental variables	**Data set**	**Description**
Altitude	DEM, Swisstopo	Average Altitude
Slope	DEM, Swisstopo	Average Slope
Artificial Surfaces	CORINE land cover	Urbanized areas, industrial zones, etc.
Non.irrigated arable land	CORINE land cover	Cultivated land parcels under rainfed agricultural use
Permanent crops	CORINE land cover	All surfaces occupied by permanent crops
Pastures	CORINE land cover	Lands that are permanently used for fodder production
Heterogeneous agricultural area	CORINE land cover	Areas of annual crops
Broad-leaved forest	CORINE land cover	Vegetation formation composed principally of trees, including shrub and bush understorey
Coniferous forest	CORINE land cover	Vegetation formation composed principally of trees, including shrub and bush understorey, where coniferous species predominate
Mixed forest	CORINE landcover	Vegetation formation composed principally of trees, including shrub and bush understorey, where neither broad-leaved nor coniferous species predominate
Natural grassland	CORINE landcover	Grasslands under no or moderate human influence
Moors heatland	CORINE landcover	Vegetation with low and closed cover, dominated by bushes, shrubs, dwarf shrubs
Transitional woodland/shrub	CORINE landcover	Transitional bushy and herbaceous vegetation with occasional scattered tress
Beaches and bare rocks	CORINE landcover	Natural non-vegetated expanses of sand and scree, cliffs, rock outcrops
Sparsely vegetated areas	CORINE landcover	Areas with sparse vegetation, covering 10-50 percent of surface
Glaciers and perpetual snow	CORINE landcover	Land covered by glaciers or permanent snowfields

Wetland	CORINE landcover	Areas such as inland marshes
Water bodies	CORINE landcover	Water areas such as lakes or marine water lands
NDVI	Swiss Data Cube	Average annual NDVI of 2016 to 2018

6.5.2 Spatial cross validation

Prior to training the algorithms, it is essential to understand how to divide the input data (also known as the learning data set) into training and testing sets, where the model is trained using the training data and the model performance is evaluated using the testing data set (previously unseen data during the training phase), a process known as cross-validation (Berrar, 2019). Cross validation methods include, but are not limited to, leave p-out cross-validation, leave-one-out cross-validation, k-fold cross-validation, and stratified k-fold cross-validation.

- Leave p-out cross validation (LpOCV): It is a cross-validation technique in which a p sample of data (out of a total n sample) is used for validation and the remaining data is used for training. This process is repeated over the entire original data set to cut it each time to a validation set of p samples and a training set. The model performance is computed in each run, and the final model performance is the average of all the performances (Liu, 2019).

- Leave one-out cross validation (LOOCV): LOOCV is a particular case of LpOCV where p=1, which means that one sample is set aside to evaluate the model while the remaining samples are used to train the model (Wong, 2015). This procedure is repeated until all samples have been used to evaluate model performance once (see Figure 6.24).

- k-fold cross-validation: In k-fold cross-validation, the original data set is divided into k equal parts known as folds. Each time, one fold is used for validation, and the model is trained on (k-1) folds, and the process is repeated k times until each fold is used as a validation set once (Wong, 2015). The final accuracy of the model is the average of the accuracy obtained on each fold ($\Sigma_{i=1}^{k} \frac{acc_i}{k}$) (See Figure 6.25).

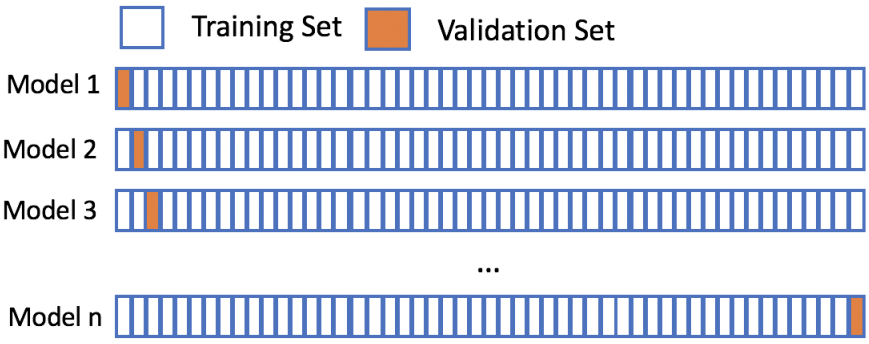


Figure 6.24: Leave-one-out cross-validation. *source: ("Cross Validation and Model Selection", n.d.)*

- Stratified k-fold cross-validation: One of the major drawbacks of k-fold cross-validation is that it does not perform well with imbalanced data (He & Ma, 2013). So when minority class is absent from a fold, the model's prediction is biased toward the majority class. The stratified k-fold cross-validation method addresses the problem of imbalanced data. The folds in this approach are designed in such a way that all folds have an equal number of target classes. For example, if we have 1000 samples with two classes of present(10) and absent(990), and we want to generate stratified 5-folds, each fold will contain two present samples and 198 absent samples.

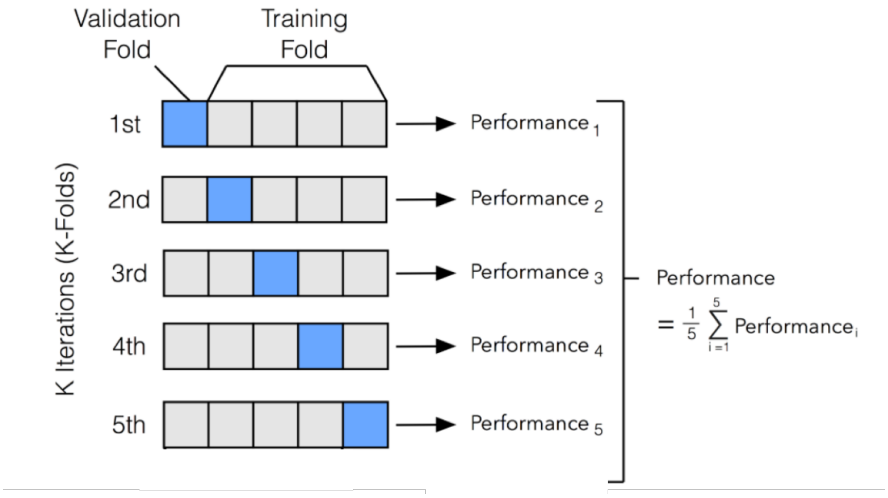


Figure 6.25: k-fold cross-validation for k=5. *source: ("Cross Validation and Model Selection", n.d.)*

When performing cross validation on spatial data, it is necessary to consider the spatial auto-correlation that exists among the variables. Considering Tobler's first law of geography (Tobler,

2004), "everything is related to everything else, but near things are more related than distant things", the training and validation data set must be spatially distinct otherwise the model would be very potential to suffer from over-fitting (Hawkins, 2004). Therefore, when working with spatial data, one approach to applying cross validation is to define spatial blocks and perform block cross-validation (Roberts et al., 2017). Roberts et al. (2017) have investigated cross validation strategies for temporal, spatial, hierarchical, or phylogenetic data. They defined three factors to consider when determining the size of a spatial block for cross-validation: auto-correlation among variables, data limits, and computational limits. Moreover, they proposed three ways to assign folds to spatial blocks: unique, systematic, and random (See figure 6.26 for a visual explanation). The size of the spatial blocks is best to be chosen larger than the range of spatial auto-correlation among the environmental variables (Roberts et al., 2017). An initial indicator for selecting the size of the spatial block would be to check the range of spatial auto-correlation among environmental variables, and thus to select a block size that is not too large (training data will be biased towards some areas, causing the model to not learn well on the data set) and not too small (the issue of spatial auto-correlation will cause over-fitting).

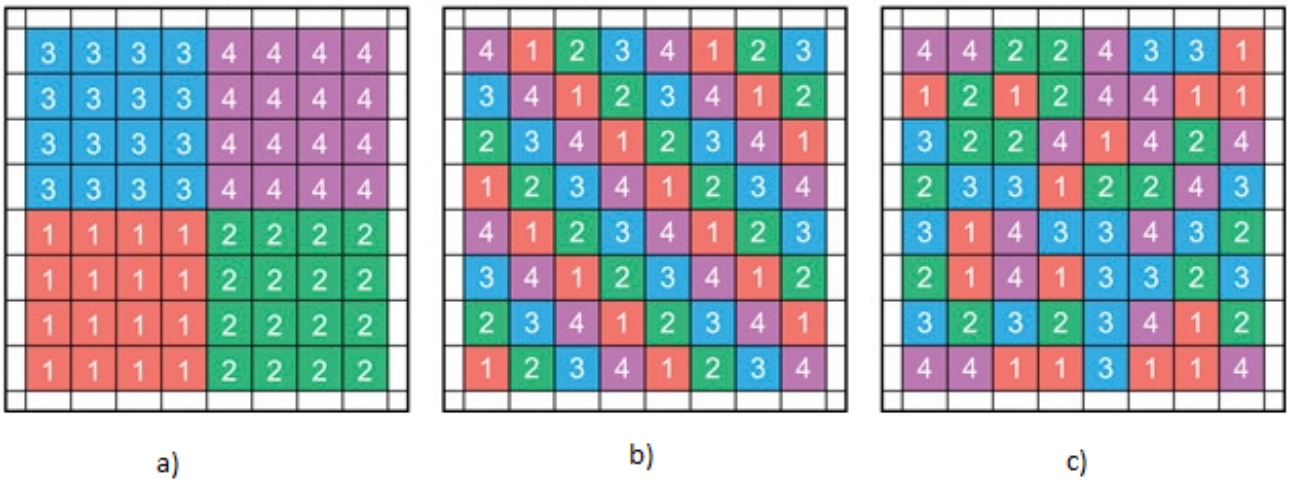


Figure 6.26: Block cross-validation: the three methods propose by Roberts et al. (2017) to arrange spatial folds, a)Unique, b)Systematic, and c)Random

We initially divided Switzerland to 6 spatial blocks and 6 unique folds. However, the models did not perform well on the large spatial block because the variations of environmental variables within each fold are rather large, causing the model not to learn well because the input features are not good representative of the entire training data set. We then chose to use the smaller

block size, taking into account the range of spatial auto-correlation among the environmental variables we chose to generate SDM. We used an R package named blockCV (Valavi et al., 2018) to obtain the auto-correlation range among the variables. The approach used in this package is to fit variograms to continuous raster data and obtain the range of spatial auto-correlation as a result. However, it is important to note that this range is not a definite value, but rather a way for users to get an initial idea of the block size to use. Figure 6.27 illustrates the proposed block size for our case study, and considering this initial indicator we defined our spatial blocks of size approximately $50km^2$.

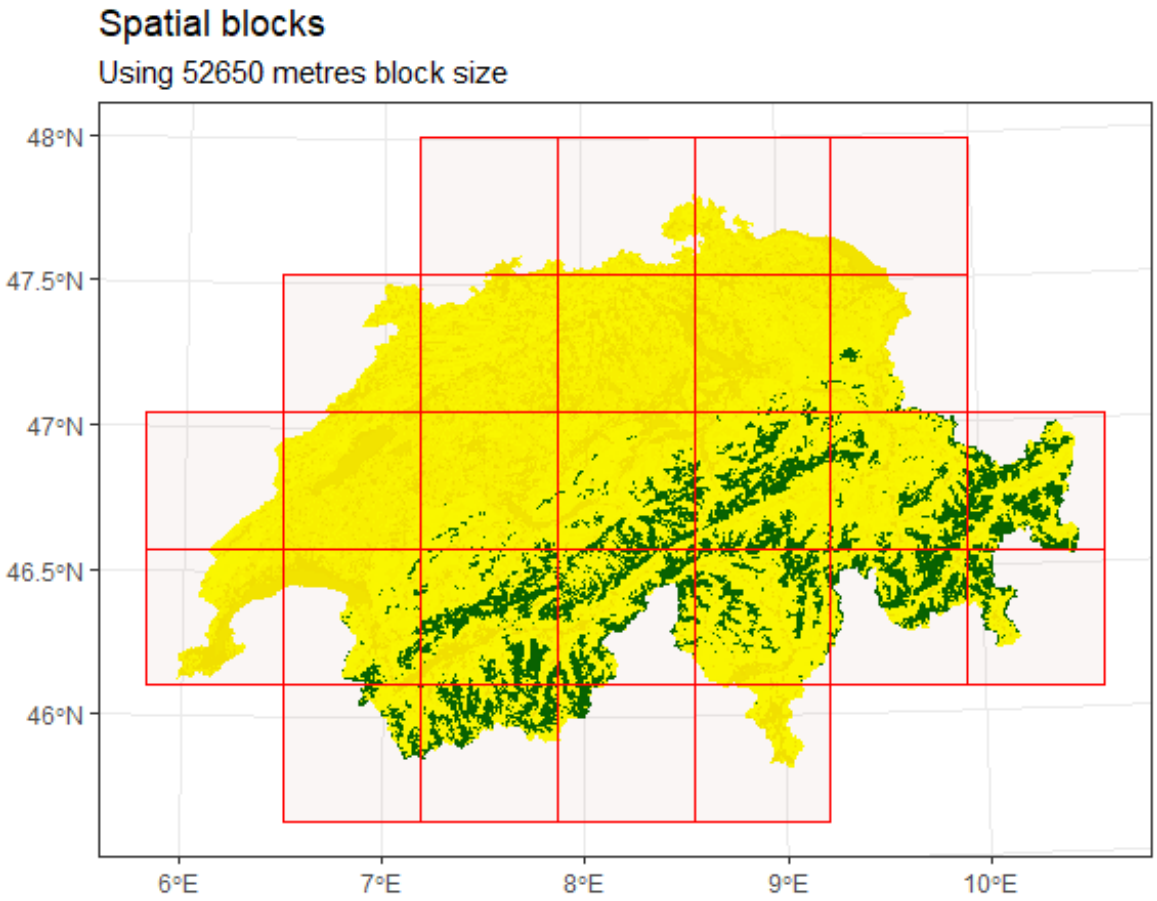


Figure 6.27: The result of blockCV package (Valavi et al., 2018) for getting the block size for spatial block cross validation taking into account the spatial auto-correlation range among the environmental variables used in our study

We used a Python package called spacv (Comber, 2021), which generates spatial blocks and folds taking into account the techniques proposed by Roberts et al. (2017). Considering the approximate chosen block size, we generated 55 spatial blocks (5 horizontal and 11 vertical)

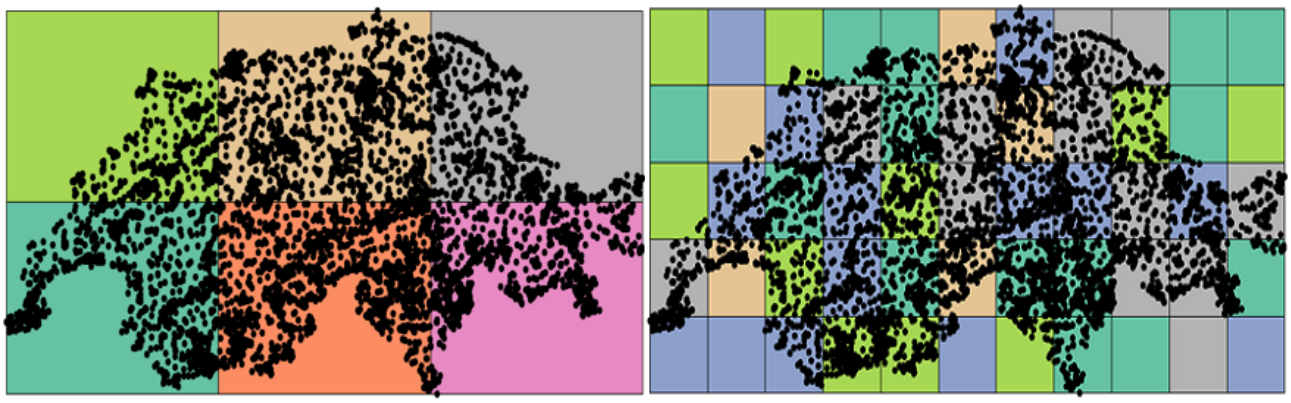
and 5 folds randomly assigned to the blocks (Figure 6.28, b). In addition to defining the spatial blocks, we had to assign the folds to the blocks in such a way that the class imbalance in each fold was minimized. To accomplish this, we wrote a Python function that ran a predetermined number of times, checking the class balance in each fold and finally selecting the best folds among these runs. Accordingly, this function takes as input variables *the number of times we want to assign random folds*, as well as *the data set of each species*, which includes the two classes of presence and absence. We checked two conditions each time we assigned random folds:

1) Each fold includes at least 15% of the sample size of each class

2) Difference of sample sizes of the two classes in each fold is minimized

To minimize class imbalance, we calculated the difference in sample sizes between the two classes in each fold, then the sum of differences was computed, and the iteration with the lowest value of this sum was chosen as the iteration with best folds. For each species, we ran *150* iterations and extracted the folds with the most balanced samples as the best folds. It took around 20 hours to obtain the best folds for all of the 101 species using a computer with 270GB of RAM, and a processor *Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz* with 14 CPU cores. We aimed to generate the spatial folds using the Google Colaboratry platform[20] at first, but due to the limitations of a free plan, such as the interruption of an "Idle" notebook after 90 minutes and the "maximum lifetime" of a running notebook limited to 12 hours, we decided to use the "super computer" that we had. However, it is still possible to reproduce the results using Google Colab, provided that the analyses are broken down for a smaller number of species in each run (e.g. instead of generating spatial blocks for all species at once, running the notebook each time for 5 species, and repeat until spatial blocks are generated for all species).

[20]https://colab.research.google.com/notebooks/intro.ipynb



(a) 6 blocks, 6 folds (b) 55 blocks, 5 folds

Figure 6.28: An example of defining spatial blocks for our study area using two different block sizes and two different methods of assigning folds: a) Unique folds for each block b) 5 folds randomly assigned to the spatial blocks

6.5.3 Comparison of algorithms to generate species distribution models for our case study

Among the ML algorithms described in this section, we applied four algorithms of DNN, RF, Balanced-RF, and NB. We trained the algorithms using the data sets generated for each species (presence-absence and environmental variables), and then compared the model performances for all 101 species.

6.5.3.1 Training the algorithms

All the algorithms were implemented in Python and the scripts for training and accuracy assessment were written in Google Colaboratory notebook. To build and train the algorithms, we used free and open-source ML libraries including *scikit-learn*(Pedregosa et al., 2011) , *imbalanced-learn* (Lemaître et al., 2017), *Tensorflow*(Abadi et al., 2016), and *Keras*. Table 6.5 summarizes the applied algorithms including their parameters, the value and description of the parameters, the training time, and the Python library used to construct the model.

Table 6.5: Algorithms trained in this study to generate SDM

Model	Parameters [i]	value	Training time (all species)	Python library
NB	all params	default	10 min	sklearn (NBGaussian)
default RF	Number of trees (nestimators)	2000	2h 30 min	sklearn (RanddomForestClassifier)
Balanced-RF	Number of trees (nestimators) replacement	2000 False	1h 40 min	imblearn (BalancedRandomForestClassifier)
DNN [ii]	Hidden layers Learning rate (lr) epochs Dropout	4 0.001 150 0.3 and 0.5	6h 30 min	TensorFlow Keras

[i] The full list of possible parameters for each model can be find in the official page of each library

The DNN was constructed with four hidden layers of sizes 50, 25, 25, and 25 neurons, and a *ReLU* activation function was applied in each hidden layer, and *sigmoid* activation was used in the final output layer. Moreover, each hidden layer was followed with a dropout layer to account for generalization and to avoid overfitting (Baldi & Sadowski, 2013). Figure 6.29 illustrates the architecture of the DNN trained in this study, however, the effect of dropout layers is not accounted for in the figure.

6.5.3.2 Evaluation and results of the algorithms

There are various metrics to evaluate ML algorithms such as classification accuracy, F1 score, Area Under (ROC) Curve (AUC), logarithmic loss, etc. To evaluate the performances of our algorithms we used AUC.

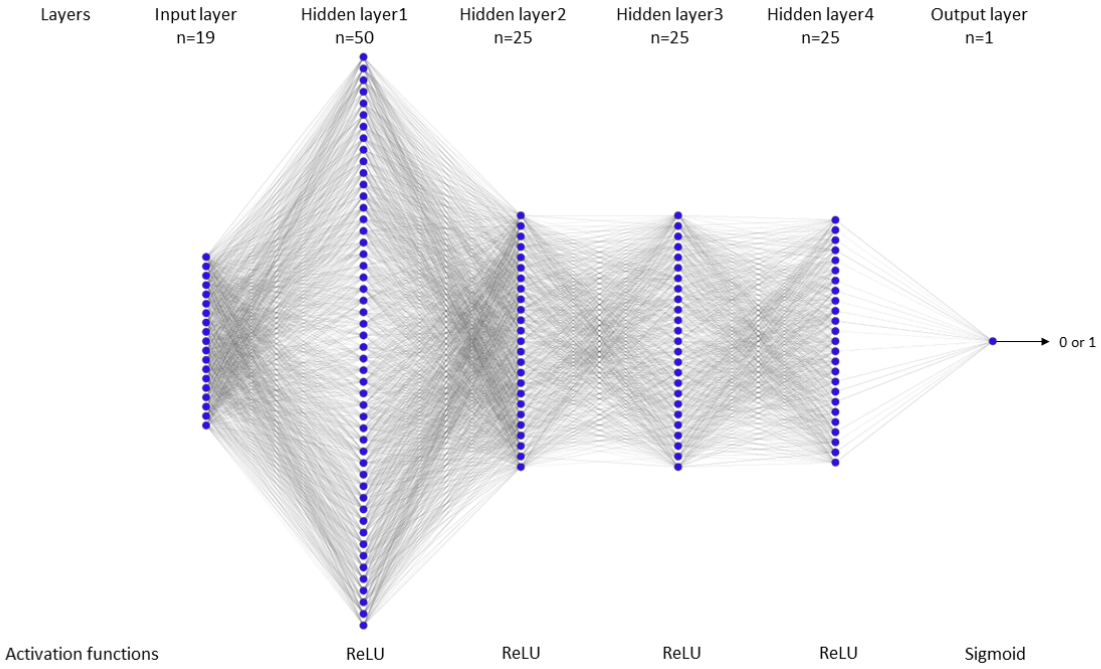


Figure 6.29: The architecture of the DNN we trained to generate SDM

- Classification accuracy: Classification accuracy is one of the most commonly used metrics to evaluate the performances of ML models. It is computed as the ratio of the total number of correctly classified samples over the total number of test sample used to evaluate the model. Usually it is computed through a confusion matrix (or error matrix). Figure 6.30 shows a confusion matrix for a binary classification problem, and equation 6.12 shows how to compute the overall accuracy. Although classification accuracy is widely used to evaluate ML algorithms, when it comes to imbalanced data, it might not give us the best indicator to measure how good the model performs (Menardi & Torelli, 2014), and it might mislead us by producing high values. Thus, other metrics such as AUC can be a better indicator for model evaluation.

$$Accuracy = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{FN} + \mathrm{FP} + \mathrm{TP}} \tag{6.12}$$

- Area Under Curve (AUC): ROC (Receiver Operating Characteristics) curve is a graph to visualize and measure a classifier's performance (Fawcett, 2006). ROC curve plots the true positive rate (y-axis) against the false positive rate (x-axis) (Fawcett, 2006).

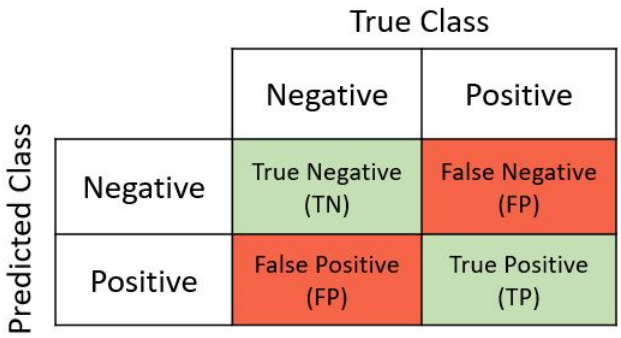


Figure 6.30: Confusion matrix to evaluate accuracy of a binary classification problem

True Positive Rate (TPR) is calculated as the ratio of true positives and the sum of true positives and false negatives (Equation 6.13). TPR is also referred to as sensitivity. False Positive Rate (FPR) is calculated as the ratio of false positives and the sum of false positives and true negatives (Equation 6.14). FPR is also referred to as the inverted specificity or $(1 - specificity)$. Specificity is calculated as the ratio of true negatives and the sum of true negatives and false positives (Equation 6.15). The AUC is then a metric that presents a summary of a model performance and is computed as the area under the ROC curve (See figure 6.31). AUC values ranges from 0 to 1, and like accuracy higher values illustrates better performance.

$$TPR = \frac{TP}{TP + FN} \tag{6.13}$$

$$FPR = \frac{FP}{FP + TN} \tag{6.14}$$

$$specificity = \frac{TN}{TN + FP} \tag{6.15}$$

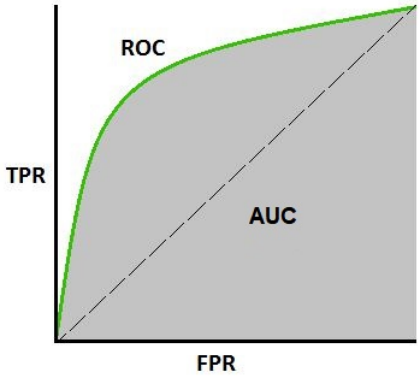


Figure 6.31: The Receiver Operating Characteristics (ROC) curve. The gray part shows the area under the curve. *Source: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5*

For each species, we computed the average AUC over the five folds for all the four algorithms. Figure 6.32 illustrates the box plots of the variations of AUC within the trained algorithms for all the species. From the box plots we can observe that DNN has a higher AUC median (0.86) compared to the other algorithms, however it's performance is not consistent through all the species. Balanced-RF, with an AUC median of 0.82, outperforms default RF (median = 0.74) and NB (median = 0.75), and performs relatively better across all species than the other three algorithms. Furthermore, for some species where the other three algorithms performed poorly (AUC less than 70%), Balanced-RF outperforms the others. Figure 6.33 depicts this variation, and it shows that default RF performs the worst for these species, NB and DNN perform similarly, and Balanced-RF performs better for all of them.

Moreover, in order to illustrate that the models' performances are statistically different, we performed a non-parametric test called Friedman's Aligned Rank for the AUC values for all the four algorithms, and the results showed statistically significant difference between the model performances with p-value$<$0.0001. Moreover, the pairwise comparison of the models' performances illustrated as well statistically significant difference. We obtained as well the model average rankings in table 6.6, and it can be seen that DNN and Balanced-RF have higher ranks than RF and NB, but DNN rank is slightly higher than Balanced-RF. One reason that DNN is ranked higher can be due to the fact that deep learning algorithms are data-hungry (Marcus, 2018; Yu et al., 2015), and their performance for larger data sets is significantly higher

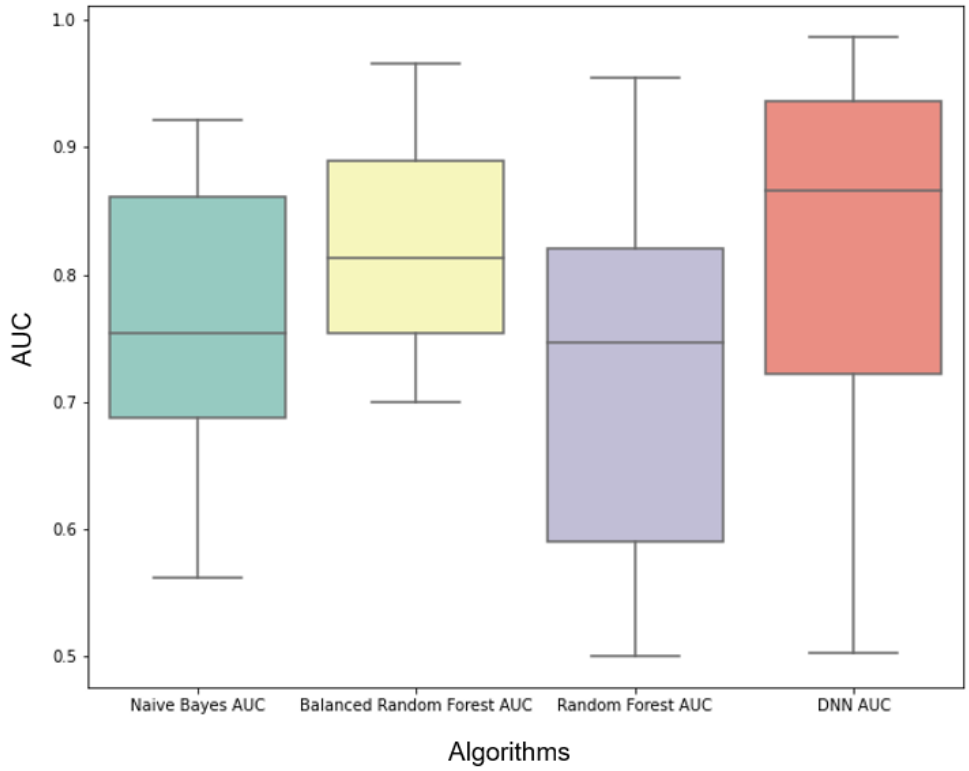


Figure 6.32: The box plots comparing the AUC among the four trained algorithms to generate SDM

Table 6.6: Average Rankings of the algorithms

Algorithm	Ranking
Naive Bayes	3.237
Balanced Random Forest	1.772
Random Forest	3.574
DNN	1.415

than standard ML algorithms, which is why when we trained DNN on species with larger presence points, the AUC was significantly higher than the other algorithms in particular NB and default RF. However, as we discussed, because the performance of Balanced-RF is more consistent across all species, this is the algorithm that we chose to validate the location of new observations. In addition, the training time for Balanced-RF, is lower than DNN, and it does not require high computing power compared to DNN.

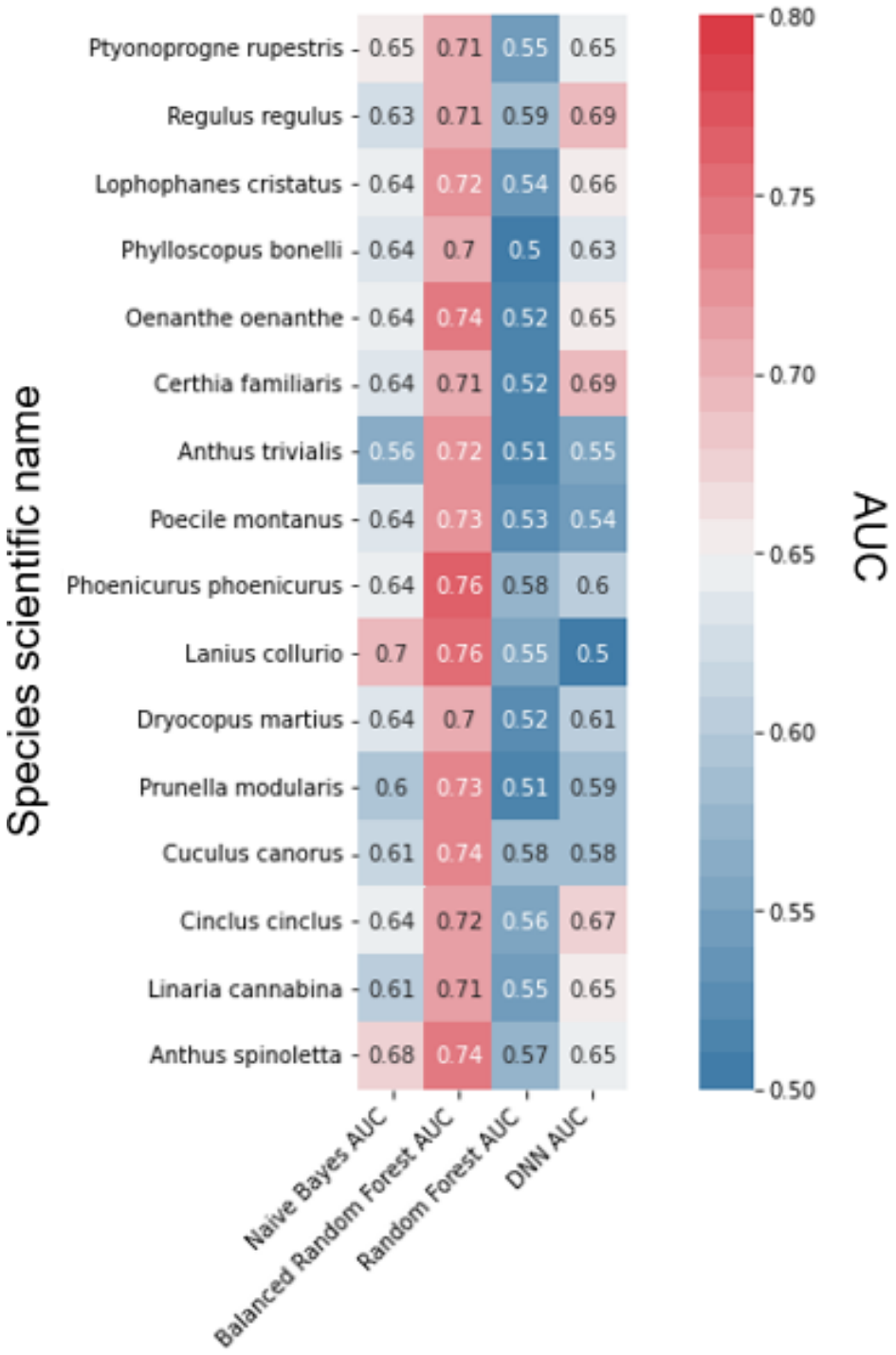


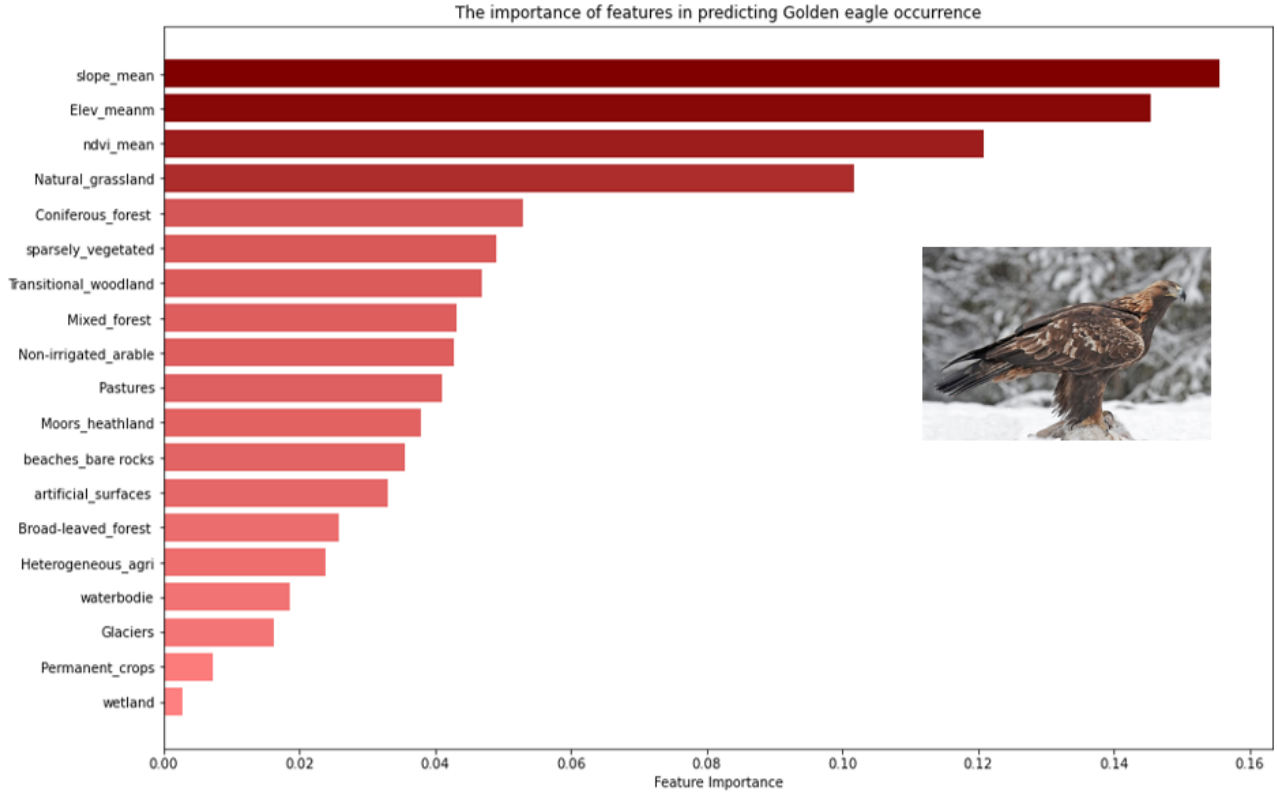
Figure 6.33: Comparison of the algorithms for the species where NB, RF, and DNN have AUC below 70%, Balanced-RF performs better for such species

After deciding on Balanced-RF, we assessed which environmental variables had the greatest influence on the model's ability to predict species occurrences. In other words, we ranked the importance of the environmental variables; the higher the importance, the greater the impact of the variable on model predictions. There are various indices for computing feature importance,

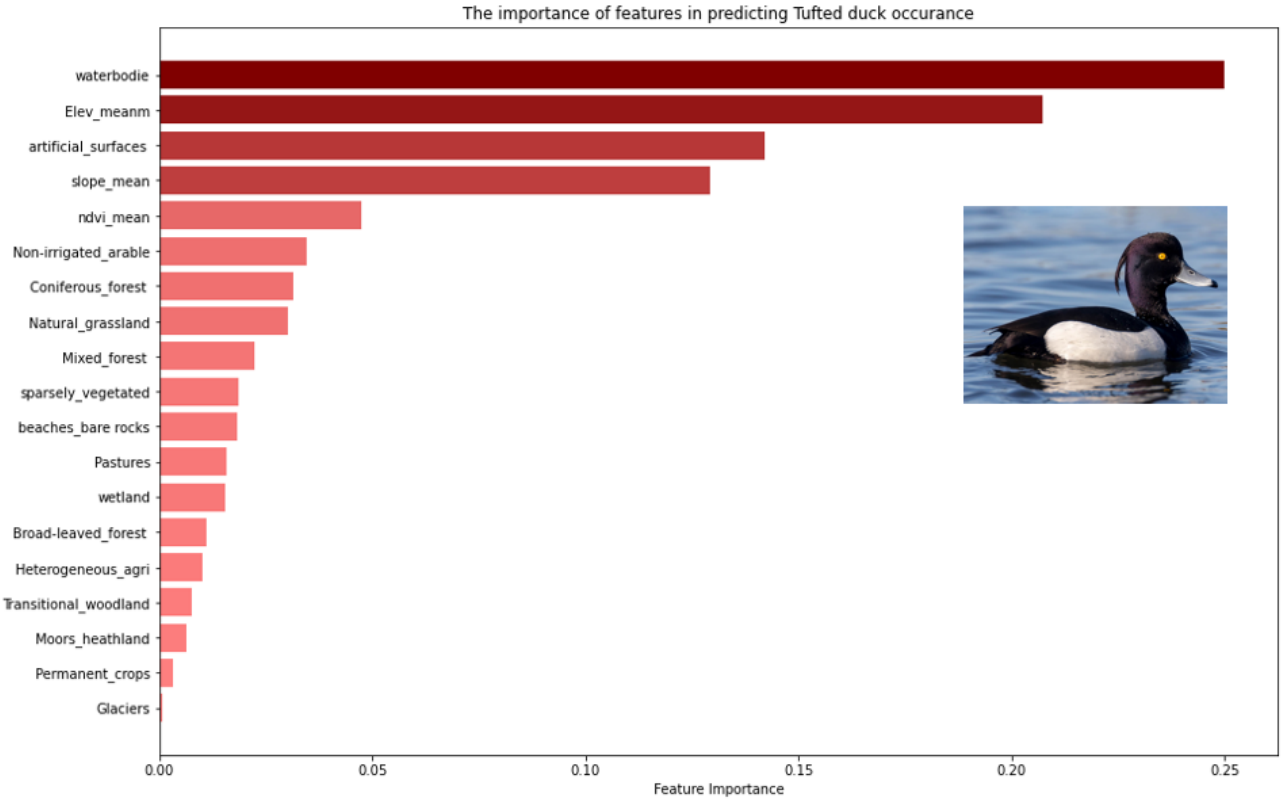
one of which is Gini-based importance, which is well-known in RF algorithms (Han et al., 2016). Gini Index, also known as Gini impurity, quantifies the likelihood of a specific feature being incorrectly classified when chosen at random. The average decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of nodes and leaves in RF trees. The greater the value of the mean decrease Gini score, the more important the variable in the model. We used the *sklearn* library to extract the variable importance from the Balanced-RF models using the *feature_importances_*[21] attribute. Figure 6.34 illustrates the ranking of the importance of environmental variables for two species of *Tufted duck* and *Golden eagle*. As expected, the importance of variables varies by species; for example, in figure 6.34, the top three variables for *Golden eagle* are *average slope, average elevation*, and *average NDVI*, whereas the top three variables for *Tufted duck* are *water bodies, average elevation*, and *artificial surfaces*. However, average elevation was an important variable in predicting the distribution of all the species, which corresponds with the available literature on elevation-based distribution patterns of bird species (Ding et al., 2021; Quintero & Jetz, 2018). Figure 6.35 illustrates the average importance of all the environmental variables for all species.

Finally, for each species, we obtained two output distributions maps: a binary classification, and a map of probability of occurrence of the species over the whole Switzerland. Figures 6.36 and 6.37 illustrates the maps of binary classification and probability of occurrence of Golden eagle and Common kingfisher species respectively. It can be illustrated from the figures that Golden eagle can be mainly observed in the alpine areas, whereas Common kingfisher can be mainly observed in the northern, and north west parts of Switzerland with high probability to be observed near lakes and water bodies.

[21] https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html



(a) Ranking of the importance of environmental variables for Golden Eagle distribution model



(b) Ranking of the importance of environmental variables for Tufted duck distribution model

Figure 6.34: Variable importance derived from Balanced-RF for two species of a) Golden eagle, and b) Tufted duck

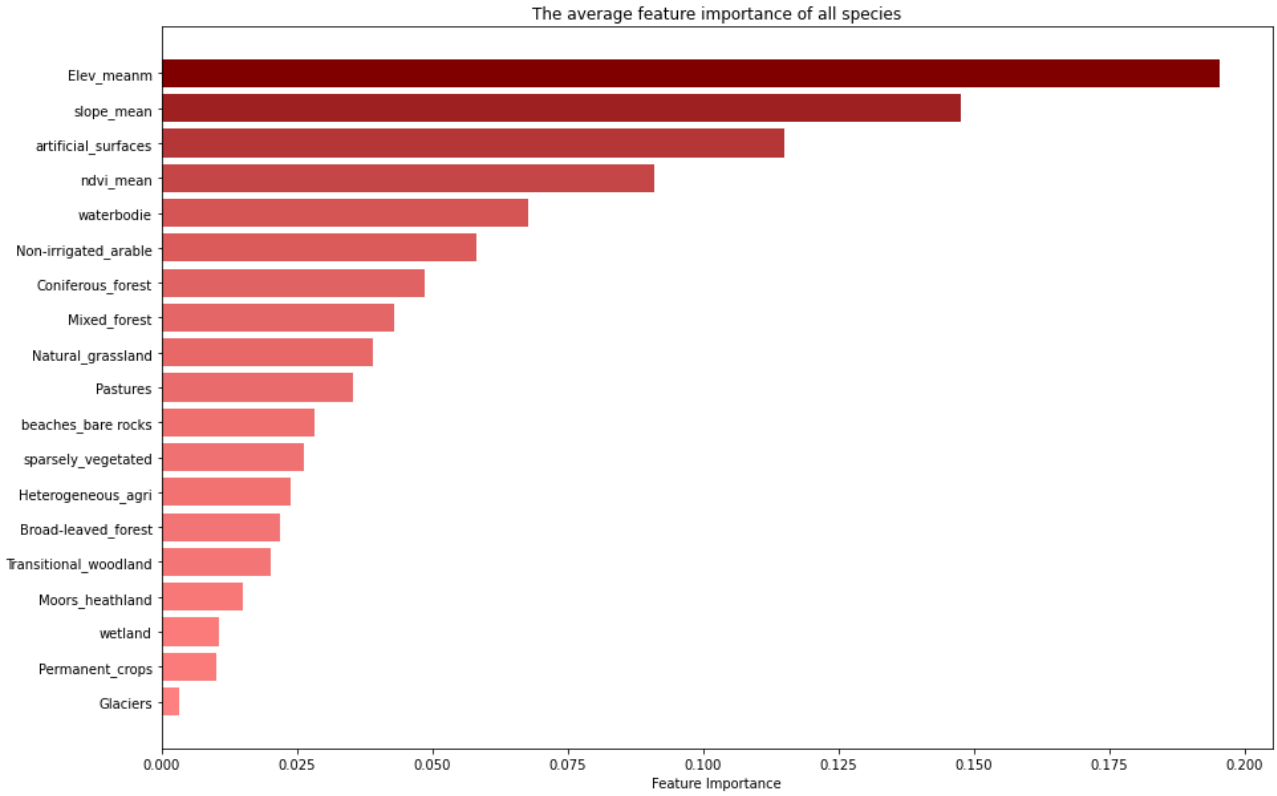
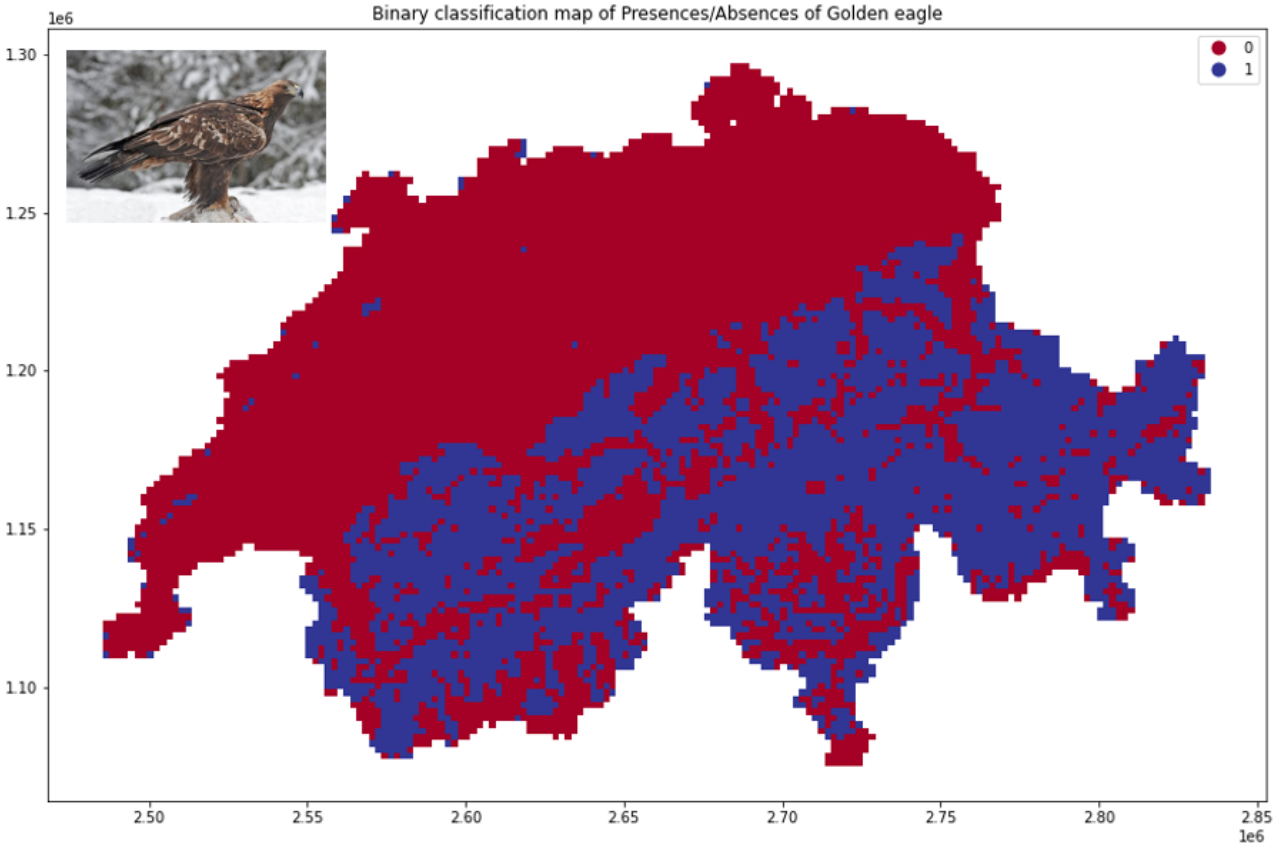
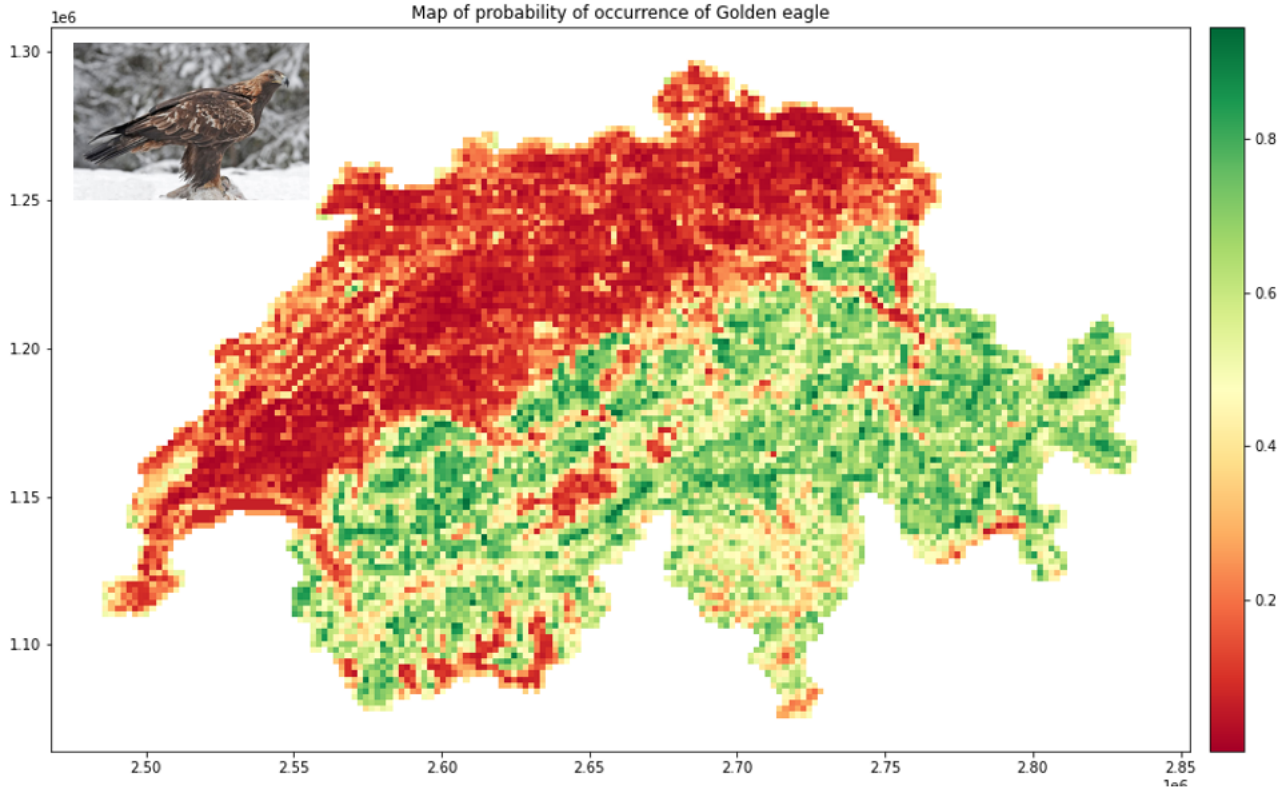


Figure 6.35: Average environmental variable importance derived from Balanced-RF for all species

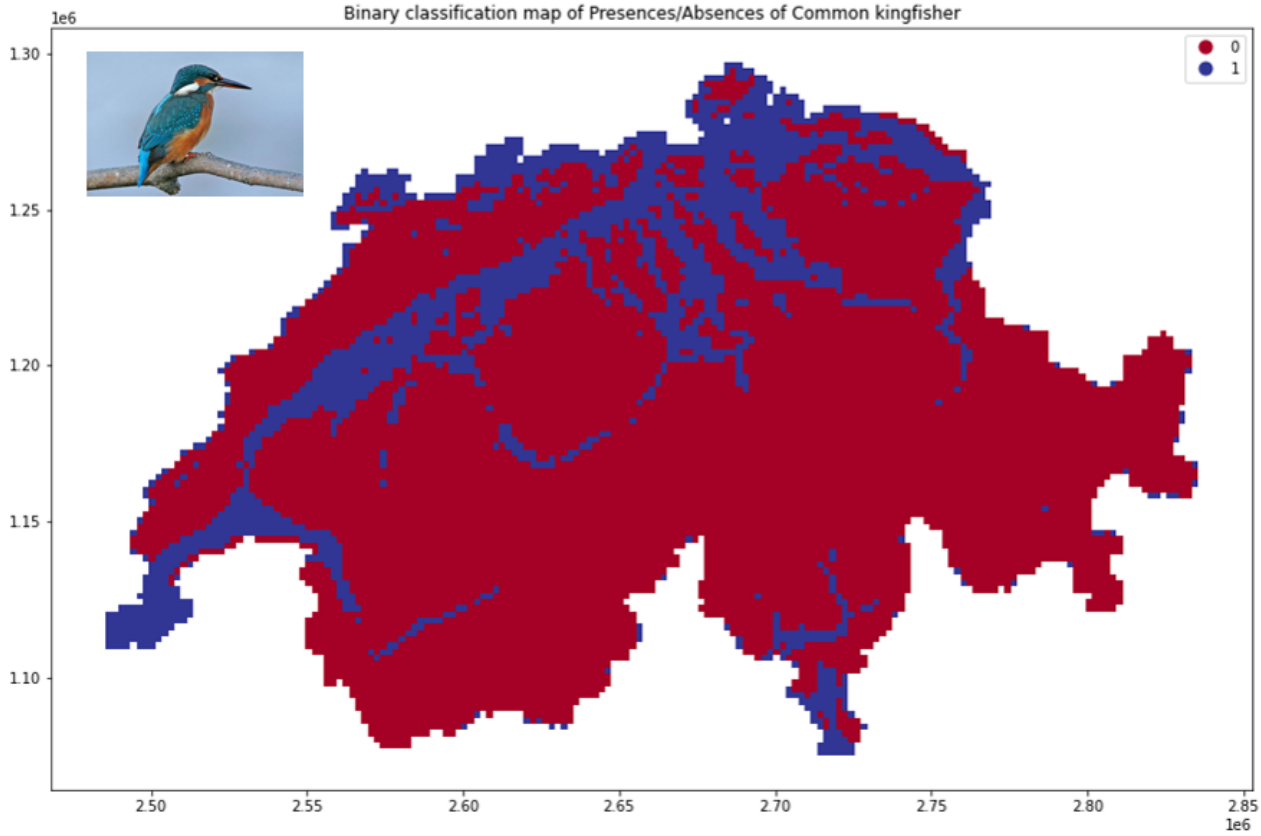


(a) Binary classification of Golden eagle

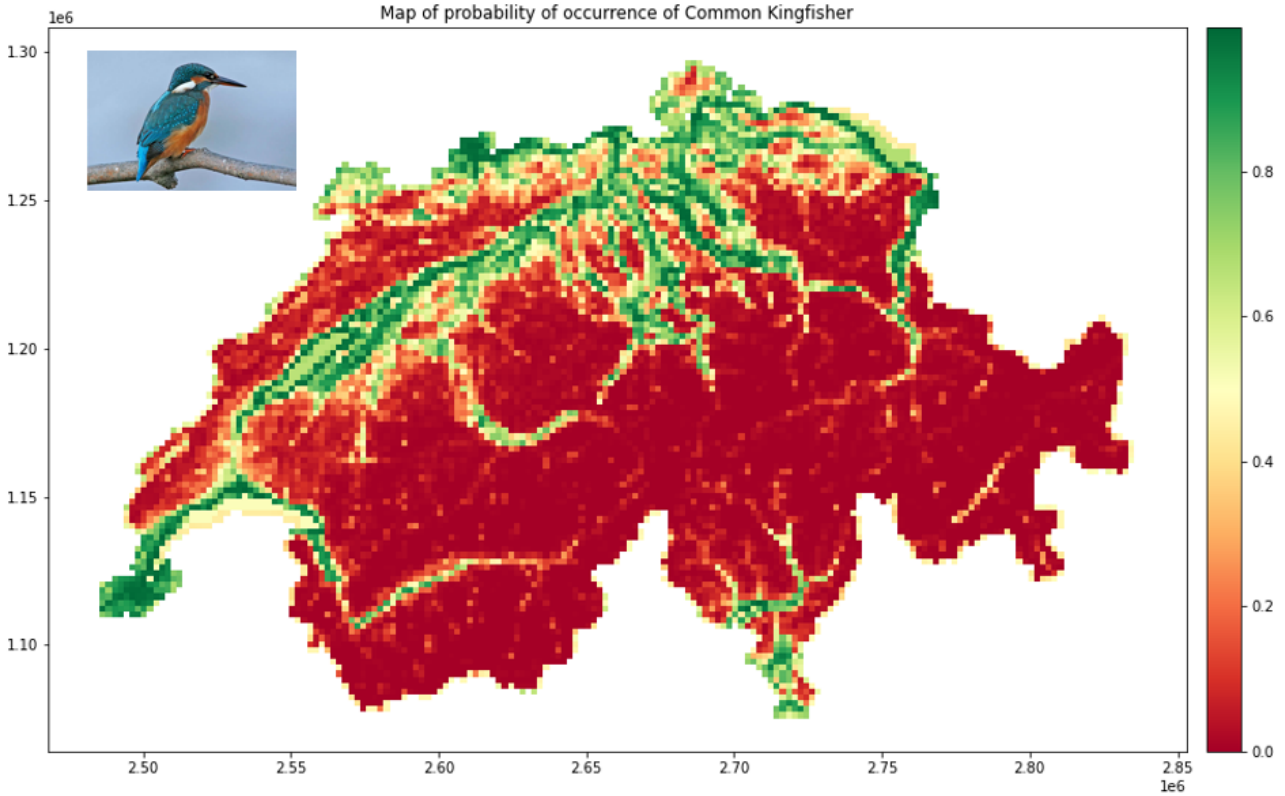


(b) Classification of probability of occurrence of Golden eagle

Figure 6.36: Maps of binary classification (a), and probability of occurrence (b) of Golden eagle within Switzerland



(a) Binary classification of Common kingfisher



(b) Classification of probability of occurrence of Common kingfisher

Figure 6.37: Maps of binary classification (a), and probability of occurrence (b) of Common kingfisher within Switzerland

6.5.4 Implementation of the API for location validation

After evaluating and comparing the algorithms, the models trained on each species dataset using the selected algorithm, in this case Balanced-RF, are saved in our server in a specific format(`.pkl`) using the Python module called *Pickle*[22]. Pickle is a standard Python object serialization method that is used to serialize ML models and save the serialized files in order to simplify reusing pre-trained models or sharing models trained on one machine with others. We loaded all 101 saved models, and according to our calculations, loading all models together required at least 32GB of RAM, so we requested a server with 64GB of RAM. As such, in order to use the models to validate new observations in the BioSenCS application, we developed an API that aims to validate new observations while also providing user-centered suggestions on the top-five high-probability species that can be observed near the user's location. To implement the API we used Flask[23], which is a micro web framework written in Python. Figure 6.38 illustrates the architecture of our API called BioLocation to validate location of new bird observations. The API includes an endpoint for obtaining species names, an endpoint for predicting species probability of occurrence, and an endpoint for suggestion, all of which are explained further below.

Validation: The overview of the location validation process is presented in figure 6.39, and details of the steps are as follows:

1) The participant selects the lactation of observation and adds species name. The location and species name are then passed as the parameters of the BioLocation API with the *predict* endpoint.

2) A neighborhood of $2km^2$ is extracted around the added location.

3) The environmental variables in that neighborhood (proportion of land cover classes, average elevation, average slope, and average NDVI) are computed and a JSON file is created (See figure 6.40).

[22]https://docs.python.org/3/library/pickle.html
[23]https://flask.palletsprojects.com/en/2.0.x/

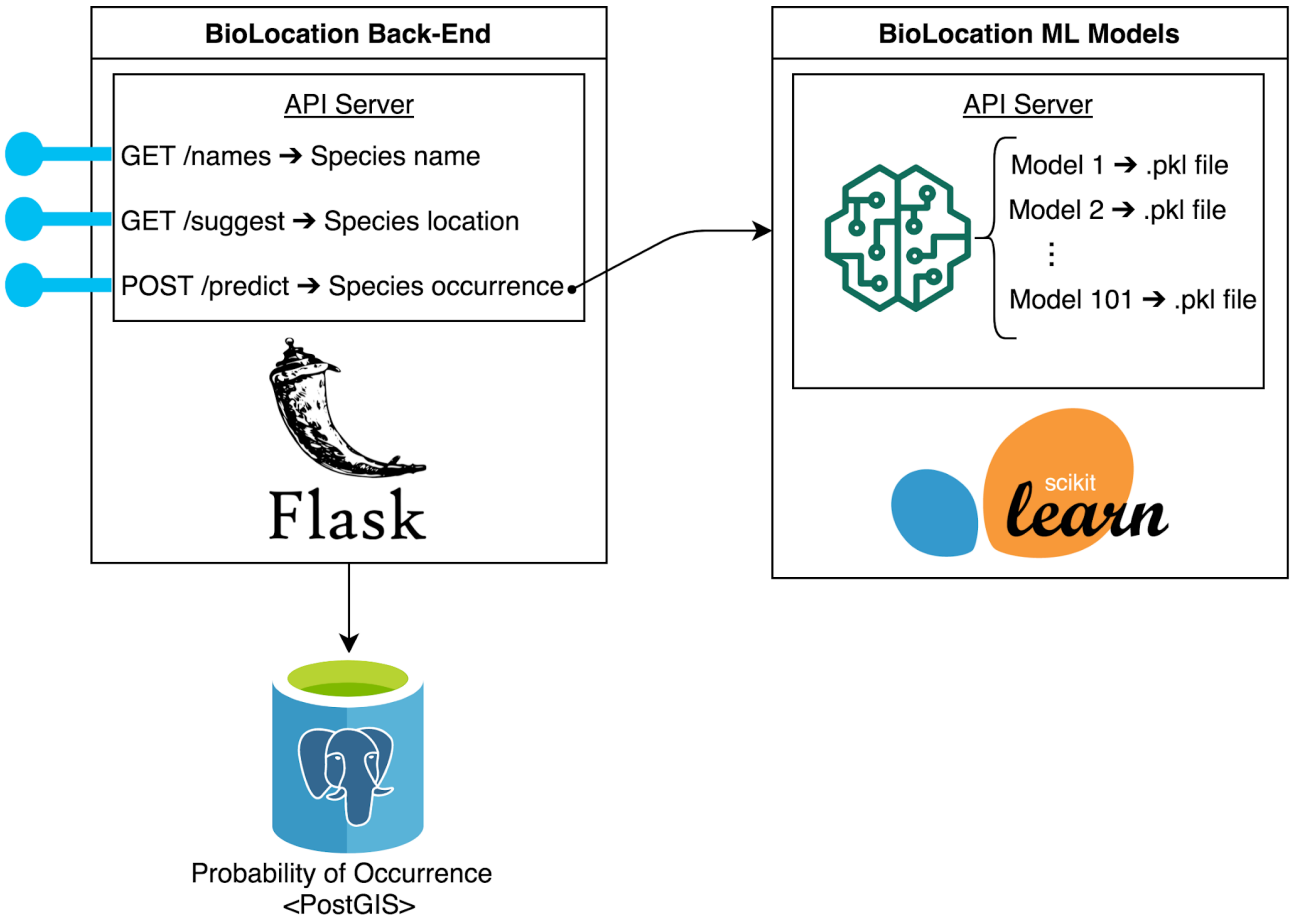


Figure 6.38: The architecture of BioLocation API

4) Based on the species name added by the participant, the environmental variables are passed to the loaded SDM model for that species.

5) The model takes the environmental variables and predicts the presence or absence of the species as well as the likelihood of observing the species in the defined neighborhood.

6) The prediction of the model is then used to validate the observation and provide the participant with real-time feedback on the predicted probability of occurrence as well as information about the species' habitat characteristics.

The generated feedback is intended to either simply provide additional information to the participant if the probability of occurrence of species in the added location is higher that 50 percent, or to propose to the participant to confirm the validity of an observation if the probability of occurrence is less than 50 percent (See figure 6.41) and in this case to flag the observation in BioSenCS database in a Boolean attribute named `FlagLocation`. Once the

participant receives the feedback, he/she decides whether or not to alter the observation.

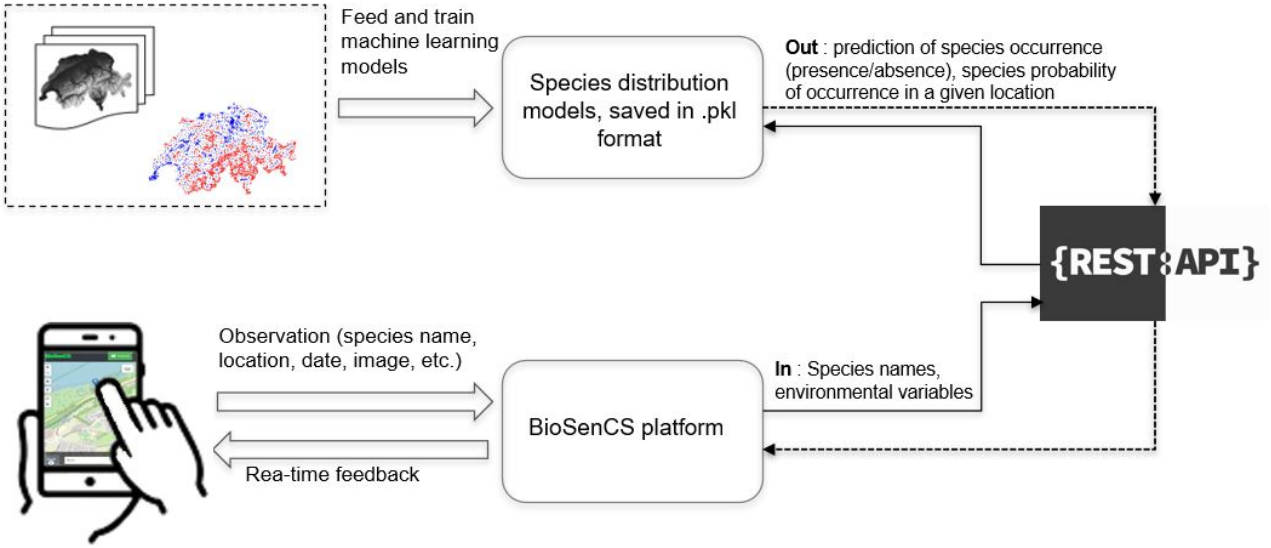


Figure 6.39: The process of automatic location validation and real-time feedback generation

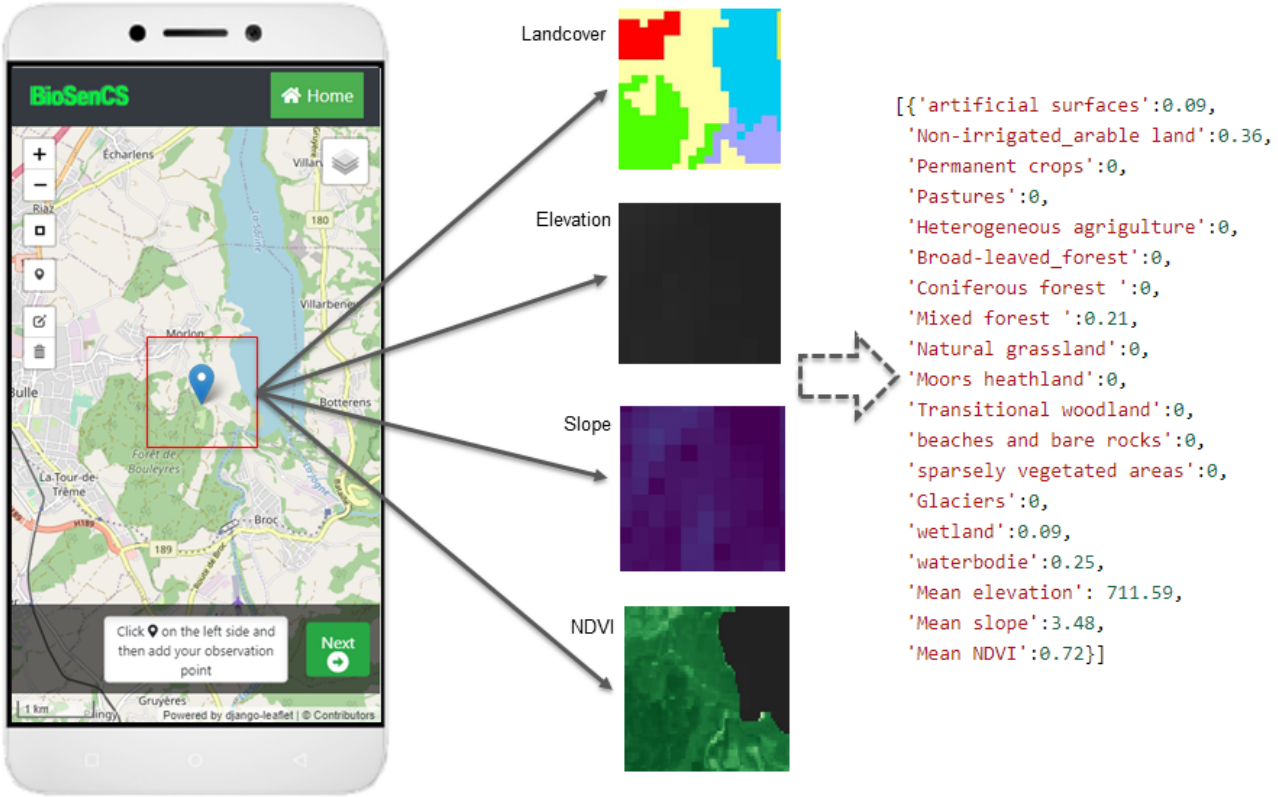


Figure 6.40: Extraction of environmental variables in a neighbourhood of 1km around the location of observation added by the participant

Suggestion: The API also provides suggestions to participants based on their location, such as possible species that can be observed within a 1km radius of the participant's location. To

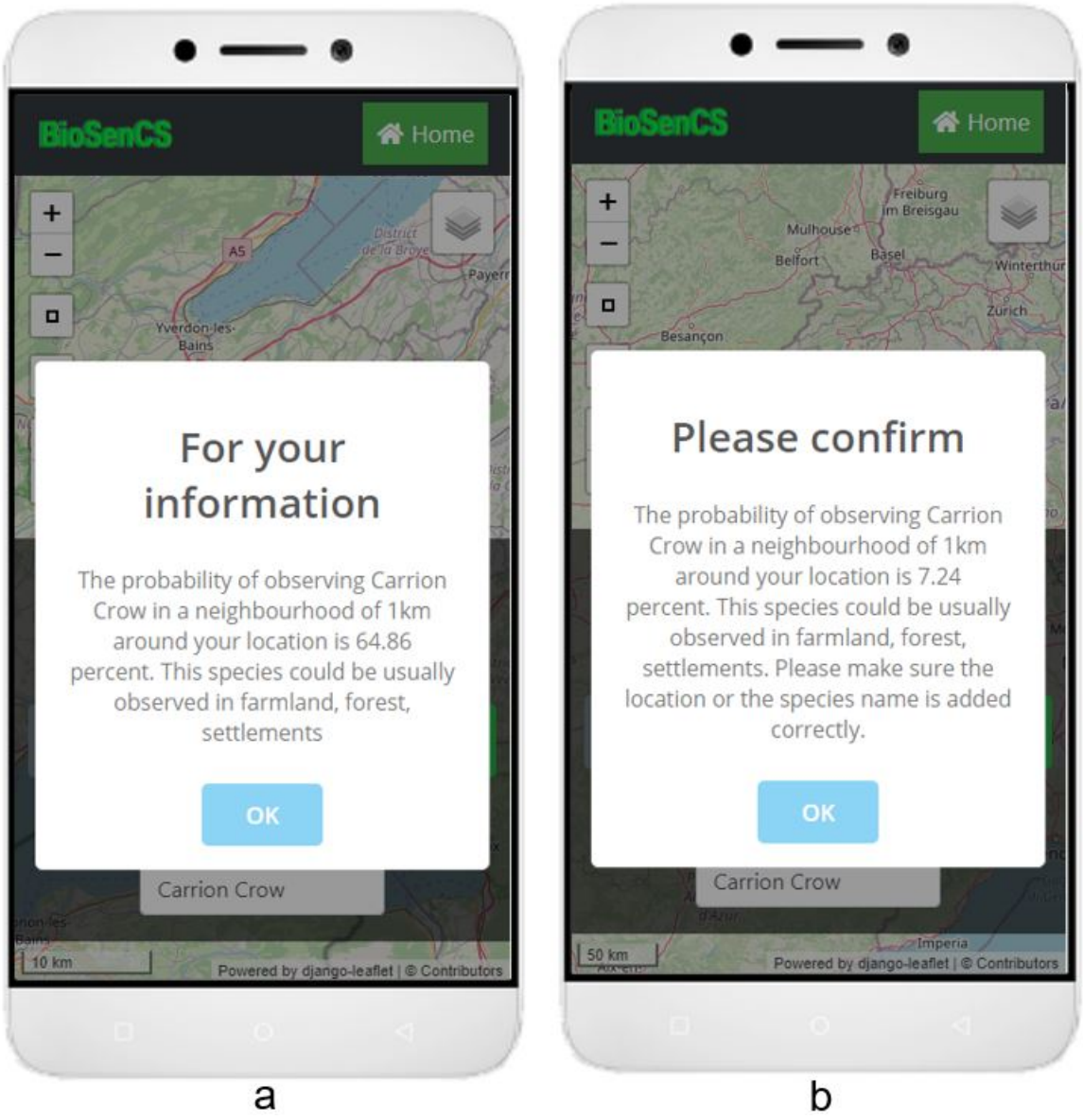


Figure 6.41: Location feedback if probability of occurrence of species is higher (a) and lower (b) than 50%

accomplish this, the predicted probabilities of occurrences of all 101 species across Switzerland are saved in the BioLocation database with a resolution of $2km^2$. Thus, whenever the participant's location is passed to the API's `suggest` endpoint, the top five species with the highest probability in that neighbourhood are queried from the database, and the results are passed to the participant as a list of species names (See figure 6.42 b), with clicking on each name opening the species' information in VogelWarte website[24]. After receiving the suggestion, the participant can either search for suggested species in that neighbourhood or see if any of the

[24]https://www.vogelwarte.ch/

suggestions include the species they have observed so that they can obtain the species' name from the suggestion list.

Name: When submitting a bird observation, the participant has three options: either the participant does not know the name of the species, the participant is unsure and checks the suggestion list for the name of the species, or the participant knows the names and writes the name in a textbox with an auto-complete text function that includes species common names in English taken from the BioLocation database. Figure 6.42 illustrates the three options to add the name of the bird species.

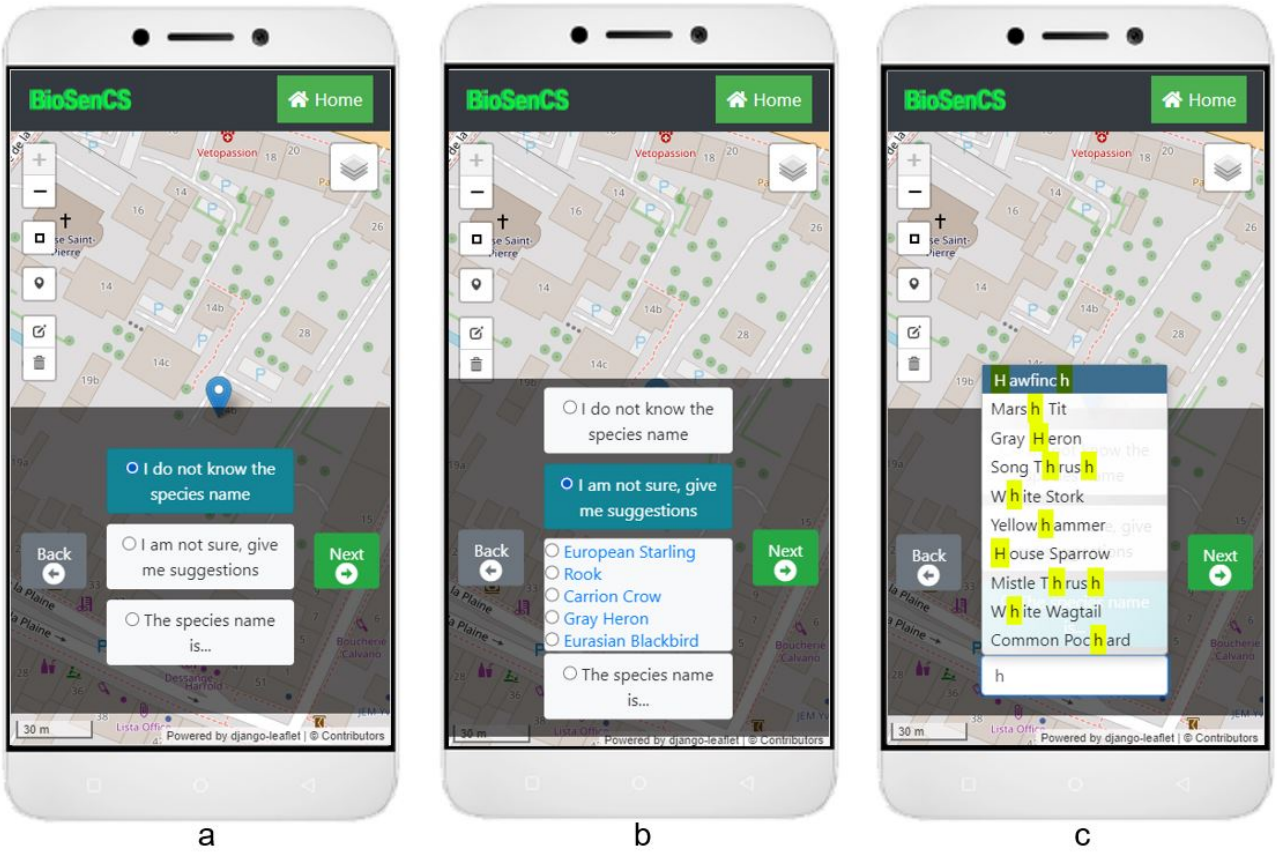


Figure 6.42: Adding bird species name in BioSenCS application: User does not know the species name (a), participant ask for suggestions to add the species name (b), and participant knows the species name and type the name and an auto-complete of bird names is offered (c)

6.6 User experiment

Finally, we tested the BioSenCS application within a three weeks period to collect user feedback about the application interface and to explore the view of the participants regarding receiving automatic feedback. We advertised for the application through direct e-mails to students and colleagues of our university, through social networks (Facebook, and LinkedIn), and simply through word of mouth. During this three weeks, 224 users visited the application, and figure 6.43 illustrates the frequency of the number of visitors (The peaks represent each time we shared the application on social media.).

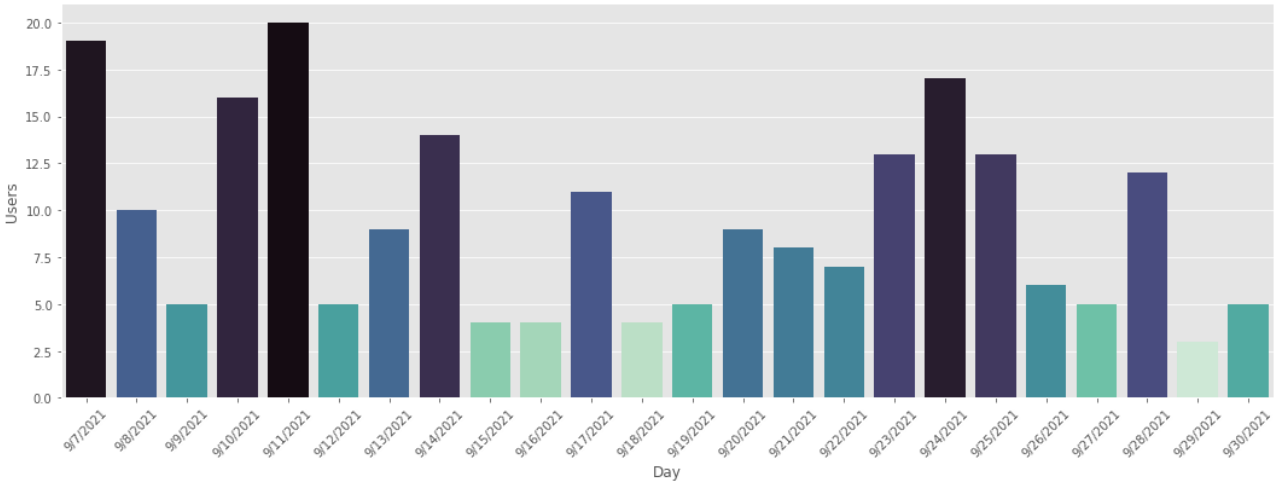


Figure 6.43: Number of users who visited BioSenCS application through the three weeks experiment

Among the 224 users who visited BioSenCS application, only 38 users created a user account, and only 14 out of these 38 users contributed to the project. In addition, among the 14 contributors, three of them were very active each collecting at least 40 observations, four participants were contributing from time to time between 10 to 20 total observations each, and the rest of the contributors contributed mostly only one day or maximum two days during the experiment period with less than 10 observations each. As mentioned earlier in Chapter 2, this participation pattern is very known in VGI and CS projects with participation pattern to OpenStreetMap being among one of the most known examples (See (Wood, 2014) for *The Long Tail of OpenStreetMap*). Figure 6.44 illustrates the pattern of participation in BioSenCS application during

the three weeks of user testing. One reason for the small number of participants was that we had a limited amount of time to first advertise the application (about two weeks) and then test it (three weeks), and time constraints can be a barrier to participation. To build a large community of participants in CS projects, time is important, but so are the methods of advertising the project, some of which are more traditional, such as word of mouth, newsletter publications, social media posts, or, as we proposed in Chapter 3, a new approach would be to use ML to reach out to the public based on their interest and recommend the appropriate project. In future work, we will use other approaches to reach a larger group of participants, but for the purposes of this chapter's findings and discussion, we focused on the 14 participants. In addition to the number of participants, during the user testing period 230 observations were collected (Figure 6.45), with 160 of them being birds, 36 flowers, 19 trees, and 15 butterflies.

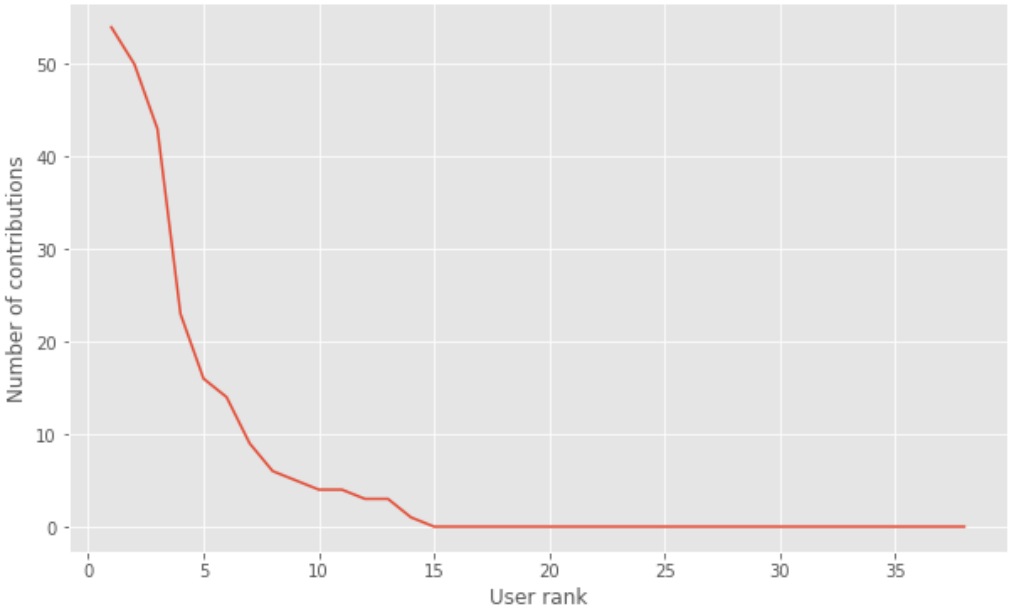


Figure 6.44: The contributions of BioSenCS' participants during three weeks of user testing

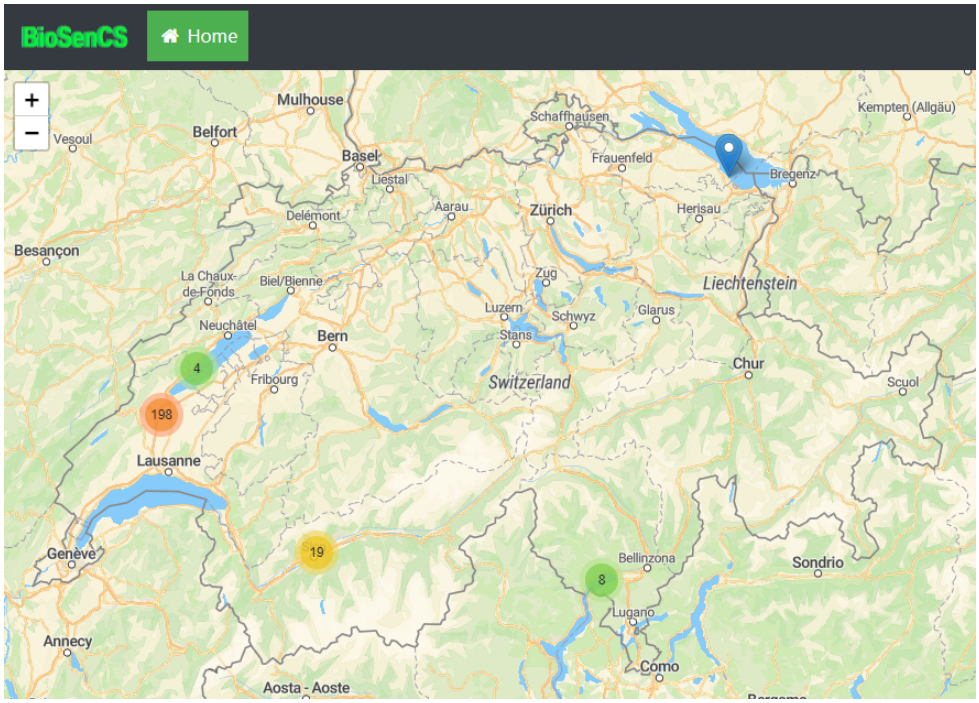


Figure 6.45: Map of all observations collected in Switzerland

After the testing period, a questionnaire was sent to the participants to obtain first some general information about their past experience contributing to VGI and CS projects and their views on data sharing in such projects, and second to obtain feedback regarding their experience with the BioSenCS application in terms of their frequency and number of contributions to the app, their views on user friendliness of the interface, the real-time feedback, their motivations to contribute, and finally their feedback on how to improve the application. The questionnaire was created with the Sphinx software[25] and included a variety of question types, including Likert scale, free text, ranking, and multiple choice questions (link to the questionnaire in the footnote[26]). Among the contributors, 10 people answered to the questionnaire, while a small sample for statistically testing the validity of the answers, provided us with the necessary information to initially understand how the feedback was affecting the participants' motivation, and what the participants needed to be improved at this version of the application. The survey was sent only to those who had made at least one contribution, not to all registered participants. In the future, other approaches, such as conducting interviews, organizing workshops or focus groups, and so on, could be used in addition to the questionnaire to elicit more feedback from

[25]https://en.lesphinx-developpement.fr/sphinx-logiciels-2/sphinx-declic/
[26]https://enquete.heig-vd.ch/SurveyServer/s/INSIT/BioSentiers-CS/questionnaire.htm

participants.

View on VGI and CS projects: We asked respondents to name the CS/VGI projects they have heard of, and they all named OpenStreetMap, with some mentioning other projects like eBird and Galaxy Zoo. Three of the respondents had never contributed to CS/VGI projects, while others had contributed multiple times. Two of the contributors stated that they primarily contribute to CS/VGI projects to fill their spare time, while the remaining respondents stated that they would like to learn new things that are interesting to them but are outside their field of expertise. We asked the contributors about their thoughts on data sharing in such projects; two said they would share the data as long as it remains anonymous, two said they would not share sensitive information such as the location of endangered species, and the rest said they would share the data because that is the main goal of CS projects: to contribute data to help science progress. Finally, we asked if they would contribute data to such projects if they were unsure about the quality of their contributions. Two of the participants stated that they would contribute the data even if they were unsure and leave the validation to the experts, but all of the other respondents stated that they would share the data but would prefer to include a note/label indicating their level of confidence in the accuracy of their contributions.

Views on BioSenCS application: The first two questions concerned the frequency with which the participants contributed and the number of contributions they made over the course of three weeks. The goal was to see if there was a difference in opinion about the app and automatic feedback between people who contributed more and those who contributed more casually over the course of three weeks. We then asked them about their thoughts on the user interface, particularly how easy it was to add an observation, whether the app's language of being only in English made it difficult for them to contribute, and whether the fact that the auto-complete text field for the names of the species was only offered as the species common names in English was a demotivating factor for them to contribute.

We created a 5-point Likert scale for each of the mentioned questions, as well as a text box for each question if the respondents had a more detailed comment. The average scores for user interface (1: extremely difficult to add observation, 5: extremely easy to add observation),

application language (1: extremely difficult to use the app in English, 5: extremely easy to use the app in English), and species names (1: extremely demotivating to have only the species common names in English, 5: not at all demotivating) were 3.56, 3.44, and 3.33 respectively (6.46). Two participants who gave the lowest score to the user interface added the comment that the number of steps before one could submit an observation were a lot, and thus tiring. Regarding the application language, two of the respondents mentioned that although they did not have problem with the language of the app being in English, they think that if other languages or at least the local language was offered, the app could attract more people. We completely agree with this feedback, and we believe that if the application was available at least also in French, we could have more participants. Regarding the species names, there were various comments. Two respondents mentioned that it was interesting to learn the species common names in English, one participant proposed us to include a website in the app which helps in translating the species names from French to English such as *Oiseaux.net*[27], and two other respondents mentioned that they prefer that only the species scientific names (Latin names) be offered.

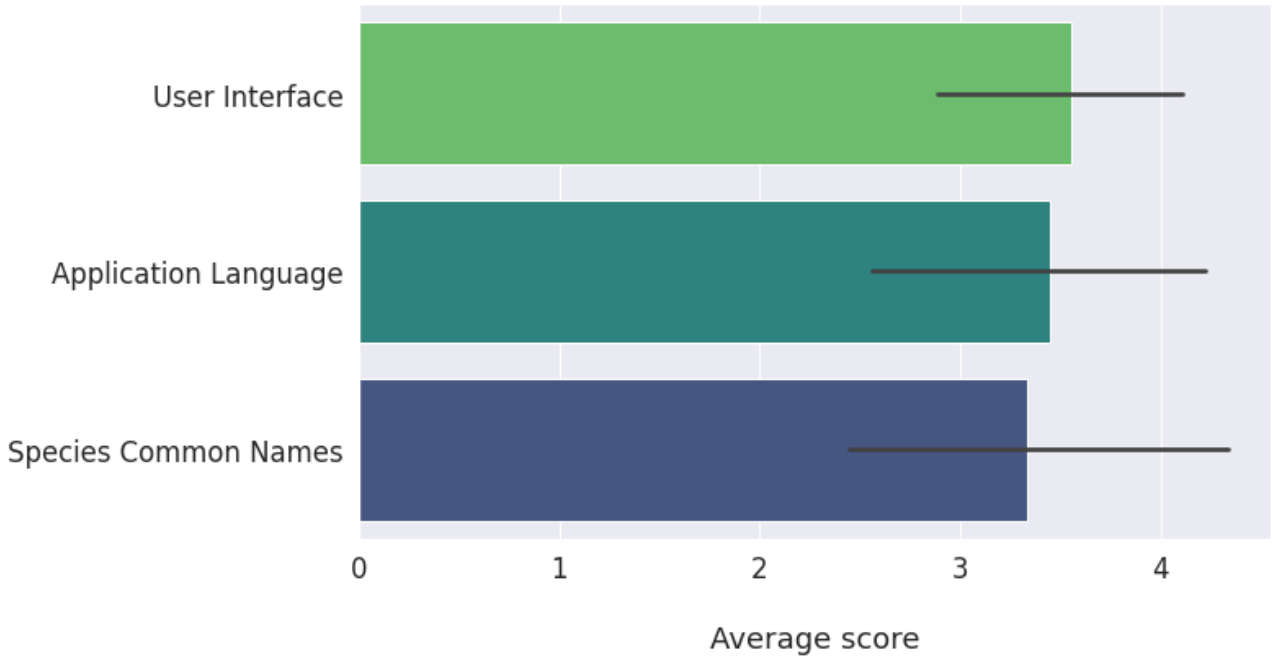


Figure 6.46: Average scores given to the three questions regarding user interface, application language, and species common names proposed in English

[27]https://www.oiseaux.net/oiseaux/france.html

The following question we posed to our participants was about their thoughts regarding receiving automatic feedback. We asked to what extent they found the information in the feedback useful, and whether receiving feedback increased their motivation to contribute to the project. The questions were also asked on a 5-point Likert scale, with the average score for the usefulness of the feedback (1: not at all useful, 5: very useful) and the role of feedback in increasing motivation (1: not at all motivating, 5: very motivating) being 3.33 and 3.5, respectively (See Figure 6.47). Furthermore, because the species common name in English must be passed to the BioLocation API in order to receive location feedback, three respondents stated that they were attempting to add the names in English in order to receive feedback on the likelihood of observing that species. Figure 6.48 depicts the top 20 observed bird species, and it is indicated that, with the exception of the cases where the participants did not know the name of the species (unknown), and one species where the common name in French was added (Loriot d'Europe), the rest of the cases the species names were added in English, indicating that the participants were interested in receiving location feedback.

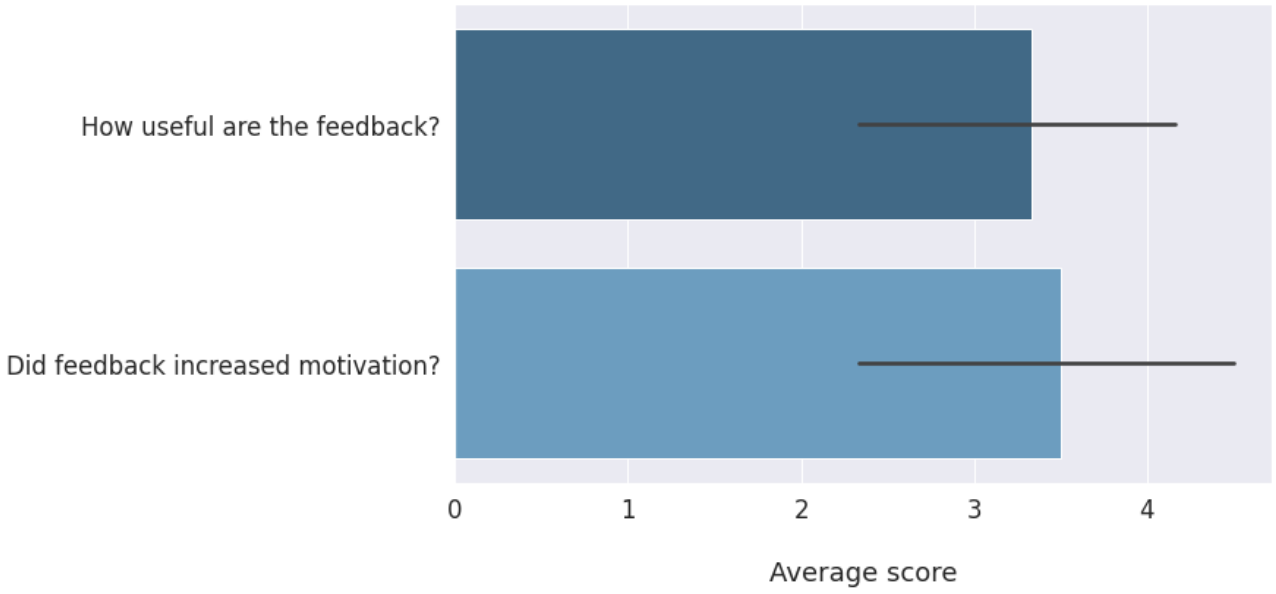


Figure 6.47: Average scores given to two questions regarding receiving automatic feedback. minimum score 1, and maximum score 5

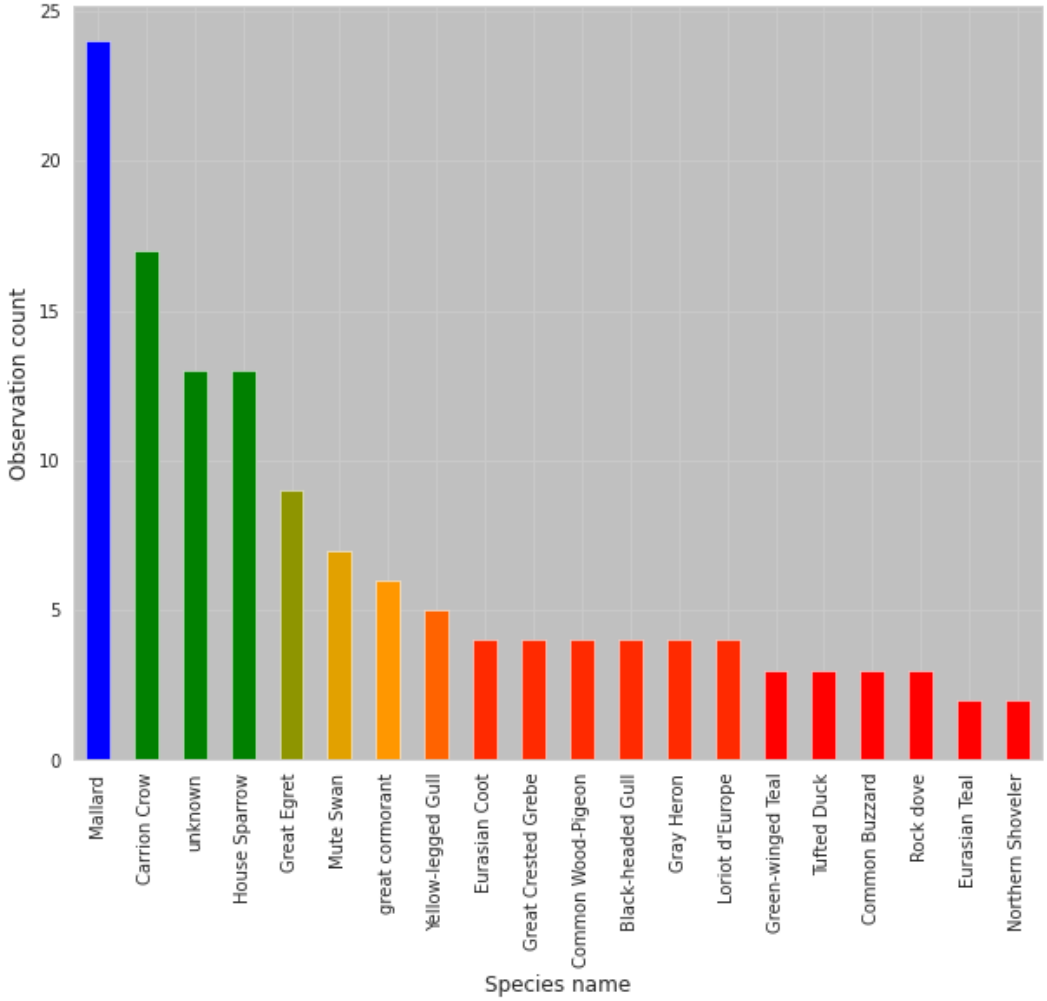


Figure 6.48: Top 20 bird species observed during the three-week user experiment

Moreover, when we asked if the feedback was useful in learning about the distribution of species, two of the three participants who responded said that it did not help them learn about the distribution of species but rather about the characteristics of species habitat. Besides that, we investigated whether the frequency with which the application is used is related to the scores assigned to the two questions about automatic feedback. We assigned scores to the frequency of app use: only once during the three weeks: 1, once a week on average during the three weeks: 2, and twice a week on average during the three weeks: 3. To explore the correlation, we ran a Pearson test using *SciPy*[28], a free and open source library in Python. The correlation coefficients between the frequency of app use and the usefulness of feedback and the role of feedback in increasing motivation were **0.79** and **0.49**, respectively, indicating a strong

[28]https://www.scipy.org/

correlation. However, due to the small sample size, the test was not statistically significant, with p-values of 0.059 for the correlation between frequency of use and usefulness of feedback and 0.34 for the correlation between frequency of use and role of feedback in increasing motivation.

Finally, similar to BioPocket survey in Chapter 5, we asked the respondents about their motivations to contribute to this project. Similar to BioPocket project, and as illustrated in the motivation framework for the classic CS projects (Chapter 5, Section 5.4), the top three motivations in this project were *contribution to science*, *learning about biodiversity*, and *helping the biodiversity*. Likewise, *ego enhancement* and *receiving awards* were the two weakest motivations. Figure 6.49 illustrates the ranking of the motivations based on the average score given to each motivation.

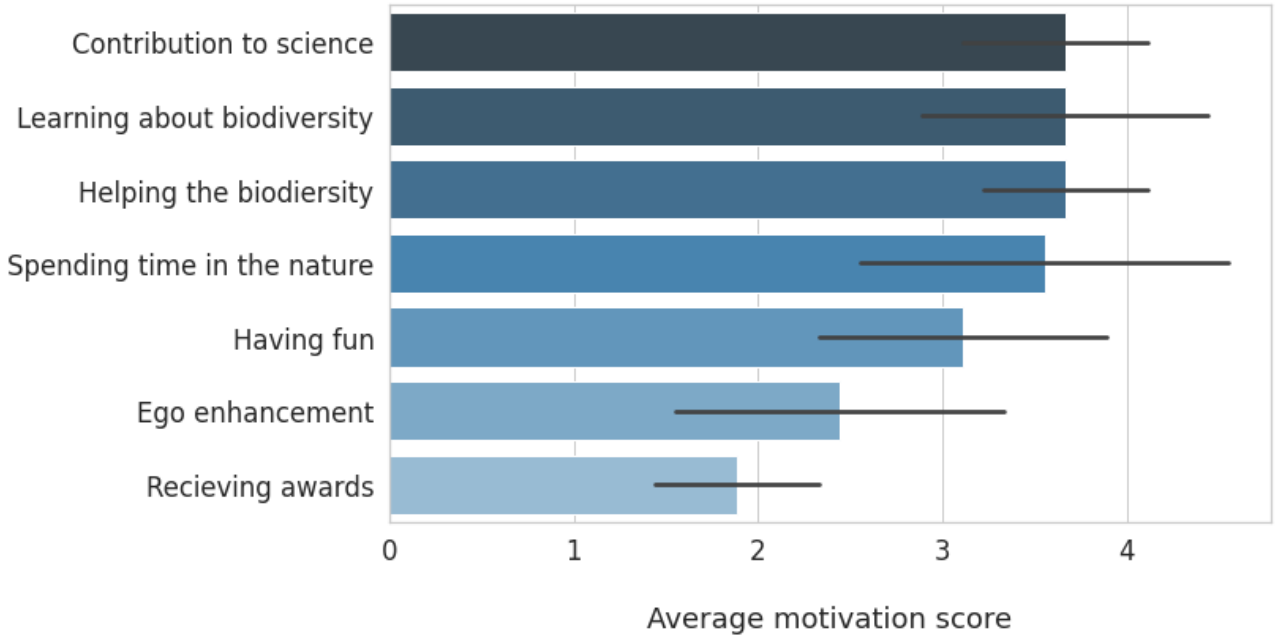


Figure 6.49: Average scores given to each motivation, 1 indicates very weak motivation, and 5 indicates very strong motivation

6.7 Discussion and evaluation of hypotheses 4 to 6

After testing the application with participants and automatically validating bird observations using ML algorithms, it is time to discuss and evaluate the hypotheses presented in Chapter 4. Some of the hypotheses were discussed in Chapter 5, and here we discuss hypotheses 4, 5,

and 6 concerning the impact of automatic feedback on public engagement and the impact of automatic data validation on data quality. As such, based on the analysis and results obtained in this chapter, we go over each hypothesis and discuss it.

H4: Giving real-time feedback to the participants can result in increasing engagement.

As previously stated, we asked volunteers whether they found the real-time feedback useful and whether the feedback increased their motivation to continue contributing to the project. Figure 6.47 illustrated that the average scores for each of these questions were greater than 3 out of a maximum 5 points. Furthermore, the findings indicated a strong correlation between the frequency at which the application was used and the scores assigned to the feedback questions. The correlation between frequency of contribution and feedback usefulness was 0.79, and the correlation between frequency of contribution and feedback impact on increasing motivation was 0.49. These correlations indicated that the participants who contributed more frequently during the three-weeks experiment period found the feedback to be more useful and even a motivator to contribute more. However, as previously stated, the obtained correlations were not statistically significant due to the small sample size, with p-values of 0.059 for the correlation between frequency of use and usefulness of feedback and 0.34 for the correlation between frequency of use and role of feedback in increasing motivation. As a consequence, the results do not allow us to statistically verify the hypothesis. However, based on the strong correlations, we can conclude that the impact of feedback on public engagement was observed positively for the small group of our participants, and we can preliminary verify the hypothesis based on the observed correlations, with the condition that the application be tested with a larger group of participants.

H5: Automatic data filtering can simplify and accelerate data validation task, and improve data quality.

As previously stated, we obtained 230 observations over a three-weeks period of user experiment. Given the short testing period and the small number of participants, this number of observations collected for a new platform is reasonable. Given a larger number of participants and a longer contribution period, this number is expected to be quite large, and validating

such a large amount of data through expert review can be time consuming. Among the 230 observations, 24 were flagged to be verified by experts, with image flags accounting for 16 of the 24 observations, followed by location and time flags with 5 and 3 observations, respectively. The image filters were done with Clarifai, as mentioned in section 6.4, and the images were flagged if the probability of predicting a tag of species type in the image was less than 85%. In most cases, where the species was very small or not very clear in the entire photo, the photo was flagged (false negatives), and some of these cases are shown in the figure 6.50. All of the non-flagged observations successfully passed the validation filters and are considered valid without further expert review. Although it is difficult to statistically support this hypothesis (with the data we have for our experimental test), the evidence suggests that by considering the observed number of flagged observations out of the total number of observations, the validation task can be simplified and less time-consuming for the experts because they only need to control the flagged observations. This is, of course, assuming that the models used for automatic validation have high performance and are retrained on a regular basis using new input training data or additional features (e.g., new environmental variables).

H6: Giving real-time feedback to volunteers increases their knowledge about scientific domain of the project (e.g., biodiversity) and results in collection of higher quality data.

We looked at two factors to evaluate this hypothesis. First, in terms of increasing biodiversity knowledge (through location feedback), we looked at how participants found the feedback useful in relation to their frequency of contribution. As previously stated, participants received location feedback whether an observation was flagged or not; if the observations did not successfully pass the automatic filter, the feedback besides giving the information (e.g., probability of of observing an species in the user location), requested the participants to verify their observation; otherwise, they received the feedback only for their information about that specific species. To support the first factor, as we discussed in Hypothesis 4, there was a strong correlation (though not statistically significant due to the small sample size) between usefulness of feedback and frequency of contribution. That is, active participants found the feedback to be a useful piece of information, particularly the location feedback, because participants who gave a higher score to feedback being useful were also the ones who gave a higher score to location feedback assisting

Figure 6.50: Examples of images of bird observations that were flagged. Meaning that the probability of the presence of bird tag in the image was less that 85%.

them in learning more about species habitat.

Second, we investigated whether there was any correlation between the number of flagged observations (O_F) and total number of contributed observations (O_T) per user to see if feedback (location, image, or time) was a factor in encouraging volunteers to collect higher quality data.

The correlation between the ratio of flagged observations to total number of observations (O_F/O_T) and the total number of observations (O_T) per user was examined. The correlation indicated a statistically significant negative correlation with a value of -0.63 and a p-value of 0.036, indicating that participants who contributed more had fewer flagged observations. This is to say that the participants either used the feedback to improve their observation before submitting it (for example, checking to see if the location pin was correctly added) or they learned to provide higher quality data (for example although the image filter in our platform was not at species level, but the feedback could help them to know that when the species is very

small in the photo the image is flagged, so they were trying to provide images where species is more clear in the image). Although we did not receive any clear evidence to statistically prove that there is a correlation between contribution over time and data quality, we can say that the above evidence suggests that the more contributions by a participant, the fewer the number of flagged observations, or in other words, higher quality data. As with other hypotheses, if the application is tested over a longer period of time, it will be easier to see how the feedback affected data quality over time.

6.8 Summary

As discussed in Section 2.4.2 of Chapter 2, the majority of CS projects rely on expert data validation to ensure the quality of data collected from participants. We discussed how, while experts play an important role in CS projects, relying on this method as the only data validation approach can be time consuming and complex, and can also demotivate volunteers from continuing their contribution because it takes a long time for the expert to validate the data and provide feedback to volunteers, which can be a demotivating factor for participants if they do not receive any updates regarding their contribution. As a result, other methods of data validation, such as model-based and automatic quality assessments, are proposed in the same section of Chapter 2 (Balázs et al., 2021). Furthermore, as we discussed in Chapter 3, combining CS and ML can be a way to address some of the challenges of CS, such as data validation and public engagement. One of our proposed approaches was to use data from existing CS projects to train ML algorithms and then save the trained models to validate new observations in the same or other CS projects (in the same or similar field of science).

Furthermore, another approach that we discussed using a combination of CS and ML to increase participation in CS projects is to provide participants with automatic machine generated feedback. Besides that, in the motivational framework we developed in Chapter 5, it was demonstrated that *interest in learning about science* was a strong motivation in almost all projects, for example, in classic CS projects, *interest in learning about environment* was among the pri-

mary motivations, which was also stated as a strong motivation in our BioPocket survey (See section 5.5). As a result, machine generated feedback can be used to not only keep volunteers involved in the project, but also to provide useful information to volunteers and assist them in learning about science (depending on the project field, e.g. learning about biodiversity, or learning about galaxies).

Taking into account all of the aforementioned elements, we implemented a CS project in this chapter with the goal of collecting biodiversity observations and, more importantly, of applying an automatic validation on the data as well as providing real-time feedback to the participants. We focused on validating the location of bird observations in this approach, and we used species distribution models to validate the location of newly added bird observations. We chose Balanced RF algorithm to model the distribution of bird species in relation to environmental variables after analyzing various algorithms. We then implemented an API to use the trained models to automatically validate the location of newly added bird observations, as well as to provide real-time feedback to participants based on the model prediction of the likelihood of observing a species in a given location. Another goal of implementing this API was to simplify sharing our trained models with other CS practitioners.

As a result of testing this application with participants, the automatic feedback was reported to be useful and a way to motivate volunteers to continue contributing, as volunteers with higher frequency of contribution were more interested in receiving feedback. However, as previously stated, in order to have statistically significant proof that the automatic feedback resulted in increasing participant motivation, the application must be tested with a larger group of participants over a longer period of time. Consequently, in the future, we aim to use other approaches to engage people, such as organizing workshops or reaching out to potential participants based on their interests, and to obtain feedback from participants, other methods other than questionnaires can be used, such as conducting interviews. Furthermore, in the current version of the application, the feedback is only generated by the machine for the participants, but the participants are unable to respond to this feedback and must instead use it as a piece of information to either modify their observation or simply learn about species.

This chapter used one of the methods discussed in this thesis for combining ML and CS to improve data validation and to encourage public engagement. Although the use of ML in biodiversity CS projects is primarily used to identify species in images, to the best of our knowledge, using SDM to automatically validate new observations and provide real-time feedback on the probability of observing a certain species in a particular neighbourhood is not used in CS biodiversity projects (There are density maps, but no real-time feedback based on model prediction). Finally, one method for enhancing this application is to use a combination of images and environmental variables to train ML algorithms and thus validate biodiversity observations. In the following chapter, we present a summary of the discussion of all hypotheses, as well as some recommendations for other CS practitioners.

Chapter 7

Discussion of Hypotheses and Recommendations for Citizen Science Practitioners

7.1 Introduction

In Chapters 5 and 6, we presented two case studies of BioPocket and BioSenCS to discuss the two challenges of public engagement and data quality in CS. In Chapter 4, we presented our hypotheses and research questions, and in Chapters 5 and 6, we were able to fully or partially support these hypotheses using data from the aforementioned case studies. Thereby, the goal of this chapter is to summarize the main findings and hypotheses discussed in the previous two chapters.

Furthermore, we intend to use the hypotheses discussed to make recommendations on the design and development of a CS project aimed at increasing public engagement and improving data quality. These recommendations are intended to serve as a guide for other CS practitioners before designing a project or improving an existing one.

7.2 Discussion of Hypotheses

H1: The design, objective, and type of a CS project are correlated with the participant's motivations to contribute to that project.

This hypothesis was supported by our established conceptual framework, which categorizes participants' motivations based on the typology of CS projects. One major finding from our motivational framework was that projects with similar design and objectives reported similar motivations for their participants, allowing us to classify these motivations based on project typology as well as the strength of the motivations. This categorization was evident not only among classic CS projects, but also among various categories of online CS projects, such as the motivations of participants in gamified projects like Foldit (Curtis, 2015a) differed from those of participants in VC projects like SETI@home (Nov et al., 2010).

Furthermore, using the BioPocket case study, we evaluated and confirmed the hypothesis by comparing the motivations obtained from a survey related to BioPocket to the motivations of classic CS typology in the framework. The results of the BioPocket survey showed that the reported motivations were similar to those of other classic CS projects, with intrinsic nature-related motivations being the primary ones and extrinsic motivations being mainly stated as secondary ones by our survey respondents. Furthermore, we found a statistically significant negative correlation between intrinsic and extrinsic motivations to contribute to the BioPocket project (See figure 5.4 Chapter 5). In addition to the BioPocket case study, we demonstrated in Chapter 6 that the motivations of BioSenCS project participants followed the same trend as BioPocket and other classic CS projects in our motivational framework.

H2: Volunteers' motivation to contribute to a CS project is related to where they live.

We used evidence from the BioPocket case study to support this hypothesis. It is worth noting that, as previously discussed, this hypothesis is more clear for classic CS projects where the

contribution typically involves outdoor activities. According to our findings from the BioPocket case study, there was a difference in the average scores given to the intrinsic motivation "spending time in nature" and the extrinsic motivation "gaining recognition among others" among respondents living in villas (sparsely built areas with vegetation) and those living in apartments (densely built areas).

Although the majority of respondents in the BioPocket survey gave higher scores to intrinsic motivations, particularly *spending time in nature*, the mean of scores given to this motivation was higher among people living in apartments than those living in villas, which, as previously discussed, could be because this group of participants prefers to contribute to biodiversity CS projects in order to spend some time in nature and be outside of urbanized areas. On the other hand, while the majority of respondents in BioPocket gave low scores to extrinsic motivations, the extrinsic motivation *gaining recognition among others* received a higher score among people living in villas than those living in apartments, and thus as previously discussed, could be because this group of participants is motivated to contribute to such projects not only to help the biodiversity but also to get involved in group activities and to increase their social interactions.

H3: The socio-demographic information of participants is correlated with their motivations and perspectives on contributing to a CS project.

Using the Biopocket case study, we were unable to confirm part of the hypothesis that the socio-demographic background of the participants correlated with their motivations. However, we have found that the socio-demographic background such as age or level of education correlates with the behavior of people when using a CS mobile / web application. As mentioned in Chapter 2, most CS projects, especially classic CS, overlap with VGI because the geolocated data needs to be collected, so an important element to consider is whether the participants share their location with the application (location access authorization). We showed that there was a statistically significant negative association between level of education and willingness to grant location access, suggesting that people with higher levels of education have more concerns

about location access authorization when using mobile / web applications.

Furthermore, one method for making it easier to track participants' activities and connect with them is to ask them to register as users, either by creating a new account or by using existing accounts (such as Google or Facebook). Thus, in the BioPocket survey, we asked whether people would rather login with their existing accounts or create new ones, and people with higher education levels were more willing to create new accounts rather than use their existing ones. Although the majority of respondents in the BioPocket survey stated that they prefer to create a new account, it is important to note that user registration should not be a barrier to participation and should be made flexible so that participants are not discouraged from contributing to a project (Jay et al., 2016).

H4: Giving real-time feedback to the participants can result in increasing engagement.

This hypothesis was tested using the BioSenCS case study, in which participants received machine-generated feedback based on ML models' predictions about the likelihood of observing a species in a specific location. The participants were asked whether they found the feedback useful and whether it increased their motivation to contribute to the project in the long term. One of the most important findings was that there was a strong positive correlation between the frequency of contribution and the views of participants on the usefulness of feedback (cor = 0.79), indicating that more active participants found the feedback more useful. Furthermore, there was a positive correlation (cor = 0.49) between frequency of contribution and the impact of feedback on increasing motivation to continue contributing, suggesting that more active participants were more motivated by receiving real-time feedback. Although the correlations were not statistically significant due to the small sample size (14 participants and 10 respondents to our questionnaire), there was major evidence to support the role of automatic feedback in increasing engagement, and we recommend further research with a larger sample size.

H5: Automatic data filtering can simplify and accelerate data validation task, and improve data quality.

This hypothesis was also tested using the BioSenCS case study, in which we automatically filtered image, date, and location of observations, with our focus, as mentioned in Chapter 6, on location filtering using ML algorithms trained to model species distribution. As stated in our thesis objectives and discussed in Chapter 3, one of the primary goals of using ML to automatically validate data is to reduce the number of erroneous observations and to simplify expert data validation. We were able to support this hypothesis by demonstrating that during our three-weeks application experiment, only 24 of the 230 observations were flagged to be verified by experts.

Although the application needs to be tested over a longer period of time and with more participants to more strongly support this hypothesis, current evidence supports the benefit of integrating ML in CS for data quality assurance as discussed in Chapter 3 (See Chapter 3 Section 3.3.3 for details) by using a model-based validation approach rather than relying solely on expert review. Another interesting point here is that when using automatic validation, the experts focus only on the flagged observations, resulting in higher quality data by combining human (expert) and machine power.

H6: Giving real-time feedback to volunteers increases their knowledge about scientific domain of the project (e.g., biodiversity) and results in collection of higher quality data.

This hypothesis was also assessed using the BioSenCS case study. We discussed the roles of automatic validation and real-time feedback in combination. In other words, we wanted to explore whether or not receiving informative feedback resulted in participants contributing higher-quality data. In order to do so, we explored the correlation between the total number of contributions per participant and the ratio of flagged observations to total contributions. Our results indicated a statistically significant negative correlation between the number of contributions per participant and the flagged observations (cor = -0.63, p-value = 0.036).

As a result of this evidence, we were able to support the hypothesis by indicating that active contributors (those who contributed more frequently and had a greater number of collected observations) had a lower number of flagged observations. Furthermore, as we discussed in H4, the more active the participants were, the more useful the real-time feedback was to them. As a result, we can support the hypothesis that, because active participants found the feedback more useful and had fewer flagged observations than other participants, the automatic data validation and thus the real-time feedback assisted the participants in contributing higher quality data.

Furthermore, we discovered evidence that participants who gave a higher score to feedback being useful also gave a higher score to feedback assisting them in learning about species habitat characteristics. Although we were able to provide preliminary evidence to support the hypothesis, further investigation is required to more strongly verify the hypothesis, for example, as previously mentioned by running experiments on the BioSenCS application over a longer period of time and with more participants.

7.3 Recommendations for Citizen Science Practitioners

Based on the hypotheses discussed and the findings of our BioPocket and BioSenCS case studies, we propose some recommendations that CS practitioners can use to increase public engagement and improve data quality. Some of our suggestions are as follows:

1. Prior to designing a new project, consult the participants' motivations from previous projects with similar goals: As discussed in this thesis, participants in projects with similar objectives and designs have similar motivations to contribute; thus, one important step before designing a project is to identify existing successful projects with similar objectives and investigate the elements that motivate participants to engage in those projects. Our motivational framework can be used as a guideline for CS practitioners to see the motivational factors based on a CS project's typology.

2. Propose a variety of activities that take into account the diversity of participants: Although this is not a new point, it is worth noting that participants come from a variety of

backgrounds, and depending on their demographic, there are a range of reasons why they contribute to a project. As a result, proposing a variety of activities that take into account this diversity can aid in sustaining the participation of various groups. Some participants, for example, are interested in social interactions; therefore, providing opportunities for them to interact with others through social forums, chatting functionality in the project application, and organizing events such as field activities for group data collection can be among the proposed activities.

3. Before designing a project, get to know your potential participants: As we discussed in H3, the demographics of participants can influence their desire to contribute to a CS application. Because the majority of CS projects are now in the form of a web/mobile application, it is critical that the user interface be adapted for different groups of participants (e.g., age groups, education level, etc.). Furthermore, as discussed in H3, participants from various educational groups have different perspectives on creating a new account in order to contribute to a CS project. As a result, one suggestion is to make user registration more flexible by offering options like creating a new account, using existing accounts, or even contributing without creating a new account. However, for the latter, some limitations can be imposed in order to encourage participants to register as users; for example, unregistered participants cannot view the contributions made by others.

4. Using ML algorithms, make user-centered suggestions to participants: This recommendation is primarily applicable to classic CS projects that collect geolocated data, such as biodiversity observations. There are situations when participants would like to contribute to a CS project but need more information on what to contribute, and thus automatically giving suggestions to participants based on their location can keep them engaged in the project while also helping them learn about their surroundings. For example, the suggestions can assist them in determining what types of species are more likely to be observed around them, or what dangerous phenomena impact the quality of water bodies near where they live.

5. Give participants real-time, informative, and user-centered feedback: Giving feedback to participants has been shown to sustain their participation and increase their motivation,

as discussed in some of the available research (Ingensand et al., 2015; Kelling et al., 2011). It is important to note, however, that the feedback should assist volunteers in learning about science and gaining useful knowledge. Furthermore, rather than providing general feedback to participants, feedback should be tailored to the data contributed by each participant in order to maximize the impact of feedback in increasing public engagement. As seen in our BioSenCS case study and discussed in H4, we propose using ML algorithms in the project to consider volunteers' contributions and provide participants with informative feedback as a result.

6. Improve data quality by providing participants with informative feedback: As previously discussed, providing informative feedback to participants demonstrated to be beneficial in terms of acquiring new knowledge that enables them in the improvement of their contribution. We indicated in H6 that active BioSenCS participants found the machine-generated feedback more useful and had fewer flagged observations compared to other casual participants. As a result, we recommend that participants' learning curves can be improved through real-time informative feedback, resulting in higher quality data.

7. To expedite data validation and improve data quality, use automatic data quality assurance: As demonstrated in the BioSenCS case study and discussed in H5, integrating ML in CS for data validation can noticeably reduce the amount of data that must be verified by experts, thereby speeding up the data validation process. Furthermore, automatically filtering the data and considering expert verification only for the flagged contributions allows experts to focus on more challenging tasks rather than screening all of the contributed data, which machines can do. As a result, we advise CS practitioners to use a combination of human and machine power for data validation tasks in order to produce high-quality data that can be used in further scientific analyses without the need for extensive pre-processing.

8. When using ML, use caution; consider biases, transparency, and trustworthiness: Although we discussed the benefits of integrating ML in CS throughout this thesis, we also discussed the potential challenges that may arise if ML is not used with caution in CS projects. A critical aspect is to create a transparent project by informing participants about how CS data is used in ML algorithms as well as communicating with participants about which tasks

in a project are automated. Furthermore, it is critical to avoid overestimation of machine intelligence over human intelligence. It may result in inaccurate data quality assurance, or in losing participants' interest in a project as a result of too much automation. Overall, if ML is integrated into CS, it must be clear how the models are trained (training data, algorithm performance, etc.), and the perspectives of the participants on integrating ML and automating tasks must be considered.

In the following and final chapter, we present the conclusions as well as perspectives and recommendations for how this research can be expanded.

Chapter 8

Conclusions and Future Work

8.1 Summary and Main Findings

Recent technological advancements have resulted in an increase in the number of CS projects in a variety of scientific areas (Land-Zandstra et al., 2016; Schade & Tsinaraki, 2016; Schade et al., 2020). Despite the large number of CS projects, not all of them are successful (Conroy, n.d.; Cox et al., 2015); therefore, keeping a CS project successful necessitates cautious consideration of the challenges that exist in this field, the two main challenges being increasing public engagement and improving data quality. Several studies have focused on these two challenges, conducting surveys to understand the motivations of volunteers to contribute to CS and evaluating data quality criteria (Kosmala et al., 2016a; Leocadio et al., 2021). Notwithstanding the existing literature, there is still a need to develop new approaches based on new technologies to address these two challenges and lead to more successful CS projects.

Therefore, the objective of this thesis was to focus on addressing the aforementioned challenges using the integration of ML in CS projects. The combination of ML in CS projects has recently become a focus among researchers (Ceccaroni et al., 2019; Franzen et al., 2021; Lotfian et al., 2021; McClure et al., 2020). This integration has primarily been focused on biodiversity CS projects (McClure et al., 2020), with species identification in images (e.g., in the iNaturalist project) being the most well-known application of ML in such projects. However, as the focus

on combining CS and ML intensifies, more ways of integrating ML and CS can be proposed, not only to automate object identification in images, but also to involve machines as a partner in such projects to help CS practitioners and citizens. For example by using natural language processing algorithms to simplify the interactions of citizens with CS application, such as through the use of chat-bots (Adriaens et al., 2021).

Thereby, we conducted an exemplary literature review of successful CS projects that combined CS and ML, and as a result, we defined a taxonomy of various ways that CS and ML can be integrated to help CS projects, from public engagement to data collection and data validation (See Chapter 3, Section 3.3). One of our key findings from this literature review was that **the majority of existing projects combining ML and CS invite volunteers to collect/label data to feed and train ML algorithms and as a result fully automating tasks, rather than using ML to benefit CS participants**. Another important finding was that **the integration of ML in CS is not so much focused on public engagement as it is on automatically detecting objects in images**. Hence, in light of our proposed taxonomy of combining CS and ML, we concentrated on the two branches of public engagement and data quality.

Prior to evaluating the role of ML on increasing public engagement, we identified the elements that motivate people to contribute to CS projects. To do so, we conducted a thorough literature review to identify and categorize participants' motivations for contributing to CS taking into account the project typology (See Chapter 5, Section 5.4). As a result, **we established a motivational framework that, in addition to the typology of CS projects, categorizes motivations to primary and secondary classes based on the role of motivations in recruiting volunteers or sustaining their participation.**

One of our main findings as a result of establishing the motivational framework was that, while the motivations of participants with similar design and objectives tend to be similar, some motivational factors are common in almost all projects, such as **interest in learning about science** and **the desire to stay informed about the project and receive feedback**. One approach for satisfying these two motivational factors is to provide participants with

informative feedback. To do so, one common method is for experts to provide feedback for participants after evaluating their contributions; however, because experts must verify large amounts of data, providing feedback for all participants is time-consuming. Accordingly, we proposed providing volunteers with real-time machine-generated feedback, and we investigated the role of such feedback in increasing participants' motivation.

Since the focus of this thesis is both on public engagement and data quality, we investigated the role of ML in automatically filtering and validating citizens' contributed data in addition to providing machine-generated feedback to participants. To that end, we developed BioSenCS, a CS project that invites participants to collect biodiversity observations with the objective of automatically validating collected biodiversity data using ML algorithms. Because our objective was to automatically validate the location of biodiversity observations (bird species in our case), we generated species distribution models and used them to automatically verify the location of an observation based on the likelihood of observing a species in a specific location. The location validation was done in real-time, and the participants received real-time feedback on the probability of observing the species as well as information on species habitat characteristics based on the model's predictions. Furthermore, we implemented an API to save and use our trained ML algorithms, which this API can be used to share our trained models with other CS practitioners working on biodiversity data validation without having to start training algorithms from scratch.

Following an experiment of the BioSenCS project with participants, we obtained several preliminary results that support the integration of ML in our project. The first key result was that **the active participants found the real-time informative feedback useful in learning about biodiversity, and that it increased their motivation to contribute to the project**. Another important finding was that **the feedback was effective in reducing the number of erroneous observations contributed by participants**, as we indicated that the more active the participants were, the fewer the number of flagged observations were. Furthermore, we demonstrated that **by automatically filtering observations, we significantly reduced the number of erroneous observations that required expert verification.** Moreover, we illustrated that combining human and machine power not only speeds up data

validation process, but also improves data quality because the machine controls and performs a full screening of the data while experts focus on verifying the most challenging ones.

Furthermore, based on the discussions and findings of this research, we made some recommendations for CS practitioners to consider before designing a project or improving an existing one:

1. Prior to designing a new project, consult the participants' motivations from previous projects with similar goals.

2. Propose a variety of activities that take into account the diversity of participants.

3. Before designing a project, get to know your potential participants.

4. Using ML algorithms, make user-centered suggestions to participants.

5. Give participants real-time, informative, and user-centered feedback.

6. Improve data quality by providing participants with informative feedback.

7. To expedite data validation and improve data quality, use automatic data quality assurance.

8. Finally, while there are benefits to combining ML and CS, as we discussed with the findings of our BioSenCS project, there are challenges that should be carefully considered when integrating ML and CS.

In this thesis, we discussed the potential benefits and challenges of combining ML and CS, and we focused on some key factors to consider, such as:

- Overestimation of machine power over human power may result in demotivating participants.
- When integrating ML in CS, participants must be informed of where and how their contributed data is being used in the algorithms.
- When using ML models for data validation, be cautious because if the models are trained with biased input, they may make false positive or false negative predictions, resulting in

inaccurate data validation.

- Considering the potential ethical issues that may arise from the use of ML in CS, such as predicting and sharing the location of endangered species.

8.2 Future Work

As a future direction of this research, we aim to investigate biodiversity data validation by using both images and environmental variables to train ML models in order to improve the accuracy of prediction of observing a species. Furthermore, we intend to test our application with a larger community of participants (e.g., by arranging workshops/focus groups, or conducting interviews) in order to strengthen our preliminary findings in this study and to better assess the impact of real-time feedback on sustaining participation. Besides that, in the future, we aim to enhance our application by allowing participants to interact with the machine-generated feedback, reducing the number of erroneous observations even further. For example, if the machine flags an observation and the participant receives feedback on why the observation was flagged, the participant can provide reasons or more details for the machine, and thus the observation is evaluated again, and given the provided details, the observation can pass the automatic filter successfully, and thus not be flagged. Our goal for the future investigation is to see if this human-machine interaction helps not only the participants, but also our models to make more accurate predictions based on the information provided by the participants.

Aside from future enhancements to our BioSenCS application, we propose some potential future research challenges where ML and CS can be combined. Some potential challenges and future ideas are as follows:

1. The focus of the use of ML in CS is currently more on automatic identification and less on public engagement; thus, exploring the use of ML in increasing engagement and sustaining participation remains an area for future investigation. For instance, one potential approach to be explored is the use of gamified AI in CS towards attracting more volunteers as well as sustaining participation (Bowser et al., 2013; Franzen et al., 2021).

2. Training participants has been shown in studies to improve data quality; however, providing training is not always simple and requires both human and financial resources. A possible suggestion will be to use AI to provide training prior to/while data collection; although this has been achieved in the case of machine-generated feedback in BioSenCS project, AI can be used to provide training in a variety of ways, such as through interactive courses managed by AI (Cope et al., 2021).

3. CS data are primarily based on the collection of images/videos or textual data, but with emerging technology, the types of data collected can be extended. For example, some of the most recent smartphones support sensors that acquire LiDAR (Light Detection And Ranging) data, and while this is currently a device-specific feature, given the rapid pace of technological development, we would expect it to be included in many future smartphones. Thus, LiDAR data can be a potential data type obtained in CS projects, and although some studies have been performed to identify objects from point clouds using deep learning (Engels et al., 2020; Wu et al., 2021), applying such techniques to LiDAR data collected by citizen scientists is a very interesting challenge towards the combination of ML and CS.

4. In this thesis, we discussed the use of ML and CS for three major CS phases: public engagement, data collection, and data validation. Participants can, however, be involved in problem definition and project design, as stated by Haklay (Haklay, 2013) and Bonney (Bonney et al., 2009a). Considering the rapid advances in AI where articles can be written by machines (Dale, 2021) or where machines can help in software development e.g., in GitHub Co-pilot (Sawhney, 2021), one potential research could be to verify whether AI can assist humans in the problem definition step and project design in CS, for example by learning from successful CS use cases.

5. In this thesis, we trained our algorithms on existing expert-verified data from eBird and then used them to validate new observations. However, there are cases where there is insufficient validated data, such as for rare species in biodiversity projects or in other domains where large amounts of "verified" data are not available. As a result, the next challenge will be to demonstrate how to perform automatic data validation in the absence of sufficient expert-

verified data. Accordingly, future research should look into approaches for amplifying expert-verified data in CS projects where there is not enough data.

The integration of ML and CS is still in its early stages, and more research is needed to evaluate various aspects of this integration. As previously stated, the primary focus where ML and CS are combined has been on the field of biodiversity, but as a future continuation of our research, we would like to investigate how our proposed approach here can be expanded to other CS projects besides the field of biodiversity. Finally, an important aspect that we discussed in this thesis is the importance of transparency in using AI in CS projects, as well as being clear with volunteers about how and where their data is being used in the algorithms. The most important conclusion to be drawn from this thesis is that citizens are at the heart of CS projects, and the integration of ML and CS must not be used to replace the role of humans, but rather to assist both citizens and CS practitioners.

Appendix A

Using Flickr Data to Perform Species Distribution Modeling: An Early Proposal

Before using the eBird data set, we investigated whether Flickr data could be a useful source for modeling species distribution (Lotfian & Ingensand, 2021). Since it is possible to obtain various types of information along with images using the Flickr API, such as date-time, geolocation, EXIF (Exchangeable Image File) data, and photo tags (user or machine generated), there have been several studies that have used Flickr data, such as natural disaster monitoring (Sun et al., 2015) or location-based behavioral analysis (Chassin & Ingensand, 2021; Kisilevich et al., 2010). Although some researchers are opposed to using Flickr data in ecological studies, it has been suggested as a supplement to more standard datasets such as data from CS projects (ElQadi et al., 2017).

As a result, we obtained bird images from Flickr in order to investigate how they are distributed across Switzerland and to explore the association between species distribution and land cover classes. Furthermore, by comparing Flickr to the eBird dataset, we aimed to examine the level of validity at which Flickr images can be used for analyses of biodiversity observations, particularly with regard to the generation of SDM.

Thus, we used the Flickr API to obtain images in Switzerland for 2018, and in order to obtain all possible data, we searched for the "bird" tag in four languages: English, French, German, and Italian. Figure A.1 illustrates the distribution of obtained Flickr data points within Switzerland.

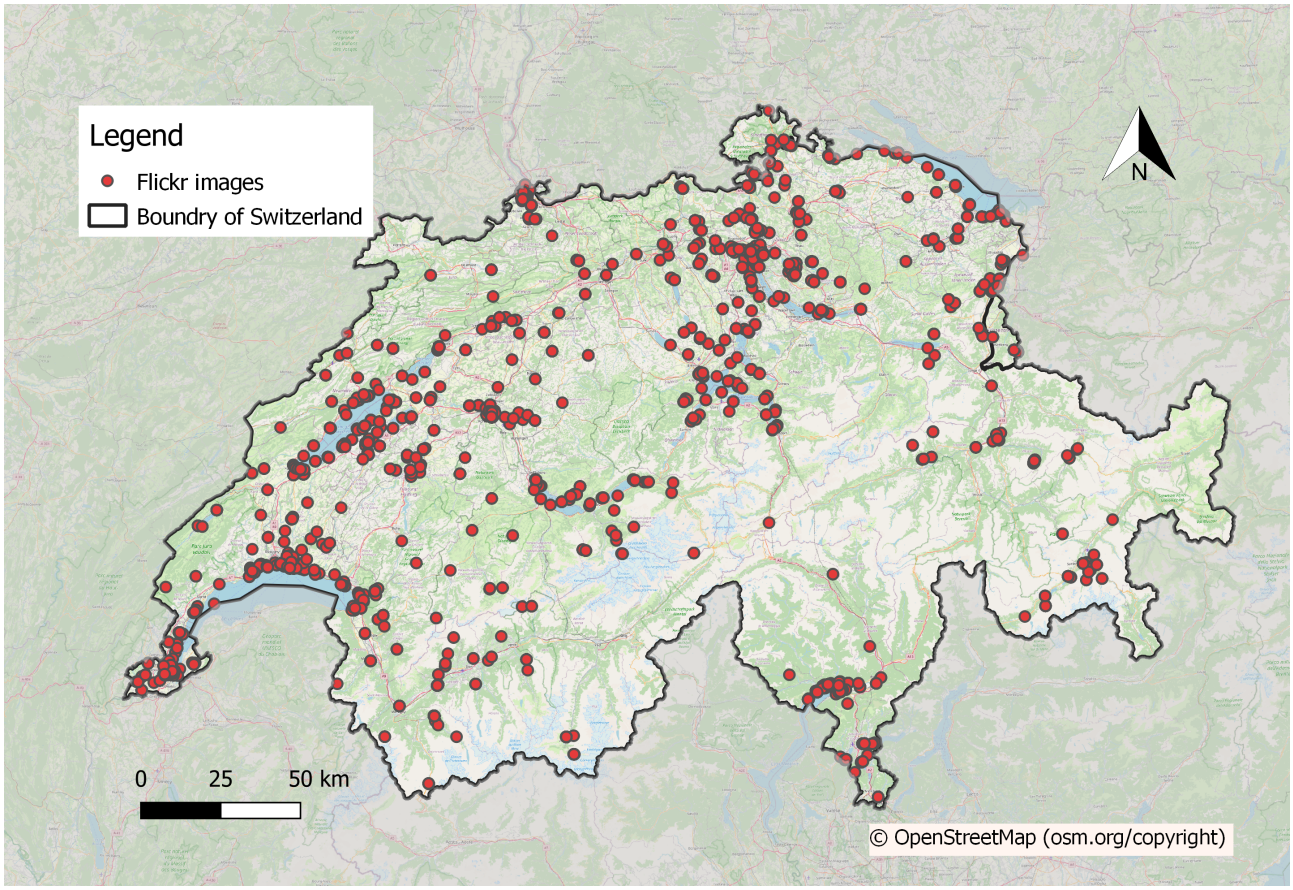


Figure A.1: Location of Flickr images with bird tags in Switzerland

We had to apply two image and text filters to the Flickr data in order to get the data set we needed for our analyses:

- Image filter: Despite the fact that the search was conducted using the "bird" tag, there are some images that have the same tag but are not birds (for example, statues or drawings of birds that had the bird tag, or images where the presence of a bird was not clear enough, See Figure A.2). As a result, we used the Clarifai general model to filter the images and get only images of bird species. We excluded an image if the likelihood of a bird being present (prediction of the tag "bird") was less than 90%.

- Text filter: We used a dataset of bird species names in Switzerland provided by The Swiss

Ornithological Institute[1] to filter the tags. This dataset includes bird species names in all four official languages of Switzerland (German, French, Italian, and Rumantsch), as well as English. As a result, we implemented a matching string function to filter out tags that are in close match with the list of bird species names. We excluded tags that had a match rate of less than 85% with the species names. After the automatic tag filtering was completed, we manually verified the filtered tags to remove any possible duplicates as well as tags that had a close match with bird names but were not a bird species (for example, the species name Verdone (European greenfinch) was matched with the city name Yverdon).

Figure A.2: An example of an image with bird tag, which was filtered out using Clarifai

After obtaining the filtered dataset, we used Kernel Density Analysis (KDE) to visualize the density of distribution of bird observations in our study area. Furthermore, an additional dataset containing the CORINE land cover values for each observation point was created to investigate the distribution of the data within various land cover classes. Thus, the frequency of bird observations within different land cover types was determined, and a chi-square test of independence was used to explore the association between bird species and land cover types.

[1]https://www.vogelwarte.ch/

Figure A.3 depicts the workflow we applied, from obtaining photos to filtering them and finally evaluating the relationship between species distribution and land cover classes.

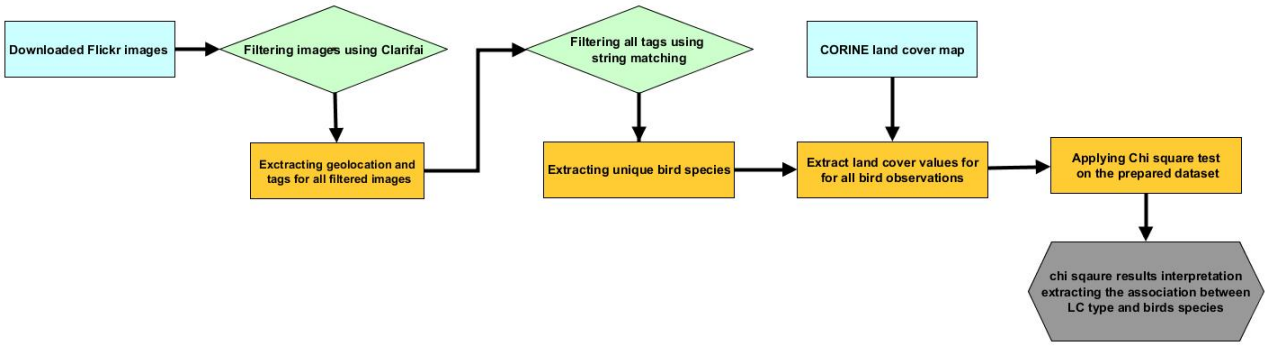


Figure A.3: The workflow of evaluating the correlation between distribution of Flickr bird images and CORINE land cover classes

Finally, in order to compare the Flickr and eBird data sets, the SDMs for a bird species called Common Kingfisher[2] were generated for both. The Common Kingfisher datasets from eBird and Flickr each had 239 and 51 unique observation points, respectively, and only the CORINE land cover map was used as an input environmental variable to generate the models. We used the MaxEnt algorithm (Phillips & Dudík, 2008) to generate the SDMs, and the AUC metric (Bradley, 1997) to compare the performance of the two models. In addition, the correlation between the two raster maps was calculated using their pair pixel values to assess the similarity of the two species distribution maps.

We obtained 7719 images, which were then reduced to 4610 images after image and text filtering, with 2604 unique geolocations. The majority of user-generated tags include the location of the image, the camera model, or general tags. Few tags, however, include species names, and in most cases, the names are added with a shortened version of the species common names, or misspelled, which is why using Flickr images to conduct distribution analysis should be done with caution. The most frequently used tags in this study's downloaded images are depicted in Figure A.4. Following tag filtering, the final dataset contained 170 distinct species with at least five observation points.

Figure A.5 depicts the KDE analysis used to visualize the density of the distribution of bird

[2]https://en.wikipedia.org/wiki/Common_kingfisher

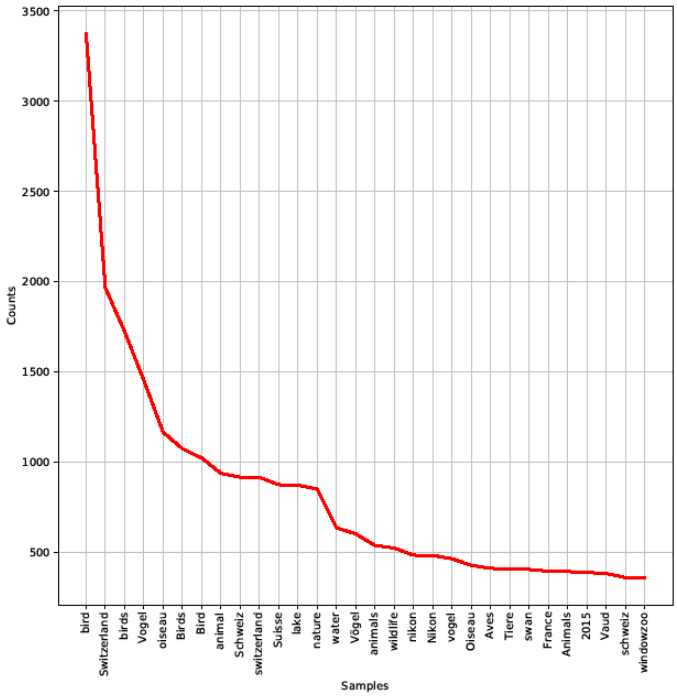


Figure A.4: The most frequent tags from the downloaded Flickr images (x-axis: Flickr tags, y-axis: frequency of tags)

images. The density map shows that the majority of images are concentrated near lakes or in the vicinity of major cities. This is a common pattern in data from CS projects such as eBird (Strimas-Mackey et al., 2020b), and it is due to the majority of contributors' tendency to gather observations in places close to where they live or in more accessible regions, resulting in a spatial bias in such datasets (both in Flickr and CS data).

In addition, the chi-square test was used to assess the association between land cover types and bird species, and the Cramer's V metric was calculated as a result of the test. Cramer's V is a metric for determining the degree of association between two variables. It ranges from 0 to 1, with values greater than 0.5 indicating a strong association. With Cramer's V = 0.5209 and p-value<0.0001, the chi-square test resulted in a statistically significant association between land cover types and bird species.

The SDM maps obtained for both datasets are illustrated in Figure A.6. The model generated using eBird data performed better with AUC=0.86 compared to the one generated using Flickr data with AUC=0.7, which is reasonable given the number of records in Flickr which was nearly

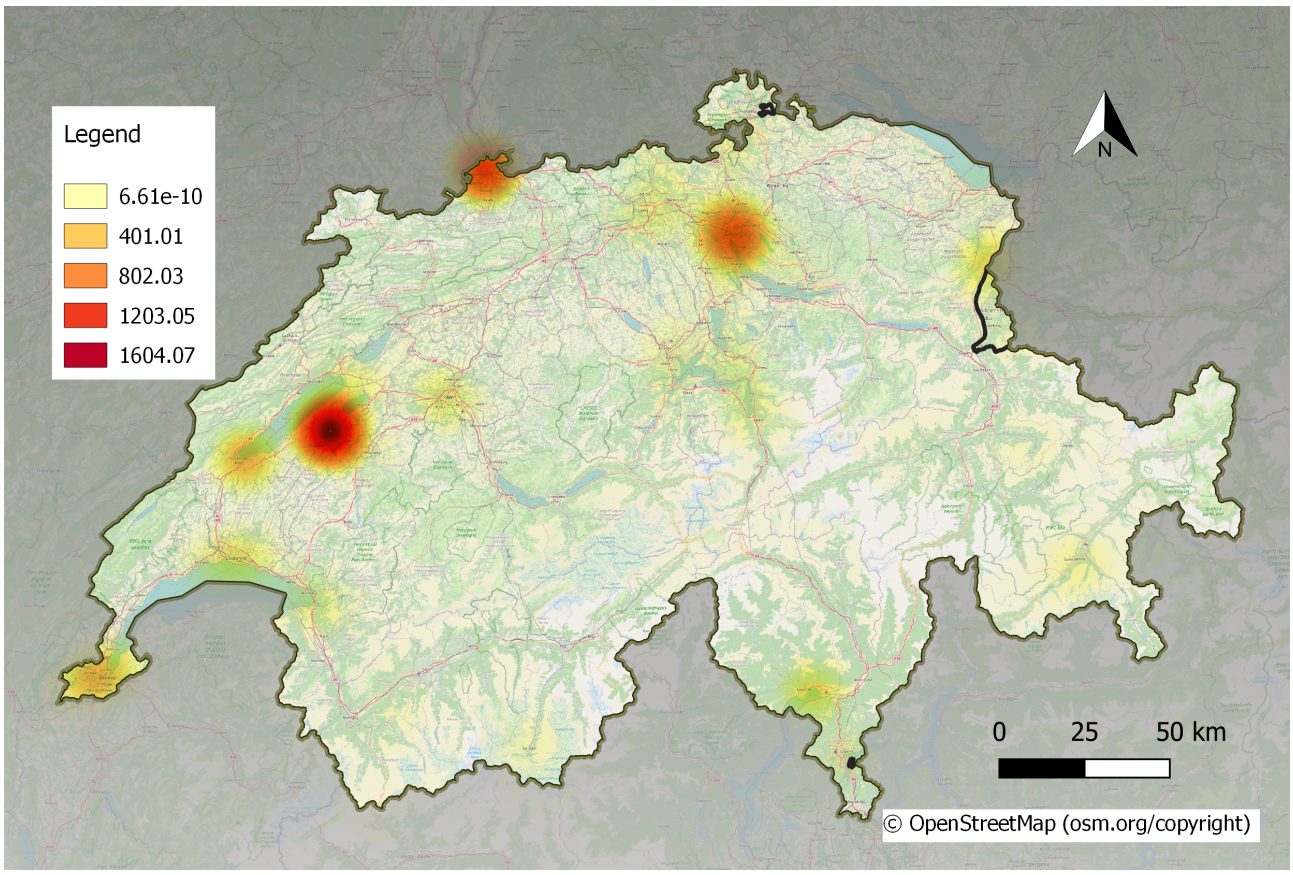


Figure A.5: Density map of the Flickr images with bird tags

four times less than eBird. While the distribution patterns in both maps look similar, the distribution from Flickr illustrates higher probability of occurrence in areas with discontinuous urban zones compared to eBird. Figure A.7 illustrates the statistics comparing the two raster maps, and it shows a very high correlation among the pixel values, supporting the similarity of the distribution between the two maps. From these analyses it can be discussed that Flickr data might be a potential source to address the issue of lack of occurrence species data particularly in SDM studies, given that necessary filtering steps are applied to the data. However, it is essential to note that a large number of species had few data points (less than 5), and thus we could not evaluate or make any comparisons of such data with eBird observations, and it remains a point for future investigations.

We concluded that because the Flickr data are not expert-verified, they cannot be used solely (in the absence of other structured datasets) to generate SDMs. The findings indicate that for common species, the SDM can produce results that are comparable to CS data; however, because many species had very few observations, the evaluation of Flickr data for those species

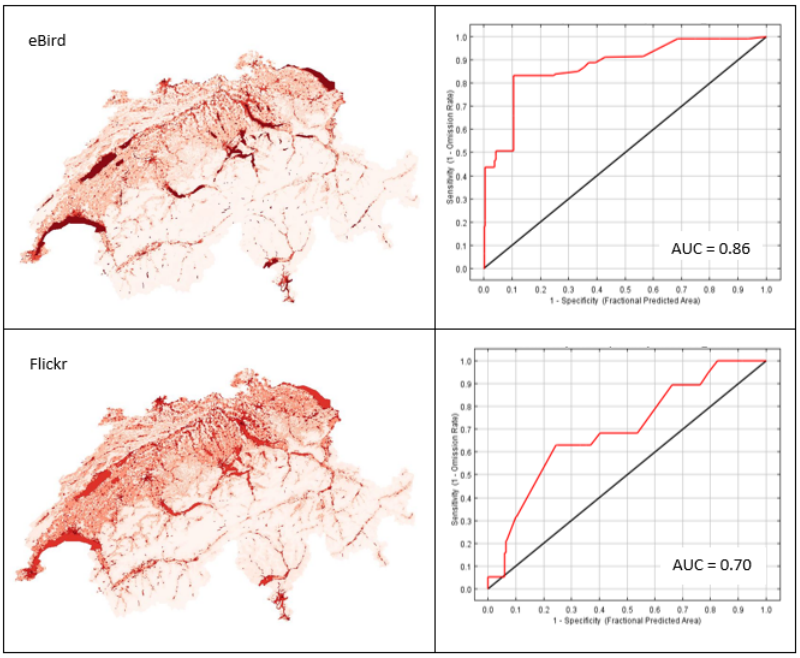


Figure A.6: Species distribution maps and the models' performances generated using Maxent for Common Kingfisher using the datasets of eBird (top), and Flickr (down)

remains unclear, and no comparison could be made. This will be looked into further in the future. Another compelling argument for future research is to look for alternative approaches to tag filtering and extracting useful information from Flickr tags, such as using CNN to predict species names from images and then comparing them to Flickr tags.

Statistics of each species distribution model

Layer	MIN	MAX	MEAN	STD
SDM_eBird	0.0923	1.0000	0.2146	0.2091
SDM_Flickr	0.4281	1.0000	0.5331	0.1336

COVARIANCE MATRIX

Layer	SDM_eBird	SDM_Flickr
SDM_eBird	0.00847	0.01222
SDM_Flickr	0.01222	0.02076

CORRELATION MATRIX

Layer	SDM_eBird	SDM_Flickr
SDM_eBird	1.00000	0.92135
SDM_Flickr	0.92135	1.00000

Figure A.7: Statistical comparison of the species distribution maps generated using eBird and Flickr datasets

Appendix B

Other case studies of combining CS and ML

We have worked on other CS projects combining CS and ML, in addition to the BioSenCS case study discussed in Chapter 6, and we briefly introduce two of these case studies in this appendix section.

B.1 AId Lakes: AI for water quality monitoring

This project began as an extension to another project called SIMILE (Brovelli et al., 2019), which aims to preserve water quality by inviting citizens to share their observations of the lake environment through geo-referenced images of algae, foams, and litter, as well as water parameter measurements (transparency, temperature, pH, etc.). As a result, the idea is to use AI to automatically filter our observations by identifying water quality phenomena in images and distinguishing clean water observations from those that include phenomena such as algae or foam (the two phenomenon considered to be detected) (Biraghi et al., 2021).

Data set preparation: The first step was to construct a data set of labeled images. To do so, we first used search engines like Google and Bing to find images of water bodies that

contained the keywords for the two phenomena (algae; foams), as well as their synonyms (algal bloom and scum; froth and spume). After downloading the images, we used a function to remove duplicates, and the final data set contained 358 and 409 images of algae and foam, respectively. The data set was used to train two different algorithms: a CNN and an object-detection algorithm called faster Region-based Convolutional Neural Network (R-CNN).

Prior to training algorithms, the images were labeled. When using the CNN algorithm, the images could be placed in a folder with the name of the phenomenon in order for the labels to be generated from the directory structure. This has been done by using the *flow_from_directory* function in Keras[1]. However, in order to label the images for the object-detection algorithm, the phenomenon within each image had to be identified by drawing bounding box(es) around it and assigning a label to each phenomenon (algae or foam). In order to perform this action, a tool for image annotation written in Python, called LabelImg[2], was used. The output consisted of the image and a homonymous XML file (Extensible Markup Language) containing the coordinates of the bounding boxes and the label name. The labelled images were then uploaded on a platform called Roboflow[3], which helped augmenting the images (increasing the input data by applying rotation, flip, etc. to the images), and exporting the required output format for the faster R-CNN algorithm. Figure B.1 illustrates a sample of labelled data set.

[1]https://keras.io/api/preprocessing/image/
[2]https://github.com/tzutalin/labelImg
[3]https://roboflow.com/

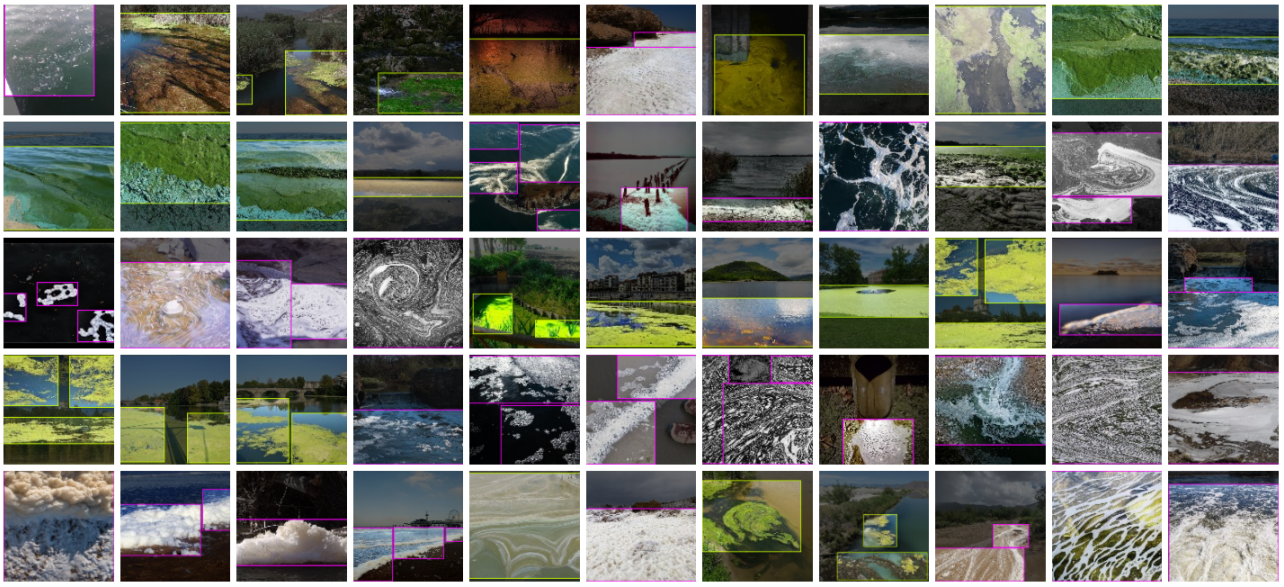


Figure B.1: Sample of dataset images after data augmentation

CNN: We initially trained a CNN to classify the images. Therefore, a CNN model composed of three hidden convolutional layers (each followed by a two-dimensional max pooling) and a fully connected layer has been trained. The ReLU activation function (Check Chapter 6 section for more details of activation functions) was also applied in the fully connected layer. A summary of the specific architecture of the model used is shown in Figure B.2.

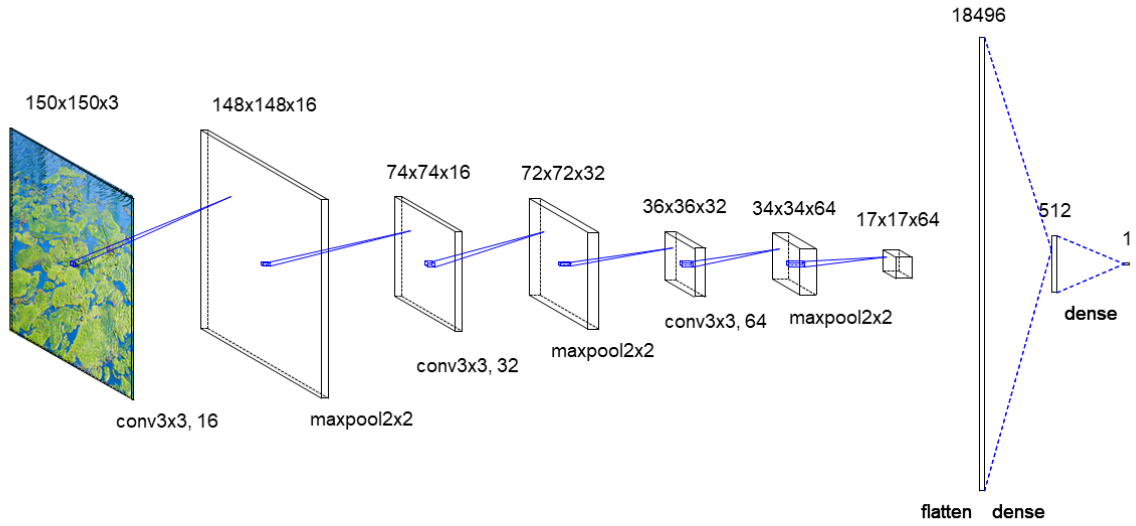


Figure B.2: Architecture of the implemented CNN model

The results obtained by using CNN show that the model has learned relatively well on the training data set. However, in the validation data set we can observe some noises in the accuracy with values fluctuating between over 90% and 50% (Figure B.3). Due to the observed gap between training and validation accuracy/loss it can be concluded that the model is suffering from high variance and thus overfitting. This behaviour could be due to the relatively small amount of data used to train the CNN or to the heterogeneity in the data set. Furthermore, while CNN could only predict whether a particular image represented algae or foam, the model provided no information on the location of the phenomenon within the image, nor on the possibility of both phenomena coexisting in a single image. As a consequence, we decided to train an object detection model which uses CNN to predict both the phenomenon and its precise location in an image.

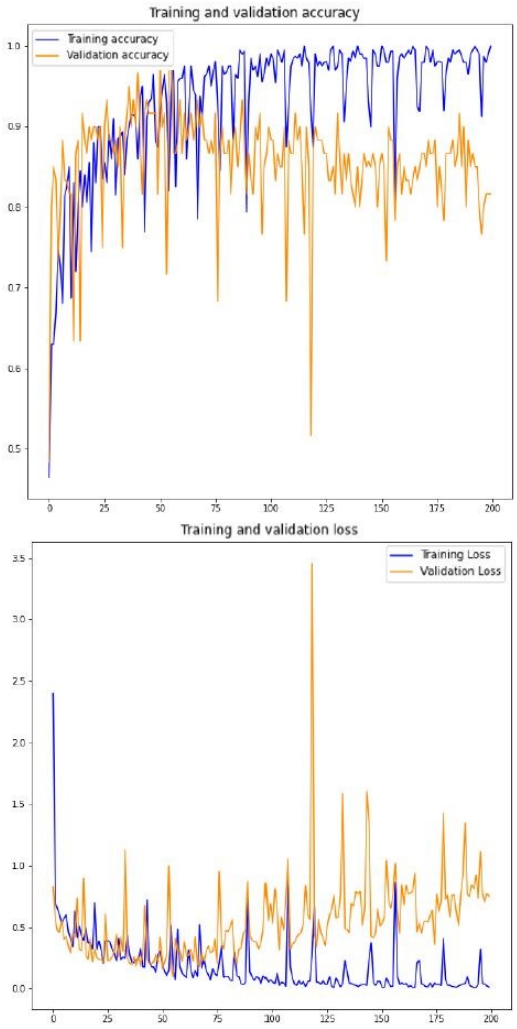


Figure B.3: CNN performance. x axis: number of epochs, y axis Top: Accuracy; Bottom: Loss.

Faster R-CNN: While image classification models predict the likelihood of an object being present in an image, object detection algorithms predict both the presence and location of objects in an image (e.g. using bounding boxes). R-CNN, developed by Ross Girshick et al. (Girshick et al., 2014), is one well-known object detection algorithm. (2014). This approach uses an algorithm for image segmentation called "selective search" in order to determine the possible regions where an object may be located (approximately 2000 region proposals per image). The regions are then passed to a CNN model which generates a feature vector from each region proposal. Finally, a SVM model performs a classification of the objects found and identifies the location of the objects in the image. Training this kind of model is computationally expensive and the test data take a long time to be predicted (approximately 49 seconds per image). As a consequence, a version of R-CNN was proposed – faster R-CNN (Ren et al., 2015) – which uses convolutional networks to propose regions, rather than an external algorithm of region proposal. Faster R-CNN requires both less training time and less time to detect test images (0.2 seconds per image), which makes it suitable for integration in applications of real-time object detection. Figure B.4 illustrates all the steps we followed to detect water quality phenomena using faster R-CNN algorithm.

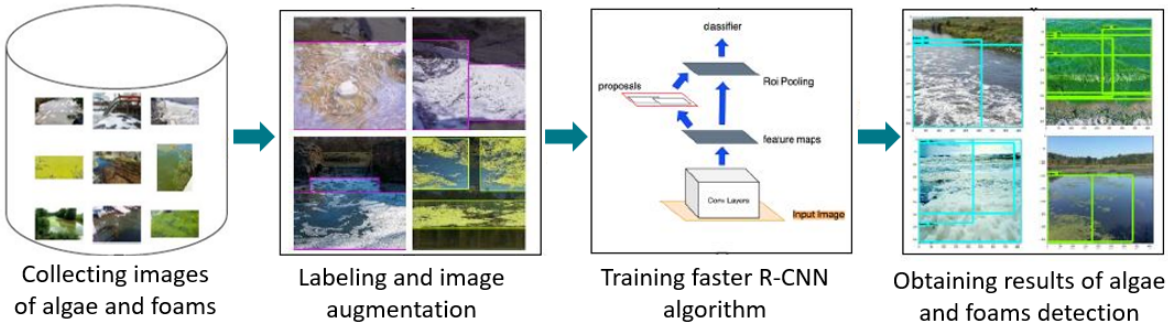


Figure B.4: Steps of detection of water quality phenomena using faster R-CNN algorithm

We trained the algorithm using 1000, 10,000, and 20,000 training steps, and the three runs were compared with regard to model performance in order to consider the ability of the model to classify the object and to define its location correctly. A possible index for model evaluation is a measurement of the overlap between the predicted bounding box and the ground truth bounding box, which is called IoU (Intersection over Union, see Figure B.5).

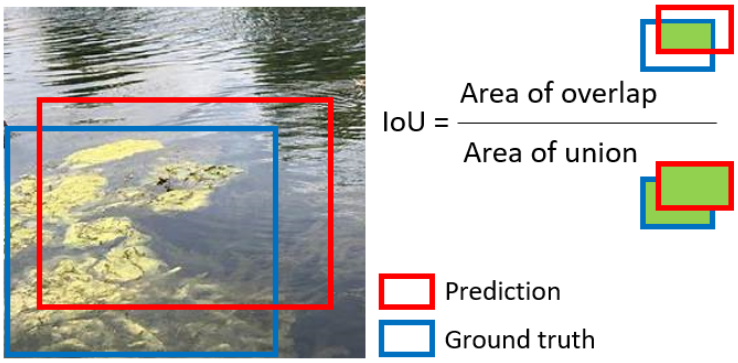


Figure B.5: Intersect over Union, visual explanation

Therefore, IoU can be set as a threshold in order to consider whether a prediction is true or whether it is a false positive. For instance, setting IoU threshold to 0.5, if the IoU of the prediction is above this threshold, it is considered as true positive, if below it is considered as a not precise detection. The IoU threshold value was set to 0.5, then the average recall and mean average precision (mAP) were compared.

Table B.1 results show that with 20'000 steps, the model performs better with a higher mAP. Despite the fact that the results of the recall with 1000 steps are similar to those of the recall with 20000 steps – and also higher than the recall with 10000 steps – the mAP increased as we trained the model with more steps. That is, the model learns to predict phenomena early on, but the location of detected phenomena improves as the number of steps increases. The findings can be compared to the few existing studies on harmful algae bloom object detection (Kumar & Bhandarkar, 2017; Samantaray et al., 2018). Even though the recall and mAP for faster R-CNN are lower than what they achieved, given that the objects of identification were two phenomena, as well as the heterogeneous and small data set, the results are still promising for the detection of more than one phenomenon, provided that the model is trained on a larger data set.

The model performs better in the detection of algae than foams. Several cases of false positives have been detected, in which clouds, stones, or light reflected on water were mistakenly identified as foams. However, the positive aspect is that no false negative was detected for foams. It is interesting to note that no cases of false positives for algae were detected. However, there were

Table B.1: Indicators of R-CNN model performance

Training Steps	Average Recall	mAP	Training Time
1000	44.5 %	12.5	Approx. 30 min
10000	41.6 %	17.6	Approx. 5 hrs
20000	44.6 %	19.9	Approx. 8 hrs

cases where algae were not detected, especially in those images where the phenomenon was particularly extended and covered the whole image and a contrast with clean water or other elements was not available. Some examples of model detections of the two phenomena can be indicated in Figures B.6, B.7, B.8, and B.9.

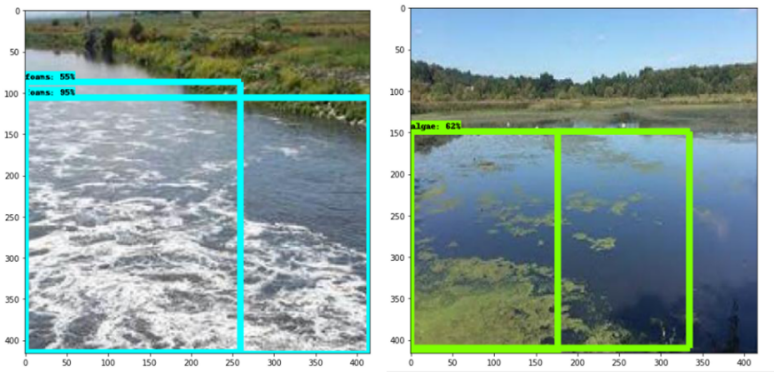


Figure B.6: True positives for foams and algae (correctly detected and located)

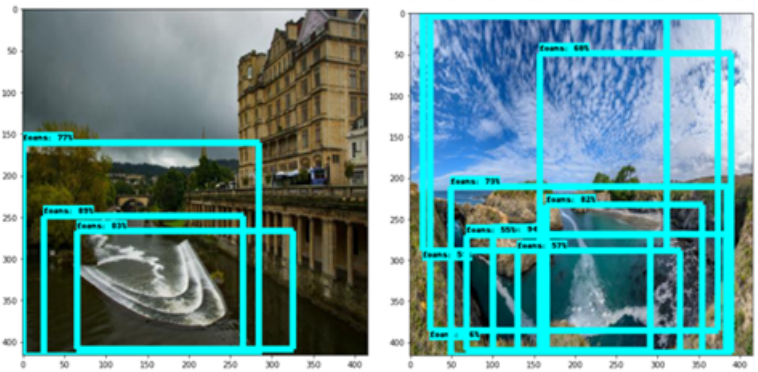


Figure B.7: True and false positives (clouds, right) for foams

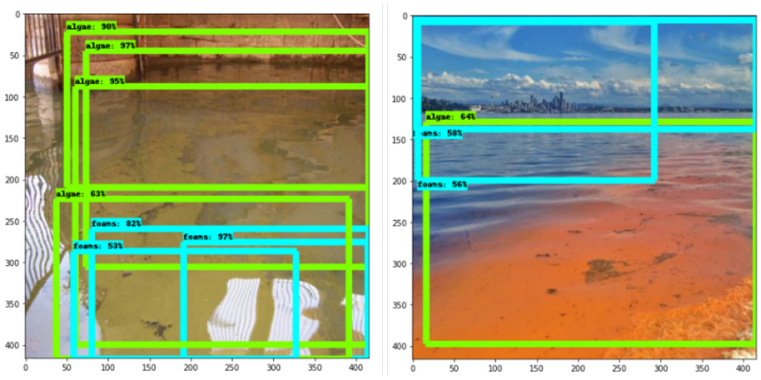


Figure B.8: True positives for algae and false positive for foams (reflection of light, left and clouds, right)

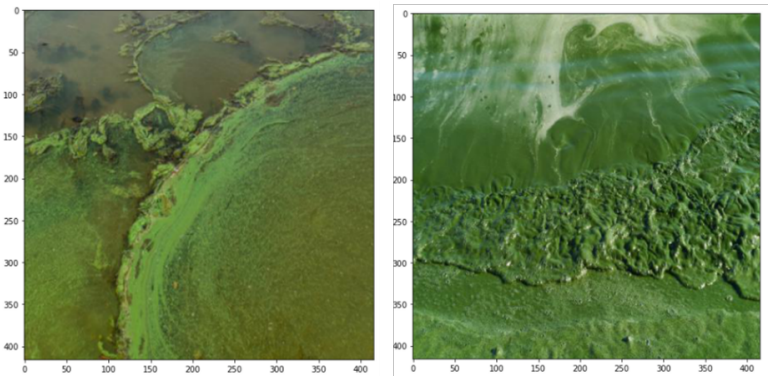


Figure B.9: False negatives for algae (not detected)

Finally, this CS/ML research is the first step in a larger future project aimed at automatically identifying and classifying water quality phenomena in images of water bodies collected by citizens. An important aspect to be investigated further in this research is training models based on geolocated images and satellite images in that location to perform predictions taking into account the location of an image, rather than just its content, with the goal of increasing the model prediction's reliability.

B.2 MoDoS

Another case study of combining CS and ML in which I was partially involved is the project MoDoS. MoDoS aims to recommend the best walking path, with the fewest obstacles, for disabled or senior citizens based on their user profile and preferences (e.g., type of disability).

With this goal in mind, this project invites citizens to collect images in the city that include an obstacle and assign a type of obstacle to the image, either from a proposed list in the MoDoS application or by adding the details of the obstacles. These labelled images are then used to feed and train a CNN algorithm to automatically detect obstacles and thus propose various paths indicating the one with least number of obstacles. Figure B.10 illustrates an overview of the workflow of this project.

Figure B.10: Summary diagram of the entire methodology

This project, despite BioSenCS, is more close to the majority of discussed use cases of combination of CS and ML in Chapter 3 where the objective is mainly to use citizens' contributions to collect and label data to feed and train ML algorithms.

List of Acronyms

AI Artificial Intelligence

ANN Artificial Neural Network

API Application Programming Interface

AR Augmented Reality

AUC Area Under (ROC) Curve

CCS Citizen Cyber-Science

CLC CORINE Land Cover

CNN Convolutional Neural Network

CORINE Co-ordination of Information on the Environment

CPU Central Processing Unit

CS Citizen Science

CV Computer Vision

DEM Digital Elevation Model

DIL Digital Internship and Leadership

DNN Deep Neural Network

EBD eBird Basic Dataset

EXIF Exchangeable Image File

fMRI Functional Magnetic Resonance Imaging

FN False Negative

FOSS Free and Open Source Software

FP False Positive

FPR False Positive Rate

GAMs Generalized Additive Models

GB Gigabyte

GCS Geographic Citizen Science

GIS Geographic Information System

GLMs Generalized Linear Models

GNSS Global Navigation Satellite System

GPS Global Positioning System

HCA Hierarchical Cluster Analysis

IoU Intersection over Union

IT Information Technology

JSDMs Joint Species Distribution Models

JSON JavaScript Object Notation

KDE Kernel Density Analysis

KNN k-Nearest Neighbors

LiDAR Light Detection And Ranging

LOOCV Leave one-out cross validation

LpOCV Leave p-out cross validation

LR Linear Regression

LS Least Squares

LTU Linear Threshold Unit

mAP mean average precision

MARS Multivariate Adaptive Regression Splines

MaxEnt Maximum-entropy

ML Machine Learning

MLE Maximum Likelihood Estimation

MLWIC Machine Learning for Wildlife Image Classification

MRI Magnetic Resonance Imaging

MSE Mean Squared Error

MVT Model-View-Template

NASA National Aeronautics and Space Administration

NB Naive Bayesian

NDVI Normalized Difference Vegetation Index

NLG Natural Language Generation

NLP Natural Language Processing

NN Neural Network

OOB Out-Of-Bag

PCA Principal Component Analysis

POI Points of Interest

PWA Progressive Web Application

R-CNN Region-based Convolutional Neural Network

RAM Random-access memory

ReLU Rectified Linear Unit

RF Random Forest

ROC Receiver Operating Characteristics

SDM Species Distribution Modeling

SED Sampling Event Data

SMOTE Synthetic Minority Oversampling Technique

SVM Support Vector Machines

TanH Tangent Hyperbolic

TN True Negative

TP True Positive

TPR True Positive Rate

UGC User Generated Content

URL Uniform Resource Locator

VC Volunteered Computing

VFI Volunteer Functions Inventory

VGI Volunteered Geographic Information

VT Volunteered Thinking

XML Extensible Markup Language

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In: *12th {usenix} symposium on operating systems design and implementation ({osdi} 16)*. 2016, 265–283.

Abdolmajidi, E., Mansourian, A., Will, J., & Harrie, L. (2015). Matching authority and vgi road networks using an extended node-based matching algorithm. *Geo-Spatial Information Science*, *18*(2-3), 65–80. https://doi.org/10.1080/10095020.2015.1071065

Abraham, A. (2005). Artificial neural networks. *Handbook of measuring system design*.

Adriaens, T., Tricarico, E., Reyserhove, L., De Jesus Cardoso, A, Gervasini, E., Lopez Canizares, C, Mitton, I., Schade, S., Spinelli, F.-A., & Tsiamis, K. (2021). Data-validation solutions for citizen science data on invasive alien species. *Publications Office of the European Union: Luxembourg*. https://doi.org/10.2760/694386

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Alender, B. (2016). Understanding volunteer motivations to participate in citizen science projects: A deeper look at water quality monitoring. *Journal of Science Communication*, *15*(3), A04.

Altartouri, A., & Jolma, A. A naive bayes classifier for modeling distributions of the common reed in southern finland. In: *Modsim2013, proceedings of the 20th international congress on modelling and simulation, adelaide, australia*. 2013, 1–6.

Anderson, D. P., Cobb, J., Korpela, E., Lebofsky, M., & Werthimer, D. (2002). Seti@ home: An experiment in public-resource computing. *Communications of the ACM*, *45*(11), 56–61.

Antoniou, V. (2011). *User generated spatial content: An analysis of the phenomenon and its challenges for mapping agencies* (Doctoral dissertation). UCL (University College London).

Antoniou, V., Fonte, C. C., Minghini, M., See, L., & Skopeliti, A. A guidance tool for vgi contributors. In: *Collection of open conferences in research transport*. 2016.

Antoniou, V., Morley, J., & Haklay, M. (2010). Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon. *GEOMATICA*, *64*(1), 99–110. https://doi.org/10.5623/geomat-2010-0009

Antunes, F., Fonte, C. C., Brovelli, M. A., Minghini, M., Molinari, M. E., Mooney, P., et al. Assessing osm road positional quality with authoritative data. In: *Viii conferência nacional de cartografia e geodésia*. PRT. 2015, 1–8.

Araujo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of biogeography*, *33*(10), 1677–1688.

Aristeidou, M., & Herodotou, C. (2020). Online citizen science: A systematic review of effects on learning and scientific literacy. *Citizen Science: Theory and Practice*, *5*(1), 1–12. https://doi.org/http://doi.org/10.5334/cstp.224

Aristeidou, M., Scanlon, E., & Sharples, M. (2017). Profiles of engagement in online communities of citizen science participation. *Computers in Human Behavior*, *74*, 246–256. https://doi.org/https://doi.org/10.1016/j.chb.2017.04.044

Arthur, C. (n.d.). What is the 1% rule. Retrieved October 20, 2019, from https://www.theguardian.com/technology/2006/jul/20/guardianweeklytechnologysection2

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological review*, *64*(6p1), 359.

Baker, E., Drury, J. P., Judge, J., Roy, D. B., Smith, G. C., & Stephens, P. A. (2021). The verification of ecological citizen science data: Current approaches and future possibilities. *Citizen Science: Theory and Practice*, *6*(1). https://doi.org/http://doi.org/10.5334/cstp.351

Balázs, B., Mooney, P., Nováková, E., Bastin, L., & Arsanjani, J. J. (2021). Data quality in citizen science. *The Science of Citizen Science*, 139.

Baldi, P., & Sadowski, P. J. (2013). Understanding dropout. *Advances in neural information processing systems*, *26*, 2814–2822.

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*, 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Barnard, T. C. (2012). *User profiling using machine learning* (Doctoral dissertation). University of Southampton.

Basic Steps for Your Project Planning | CitizenScience.gov. (n.d.). Retrieved May 20, 2021, from https://www.citizenscience.gov/toolkit/howto/

Batson, C. D., Ahmad, N., & Tsang, J.-A. (2002). Four motives for community involvement. *Journal of social issues*, *58*(3), 429–445.

Bayes' theorem [Page Version ID: 1042113059]. In: In *Wikipedia*. Page Version ID: 1042113059. 2021, September 3. Retrieved September 3, 2021, from https://en.wikipedia.org/w/index.php?title=Bayes%27_theorem&oldid=1042113059

Beaumont, C. N., Goodman, A. A., Kendrew, S., Williams, J. P., & Simpson, R. (2014). The milky way project: Leveraging citizen science and machine learning to detect interstellar bubbles. *The Astrophysical Journal Supplement Series*, *214*(1), 3.

Beberg, A. L., Ensign, D. L., Jayachandran, G., Khaliq, S., & Pande, V. S. Folding@ home: Lessons from eight years of volunteer distributed computing. In: *2009 ieee international symposium on parallel & distributed processing*. IEEE. 2009, 1–8.

Beberg, A. L., Ensign, D. L., Jayachandran, G., Khaliq, S., & Pande, V. S. Folding@home: Lessons from eight years of volunteer distributed computing. In: *2009 ieee international symposium on parallel distributed processing*. 2009, 1–8. https://doi.org/10.1109/IPDPS.2009.5160922.

Berger, A., Della Pietra, S. A., & Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, *22*(1), 39–71.

Berger-Wolf, T. Y., Rubenstein, D. I., Stewart, C. V., Holmberg, J. A., Parham, J., Menon, S., Crall, J., Van Oast, J., Kiciman, E., & Joppa, L. (2017). Wildbook: Crowdsourcing, computer vision, and data science for conservation. *arXiv preprint arXiv:1710.08880*.

Berrar, D. Cross-validation (S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach, Eds.). In: In *Encyclopedia of bioinformatics and computational biology* (S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach, Eds.). Ed. by Ranganathan, S., Gribskov, M., Nakai, K., & Schönbach, C. Oxford: Academic Press, 2019, pp. 542–545. ISBN: 978-0-12-811432-2. https://doi.org/https://doi.org/10.1016/B978-0-12-809633-8.20349-X.

Biraghi, C. A., Lotfian, M., Carrion, D., & Brovelli, M. A. (2021). Ai in support to water quality monitoring. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLIII-B4-2021*, 167–174. https://doi.org/10.5194/isprs-archives-XLIII-B4-2021-167-2021

Boakes, E. H., Gliozzo, G., Seymour, V., Harvey, M., Smith, C., Roy, D. B., & Haklay, M. (2016). Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. *Scientific reports*, *6*(1), 1–11.

Bonafilia, D., Gill, J., Basu, S., & Yang, D. Building high resolution maps for humanitarian aid and development with weakly- and semi-supervised learning. In: *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr) workshops*. 2019.

Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., & Wilderman, C. C. (2009a). Public participation in scientific research: Defining the field and assessing its potential for informal science education. a caise inquiry group report. *Online Submission*.

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009b). Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, *59*(11), 977–984. https://doi.org/10.1525/bio.2009.59.11.9

Bonter, D. N., & Cooper, C. B. (2012). Data validation in citizen science: A case study from project feederwatch. *Frontiers in Ecology and the Environment*, *10*(6), 305–307.

Borji, A., & Itti, L. Human vs. computer in scene and object recognition. In: *Proceedings of the ieee conference on computer vision and pattern recognition*. 2014, 113–120.

Botella, C., Joly, A., Bonnet, P., Monestiez, P., & Munoz, F. A deep learning approach to species distribution modelling. In: *Multimedia tools and applications for environmental*

& biodiversity informatics. Springer, 2018, pp. 169–199. https://doi.org/10.1007/978-3-319-76445-0_10.

Bowser, A., Hansen, D., He, Y., Boston, C., Reid, M., Gunnell, L., & Preece, J. Using gamification to inspire new citizen science volunteers. In: Gamification '13. Association for Computing Machinery, 2013, 18–25. ISBN: 9781450328159. https://doi.org/10.1145/2583008.2583011.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/https://doi.org/10.1016/S0031-3203(96)00142-2

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Routledge. https://doi.org/https://doi.org/10.1201/9781315139470

Brovelli, M., Cannata, M, & Rogora, M. (2019). Simile, a geospatial enabler of the monitoring of sustainable development goal 6 (ensure availability and sustainability of water for all). *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*.

Brovelli, M. A., & Zamboni, G. (2018). A new method for the assessment of spatial accuracy and completeness of openstreetmap building footprints. *ISPRS International Journal of Geo-Information*, *7*(8). https://doi.org/10.3390/ijgi7080289

Buck, J. L., & Finner, S. L. (1985). A still further note on freeman's measure of association. *Psychometrika*, *50*(3), 365–366.

Budhathoki, N. R., & Haythornthwaite, C. (2013). Motivation for open collaboration: Crowd and community models and the case of openstreetmap. *American Behavioral Scientist*, *57*(5), 548–575.

Budhathoki, N. R. (2010). *Participants' motivations to contribute geographic information in an online community*. University of Illinois at Urbana-Champaign.

Butcher, G. S., Fuller, M. R., McAllister, L. S., & Geissler, P. H. (1990). An evaluation of the christmas bird count for monitoring population trends of selected species. *Wildlife Society Bulletin (1973-2006)*, *18*(2), 129–134.

Büttner, G. Corine land cover and land cover change products. In: *Land use and land cover mapping in europe*. Springer, 2014, pp. 55–74.

Carley, K. M., Malik, M., Landwehr, P. M., Pfeffer, J., & Kowalchuck, M. (2016). Crowd sourcing disaster management: The complex nature of twitter usage in padang indonesia. *Safety science*, *90*, 48–61.

Ceccaroni, L., Bibby, J., Roger, E., Flemons, P., Michael, K., Fagan, L., & Oliver, J. L. (2019). Opportunities and risks for citizen science in the age of artificial intelligence. *Citizen Science: Theory and Practice*, *4*(1).

Chassin, T., & Ingensand, J. (2021). ARE CITY FEATURES INFLUENCING THE BEHAVIOR OF PHOTOGRAPHERS? AN ANALYSIS OF GEO-REFERENCED PHOTOS SHOOTING ORIENTATION. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLIII-B4-2021*, 353–359. https://doi.org/10.5194/isprs-archives-xliii-b4-2021-353-2021

Chatfield, A. T., & Brajawidagda, U. Twitter early tsunami warning system: A case study in indonesia's natural disaster management. In: *2013 46th hawaii international conference on system sciences*. IEEE. 2013, 2050–2060.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chen, X., Elmes, G., Ye, X., & Chang, J. (2016). Implementing a real-time twitter-based system for resource dispatch in disaster management. *GeoJournal*, *81*(6), 863–873.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51–89.

Clary, E. G., Snyder, M., Ridge, R. D., Copeland, J., Stukas, A. A., Haugen, J., & Miene, P. (1998). Understanding and assessing the motivations of volunteers: A functional approach. *Journal of personality and social psychology*, *74*(6), 1516.

Classification Tree. (2021). Retrieved December 31, 2021, from https://support.bccvl.org.au/support/solutions/articles/6000083204-classification-tree

Coetzee, S. M., Minghini, M., Solis, P., Rautenbach, V., & Green, C. (2018). Towards understanding the impact of mapathons-reflecting on youthmappers experiences.

Cohn, J. P. (2008). Citizen Science: Can Volunteers Do Real Research? *BioScience*, *58*(3), 192–197. https://doi.org/10.1641/B580303

Comber, S. (2021, August 14). *Spacv: Spatial cross-validation in python* [original-date: 2020-06-12T08:07:15Z]. Retrieved September 9, 2021, from https://github.com/SamComber/spacv

Conroy, G. (n.d.). How to run a successful citizen science project. Retrieved December 8, 2021, from https://www.natureindex.com/news-blog/how-to-run-successful-citizen-science-project

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z., et al. (2010). Predicting protein structures with a multiplayer online game. *Nature*, *466*(7307), 756–760.

Cope, B., Kalantzis, M., & Searsmith, D. (2021). Artificial intelligence for education: Knowledge and its assessment in ai-enabled learning ecologies. *Educational Philosophy and Theory*, *53*(12), 1229–1245. https://doi.org/10.1080/00131857.2020.1728732

Cox, J., Oh, E. Y., Simmons, B., Lintott, C., Masters, K., Greenhill, A., Graham, G., & Holmes, K. (2015). Defining and measuring success in online citizen science: A case study of zooniverse projects. *Computing in Science Engineering*, *17*(4), 28–41. https://doi.org/10.1109/MCSE.2015.65

Cross Validation and Model Selection. (n.d.). Retrieved November 30, 2021, from https://biol607.github.io/lectures/crossvalidation.html#20

Curtis, V. (2015a). Motivation to participate in an online citizen science game: A study of foldit. *Science Communication*, *37*(6), 723–746.

Curtis, V. (2015b). *Online citizen science projects: An exploration of motivation, contribution and participation*. Open University (United Kingdom).

Curtis, V. (2018). Patterns of participation and motivation in folding@ home: The contribution of hardware enthusiasts and overclockers. *Citizen Science: Theory and Practice*, *3*(1), 1–14.

Dalby, O., Sinha, I., Unsworth, R. K. F., McKenzie, L. J., Jones, B. L., & Cullen-Unsworth, L. C. (2021). Citizen science driven big data collection requires improved and inclusive societal engagement. *Frontiers in Marine Science*, *8*, 432. https://doi.org/10.3389/fmars.2021.610397

Dale, R. (2021). Gpt-3: What's it good for? *Natural Language Engineering*, *27*(1), 113–118. https://doi.org/10.1017/S1351324920000601

De la Calzada, G., & Dekhtyar, A. On measuring the quality of wikipedia articles. In: *Proceedings of the 4th workshop on information credibility*. WICOW '10. Association for Computing Machinery, 2010, 11–18. ISBN: 9781605589404. https://doi.org/10.1145/1772938.1772943.

De Moor, T., Rijpma, A., & Prats López, M. (2019). Dynamics of engagement in citizen science: Results from the "yes, i do!" project. *Citizen Science: Theory and Practice*, *4*(1), 1–17. https://doi.org/http://doi.org/10.5334/cstp.212

Debeljak, M., Cortet, J., Demšar, D., Krogh, P. H., & Džeroski, S. (2007). Hierarchical classification of environmental factors and agricultural practices affecting soil fauna under cropping systems using bt maize. *Pedobiologia*, *51*(3), 229–238.

Deep learning [Page Version ID: 1060417519]. (2021, December). Retrieved December 21, 2021, from https://en.wikipedia.org/w/index.php?title=Deep_learning&oldid=1060417519

Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., & Joly, A. (2021). Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS computational biology*, *17*(4), e1008856. https://doi.org/https://doi.org/10.1371/journal.pcbi.1008856

Deng, D.-P., Chuang, T.-R., Shao, K.-T., Mai, G.-S., Lin, T.-E., Lemmens, R., Hsu, C.-H., Lin, H.-H., & Kraak, M.-J. Using social media for collaborative species identification and occurrence: Issues, methods, and tools. In: *Proceedings of the 1st acm sigspatial international workshop on crowdsourced and volunteered geographic information*. 2012, 22–29.

Devaraj, A., Murthy, D., & Dontula, A. (2020). Machine-learning methods for identifying social media-based requests for urgent help during hurricanes. *International Journal of Disaster Risk Reduction*, *51*, 101757.

Di Minin, E., Fink, C., Hiippala, T., & Tenkanen, H. (2019). A framework for investigating illegal wildlife trade on social media with machine learning. *Conservation Biology*, *33*(1), 210.

Ding, Z., Hu, H., Cadotte, M. W., Liang, J., Hu, Y., & Si, X. (2021). Elevational patterns of bird functional and phylogenetic structure in the central himalaya. *Ecography*, *44*(9), 1403–1417.

DiNucci, D. (1999). Fragmented future. print magazine, 53 (4), 32, 221-222.

Django's Structure – A Heretic's Eye View - Python Django. (n.d.). Retrieved October 21, 2021, from https://djangobook.com/mdj2-django-structure/

Domroese, M. C., & Johnson, E. A. (2017). Why watch bees? motivations of citizen science volunteers in the great pollinator project. *Biological Conservation*, *208*, 40–47.

Drake, J. M., Randin, C., & Guisan, A. (2006). Modelling ecological niches with support vector machines. *Journal of applied ecology*, *43*(3), 424–432.

Dutta, P., Aoki, P. M., Kumar, N., Mainwaring, A., Myers, C., Willett, W., & Woodruff, A. Common sense: Participatory urban sensing using a network of handheld air quality monitors. In: *Proceedings of the 7th acm conference on embedded networked sensor systems*. 2009, 349–350.

Džeroski, S., & Drumm, D. (2003). Using regression trees to identify the habitat preference of the sea cucumber (holothuria leucospilota) on rarotonga, cook islands. *Ecological modelling*, *170*(2-3), 219–226.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of maxent for ecologists. *Diversity and distributions*, *17*(1), 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

ElQadi, M. M., Dorin, A., Dyer, A., Burd, M., Bukovac, Z., & Shrestha, M. (2017). Mapping species distributions with social media geo-tagged images: Case studies of bees and flowering plants in australia. *Ecological Informatics*, *39*, 23–31. https://doi.org/https://doi.org/10.1016/j.ecoinf.2017.02.006

Engels, G., Aranjuelo, N., Arganda-Carreras, I., Nieto, M., & Otaegui, O. (2020). 3d object detection from lidar data using distance dependent feature extraction. *Proceedings of the 6th International Conference on Vehicle Technology and Intelligent Transport Systems*. https://doi.org/10.5220/0009330402890300

Estellés-Arolas, E., & González-Ladrón-de Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information science*, *38*(2), 189–200.

Fan, G.-Z., Ong, S. E., & Koh, H. C. (2006). Determinants of house price: A decision tree approach. *Urban Studies*, *43*(12), 2301–2315.

Farnaghi, M., Ghaemi, Z., & Mansourian, A. (2020). Dynamic spatio-temporal tweet mining for event detection: A case study of hurricane florence. *International Journal of Disaster Risk Science*, *11*, 378–393. https://doi.org/10.1007/s13753-020-00280-z

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Franzen, M., Kloetzer, L., Ponti, M., Trojan, J., & Vicens, J. (2021). Machine learning in citizen science: Promises and implications. *The Science of Citizen Science*, 183.

Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, *61*(3), 399–409.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 1–67.

Frisch, M. B., & Gerrard, M. (1981). Natural helping systems: A survey of red cross volunteers. *American Journal of Community Psychology*, *9*(5), 567.

Fritz, S., See, L., & Brovelli, M. (2017). Motivating and sustaining participation in vgi.

Gamification - OpenStreetMap Wiki. (n.d.). Retrieved September 30, 2020, from https://wiki.openstreetmap.org/wiki/Gamification

Ganzevoort, W., van den Born, R. J., Halffman, W., & Turnhout, S. (2017). Sharing biodiversity data: Citizen scientists' concerns and motivations. *Biodiversity and Conservation*, *26*(12), 2821–2837.

Gao, B., & Pavel, L. (2017). On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*.

Geller, E. H. (n.d.). Bayes' Rule and Bomb Threats. Retrieved November 29, 2021, from https://www.psychologyinaction.org/psychology-in-action-1/2012/10/22/bayes-rule-and-bomb-threats

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the ieee conference on computer vision and pattern recognition*. 2014, 580–587.

Göbel, C., Nold, C., Berditchevskaia, A., & Haklay, M. (2019). How does citizen science" do" governance? reflections from the ditos project. *Citizen Science: Theory and Practice*, *4*(1). https://doi.org/http://doi.org/10.5334/cstp.204

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221.

Green, C., Rautenbach, V., & Coetzee, S. M. (2019). Evaluating student motivation and productivity during mapathons.

Green, S. E., Rees, J. P., Stephens, P. A., Hill, R. A., & Giordano, A. J. (2020). Innovations in camera trapping technology and approaches: The integration of citizen science and artificial intelligence. *Animals*, *10*(1), 132.

Grinnell, J. (1917). Field tests of theories concerning distributional control. *The American Naturalist*, *51*(602), 115–128.

Grinnell, J. (1924). Geography and evolution. *Ecology*, *5*, 225–229. https://doi.org/10.2307/1929447

Guillaume, G., Can, A., Petit, G., Fortin, N., Palominos, S., Gauvreau, B., Bocher, E., & Picaut, J. (2016). Noise mapping based on participative measurements. *Noise Mapping*, *3*(1).

Guisan, A., Edwards Jr, T. C., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological modelling*, *157*(2-3), 89–100.

Haklay, M. E. Why is participation inequality important? In: Ubiquity Press, 2016.

Haklay, M. (2013). Citizen science and volunteered geographic information: Overview and typology of participation. *Crowdsourcing geographic knowledge*, 105–122.

Haklay, M., Fraisl, D., Greshake Tzovaras, B., Hecker, S., Gold, M., Hager, G., Ceccaroni, L., Kieslinger, B., Wehn, U., Woods, S., et al. (2021). Contours of citizen science: A vignette study. *Royal Society open science*, *8*(8), 202108. https://doi.org/http://doi.org/10.1098/rsos.202108

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), 8–12.

Han, H., Guo, X., & Yu, H. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In: *2016 7th ieee international conference on software engineering and service science (icsess)*. IEEE. 2016, 219–224.

Han, J., & Moraga, C. The influence of the sigmoid function parameters on the speed of backpropagation learning. In: *International workshop on artificial neural networks*. Springer. 1995, 195–201.

Hara, Y. (2015). Behaviour analysis using tweet data and geo-tag data in a natural disaster [Transport Survey Methods: Embracing Behavioural and Technological Changes Selected contributions from the 10th International Conference on Transport Survey Methods 16-21 November 2014, Leura, Australia]. *Transportation Research Procedia*, *11*, 399–412. https://doi.org/https://doi.org/10.1016/j.trpro.2015.12.033

Hars, A, & Ou, S. Working for free?–motivations of participating in open source projects; 2001. In: *34th annual hawaii international conference on system sciences (hicss-34), havaí*, 25–39.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, *44*(1), 1–12.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263–1284.

He, H., & Ma, Y. (2013). Imbalanced learning: Foundations, algorithms, and applications.

Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., & Bonn, A. (2018). *Citizen science: Innovation in open science, society and policy*. UCL Press.

Hill, W. C., Hollan, J. D., Wroblewski, D., & McCandless, T. Edit wear and read wear. In: *Proceedings of the sigchi conference on human factors in computing systems*. 1992, 3–9.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(02), 107–116.

Holliman, R., & Curtis, V. (2015). Online media.

Horning, N. et al. Random forests: An algorithm for image classification and generation of continuous fields data sets. In: *Proceedings of the international conference on geoinformatics for spatial infrastructure development in earth and allied sciences, osaka, japan. 911*. 2010.

Hossain, M. Users' motivation to participate in online crowdsourcing platforms. In: *2012 international conference on innovation management and technology research*. IEEE. 2012, 310–315.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. Searching for mobilenetv3. In: *Proceedings of the ieee/cvf international conference on computer vision*. 2019, 1314–1324.

Howe, J. et al. (2006). The rise of crowdsourcing. *Wired magazine*, *14*(6), 1–4.

Hristova, D., Quattrone, G., Mashhadi, A., & Capra, L. The life of the party: Impact of social mapping in openstreetmap. In: *Seventh international aaai conference on weblogs and social media*. 2013.

Hsing, P.-Y., Bradley, S., Kent, V. T., Hill, R. A., Smith, G. C., Whittingham, M. J., Cokill, J., Crawley, D., Volunteers, M., & Stephens, P. A. (2018). Economical crowdsourcing for camera trap image classification. *Remote Sensing in Ecology and Conservation*, *4*(4), 361–374.

Hubbell, S. P. (2005). Neutral theory in community ecology and the hypothesis of functional equivalence. *Functional ecology*, *19*(1), 166–172.

Iacovides, I., Jennett, C., Cornish-Trestrail, C., & Cox, A. L. Do games attract or sustain engagement in citizen science? a study of volunteer motivations. In: *Chi'13 extended abstracts on human factors in computing systems*. 2013, pp. 1101–1106.

Ingensand, J., Lotfian, M., Ertz, O., & Piot, D. (2018). Augmented reality technologies for biodiversity education – a case study, 5.

Ingensand, J., Nappez, M., Joost, S., Widmer, I., Ertz, O., & Rappo, D. The urbangene project: Experience from a crowdsourced mapping campaign. In: *2015 1st international conference on geographical information systems theory, applications and management (gistam)*. IEEE. 2015, 1–7.

Introduction to Species Distribution Models. (n.d.). Retrieved August 9, 2021, from https://support.bccvl.org.au/support/solutions/articles/6000127048-introduction-to-species-distribution-models

Irwin, A. (1995). *Citizen science: A study of people, expertise and sustainable development*. Routledge. https://doi.org/https://doi.org/10.4324/9780203202395

Jay, C., Dunne, R., Gelsthorpe, D., & Vigo, M. To sign up, or not to sign up? maximizing citizen science contribution rates through optional registration. In: CHI '16. Association for Computing Machinery, 2016, 1827–1832. ISBN: 9781450333627. https://doi.org/10.1145/2858036.2858319.

Jiménez, M., Torres, M. T., John, R., & Triguero, I. (2020). Galaxy image classification based on citizen science data: A comparative study. *IEEE Access*, *8*, 47232–47246.

Johnson, B. A. How an augmented reality game (pokémon go) affected volunteer contributions to openstreetmap. In: *Proc. ica. 2*. 2019, 1–4.

Johnston, A., Fink, D., Hochachka, W. M., & Kelling, S. (2018). Estimates of observer expertise improve species distributions from citizen science data. *Methods in Ecology and Evolution*, *9*(1), 88–97.

Joppa, L. N. (2017). The case for technology investments in the environment.

Joshi, S., Randall, N., Chiplunkar, S., Wattimena, T., & Stavrianakis, K. 'we'-a robotic system to extend social impact of community gardens. In: *Companion of the 2018 acm/ieee international conference on human-robot interaction*. 2018, 349–350.

Kampichler, C., Džeroski, S., & Wieland, R. (2000). Application of machine learning techniques to the analysis of soil ecological data bases: Relationships between habitat features and collembolan community characteristics. *Soil Biology and Biochemistry*, *32*(2), 197–209.

Kamptner, E., & Kessler, F. (2019). Small-scale crisis response mapping: Comparing user contributions to events in openstreetmap. *GeoJournal*, *84*(5), 1165–1185.

Kanoje, S., Mukhopadhyay, D., & Girase, S. (2016). User profiling for university recommender system using automatic information retrieval. *Procedia Computer Science*, *78*, 5–12.

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)*, *52*(4), 1–36.

Kelling, S., Johnston, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Bonn, A., Fernandez, M., Hochachka, W. M., Julliard, R., Kraemer, R., et al. (2018). Finding the signal in the noise of citizen science observations. *bioRxiv*, 326314.

Kelling, S., Johnston, A., Hochachka, W. M., Iliff, M., Fink, D., Gerbracht, J., Lagoze, C., La Sorte, F. A., Moore, T., Wiggins, A., et al. (2015). Can observation skills of citizen scientists be estimated using species accumulation curves? *PloS one*, *10*(10), e0139600. https://doi.org/10.1371/journal.pone.0139600

Kelling, S., Yu, J., Gerbracht, J., & Wong, W.-K. Emergent filters: Automated data verification in a large-scale citizen science project. In: *2011 ieee seventh international conference on e-science workshops*. IEEE. 2011, 20–27.

Kendrew, S., Simpson, R, Bressert, E., Povich, M. S., Sherman, R., Lintott, C., Robitaille, T. P., Schawinski, K., & Wolf-Chase, G. (2012). The milky way project: A statistical study of massive star formation associated with infrared bubbles. *The Astrophysical Journal*, *755*(1), 71.

Keshavan, A., Yeatman, J. D., & Rokem, A. (2019). Combining citizen science and deep learning to amplify expertise in neuroimaging. *Frontiers in neuroinformatics*, *13*, 29.

Khanal, K., Budhathoki, N. R., & Erbstein, N. (2019). Filling openstreetmap data gaps in rural nepal: A digital youth internship and leadership programme. *Open Geospatial Data, Software and Standards*, *4*(1), 1–10.

Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popović, Z., Baker, D., & Players, F. (2011). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*, *108*(47), 18949–18953. https://doi.org/10.1073/pnas.1115898108

Kienast, F, Bolliger, J, & Zimmermann, N. E. (2012). *Species distribution modeling (sdm) with glm, gam and cart dependent vs. independent variables: A conceptual ecological view.*

Kirch, W. (2008). Pearson's correlation coefficient. *Encyclopedia of Public Health*, 1090–1091.

Kisilevich, S., Krstajic, M., Keim, D., Andrienko, N., & Andrienko, G. Event-based analysis of people's activities and behavior using flickr and panoramio geotagged photo collections. In: *2010 14th international conference information visualisation*. IEEE, 2010, July. https://doi.org/10.1109/iv.2010.94.

Kobler, A., & Adamic, M. Brown bears in slovenia: Identifying locations for construction of wildlife bridges across highways. In: *Proceedings of international conference on ecology and transportation. available from http://www. icoet. net/icowet/99proceedings. asp (accessed january 2008)*. 1999.

Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016a). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, *14*(10), 551–560. https://doi.org/https://doi.org/10.1002/fee.1436

Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016b). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, *14*(10), 551–560.

Krumm, J., Davies, N., & Narayanaswami, C. (2008). User-generated content. *IEEE Pervasive Computing*, *7*(4), 10–11. https://doi.org/10.1109/MPRV.2008.85

Kuhn, M., Johnson, K. et al. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Kullenberg, C., & Kasperowski, D. (2016). What is citizen science?–a scientometric meta-analysis. *PloS one*, *11*(1), e0147152.

Kumar, A. C., & Bhandarkar, S. M. A deep learning paradigm for detection of harmful algal blooms. In: *2017 ieee winter conference on applications of computer vision (wacv)*. 2017, 743–751. https://doi.org/10.1109/WACV.2017.88.

Land-Zandstra, A. M., Devilee, J. L. A., Snik, F., Buurmeijer, F., & van den Broek, J. M. (2016). Citizen science on a smartphone: Participants' motivations and learning [PMID: 26346340]. *Public Understanding of Science*, *25*(1), 45–60. https://doi.org/10.1177/0963662515602406

Langenkämper, D., Simon-Lledó, E., Hosking, B., Jones, D. O., & Nattkemper, T. W. (2019). On the impact of citizen science-derived data quality on deep learning based classification in marine images. *Plos one*, *14*(6), e0218086.

Leibold, M. A. (1995). The niche concept revisited: Mechanistic models and community context. *Ecology*, *76*(5), 1371–1382.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, *18*(1), 559–563.

Leocadio, J. N., Ghilardi-Lopes, N. P., Koffler, S., Barbiéri, C., Francoy, T. M., Albertini, B., & Saraiva, A. M. (2021). Data reliability in a citizen science protocol for monitoring stingless bees flight activity. *Insects*, *12*(9). https://doi.org/10.3390/insects12090766

Leung, K. M. (2007). Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, *2007*, 123–156.

Li, X., & Wang, Y. (2013). Applying various algorithms for species distribution modelling. *Integrative Zoology*, *8*, 124–135. https://doi.org/10.1111/1749-4877.12000

Lintott, C, Forston, L, Smith, A, et al. (2013). Zooniverse.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., et al. (2008). Galaxy zoo: Morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, *389*(3), 1179–1189.

Liu, S. (2019). Leave-p-out cross-validation test for uncertain verhulst-pearl model with imprecise observations. *IEEE Access*, *7*, 131705–131709.

Lotfian, M., & Ingensand, J. (2021). Using geo geo-tagged flickr images to explore the correlation between land cover classes and the location of bird observations. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLIII-B4-2021*, 189–194. https://doi.org/10.5194/isprs-archives-XLIII-B4-2021-189-2021

LOTFIAN, M. (2016). *Urban climate modeling: Case study of milan city* (Master's thesis). Politecnico di Milano.

Lotfian, M., Ingensand, J., & Brovelli, M. A. (2020). A framework for classifying participant motivation that considers the typology of citizen science projects. *ISPRS International Journal of Geo-Information*, *9*(12), 704.

Lotfian, M., Ingensand, J., & Brovelli, M. A. (2021). The partnership of citizen science and machine learning: Benefits, risks, and future challenges for engagement, data collection, and data quality. *Sustainability*, *13*(14). https://doi.org/10.3390/su13148087

Lotfian, M., Ingensand, J., Ertz, O., Composto, S., Oberson, M., Oulevay, S., Campisi, D., & Joerin, F. (2018). *Participants' motivations to contribute to biodiversity citizen science projects* (tech. rep.). PeerJ Preprints.

Lotfian, M., Ingensand, J., Ertz, O., Oulevay, S., & Chassin, T. Auto-filtering validation in citizen science biodiversity monitoring. In: *Proceedings of the ica; proceedings of 29th international cartographic conference (icc 2019), 15–20 july 2019, tokyo, japan.* (CONFERENCE). 15-20 July 2019. 2019.

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models.

Mac Aodha, O., Gibb, R., Barlow, K. E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G. R., Newson, S. E., et al. (2018). Bat detective—deep learning tools for bat acoustic signal detection. *PLoS computational biology*, *14*(3), e1005995.

MacKerron, G., & Mourato, S. (2013). Happiness is greater in natural environments. *Global environmental change*, *23*(5), 992–1000.

MacLeod, C. J., & Scott, K. (2021). Mechanisms for enhancing public engagement with citizen science results. *People and Nature*, *3*(1), 32–50. https://doi.org/https://doi.org/10.1002/pan3.10152

Maisonneuve, N., Stevens, M., & Ochab, B. (2010). Participatory noise pollution monitoring using mobile phones. *Information polity*, *15*(1, 2), 51–71.

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Maxent [ARDC support]. (n.d.). Retrieved September 22, 2021, from https://support.bccvl.org.au/support/solutions/articles/6000083216-maxent

McClure, E. C., Sievers, M., Brown, C. J., Buelow, C. A., Ditria, E. M., Hayes, M. A., Pearson, R. M., Tulloch, V. J., Unsworth, R. K., & Connolly, R. M. (2020). Artificial intelligence meets citizen science to supercharge ecological monitoring. *Patterns*, *1*(7), 100109.

McShea, W. J., Forrester, T., Costello, R., He, Z., & Kays, R. (2016). Volunteer-run cameras as distributed sensors for macrosystem mammal research. *Landscape Ecology*, *31*(1), 55–66.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, *28*(1), 92–122. https://doi.org/10.1007/s10618-012-0295-5

Mohammed, R., Rawashdeh, J., & Abdullah, M. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In: *2020 11th international conference on information and communication systems (icics)*. IEEE. 2020, 243–248.

Mondzech, J., & Sester, M. (2011). Quality analysis of openstreetmap data based on application needs. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *46*(2), 115–125. https://doi.org/https://doi.org/10.3138/carto.46.2.115

Monti, L., Vincenzi, M., Mirri, S., Pau, G., & Salomoni, P. (2020). Raveguard: A noise monitoring platform using low-end microphones and machine learning. *Sensors*, *20*(19), 5583.

Mooney, P., & Corcoran, P. (2012). Characteristics of heavily edited objects in openstreetmap. *Future Internet*, *4*(1), 285–305.

Moradi, M. (2020). *Evaluating the quality of osm roads and buildings in the québec province* (Master's thesis). Université Laval.

Nielsen, J. (n.d.). Participation Inequality: The 90-9-1 Rule for Social Features. Retrieved October 20, 2019, from https://www.nngroup.com/articles/participation-inequality/

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 25). Determination press San Francisco, CA.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, *115*(25), E5716–E5725.

Nov, O., Anderson, D., & Arazy, O. Volunteer computing: A model of the factors determining contribution to community-based scientific research. In: *Proceedings of the 19th international conference on world wide web*. WWW '10. Association for Computing Machinery, 2010, 741–750. ISBN: 9781605587998. https://doi.org/10.1145/1772690.1772766.

Nov, O., Arazy, O., & Anderson, D. Dusting for science: Motivation and participation of digital citizen science volunteers. In: *Proceedings of the 2011 iconference*. 2011, pp. 68–74.

Ogris, N, & Jurc, M. Potential changes in the distribution of maple species (acer pseudoplatanus, a. campestre, a. platanoides, a. obtusatum) due to climate change in slovenia. In: *Proceedings of the symposium on climate change influences on forests and forestry*. 2007.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika*, *47*(3), 337–347.

O'reilly, T. (2005). What is web 2.0: Design patterns and business models for the next generation of software. Retrieved November 26, 2021, from https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

Paeglis, A., Strumfs, B., Mezale, D., & Fridrihsone, I. A review on machine learning and deep learning techniques applied to liquid biopsy. In: *Liquid biopsy*. IntechOpen, 2018.

Panou, D. N., & Reczko, M. (2020). Deepfoldit–a deep reinforcement learning neural network folding proteins. *arXiv preprint arXiv:2011.03442*.

Parham, J., Stewart, C., Crall, J., Rubenstein, D., Holmberg, J., & Berger-Wolf, T. An animal detection pipeline for identification. In: *2018 ieee winter conference on applications of computer vision (wacv)*. IEEE. 2018, 1075–1083.

Park, J., Krishna, R., Khadpe, P., Fei-Fei, L., & Bernstein, M. Ai-based request augmentation to increase crowdsourcing participation. In: *Proceedings of the aaai conference on human computation and crowdsourcing. 7.* (1). 2019, 115–124.

PARTHENOS. (n.d.). Challenges of conducting a Citizen Science project – Parthenos training. Retrieved November 25, 2021, from https://training.parthenos-project.eu/sample-page/citizen-science-in-the-digital-arts-and-humanities/what-is-citizen-science/challenges-of-conducting-a-citizen-science-project/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., Lyons, T., Manyika, J., Mishra, S., & Niebles, J. C. (2019). The ai index 2019 annual report. *AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA*.

Petersen, C., Austin, R. R., Backonja, U., Campos, H., Chung, A. E., Hekler, E. B., Hsueh, P.-Y. S., Kim, K. K., Pho, A., Salmi, L., Solomonides, A., & Valdez, R. S. (2019). Citizen science to further precision medicine: from vision to implementation. *JAMIA Open*, *3*(1), 2–8. https://doi.org/10.1093/jamiaopen/ooz060

Pettibone, L., Vohland, K., & Ziegler, D. (2017). Understanding the (inter) disciplinary and institutional diversity of citizen science: A survey of current practice in germany and austria. *PloS one*, *12*(6), e0178778.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, *190*(3-4), 231–259.

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with maxent: New extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

Pi, D. V. (2016). *Applying species distribution models in conservation biology= aplicacions dels models de distribució d'espècies en biologia de la conservació* (Doctoral dissertation). Universitat De Barcelona.

Popenici, S. A., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, *12*(1), 1–13.

Preece, J., & Shneiderman, B. (2009). The reader-to-leader framework: Motivating technology-mediated social participation. *AIS transactions on human-computer interaction*, *1*(1), 13–32.

Produit, T., & Ingensand, J. 3d georeferencing of historical photos by volunteers. In: *The annual international conference on geographic information science*. Springer. 2018, 113–128.

Pullin, A. S. (2002). *Conservation biology*. Cambridge University Press. https://doi.org/10.1017/CBO9781139051927

Quill, T. M. (2018). Humanitarian mapping as library outreach: A case for community-oriented mapathons. *Journal of Web Librarianship*, *12*(3), 160–168.

Quintero, I., & Jetz, W. (2018). Global elevational diversity and diversification of birds. *Nature*, *555*(7695), 246–250.

Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C. J., Cardamone, C., Murray, P., Schawinski, K., Szalay, A. S., & Vandenberg, J. (2013). Galaxy zoo: Motivations of citizen scientists. *arXiv preprint arXiv:1303.6886*.

Raes, N., Aguirre-Gutiérrez, J., Hoorn, C, Perrigo, A, & Antonelli, A. (2018). Modeling framework to estimate and project species distributions space and time. *Mountains, climate and biodiversity*, 309.

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Swish: A self-gated activation function. *arXiv preprint arXiv:1710.05941*, *7*, 1.

Redford, K. H., Coppolillo, P., Sanderson, E. W., Da Fonseca, G. A., Dinerstein, E., Groves, C., Mace, G., Maginnis, S., Mittermeier, R. A., Noss, R., et al. (2003). Mapping the conservation landscape. *Conservation biology*, *17*(1), 116–131.

Reed, J., Raddick, M. J., Lardner, A., & Carney, K. An exploratory factor analysis of motivations for participating in zooniverse, a collection of virtual citizen science projects. In: *2013 46th hawaii international conference on system sciences*. IEEE. 2013, 610–619.

European Commission and Directorate-General for Environment. (2018). *Citizen science for environmental policy : Development of an eu-wide inventory and analysis of selected practices*. Publications Office. https://doi.org/doi/10.2779/961304

ReLU6 — PyTorch 1.10.0 documentation. (n.d.). Retrieved November 30, 2021, from https://pytorch.org/docs/stable/generated/torch.nn.ReLU6.html

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*, 91–99.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929.

Robinson, A. J., & Voronkov, A. (2001). *Handbook of automated reasoning* (Vol. 1). Elsevier.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rotman, D., Hammock, J., Preece, J., Hansen, D., Boston, C., Bowser, A., & He, Y. (2014). Motivations affecting initial and long-term participation in citizen science projects in three countries. *IConference 2014 Proceedings*.

Roy, H., Pocock, M., Preston, C., Roy, D., Savage, J, Tweddle, J., & Robinson, L. (2012). Understanding citizen science & environmental monitoring. final report on behalf of uk-eof. nerc centre for ecology & hydrology and natural history museum. *Natural History Museum, London, UK. See http://www. ukeof. org. uk/co citizen. aspx (accessed 28/11/2012)*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, *25*(1), 54–67.

Rzanny, M., Seeland, M., Wäldchen, J., & Mäder, P. (2017). Acquiring and preprocessing leaf images for automated plant identification: Understanding the tradeoff between effort and information gain. *Plant methods*, *13*(1), 1–11.

Sagiroglu, S., & Sinanc, D. Big data: A review. In: *2013 international conference on collaboration technologies and systems (cts)*. IEEE. 2013, 42–47.

Samantaray, A., Yang, B., Dietz, J. E., & Min, B.-C. (2018). Algae detection using computer vision and deep learning. *arXiv preprint arXiv:1811.10847*.

Saponara, S., Elhanashi, A., & Gagliardi, A. (2021). Implementing a real-time, ai-based, people detection and social distancing measuring system for covid-19. *Journal of Real-Time Image Processing*, 1–11.

Sawhney, R. (2021). Can artificial intelligence make software development more productive? *LSE Business Review*.

Schade, S., & Tsinaraki, C. (2016). Survey report: Data management in citizen science projects. (LB-NA-27920-EN-N (online),LB-NA-27920-EN-E (ePub)). https://doi.org/10.2788/539115(online),10.2788/00005(ePub)

Schade, S., Tsinaraki, C., Manzoni, M., Berti Suman, A., Spinelli, F. A., Mitton, I., Kotsev, A., Delipetrev, B., & Fullerton, K. T. (2020). Activity report on citizen science ? discoveries from a five year journey. *Publications Office of the European Union, Luxembourg*. https://doi.org/10.2760/841551

See, L., Estima, J., Pődör, A., Arsanjani, J. J., Bayas, J.-C. L., & Vatseva, R. (2017). Sources of vgi for mapping. *Mapping and the Citizen Sensor*, 13–35. https://doi.org/https://doi.org/10.5334/bbf.b

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.-Y., Milčinski, G., Nikšič, M., Painho, M., Pődör, A., Olteanu-Raimond, A.-M., & Rutzinger, M. (2016). Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic in-

formation. *ISPRS International Journal of Geo-Information*, *5*(5). https://doi.org/10.3390/ijgi5050055

Senay, S. D., Worner, S. P., & Ikeda, T. (2013). Novel three-step pseudo-absence selection technique for improved species distribution modelling. *PLoS ONE*, *8*, 71218. https://doi.org/10.1371/journal.pone.0071218

Sessility (motility) [Page Version ID: 1036919706]. (2021, August). Retrieved November 29, 2021, from https://en.wikipedia.org/w/index.php?title=Sessility_(motility)&oldid=1036919706

Shinde, P. P., & Shah, S. A review of machine learning and deep learning applications. In: *2018 fourth international conference on computing communication control and automation (iccubea)*. IEEE. 2018, 1–6.

Shirk, J. L., Ballard, H. L., Wilderman, C. C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, B. V., Krasny, M. E., et al. (2012). Public participation in scientific research: A framework for deliberate design. *Ecology and society*, *17*(2). https://doi.org/http://dx.doi.org/10.5751/ES-04705-170229

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution*, *24*(9), 467–471. https://doi.org/https://doi.org/10.1016/j.tree.2009.03.017

Song, Y.-Y., & Ying, L. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.

Soulé, M. E. (1985). What is conservation biology? *BioScience*, *35*(11), 727–734.

Stowell, D., Petrusková, T., Šálek, M., & Linhart, P. (2019). Automatic acoustic identification of individuals in multiple species: Improving identification across recording conditions. *Journal of the Royal Society Interface*, *16*(153), 20180940.

Strasser, B, & Haklay, M. (2018). Citizen science: Expertise, democracy, and public participation.

Strimas-Mackey, M., Hochachka, W. M., Ruiz-Gutierrez, V., Robinson, O. J., Miller, E. T., Auer, T., Kelling, S., Fink, D., & Johnston, A. (2020a, January). *Best practices for using ebird data v1.0*. Zenodo. https://doi.org/10.5281/zenodo.3620739

Strimas-Mackey, M., Hochachka, W. M., Ruiz-Gutierrez, V., Robinson, O. J., Miller, E. T., Auer, T., Kelling, S., Fink, D., & Johnston, A. (2020b). *Best practices for using ebird data v1.0*. Zenodo. https://doi.org/10.5281/ZENODO.3620739

Strimas-Mackey, M., Miller, E., Hochachka, W., & Ornithology, C. L. o. (2021, September). Auk: eBird Data Extraction and Processing in R. Retrieved September 29, 2021, from https://CRAN.R-project.org/package=auk

Student's t-test [Page Version ID: 1060127598]. (2021, December). Retrieved December 15, 2021, from https://en.wikipedia.org/w/index.php?title=Student%27s_t-test&oldid=1060127598

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., et al. (2014). The ebird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40.

Sun, D., Li, S., Zheng, W., Croitoru, A., Stefanidis, A., & Goldberg, M. (2015). Mapping floods due to hurricane sandy using NPP VIIRS and ATMS data and geotagged flickr imagery. *International Journal of Digital Earth*, *9*(5), 427–441. https://doi.org/10.1080/17538947.2015.1040474

Sun, Y., Fan, H., Helbich, M., & Zipf, A. (2013). Analyzing human activities through volunteered geographic information: Using flickr to analyze spatial and temporal pattern of tourist accommodation. In J. M. Krisp (Ed.), *Progress in location-based services* (pp. 57–69). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34203-5_4

Swan, M., Hathaway, K, Hogg, C, McCauley, R, & Vollrath, A. (2010). Citizen science genomics as a model for crowdsourced preventive medicine research. *J Participat Med*, *2*, e20.

Syphard, A. D., & Franklin, J. (2009). Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography*, *32*(6), 907–918.

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., VerCauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., et al. (2019). Ma-

chine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, *10*(4), 585–590.

Tang, J., Zhou, X., & Yu, M. (2019). Designing feedback information to encourage users' participation performances in citizen science projects. *Proceedings of the Association for Information Science and Technology*, *56*(1), 486–490.

Taylor-Rodríguez, D., Kaufeld, K., Schliep, E. M., Clark, J. S., & Gelfand, A. E. (2017). Joint species distribution modeling: Dimension reduction using dirichlet processes. *Bayesian Analysis*, *12*(4), 939–967.

Terry, J. C. D., Roy, H. E., & August, T. A. (2020). Thinking like a naturalist: Enhancing computer vision of citizen science images by harnessing contextual data. *Methods in Ecology and Evolution*, *11*(2), 303–315.

Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. B. (2009). Biomod–a platform for ensemble forecasting of species distributions. *Ecography*, *32*(3), 369–373.

Tinati, R., Luczak-Roesch, M., Simperl, E., & Hall, W. (2017a). An investigation of player motivations in eyewire, a gamified citizen science project. *Computers in Human Behavior*, *73*, 527–540.

Tinati, R., Simperl, E., & Luczak-Roesch, M. To help or hinder: Real-time chat in citizen science. In: *Proceedings of the international aaai conference on web and social media. 11.* (1). 2017.

Tobler, W. (2004). On the first law of geography: A reply. *Annals of the Association of American Geographers*, *94*(2), 304–310.

Torney, C. J., Lloyd-Jones, D. J., Chevallier, M., Moyer, D. C., Maliti, H. T., Mwita, M., Kohi, E. M., & Hopcraft, G. C. (2019). A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. *Methods in Ecology and Evolution*, *10*(6), 779–787.

Ueda, K.-i. (2020, March). A New Vision Model! Retrieved June 26, 2021, from https://www.inaturalist.org/blog/31806-a-new-vision-model

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2018). Blockcv: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *bioRxiv*, 357798.

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC medical research methodology*, *14*(1), 1–13.

Van der Wal, R., Sharma, N., Mellish, C., Robinson, A., & Siddharthan, A. (2016). The role of automated feedback in training and retaining biological recorders for citizen science. *Conservation Biology*, *30*(3), 550–561.

Van Dyke, F., & Lamb, R. L. Biodiversity conservation and climate change. In: *Conservation biology*. Springer, 2020, pp. 125–170.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., & Belongie, S. The inaturalist species classification and detection dataset. In: *Proceedings of the ieee conference on computer vision and pattern recognition*. 2018, 8769–8778.

Van Le, D., & Tham, C.-K. Machine learning (ml)-based air quality monitoring using vehicular sensor networks. In: *2017 ieee 23rd international conference on parallel and distributed systems (icpads)*. IEEE. 2017, 65–72.

Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R., & Wagenknecht, K. (2021). Editorial: The science of citizen science evolves. In K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The science of citizen science* (pp. 1–12). Springer International Publishing. https://doi.org/10.1007/978-3-030-58278-4_1

Vos, R. A., Rademaker, M., & Hogeweg, L. (2019). Species distribution modelling using deep learning. *Biodiversity Information Science and Standards*, *3*, e38333. https://doi.org/10.3897/biss.3.38333

Wang, R., Storey, V., & Firth, C. (1995). A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*, *7*(4), 623–640. https://doi.org/10.1109/69.404034

Weinstein, B. G. (2018). A computer vision for animal ecology. *Journal of Animal Ecology*, *87*(3), 533–545.

Wessels, P., Moran, N., Johnston, A., & Wang, W. (2019). Hybrid expert ensembles for identifying unreliable data in citizen science. *Engineering Applications of Artificial Intelligence*, *81*, 200–212. https://doi.org/https://doi.org/10.1016/j.engappai.2019.01.004

West, S., Dyke, A., & Pateman, R. (2021). Variations in the motivations of environmental citizen scientists. *Citizen Science: Theory and Practice*, *6*(1). https://doi.org/10.5334/cstp.370

Westphal, A., Von Korff, J, Anderson, D., Alexander, A, Betts, B, Brownlee, D., Butterworth, A., Craig, N, Gainsforth, Z, Mendez, B, et al. Stardust@ home: Virtual microscope validation and first results. In: *37th annual lunar and planetary science conference*. 2006, 2225.

Westreicher, F., Cieslinski, M., Ernst, M., Frigerio, D., Heinisch, B., Hübner, T., & Rüdisser, J. (2021). Recognizing failures in citizen science projects: Lessons learned. *PoS*, *ACSC2020*, 007. https://doi.org/10.22323/1.393.0007

Wiggers, K. (n.d.). Google's AI can identify wildlife from trap-camera footage with up to 98.6% accuracy | VentureBeat. Retrieved May 30, 2021, from https://venturebeat.com/2019/12/17/googles-ai-can-identify-wildlife-from-trap-camera-footage-with-up-to-98-6-accuracy/

Wiggins, A., & Crowston, K. From conservation to crowdsourcing: A typology of citizen science. In: *2011 44th hawaii international conference on system sciences*. IEEE. 2011, 1–10.

Wiggins, A., Newman, G., Stevenson, R. D., & Crowston, K. Mechanisms for data quality and validation in citizen science. In: *2011 ieee seventh international conference on e-science workshops*. IEEE. 2011, 14–19.

Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., & Fortson, L. (2019). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, *10*(1), 80–91.

Wintle, B. A., Elith, J., & Potts, J. M. (2005). Fauna habitat modelling and mapping: A review and case study in the lower hunter central coast region of nsw. *Austral Ecology*, *30*(7), 719–738.

Wolf, R. (2005). Why hackers do what they do: Understanding motivation and effort in free/open source software projects. *Joseph Feller/Brian Fitzgerald/Scott A. Hissam/Karim R. Lakhani (Hg.), Perspectives on Free and Open Source Software, Boston*, 3–22.

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, *48*(9), 2839–2846.

Wood, H. (2014). The Long Tail of OpenStreetMap. Retrieved October 25, 2021, from https://harrywood.co.uk/blog/2014/11/17/the-long-tail-of-openstreetmap/

Wright, D. E., Fortson, L., Lintott, C., Laraia, M., & Walmsley, M. (2019). Help me to help you: Machine augmented citizen science. *ACM Transactions on Social Computing*, *2*(3), 1–20.

Wu, Y., Wang, Y., Zhang, S., & Ogai, H. (2021). Deep 3d object detection networks using lidar data: A review. *IEEE Sensors Journal*, *21*(2), 1152–1171. https://doi.org/10.1109/JSEN.2020.3020626

Yadav, P., Charalampidis, I., Cohen, J., Darlington, J., & Grey, F. (2018). A collaborative citizen science platform for real-time volunteer computing and games. *IEEE Transactions on Computational Social Systems*, *5*(1), 9–19.

Yap, B. W., Abd Rani, K., Abd Rahman, H. A., Fong, S., Khairudin, Z., & Abdullah, N. N. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In: *Proceedings of the first international conference on advanced data and information engineering (daeng-2013)*. Springer. 2014, 13–22.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Yu, J., Wong, W.-K., & Hutchinson, R. A. Modeling experts and novices in citizen science data for species distribution modeling. In: *2010 ieee international conference on data mining*. IEEE. 2010, 1157–1162.

Zhou, X., Tang, J., Zhao, Y. C., & Wang, T. (2020). Effects of feedback design and dispositional goal orientations on volunteer performance in citizen science projects. *Computers in Human Behavior*, *107*, 106266.