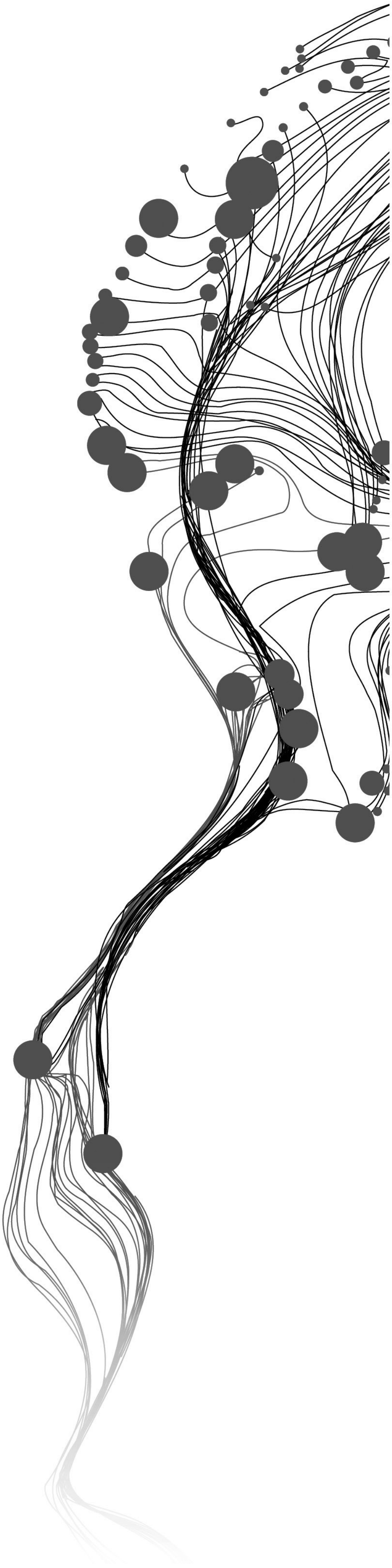
A BIG DATA APPROACH TO MODEL BIRD OCCURRENCE FROM CROWD-SOURCED DATA

VIJAYUDU KONDI
August, 2021

SUPERVISORS:

dr. ir. Rolf A. de By
dr. Frank O. Ostermann



A BIG DATA APPROACH TO MODEL BIRD OCCURRENCE FROM CROWD-SOURCED DATA

VIJAYUDU KONDI
Enschede, The Netherlands, August, 2021

Thesis submitted to the Faculty of Geo-information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: M-GEO

SUPERVISORS:

dr. ir. Rolf A. de By
dr. Frank O. Ostermann

THESIS ASSESSMENT BOARD:

dr. M. J. Kraak (chair)
dr. ir. Maurice van Keulen (External examiner)

Disclaimer

This document describes work undertaken as part of a programme of study at the Faculty of Geo-information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

There are many endangered bird species to be conserved. For conservation planning it is important to understand the species occurrence and their habitat preference through space and time. Range maps are one of the important tools to understand the species occurrence. Range maps are generally obtained by statistical models. Input data for such models is usually obtained from traditional bird surveys where the rules and regulations of the survey are predefined and strictly followed. In this study we explore the possibilities of generating seasonal and annual range maps for 213 species using presence-only type of crowd-sourced data collected within the Netherlands by Waarneming.nl (WNL) from 2010-19.

Crowd sourcing has immense potential for collecting bird observation at huge spatio-temporal extents. Unlike traditional bird surveys, observers in crowd-sourced programs are free to choose where to visit, when to visit, what to find, and what to report. Such freedom of choice leads to creation of voluminous data but brings different types of variability in the collected data. Types of variability that are common in crowd-sourced data are: variability in observer effort through space and time, variability in observer skills, variability in detectability of the species, and variability in report likelihood of the species.

The aim of this study is to account for the variability in the selected dataset and generate four seasonal range maps and one annual range map for any of the selected 213 species. Few metrics are designed to account for the variability in the data, they are: weighted observer days, weighted encounter days, and weighted observed hitrate. To generate a seasonal range map for a selected (species, season) pair, the designed model automatically selects two sets of spatial units or blocks. First set represents the blocks where the species is supposed to be present and the second set represents the blocks where the species are supposed to be absent during the selected season. Withe these two sets and 306 explanatory variables a Random Forest classifier (RF) is trained and range map for the selected (species, season) is generated. By repeating the same procedure, range maps for other three seasons are obtained. A set of conditions are determined and used to combine the seasonal range maps and obtain the annual range map.

The winter and summer range maps generated by the model are compared against the range maps obtained from Sovon Atlas. Performance of the model is assessed based on classification accuracy, precision, recall, and F1 score. Overall, accounting for variability in detectability and report likelihood from the observations data was a big challenge in this study. The model has performed better for species that occur inland compared to the species that occur along the coastal waters. Predictions for species that has limited occurrence in space were more optimistic compared to the occurrence shown in validation maps.

Keywords

Citizen science, Crowd sourcing, Range map, Presence-only, Detectability, Report likelihood, Observer days, Encounter days, Observed hitrate, Random Forest

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the sponsors of ITC excellence scholarship program for providing the financial support for many students like me. Without your support, studying at ITC would have been an unfulfilled dream of mine. I am thankful to my parents, friends, and family for their enormous support in completing my studies at ITC. I am thankful to my supervisors Dr. Ir. Rolf A. de By and Dr. Frank O. Ostermann. My research would have been impossible without the guidance and support of my supervisors. I would like to particularly thank my first supervisor Dr. Ir. Rolf A. de By for his enormous interest and guidance till the end. I appreciate their patience, interest and time spent to regularly improve my thought process and writing skills. It was an honour to work with such highly experienced subject experts. I thank all the administrative and teaching staff for taking special responsibility to make student life easy in the present COVID crisis. Finally, I would like to thank the external examiner Dr. Ir. Maurice Van Keulen for accepting the invitation to examine my research work.

TABLE OF CONTENTS

Abstract i

Acknowledgements ii

1 Introduction 1

1.1 Background 1

1.2 Problem statement 2

1.3 Research interests 3

1.4 Research objectives 3

1.5 Sub-research objectives 4

1.6 Research questions 4

1.7 Innovation and novelty of the research 4

1.8 Thesis structure 5

2 Literature review 7

2.1 Important tools for bird conservation 7

2.2 Types of species monitoring 7

2.3 Challenges in crowd-sourced biological records 8

2.4 Approaches used to account variability in crowd-sourced data 9

2.5 Methods to infer species presence from crowd-sourced data 10

3 Data 13

3.1 Observations from Waarneming.nl (WNL) 13

3.2 Exploratory data analysis 14

3.2.1 Defining seasons 14

3.2.2 Variability in spatial distribution of observations 15

3.2.3 Variability in temporal distribution of observations 16

3.2.4 Variability in observer skills 16

3.2.5 Variability in detectability of species 18

3.2.6 Variability in report likelihood of species 19

3.3 Explanatory data 19

3.4 Selection of species and season specific input data 20

4 Methodology **23**

4.1 Aims and intuition 23

4.2 Estimation of observer effort 26

4.3 Selection of sufficiently visited blocks 26

4.4 Estimation of species presence 27

4.5 Estimation of species presence relative to observer effort 27

4.5.1 Estimation of detectability and report likelihood 28

4.5.2 Separating species presence from the three-step probabilistic process . . . 30

4.6 Selection of training data 31

4.7 Training a Random Forest classifier 32

4.8 Combination of seasonal range maps into an annual range map 32

4.9 Calculation of observer weights 32

4.9.1 Definitional basis for observer weights 33

4.9.2 Weighted observer days 33

4.9.3 Weighted encounter days 34

4.9.4 Weighted observed hitrate 34

5 Results **37**

5.1 Overview of input data 37

5.2 Selection of sufficiently visited blocks 37

5.3 Selection of the classes species-is-present and species-is-absent 38

5.3.1 Selection of a threshold per season to obtain training classes 39

5.4 Results of Random Forest classification 39

6 Validation of modelling outcomes 45

6.1 Preparation of validation maps from the Sovon Atlas 45

6.2 Validation of winter and summer classifiers 46

6.2.1 Validation of range maps of other species 46

6.3 Comparison of results obtained using weighted and non-weighted metrics 47

7 Discussion of outcomes and Conclusion 55

7.1 Introduction 55

7.2 Accounting for variability in input data sources 55

7.2.1 Accounting for variability in observer effort through space 56

7.2.2 Accounting for variability in observer effort through time 57

7.2.3 Accounting for variability in observer skills 58

7.2.4 Accounting for variability in detectability and report likelihood of species 58

7.3 Summary of conclusion 58

7.4 Options for future work 59

A Other figures 61

List of References 71

LIST OF FIGURES

3.1 Map view of 1 km^2 spatial units of the Netherlands 13

3.2 The map shows the observations count per block collected over ten years. The ten year time period is aggregated into one year as explained in section 3.2.1. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841) 15

3.3 Distribution of observation counts from 2010 to 2019. The bar with value 2010 on X-axis and value around 2 on Y-axis represents 2×10^6 observations collected in the year 2010 16

3.4 Distribution of observations in winter and spring. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841) 17

3.5 Distribution of observations in summer and autumn. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841) 17

3.6 Observation counts per week. The bar with value 1 on X-axis and value around 0.8 on Y-axis represents 0.8×10^6 observations collected in the week 1 18

3.7 Scatter plot showing the variability in observer quality. Each point represents an observer with number of block day visits on X-axis and the sum of count of distinct species reported during the block day visits on Y-axis 18

3.8 Scatter plot showing the variability in report likelihood of the species. Each point is a species with annual population on X-axis and observations count on Y-axis. The values on X and Y axes are natural log transformed populations and observation counts. 19

4.1 Block set and subsets selected at different stages of the method 23

4.2 Overall workflow for generating an annual range map of s from Ω 25

4.3 Observer days in the set of blocks **B** within time period t; a, b, c and d are number of observer days 29

4.4 Workflow to select **C0(t,s)**, **C1(t,s)** from **B1(t,s)**, and **B1(t,s)** from **B** 35

5.1 Distribution of weighted observer days in the set of blocks **B** for four different seasons. A bar with value 3 on X-axis represents the $wOD(b, t)$ value natural $log(3) = 1.098$ and the respective number on Y-axis represents number of blocks with $wOD(b, t)$=1.098 38

5.2 Distribution of weighted encounter days in the set of blocks **B1**. Class 1 on X-axis of all the histograms represent the values above 0 and below 0.1, class 2 represents values between 0.1 to 0.2, and so on, class 20 represents the values greater than 1.9. The range of each class from 1 to 20 is increased by a multiple of 0.1. Values on Y-axis represents the number of blocks. SO, each bar in the histogram represents the number of blocks on Y-axis with $wED(b, t, s)$ values within respective class and season on X-axis. 40

5.3 Distribution of $wOHR(b, t, s)$ in the set of blocks **B1**. Class 1 on X-axis represents the $wOHR(b, t, s)$ values above 0 and below 0.006, class 2 represents $wOHR(b, t, s)$ values between 0.006 to 0.012, class 3 represents $wOHR(b, t, s)$ values between 0.012 and 0.018, and so on, value 20 represents $wOHR(b, t, s)$ values above 0.114. The range of each class from 1 to 20 is increased by a multiple of 0.006. So, each bar in the histogram represents the number of blocks on Y-axis with $wOHR(b, t, s)$ values within respective class and season on X-axis. 41

5.4 Seasonal range (presence) maps for species 423. The threshold for $ESP(b, t, s)$, the constant, is 0.07 and the constant $cDETREL$ is 0.3942. Top left: winter; top right: spring; bottom left: summer; bottom right: autumn. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841) 42

5.5 Predicted annual range map of species 423. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841) 43

6.1 Seasonal range maps of species 423 obtained for validation from the Sovon Atlas and predicted seasonal range maps of species 423 from our RF classifiers. Top left: winter range of species 423 obtained for validation from the Sovon Atlas; top right: predicted winter range map of species 423 by our RF classifier; bottom left: summer range of species 423 obtained for validation from the Sovon Atlas; bottom right: predicted summer range map of species 423 by our RF classifier. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841) 48

6.2 Seasonal validation and predicted range maps of species 165 and 224. Top left: validation map for species 165, winter; top right: predicted range map of species 165, winter; bottom left: validation map for species 402, winter; bottom right: predicted range map for species 224, winter. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841) 49

6.3 Seasonal validation and predicted range maps of species 402 and 639. Top left: validation map for species 402, winter; top right: predicted winter range map of species 402; bottom left: validation map for species 639, winter; bottom right: predicted range map for species 639, winter. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841) 50

6.4 Seasonal validation and predicted range maps of species 692 and 706. Top left: validation map for species 692, winter; top right: predicted winter range map of species 692; bottom left: validation map for species 706, winter; bottom right: predicted range map for species 706, winter. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841) 51

6.5 Seasonal validation and predicted range maps of species 715 and 753. Top left: validation map for species 715, summer; top right: prediction summer range map of species 715; bottom left: validation map for species 753, summer; bottom right: predicted range map for species 715, winter. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841) 52

7.1 Workflow explaining the data loss and reduction in representativeness from **3** and **4** to **5** and **6**. Assume that **1** is the set of blocks **B**, **2** is the set of sufficiently visited blocks **B1(t)**, **3** and **4** are the set of blocks labelled as species-is-absent and -present respectively, **5** and **6** are balanced training classes **C0(t,s)** and **C1(t,s)**. Spatial relation between the blocks from **2** to **6** are false. 57

A.1 Annual distribution of observations of species 423. Value on X-axis represents the week number and value on Y-axis represents the count of observations of species 423. 61

A.2 Distribution of 7 LULC classes in NL. Each class on X-axis represents area in hectares within a particular range covered by the respective LULC class. class 1 represents values within range 0 to 5, class 2 represents values within range 5 to 10, and so on, class 20 represents the values above 95. The range of each class is increased by a multiple 5 from class 1 to class 20. The value on Y-axis represents the number of blocks transformed using natural log 63

A.3 Distribution of 7 LULC classes in sufficiently visited blocks in winter. Each class on X-axis represents area in hectares within a particular range covered by the respective LULC class. class 1 represents values within range 0 to 5, class 2 represents values within range 5 to 10, and so on, class 20 represents the values above 95. The range of each class is increased by a multiple 5 from class 1 to class 20. The value on Y-axis represents the number of blocks transformed using natural log 64

A.4 Distribution of 7 LULC classes in sufficiently visited blocks in spring. Each class on X-axis represents area in hectares within a particular range covered by the respective LULC class. class 1 represents values within range 0 to 5, class 2 represents values within range 5 to 10, and so on, class 20 represents the values above 95. The range of each class is increased by a multiple 5 from class 1 to class 20. The value on Y-axis represents the number of blocks transformed using natural log 65

A.5 Distribution of 7 LULC classes in sufficiently visited blocks in summer. Each class on X-axis represents area in hectares within a particular range covered by the respective LULC class. class 1 represents values within range 0 to 5, class 2 represents values within range 5 to 10, and so on, class 20 represents the values above 95. The range of each class is increased by a multiple 5 from class 1 to class 20. The value on Y-axis represents the number of blocks transformed using natural log 66

A.6 Distribution of 7 LULC classes in sufficiently visited blocks in autumn. Each class on X-axis represents area in hectares within a particular range covered by the respective LULC class. class 1 represents values within range 0 to 5, class 2 represents values within range 5 to 10, and so on, class 20 represents the values above 95. The range of each class is increased by a multiple 5 from class 1 to class 20. The value on Y-axis represents the number of blocks transformed using natural log 67

A.7 Seasonal abundance and validation range maps of species 423 obtained from the Sovon Atlas. Blocks with abundance class above 3 are labelled under the class present and others under the class absent. Top left: winter abundance map of species 423; top right: winter validation map of species 423; bottom left: summer abundance map of species 423; bottom right: summer validation map of species 423. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841) 68

A.8 Obtained seasonal abundance maps of selected species, season pairs. Top left: 165, winter; top right: 224, winter; bottom left: 402, winter; bottom right: 639, winter. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841) 69

A.9 Obtained seasonal abundance maps of selected species, season pairs. Top left: 692, winter; top right: 706, winter; bottom left: 715, summer; bottom right: 753, summer. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841) 70

LIST OF TABLES

3.1 Start and end dates for each of the 10 years 14

3.2 Starting and ending week numbers of seasons 15

3.3 Values of observations count at different quantiles 16

3.4 LULC classes and their acronyms 20

4.1 Conditions to generate annual range map from seasonal range maps 32

5.1 Species-independent threshold for wOD per season and number of blocks obtained 37

5.2 Number of blocks per season with $wED(b, t, s)$ and $wOHR(b, t, s)$ equal to 0 . . 39

5.3 Summary statistics of $wED(b, t, s)$ per season 39

5.4 Summary statistics of $wOHR(b, t, s)$ per season 39

5.5 Number of blocks per season in different classes with threshold for $ESP(b, t, s)$ at 0.07 40

6.1 Classification report for winter and summer classifiers of species 423. 0 represents absence class and 1 represents presence class 47

6.2 Classification reports of selected species, season pairs used to determine the constants $cDETREL$ and threshold for $ESP(b, t, s)$ 47

6.3 Classification reports of weighted and non-weighted models for the eight (species,season) pairs 53

A.1 F1 scores and classification accuracies of the predicted range maps of eight (species, season) pairs when compared with validation maps obtained using different cutoffs for abundance 62

Chapter 1

Introduction

1.1 BACKGROUND

At present there are about 50 billion birds on this planet (Callaghan et al., 2021). Though the individual count of common species is comparatively higher than individuals of rare species, there are more rare species and comparatively fewer common species (Callaghan et al., 2021). Many of these rare species are on the verge of extinction and need conservation. According to the IUCN red list, 256 species were declared as either endangered or extinct (Ghiraldi & Aimassi, 2019). Studying bird behavior and their abundance across the globe can solve many long standing questions in ecology (Callaghan et al., 2021). Bird behaviour is considered as an indicator in studying climate change, and it is one of the official environmental indicators of the European Union (Gregory et al., 2005). Species monitoring can help in understanding the outcomes of human interaction with the environment (Snäll et al., 2011). For conservation planning, it is essential to understand the range of the annual avian cycle of the species (Kelling et al., 2015). Habitat Suitability Models (HSMs), Species Distribution Models (SDMs), and Range maps that show the locations where the species could possibly occur are considered important tools for conservation. These are usually derived from (a) statistical methods using presence and absence data or presence-only data of the species and (b) explanatory environmental data. The primary source for species presence and absence data are traditional bird surveys with pre-defined survey protocols, while one of the main novel sources is citizen-science. The spatial extent of annual occurrence of many birds range from nations to continents where protocolised surveys are highly expensive and challenging (Fink et al., 2019). In such conditions, citizen science is considered to be a better choice (Fink et al., 2019).

Citizen science is defined as the participation of citizens, the general public, non-scientists in collaboration with the scientific community for collecting data and creation of knowledge (Buytaert et al., 2014). This participation happens at different levels and in different fields of science. The terms used in the literature for citizen science data with spatial characteristics are Volunteered Geographic Information (Klonner et al., 2016) and Crowd-sourced Geographic Information (See et al., 2016). Disruptive technological innovations have made data collection and dissemination feasible. This innovation has facilitated the creation of 'big geodata'. As the spatial data being generated is big in size (volume), being generated rapidly (velocity) and generated in different formats (variety), crowd-sourced geospatial data can be considered big geodata. Apart from traditional sources of geospatial data, citizen science has become a novel source of data for many applications in geoinformation science.

The Christmas Bird Count (CBC) is the pioneer of citizen science projects for birds, and is the longest-running project. It started in the 20th century (Sauer & Butcher, 2014). Fink et al., 2019 successfully generated annual range maps for many bird species covering North America from the e-bird reference dataset. Bradter et al., 2018 generated HSMs for many birds, the Siberian Jay, for example, from crowd-sourced presence-only data. So, apart from traditional sources like systematic monitoring projects, professional surveys and mass participation projects, citizen science projects have become a reliable source for inferring bird occurrence and populations. To make it easy for the participants and encourage them to collect more data, crowd-sourcing usually do not have strict rules of sampling like traditional protocolised surveys, so they are collected without following many strict sampling rules. This sometimes makes crowd-sourced data unstructured and less reliable for inference of species presence or absence. But when the reported observations enter the database, they pass through data quality filters and are stored in a defined structure. This improves the data from un-structured to semi-structured. Surveys with strict surveying rules (protocolised surveys) can also become unstructured if the data collected is not stored in a proper structure.

There are advantages and disadvantages for both protocolised survey data and crowd-sourced data. The Swedish Bird Survey (SBS), a structured bird survey, has recorded less than 25 individuals per year for 43% of listed rare species (Snäll et al., 2011). As already mentioned, there are comparatively more rare species than common ones, and structured surveys are robust only for common species because observers do not frequently visit the site to capture rare occurrence of the species (Fox, 2004). Hence, there is a need to develop techniques to study rare species. In crowd-sourced data, the freedom of choice in selecting the sampling area, time interval and protocol makes it easy for the observer to report observations from anywhere and anytime (Kelling et al., 2018). Observers also spend their time during nights which is uncommon in structured surveys. Observers usually prefer to report less common to rare species than common species (Snäll et al., 2011). Though this behavior of the observers brings variability in the data, it helps in collecting comparatively more data for rare species. Crowd-sourced data can be considered as a sample of the phenomena observed, as the observers are not there all the time. Statements made for the population from its sample will always have an error and cannot fully correctly represent the population. This applies to all surveys. As the observers collect data whenever and wherever they prefer, this can be called convenience sampling, which is more prone to bias. Kelling et al., 2018 mentiones three main types of variability in crowd-sourced data into three categories:

1. Variability in observer effort over space and time,
2. Variability in detectability of species, and
3. Variability in observer skill across participants.

The other errors which exist in the data are systematic errors, and random errors. With varying observer skills and instruments used, these errors can be extremely difficult to quantify. Before making any inferences from crowd-sourced biological records, the above-mentioned types of variability have to be accounted for.

1.2 PROBLEM STATEMENT

Most of the studies that have developed techniques to handle crowd-sourced biological records are either dependent on data from structured surveys or have worked on semi-structured type of crowd-sourced data where information about the observation process is recorded (Welvaert & Caley, 2016). Therefore, combining crowd-sourced data and protocolised survey data to generate HSMs and range maps is a common practice. If the range maps have to be generated for multiple species, collecting protocolised bird survey data for all the species can be challenging and sometimes it might not provide for some species. Crowd-sourced biological records can be of two types: (1) Presence-only data and (2) Presence-absence data. The potential for producing reliable information is high for presence-absence data compared to presence-only data (Isaac & Pocock, 2015). Studies that have produced high quality range maps (Fink et al., 2019) have used the e-bird reference dataset (Munson et al., 2010), which is a sophisticated crowd-sourced dataset. These type of datasets are also referred to as 'checklists' and 'full checklists'. In a full checklist, observers report both presence and absence of species, time spent by the observers, number of observers involved and protocol followed during the observation. This information in checklists makes it easy to account for the important types of variability associated with crowd-sourced biological records. But this information is absent in the majority of crowd-sourced datasets. Techniques that work without using data from protocolised surveys but use presence-only data to model species occurrence are few and underdeveloped.

1.3 RESEARCH INTERESTS

Our interest in this research is to study whether crowd-sourced presence-only data can achieve reliable distribution maps without the use of highly protocolised techniques. Becoming less dependent of those techniques is not a purpose in itself but economy of scale is of interest and may allow specialist protocols to target more intricate and smaller survey challenges. We aim to account for different types of variability in crowd-sourced data without using any data from protocolised surveys to infer seasonal occurrence f0r 213 species. Briefly, we aim to design a machine learning model that can learn from the selected and processed presence-only type of crowd-sourced data and make reliable prediction on species occurrence. We formulate few metrics that can account for the variability in crowd-sourced data.

1.4 RESEARCH OBJECTIVES

The main objective is to create, for 213 species in the selected dataset, four seasonal range maps and derive an annual range map from them based on certain conditions.

A range map captures the expected presence of a species in both space and time (for a given species, when and where it can be expected to be found). To create a range map, the study area is divided into spatial units of equal resolution. In seasonal range map, each spatial unit is classified as species is present or absent. In the annual range map, each spatial unit is classified into one of the six presence classes colours where each presence class represents the presence of a species for a particular time period. The six classes of an annual range map are

blue non-breeding (winter) presence,

orange pre-breeding migratory presence,

red breeding summer presence,

yellow post-breeding migratory presence,

green resident (present year-round) and

transparent absent (does not or only infrequently occur).

When all spatial units are coloured in this way, for a single species, we obtain (an annual) range map.

1.5 SUB-RESEARCH OBJECTIVES

1. To account for the variability in observer effort through space and time.
2. To account for variability in observer skills.
3. To account for the variability in detectability and report likelihood of the species.
4. To estimate species presence and generate seasonal range maps.
5. To derive annual range map from the seasonal range maps of the species.

1.6 RESEARCH QUESTIONS

1. How to quantify observer effort and account for its variability through space and time?
2. How to determine skills of an observer and account for the variability?
3. How to estimate detectability and report likelihood from the selected dataset and account for their variability?
4. How to estimate seasonal and annual presence of the species?
5. How to generate annual range maps from seasonal range maps?

1.7 INNOVATION AND NOVELTY OF THE RESEARCH

In this research, we do not use any data from protocolised surveys to train our machine learning model. This study designs a novel method where seasonal range maps of the species, obtained as intermediate results, are combined to obtain the annual range map. The model automatically selects the input data and training data based on the selected (species, season), trains the classifier, predicts seasonal range maps, and derive annual range map from them. We have explored different possibilities to estimate detectability and report likelihood from presence-only type of crowd-sourced data and have estimated detectability and report likelihood together as one quantity.

1.8 THESIS STRUCTURE

The thesis consists of eight chapters. This Chapter 1 briefly discusses the background, problem statement, research objectives and questions that are stated and answered respectively by this research. Chapter 2 reviews the scientific literature and briefly discusses different types of crowd-sourced biological records, data variability associated with crowd-sourced data, and techniques to account for the variability to infer species presence and absence. Chapter 3 describes the data from Waarneming.nl, explores different types of variability in the data, and describes the ancillary data and explanatory data used for prediction. Chapter 4 explains the method used to generate seasonal and annual range maps. Chapter 5 has the seasonal and annual range maps of species 423 obtained from the proposed method. Chapter 6 validates the results and chapter 7 briefly discuses the outcomes and provides conclusions and recommendations.

Chapter 2

Literature review

2.1 IMPORTANT TOOLS FOR BIRD CONSERVATION

Habitat Suitability Models (HSMs), Environmental Niche Models (ENMs), Species Distribution Models (SDMs), and Range maps are some of the important tools used for wildlife conservation. The terms HSM, ENM and SDM are interchangeably used in the literature. Though they are closely related they are different in terms of information they provide. These are obtained by applying statistical techniques to a species' presence-absence data or presence-only data using environment variables as predictors. Though maps obtained from all the three types of models give information on species existence, their level of detail or information varies. Range maps illustrate the species presence or absence per spatial unit and time interval. Each spatial unit will be classified as species-is-present or species-is-absent. Maps derived from HSMs and ENMs show the suitability values for the species existence. Based on the environmental factors, a site suitability value for the species is calculated per spatial unit and time interval. An SDM provides the information on population density of the species: it provides the relative number of birds that can be possibly present in each spatial unit. Both HSM and SDM have continuous values, whereas range maps simply have two discrete classes. Compared to range maps, SDMs and HSMs require high quality data and involve complex calculations.

2.2 TYPES OF SPECIES MONITORING

There are five main approaches in bird montoring: comparing bird atlases of different time period, repeated monitoring, checklist programmes (presence-absence), and species specific surveys, and presence-only observations (Snäll et al., 2011). In presence-only, observers report only the name of the bird species they have encountered but not the individual count of the species. In presence-absence data, observers record both the name and individual count of the bird species they have encountered and not encountered. Usually, information about observation process like the sampling procedure followed, time spent, number of contributors involved is recorded only in presence-absence type of data.

The E-bird reference dataset is a typical and well-known example of presence-absence type of data (Munson et al., 2010). Contributors to E-bird fill out a checklist for every visit they made. In the checklist, they provide the information of name and count of individuals of species they have encountered and names of the species they did not encounter, number of contributors involved, date and time, location, protocol followed and amount of time spent. If all such information is provided it is referred to as a full-checklist. Most of the presence-absence data do have information on number of observers involved and time spent by them. From this information, the observer effort is calculated. On the other hand, the information on the amount of time spent by the observers might not be available for presence-only data. With availability of comparatively more information, presence-absence type of data provides more opportunities to infer a species' presence and absence or abundance compared to presence-only data (Isaac & Pocock, 2015). Even with enormous amounts of data it is hard to make accurate inferences if the protocol followed and observer effort are not recorded (Conrad & Hilchey, 2011).

Many techniques and methods have been developed to infer species presence and absence or abundance from crowd-sourced data, but most have chosen to work with presence-absence while techniques to work with presence-only type of data remain underdeveloped (Welvaert & Caley, 2016). Class imbalance is also a challenge in crowd-sourced data as stated by Robinson et al., 2018.If the dataset has comparatively more number of non-detections (absent) than detections (present) or vice versa the dataset has class imbalance. This imbalance is more prevalent in rare or low-detectability species because of a low number of presence records compared to a high number of absence records. Predicting the occurrence or distribution of a species covering their range with the few presence records is hard.

2.3 CHALLENGES IN CROWD-SOURCED BIOLOGICAL RECORDS

With enormous potential to infer species distribution, crowd-sourced data also comes with many challenges: data may be skewed to certain locations, to certain times of year, and to certain landscapes. This unevenness or variability brings bias in inferences made from crowd-sourced data. Bias is a systematic tendency to error caused due to selection of a sample from a non uniform probability. For example, a transport authority wants to calculate the probability for 50% seats of a public bus being empty between station A and B. A sample of 10 days is considered and the probability was calculated. If the sample has more weekdays than weekends or vice versa then the sample is biased as the number of passengers greatly varies between weekdays and weekends. Biases differ with sampling schemes. Simple random sampling and Stratified sampling are less prone to bias than convenience sampling. In crowd-sourced data observers select the sampling area and sampling period based on their convenience, so this is convenience sampling. This selective nature of the observers brings variability in observer effort through space and time. Courter et al., 2012 have found more number of observations during week ends than week days in crowd-sourced data. Kelling et al., 2018 have mentioned three types of variability in crowd-sourced biological records. They are, Variability in observer effort through space and time, Variability in detectability of species, and Variability in observer skills.

In a presense-absence type of records, any given observation can be a True positive (if the species to which the encountered bird or birds belong to is correctly identified and reported) or False positive (the bird or birds encountered is misidentified and reported) or True negative (the bird was not encountered and the observer reports it as absent) or False negative (the bird was present but not encountered so the observer reports it as absent). Similarly, in presence-only type of data, an observation can be True positive or False positive. As there are no absence records, all locations and all time periods without observations are inferred as absent and lets call them inferred negatives. All the inferred negatives can be True negatives or False negatives. False positives and False negatives are the errors in crowd-sourced data (Kelling et al., 2018). These errors are tough to be identified because the true data is not available.

The reason for false positives is that some species require expert skills to identify as they are hard to be distinguished from other species. In structured surveys, commonly the observers are skilful at identifying the species being surveyed. On the other hand, observation skills vary for different observers in citizen science projects. There are three main reasons for False negatives. One, is the varying observer effort through space and time. Two, is the varying detectability of the species. Three, is the varying report likelihood of the species. Detectability is the probability of finding a species when it is present. Detectability of a species changes with space and time. Many species are easily found in the early breeding season when they vocalize, and may become much more obscure outside that season. Most owl species are a case in point. There are many factors that influence detectability. Species abundance, behavior (flocking, roosting, singing), and habitat preference are some of the important factors. These factors greatly vary with seasons and remains almost the same within a season. So, it can be said that the detectability of a species do not vary much within a season. However, the variability in detectability among different species still exist. Surveys aimed at observing multiple species are more prone to bias due to varying detectability among different species. Finally, report likelihood is the probability of reporting a species when detected . Not all the observers report all the species and species' individuals when detected. Some observers report a species only once per visit while some report the individuals of the same species multiple times per visit. Similarly, not all the observers report all the distinct species, they only report the species that interests them. Only few observers meticulously report all the distinct species encountered. Observer behavior in reporting a species when found can vary per observer, and it can also vary per species, space, and time. Snäll et al., 2011 have found an imbalance in counts of observations for common and uncommon species when compared with independent data sources, indicating that an observer's interest varies with species. So, there is variability in report likelihood of species. The types of variability in crowd-sourced biological records to be accounted are,

1. Variability in observer effort through space and time
2. Variability in observer skills
3. Variability in detectability per species, space, and time
4. Variability in report likelihood of species per observer, species, space, and time

2.4 APPROACHES USED TO ACCOUNT VARIABILITY IN CROWD-SOURCED DATA

Kelling et al., 2018 suggests to collect enough information on the observation process to control different types of variability associated with crowd-sourced data. In most citizen science projects the observations pass through data quality filters before they enter the database. These data quality filters ensure minimal data quality. About 5% of observations annually are flagged by eBird and sent for review (Kelling et al., 2015). Their data quality filters are a combination of expert reviewers and artificial intelligence algorithms (Kelling et al., 2012). Kelling et al., 2015 proposed a big data approach to improve the data quality in citizen science projects. To quantify the variability in observer skills, they ranked observers based on the observations reported per recorded observation time. Species abundance, date and time of the year, are also considered for ranking the observers (Kelling et al., 2015). Then they generate species accumulation curves which shows the variation in number of species observed with increasing observation period and calculate observer expertise index. To address the problem of uneven detectability, Kelling et al., 2015 have tuned the spatio-temporal explanatory model (STEM) of Fink et al., 2010 to a two-stage species distribution model (SDM). At stage 1, multiple SDMs are generated for randomly selected spatial units and the predictions are made using local explanatory environmental data. At stage 2 all the local SDMs are stitched together and used to predict for the entire study area.

Bradter et al., 2018 generated HSMs for the Siberian Jay from opportunistically collected data from the Swedish Bird Survey (SBS). To account for uneven observer skills and to improve the quality of data, they conducted a survey on the observers asking about their confidence in reporting the species' presence or absence during each visit. Quantification of observer skills and separation of the observations of highly skilled observers have helped to improve the quality of data and to address the issue of varying observer skills. In the work of Geldmann et al., 2016, they used contextual spatial data on land cover and human infrastructure to account for variation in observer effort through space. As reporting the observations is a random process, the effect of the contextual spatial variables (roads, population density, and land use and land cover (LULC)) on the number of observations was captured as a Poisson intensity (Geldmann et al., 2016). However, this method is not efficient if most of the observations are contributed by only few of the observers because it violates the condition that Point Process Models (PPMs) assume independence among the observation locations (Geldmann et al., 2016). They observed that agricultural areas were oversampled compared to their species richness. On the other hand, they found forests and grasslands to be undersampled.

2.5 METHODS TO INFER SPECIES PRESENCE FROM CROWD-SOURCED DATA

Most studies have chosen to work with sophisticated crowd-sourced datasets like those of eBird. Fink et al., 2019 estimated the occurrence and abundance of the Wood Thrush, covering its entire annual avian cycle. Additionally they estimated the changes in inter-annual range, inter-annual changes in species' association with the environment, and changes in inter-annual abundance trends. To obtain this information, they modified the spatio-temporal explanatory model (STEM) of Fink et al., 2010 to an Adaptive Spatio-Temporal Model (AdaSTEM). AdaSTEM is an ensemble of local regression models. Some of the spatial units are randomly selected as spatio-temporal blocks called stixels. Each stixel is a base model. The spatial extent of the stixel is decided on the basis of density of observations within the stixel. The spatial extent of the stixel is not allowed to be less than 5° longitude by 5° latitude and stixels above 25° longitude by 25° latitude are forced to split. This leads to 100 independent regression models as base models. The estimates are obtained by averaging the estimates of the corresponding base models. Each base model is a two-step Boosted Regression Tree (BRT). As first step, a Bernoulli BRT is trained to estimate the probability of species occurrence. In the second step, a Poisson response BRT is used to estimate the counts conditional on occurrence. The predictors used in the two-step BRT are five variables explaining observer effort, day of the year, year that can be used to account for variation in observer effort, and environmental variables.

Only few studies have worked with presence-only type of data to infer species presence. Phillips et al., 2009 used environmental data to infer possible conditions where the species was possibly absent and used them as pseudo-absences. These pseudo-absences can also be obtained from spatial units where the observer effort is significant but the species was not reported. Phillips et al., 2009 suggest that the approach of using true-presence and pseudo-absence data for predicting species occurrence is applicable to regression-based models like generalised linear models, additive models and boosted regression trees. Using available presence data and inferred absences, Bradter et al., 2018 generated HSMs similar to the ones generated from protocolised survey data. They conducted a survey in which the observers were asked about their observation skills and whether they have reported the species each time they found it. In this way, they obtained the observations of highly skilled observers and pseudo-absences. With logistic regression, two versions of MaxEnt, and a Bayesian site occupancy detection model were implemented. Logistic regression with high quality presence data allowed to produce accurate HSMs. Random Forests (RF) is one of the widely used machine learning algorithms in various fields but it is comparatively little exploited in the field of ecology (Cutler et al., 2007). With RF it is easy to calculate feature importance and it can handle high-dimensional datasets effectively when properly parameterised. Cutler et al., 2007 have used RF and four other widely used classifiers to classify invasive plant species in the United States. They have found that the classification accuracy of RF is better than the other four algorithms. They also found that the feature importance calculated by the RF fairly explains the plants occurrence as explained in the literature.

Chapter 3

Data

The study area is The Netherlands. It is represented with a grid of 41,732 spatial blocks of 1 km^2 resolution. Fig 3.1 shows the map view of the spatial blocks. There are two main datasets used in this study. First is the observations data from `waarneming.nl`[1] (which we shall call WNL from here). Second is a dataset of 357 explanatory variables that include 21 LULC classes, 306 soil classes, and 30 water classes. Sections 3.1 and 3.3 provide detailed information on observations and explanatory variables respectively. In section 3.2, the observations data is explored for variability in observer effort, observer quality, detectability and report likelihood of species.

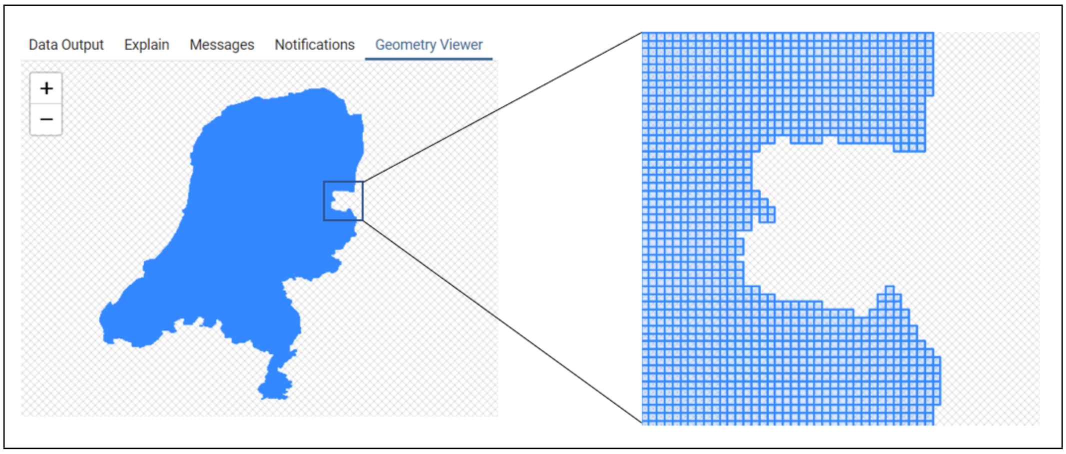


Figure 3.1: Map view of 1 km^2 spatial units of the Netherlands

3.1 OBSERVATIONS FROM WAARNEMING.NL (WNL)

The dataset (Ω) used in this study is obtained from WNL. There are around 30 million records in Ω. Each record(r) is a tuple with attributes block, observer, species, date and time representing an observation of a species (individual or group) reported in the Netherlands. The attribute block represents the 1 km^2 spatial unit of the Netherlands in which the observation was made. Each block is represented by a unique number. All the 30 million observations are distributed within in 39,688 blocks. Blocks not represented are mostly open (sea/fresh) water where no observers have visited. Observer is the person who has reported the observation. There are 40,646 such observers who have contributed to the observations in Ω. Each observer is represented by a unique

[1] https://waarneming.nl/

Table 3.1 Start and end dates for each of the 10 years

Year number	Start date	End date
1	2010-01-06	2011-01-04
2	2011-01-05	2012-01-03
3	2012-01-04	2013-01-01
4	2013-01-02	2013-12-31
5	2014-01-01	2014-12-30
6	2014-12-31	2015-12-29
7	2015-12-30	2016-12-27
8	2016-12-28	2017-12-26
9	2017-12-27	2018-12-25
10	2018-12-26	2019-12-24

number. Species represents the species number (pseudonymized name) of the species observed. There are 765 distinct species in Ω. Date and time represent the calendar date and clock time of the observation respectively. The start and end calendar dates of the observations are 2010-01-01 and 2019-12-31. Only 34% of the observations have a clock timestamp.

3.2 EXPLORATORY DATA ANALYSIS

The observations data is explored to visualise the variability in observer effort through space and time, variability in observer skills, variability in the species' detectability and report likelihood. Any inference made from the data with such variability will likely be biased. The model should account for such variability in the data before making any inferences on species occurrence.

3.2.1 Defining seasons

The calendar dates from 2010-01-01 to 2019-12-31 are adjusted in such a way that each year has exactly 52 weeks. By starting from 2010-01-06 as day 1 of week 1, year 1; 2010-01-07 as day 2 of week 1 and year 1; 2010-01-08 as day 3 of week 1, year 1 and so on, the date 2011-01-12 will be day 7 of week 52, year 1. The next date 2011-01-13 will be day 1 of week 1, year 2. The end date is 2019-12-24 that is day 7 of week 52, year 10. By using the start and end dates of each of the 10 years shown in the table 3.1, each year gets exactly 52 weeks that are different from calendar week numbers. Here after, we only use the augmented week numbers but not the calendar week numbers. Now we define the seasons using these week numbers. The days falling in weeks 49 to 52 and weeks 1 and 9 of all the ten years together is defined as winter season. Similarly, the days within weeks 10 and 22; 23 and 35; 36 and 48 of all the ten years are defined as spring, summer, and autumn seasons respectively. These adjustments aggregate the observations of ten years into one year. We cannot make use of all the observations in Ω if we choose to work with any one year. So we aggregate the observations of all ten years into one year as explained above.

Table 3.2 Starting and ending week numbers of seasons

	Winter	Spring	Summer	Autumn
Weeks	49-52, 1-9	10-22	23-35	36-48

3.2.2 Variability in spatial distribution of observations

To know the spatial distribution of observer effort, we count the number of observations per block collected over the considered ten year time period, aggregate them into one year as explained in the section 3.2.1 and plot them as shown in the figure 3.2. If the observations count per block varies substantially, it can be said that this is an effect of varying observer effort. To visualise the spatial distribution of observations, the number of observations per block are visualised in figure 3.2. From table 3.3, it can be observed that there is a huge difference between the percentile values of observation counts. There are less than 50 observations in 25% (10,386) of the blocks. About 26 million (86%) observations are within 25% (10,417) blocks. This shows the variability in observer effort through space.

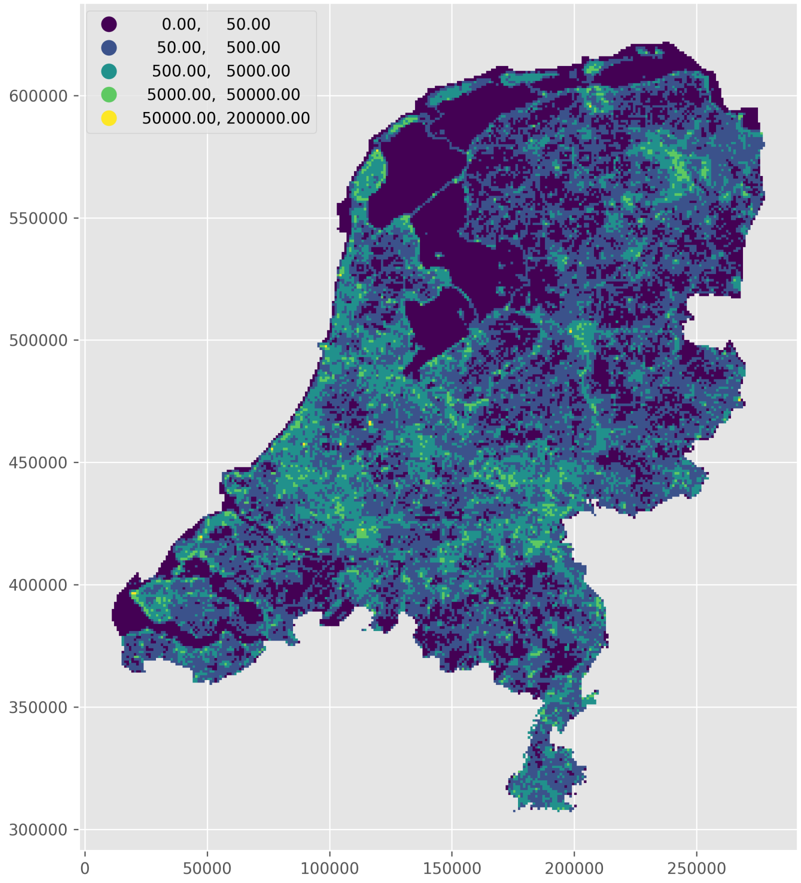


Figure 3.2: The map shows the observations count per block collected over ten years. The ten year time period is aggregated into one year as explained in section 3.2.1. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841)

Table 3.3 Values of observations count at different quantiles

Std.	Min.	25%	50%	75%	Max.
2637.63	0	36	139	479	1,84,165

3.2.3 Variability in temporal distribution of observations

If there is considerable difference in observation counts through time, this can be considered as an effect of varying observer effort through time. We count the number of observations collected per year from 2010 to 2019 and plot their distribution as shown in figure 3.3. Then we count the number of observations per block, season and visualise them in four maps. Figure 3.4,and 3.5 shows the distribution of observations per block in winter, spring; and summer, autumn respectively. Then we count the number of observations per week and plot their distribution as shown in figure 3.6. Figure 3.3 shows an increasing trend in the observation counts from 2010 to 2019. The maps 3.4, 3.5 and the distribution in figure 3.6 shows that the observation counts vary with season. There are more number of observations in winter and spring compared to summer and autumn. This is an effect of varying observer effort through time or generally species population in some season are higher than others. If an annual range map is are generated without accounting for this variability it will be biased to seasons with high observation counts. Plotting the number of observations of species 423 per week as shown in the figure A.1 explains the temporal occurrence of the species but inferring spatial occurrence of the species seems to be a big challenge.

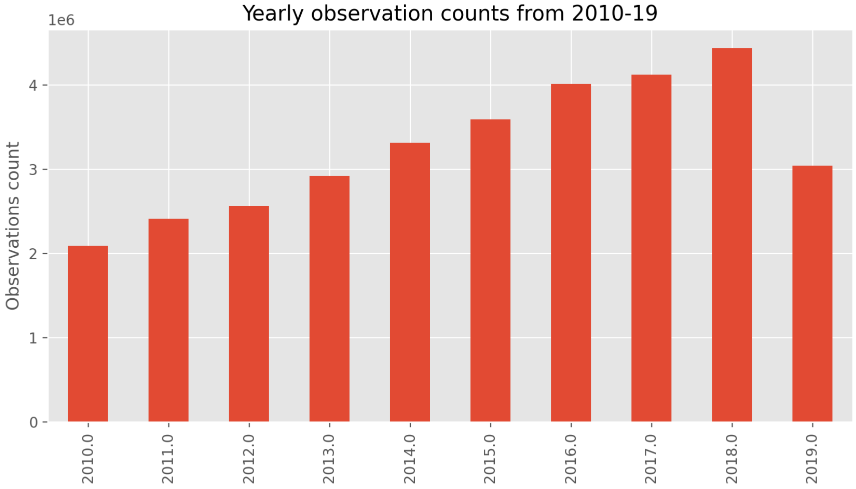


Figure 3.3: Distribution of observation counts from 2010 to 2019. The bar with value 2010 on X-axis and value around 2 on Y-axis represents 2×10^6 observations collected in the year 2010

3.2.4 Variability in observer skills

Observers who make more day visits to the blocks and meticulously report all the distinct species detected during every block visit are considered as highly skilled observers. Observers vary in terms of number of day visits made and in terms of reporting the detected species per visit. So, the observers skill also varies. To check the variability in observer quality, for each observer in Ω,

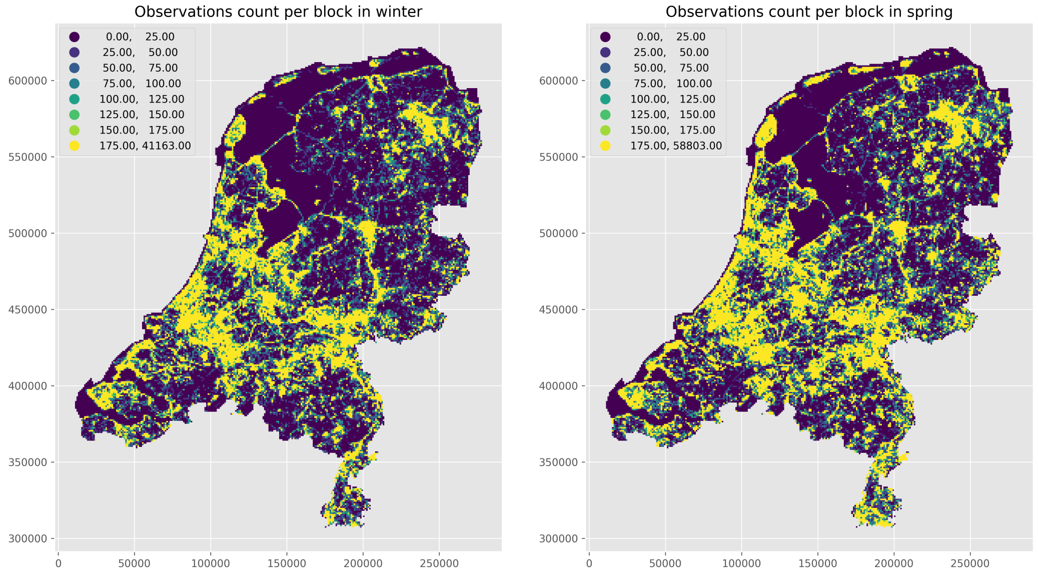


Figure 3.4: Distribution of observations in winter and spring. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841)

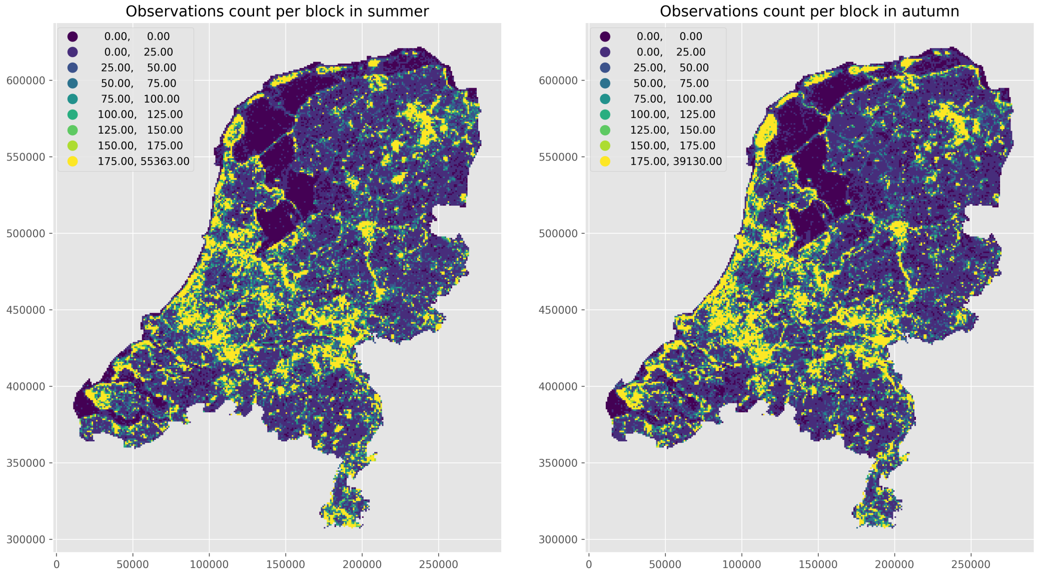


Figure 3.5: Distribution of observations in summer and autumn. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841)

we count the number of day visits made to any of the blocks and plot it against the sum of distinct species reported during the day visits to the blocks as shown in figure 3.7. From the figure 3.7, it can be understood that there is huge variation in observers skill. From the vertical spread in the figure 3.7 it is evident that, though observers make high number of block day visits they do not report all the species they detect.

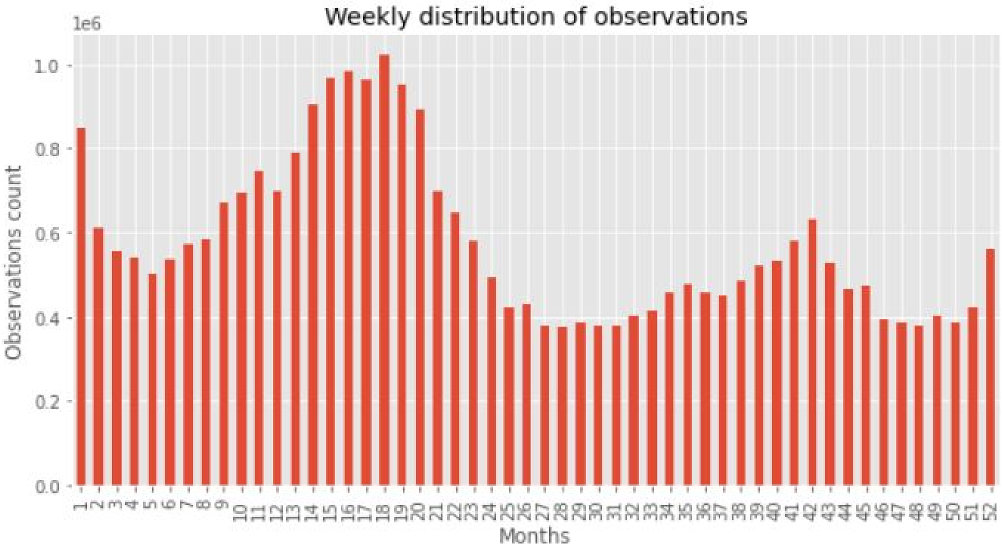


Figure 3.6: Observation counts per week. The bar with value 1 on X-axis and value around 0.8 on Y-axis represents 0.8×10^6 observations collected in the week 1

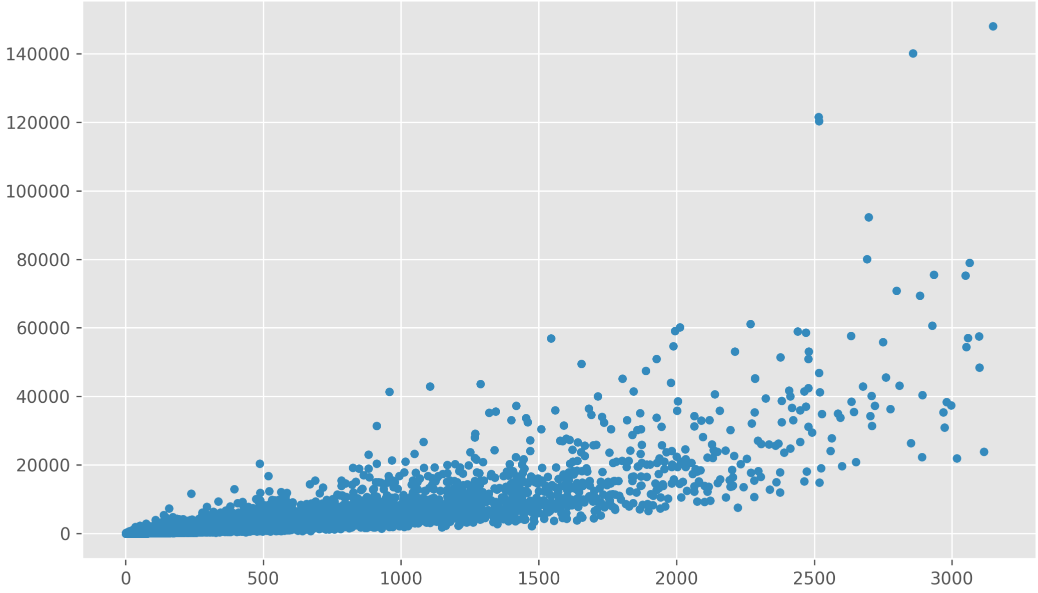


Figure 3.7: Scatter plot showing the variability in observer quality. Each point represents an observer with number of block day visits on X-axis and the sum of count of distinct species reported during the block day visits on Y-axis

3.2.5 Variability in detectability of species

Detectability is the probability of finding a species s in a season or time period t in a block b if it is present. There are many factors that influence the detectability of a species. Some of the important factors are observer effort and behavior, species' abundance, behavior and size, and its habitat preference. Detectability of a species changes with space and time. It is challenging to calculate detectability of a species with the available data because an observation reported can be considered as a three-step probabilistic process: the species must be present, the observer must find (detect) it, and observer must report the species detected. Observers do not report all the species they encounter so it is unknown which species was found by the observer but was not reported by her/him. Estimating detectability from Ω is further discussed in the section 4.5.1.

3.2.6 Variability in report likelihood of species

Report likelihood is the probability of reporting a species when it is found. This is completely dependent on observer interests, which can greatly vary per observer. Usually rare and attractive species are expected to have higher report likelihood than common ones. If observers report all the species they encounter, species with high population should have more observation counts than those with a low population, and vice versa. To check for the vriability in report likelihood of the species, we plot the count of observations of 243 species against their annual population as shown in the figure 3.8. The figure 3.8 shows that species with low population show a relatively high number of observations. This indicates that report likelihood varies with rarity of the species.

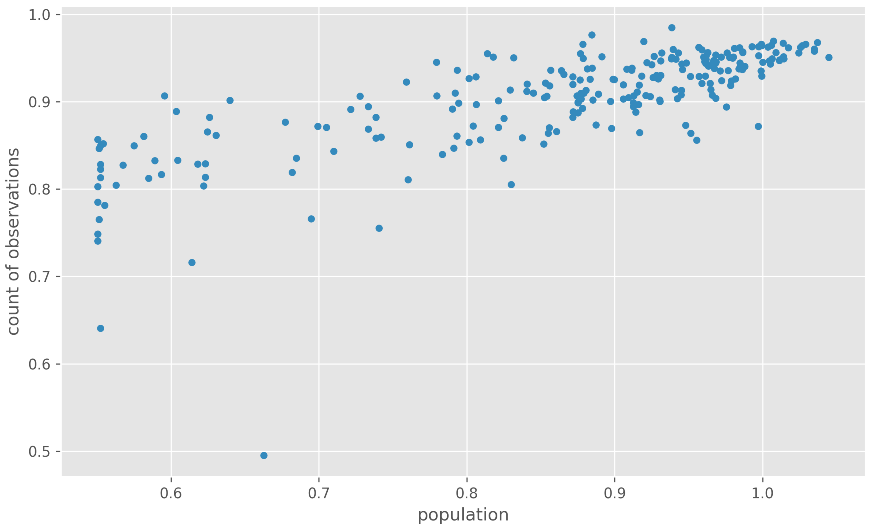


Figure 3.8: Scatter plot showing the variability in report likelihood of the species. Each point is a species with annual population on X-axis and observations count on Y-axis. The values on X and Y axes are natural log transformed populations and observation counts.

3.3 EXPLANATORY DATA

There are 357 explanatory variables that include 21 LULC classes, 30 water classes, and 306 soil classes. The amount of area covered in hectares by each of the 306 variables is calculated for all of 41,732 spatial blocks. There can be an overlap between LULC and soil classes. Table 3.4 shows the 21 LULC classes and their acronyms. Other explanatory variables that can be useful in understanding the habitat preference of species are average seasonal NDVI, and number of old buildings per block as some species prefer to breed in old buildings.

Table 3.4 LULC classes and their acronyms

	LULC	Acronym
1	jetty	JE
2	cropland	CR
3	basalt blocks	BA
4	built-up area	BU
5	orchard	OR
6	tree nursery	NT
7	mixed forest	FM
8	griend" forest, ie, riparian willows"	FG
9	deciduous forest	FD
10	coniferous forest	FC
11	graveyard	GY
12	forested graveyard	GF
13	dune	DU
14	fruit nursery	NF
15	grassland	GR
16	heathland	HE
17	other	OT
18	poplar (rows)	PO
19	railroad footprint	RR
20	sand	SA
21	bare	BR

3.4 SELECTION OF SPECIES AND SEASON SPECIFIC INPUT DATA

In this section we describe how some species are treated specially because of their unique occurrence behavior. It also explains why some categories of blocks are treated differently depending on the species, and time period of interest.

All the species considered are categorized into floating and non-floating species. These categories are obtained from expert ornithologists. Species that are divers, grebes, coromorants, duck and geese, gulls, terns and skuas, auks, gannets, petrels and shearwaters are categorized as floating and others as non-floating species. If **A** is the set of blocks covering the Netherlands, all the blocks in **A** are categorized into three groups, **A1**, **A2**, **A3**, based on their water cover. **A1** is the set of blocks in **A** that are completely covered with water throughout the year. **A2** is the set of blocks in **A** that are mudflats that occasionally go dry. **A3** is the set of blocks in **A** that are not in either **A1** or **A2**. The set of blocks **A1** is masked from **A** before they are sent into the classifier. All the masked blocks are by default labelled as species-is-absent because we do not expect species to occur on water, but there are exceptions. Some of the floating species winter on open waters. So, there are chances for floating species to be present on open waters. The set of blocks **A1** are by default labelled as species-is-absent in the range map of species s and time period t except for cases where the s is a floating species and t is winter. The set of blocks **A2** that occasionally go dry are not masked because species like waders feed on micro organisms that are exposed when the mudflats go dry. So there are chances of species occurrence on mud flats. As the time period between the mudflats going dry and wet is very short we consider them in the input data of the classifier for all the species and time periods. Depending on the species and time period of interest, the blocks in **A** are selected as input data for the model.

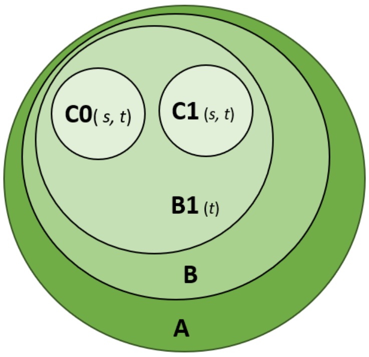
Chapter 4

Methodology

4.1 AIMS AND INTUITION

The aim of this project is to generate four seasonal range maps for a given species s and combine them to generate an annual range map for s. The method described to generate a seasonal range map is designed in such a way that it is applicable for any given species in the set of observations Ω, and any time period or season. As the method to generate a seasonal range map is the same for any species and time period, it is explained by considering an abstract species s, and time period t. Where s can be any species in Ω and t can be any season.

A range map represents the inferred presence and absence of a species in both space and time. Each block in the seasonal range map will have a binary class value of either 0 or 1. These values represent the species absence and presence respectively. Such binary class values for all the blocks are obtained as an output of some supervised classifier like Random Forest (RF). Like all supervised machine learning methods, RF needs training data, i.e., a certain number of already labelled blocks. Ideally, the labelled training data represents the ground truth with high certainty, in this case meaning that we have true absence and presence information of bird occurrence. Unfortunately, we do not have such ground truth data. So, there is a need to generate a unique training dataset for the pair (s, t) from Ω. After some pre-processing, an RF classifier is trained using the training data for (s, t), and explanatory variables mentioned in section 3.3. The trained classifier is used to generate the range map for s, t. The same method is repeated for the other three seasons. Finally, each combination of four seasonal occurrences of the species s per block b is translated into one of six classes determined by conditions discussed in section 4.8, and resulting in annual range map for s, as shown in figure 4.2.



A : All blocks of NL
B : Blocks in **A** for which we have observations
B1(t) : Sufficiently visited blocks in **B** in time period t
C0(s, t) : Blocks in **B1**(t) classified as negative, i.e. species s is absent during period t
C1(s, t) : Blocks in **B1**(t) classified as positive, i.e. species s is present during period t
s denotes a species, and t a time period

Figure 4.1: Block set and subsets selected at different stages of the method

Let **A** be the total blocks of the Netherlands represented as a set in figure 4.1. **B** is the set the blocks in **A** that have at least one observation. Let **B1(t)** is the set of blocks in **B** that are sufficiently visited during time period t, which is determined by some condition described below.

C0(t,s) represents the set of blocks where species s is present during t and **C1(t,s)** represent the set of blocks where species s is absent. How we choose blocks to be in these sets is again discussed below. **C0(t,s)** and **C1(t,s)** are obtained from **B1(t)** and used as training data for the classifier to generate range map of species s during time period t. The set of blocks **B1(t)** depends on the time period t because of varying observer effort through time as we observed in section 3.2.3. Similarly **C0(t,s)** and **C1(t,s)** varies with species and time because species occurrence varies within species and seasons.

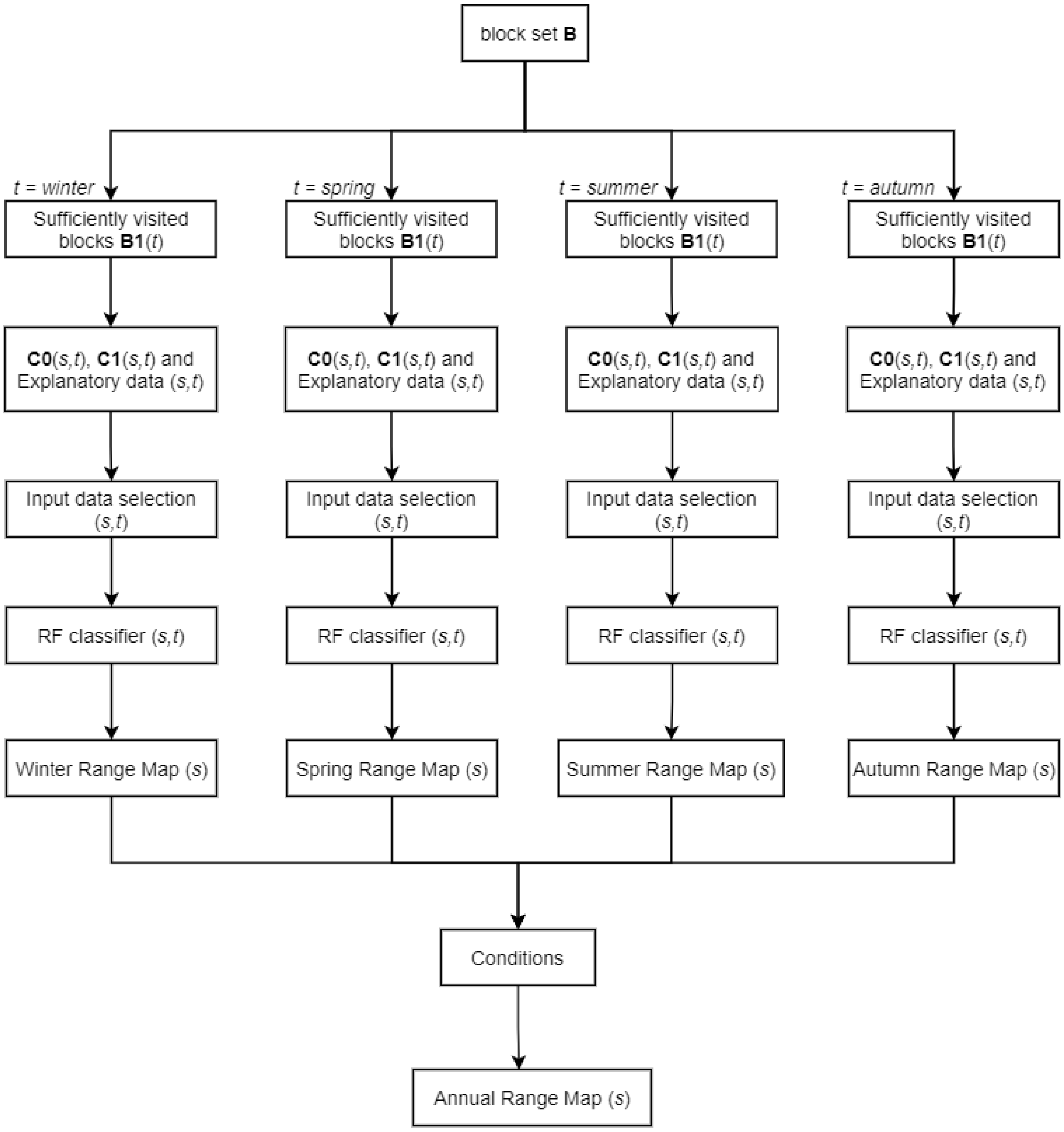
There are four major steps in the method. First, to select the sufficiently visited blocks **B1(t)** for time period t from **B**. Second, to determine sets **C0(t,s)** and **C1(t,s)** for species s and time period t, picked from **B1(t,s)**. Third, after selecting the input data for the given (species, season) combination , we train an RF classifier using **C0(t,s)** and **C1(t,s)** and generate the seasonal range map for (s, t). Fourth, we repeat the three stages for the other three seasons and thus generate four seasonal range maps. These are combined to obtain the annual range map for s. This last step in described in section 4.8. A workflow with sequence from stage one to stage four is illustrated in the figure 4.2.

The challenge that we face here is the construction of representative sets for blocks that can be used as training for absence and presence of the species. We propose to use the observation set Ω to this end, and describe here intuitively the approach that we take. In sections 4.2 to 4.6, we will describe the foundation to this approach more formally.

Species absence and presence (in a location, in a season) is expected to correlate somehow with how often the species is reported (in a location, in season). Instead of looking at absolute numbers of species encounters, which would be highly skewed by the number of observers visiting, we can determine the number of encounters of a species against the number of visits by observers. These two notions we shall call $ObserverDays$ and $EncounterDays$, and they are defined in sections 4.2 and 4.4. Their ratio we shall call $observedHitRate$, and it is defined in section 4.5. These three variables are denoted by $OD(b, t)$, $ED(b, t, s)$ and $OHR(b, t, s)$.

We would be keen to use $observedHitRate$ directly as a proxy for presence, but this will not work well. The reason is that each encounter is the result of a three-step probabilistic process: the bird must be present, the observer must find (detect) it, and the observer must think the observation relevant enough to report it to the WNL platform. While the first of these steps represents species presence, which is what we aim to classify and map, *detectability* and *report likelihood* are parts of the process that we do not want to map. These two notions are explained in the section 4.5.1. In other words, we should compensate their effects in $observedHitRate$ to obtain an unbiased estimator for presence. That unbiased estimator is called ESP and it is defined in section 4.5.2.

Once we have defined, and are able to compute, the estimator ESP, we can postulate a threshold value above which a species is assumed present, and below which a species is assumed to be absent. How this threshold is determined is explained in the section 4.5.2. The threshold will not be used on arbitrary blocks, but only on blocks that are sufficiently visited. The threshold value for ESP can then be used to determine two similarly sized sets of blocks **C0(t,s)** and **C1(t,s)**. How we do this is discussed in section 4.6.



- *s* = species
- *t* = season or time period
- *b* = block
- **C0** = set of blocks classified as species-is-absent, varies with *s* and *t*
- **C1** = set of blocks classified as species-is-present, varies with *s* and *t*
- **B1** = set of sufficiently visited blocks, varies with *t*
- **B** = set of blocks with at least one observation

Figure 4.2: Overall workflow for generating an annual range map of s from Ω

It turns out that the data that we have in Ω does not allow an easy estimation of either species detectability or species report likelihood, for that matter. This is because we have no knowledge about observers that find species but do not report them. Fortunately, the model that we propose does not need isolated estimators for either factor, and only needs an estimator for the two factors combined. That estimator is introduced as $DETREL$ in section 4.5.1.

Now, it happens to be the case that we know that not all observers are equally good at detecting species, nor are they equally likely to report their findings. Some also report insistently on almost each of their visits, while others cherry-pick and report only the most special of their finds. In short, we have high skilled observers and low skilled observers, and one may suggest that such differences in skills are quantified and play a role in how we handle the associated observations. This intuition is followed in section 4.9. The inclusion of observer weights requires a revisit of notions defined previously.

4.2 ESTIMATION OF OBSERVER EFFORT

Observer effort would ideally be measured by the amount of time (hours/minutes) that an observer devotes to visiting the block under study within the time period. A summation of such durations over all observers would inform us of observer effort. But we do not have sufficiently rich data to determine effort in this way. Thus, we choose to make use of a more coarse-grain measure for estimated observer effort, which we call *observer day*. An observer day is recorded for a block and time period per unique observer if we have at least one observation by the observer on a day in the block and time period. In other words, multiple observations by the same observer in a block on a day count for only one observer day, but observations from different observers or on different days contribute different observer days.

We thus define the observer days in a block and time period as follows:

$$\begin{aligned} &\forall b \in Blocks, t \in TimePeriods : \\ &\qquad OD(b,t) = \\ &\qquad count(unique(\mathbf{S}\; r.observer, r.date\; \mathbf{F}\; r \in \Omega\; \mathbf{W}\; r.block = b\; and\; r.date\; in\; t)) \end{aligned} \tag{4.1}$$

The definition is based on the complete set of observations Ω. The notation $\mathbf{S} \ldots \mathbf{F} \ldots \mathbf{W} \ldots$ is best understood as a SQL select–from–where expression. Here we select those observations within the given block and time period, and we collect $(observer, date)$ pairs. The $unique()$ operator removes any duplicate pairs, and $count()$ just gives us the number of pairs remaining.

4.3 SELECTION OF SUFFICIENTLY VISITED BLOCKS

Blocks with high observer effort are more reliable for inferring species occurrence than blocks with less observer effort. As observer effort is quantified as observer days, blocks with accumulated observer days above the 75th percentile are classified as sufficiently visited and others as insufficiently visited. The value on 75th percentile of observer days changes with time period, so the threshold value also changes accordingly. Using threshold for observer days (value on 75th percentile) we get around 25% blocks that decently represents the study area.

The block sets **C0(t,s)** and **C1(t,s)** are selected from the sufficiently visited blocks **B1(t)**. As the block sets **C0(t,s)** and **C1(t,s)** are used for training the classifier, they should fairly represent the study area. But **C0(t,s)** and **C1(t,s)** changes with species and time period. It is difficult to check their representation each time. So we ensure the representativeness in **B1(t)** as **C0(t,s)** and **C1(t,s)** are selected from **B1(t)**.

Representativeness of blocks in **C0** and **C1** can be assessed from the distribution of LULC classes that are highly related to species occurrence. The relation between LULC and species occurrence varies with species and time. So, we consider 7 out of 22 LULC classes that are generally important for occurrence of any species and time. They are Mixed, Deciduous, and Evergreen forests; Cropland; Grassland; Built-up; and Others. If the distribution of these seven LULC classes in **C0(t,s)**, and **C1(t,s)** is similar to the distribution of same LULC classes all over the Netherlands, then it can be said that the block sets **C0(t,s)**, and **C1(t,s)** fairly represent the study area. The distribution of the seven LULC classes in the Netherlands is shown in figure A.2 and the distribution of the same LULC classes in sufficiently visited blocks during winter, spring, summer, and autumn are shown in the figures A.3, A.4, A.5, and A.6 respectively.

4.4 ESTIMATION OF SPECIES PRESENCE

The presence of a species s in block b within time period t is quantified as encounter days $ED(b, t, s)$. It tells us how often s is found and reported in b within t. The definitional restriction is that multiple observations of s in block b on day d (with $(d \in t)$ by the same observer count as one encounter day. This restriction allows us to define a useful ratio with observer days that expresses how common the species was in the block b and time period t.

$$\begin{aligned}&\forall b \in Blocks, t \in TimePeriods, s \in Species :\\ &\qquad ED(b,t,s) =\\ &\qquad count(unique(\mathbf{S}\ r.observer, r.date, r.species\ \mathbf{F}\ r \in \Omega\\ &\qquad\qquad \mathbf{W}\ r.block = b\ and\ r.date\ in\ t\ and\ r.species = s))\end{aligned} \tag{4.2}$$

Here we select those observations of species s within the given block b and time period t, and we collect $(observer, date, species)$ triplets. The $unique()$ operator removes any duplicate triplets, and $count()$ just gives us the number of triplets remaining.

4.5 ESTIMATION OF SPECIES PRESENCE RELATIVE TO OBSERVER EFFORT

The ratio of encounter days $ED(b, t, s)$ and observer days $OD(b, t)$ is called observed hitrate $OHR(b, t, s)$. As mentioned above, $OHR(b, t, s)$ can be used as a proxy for species presence, but this will not work well because each encounter is the result of a three-step probabilistic process:the bird must be present, the observer must find (detect) it, and report it. We aim to map only the first part, species presence, but not the second and third parts of the process. The process of seperating species presence from the three step process is explained in the section 4.5.2.

4.5.1 Estimation of detectability and report likelihood

In this section, we first define the notion of species detectability and that of species report likelihood. Next, we explain our attempt to estimate detectability and report likelihood from Ω. Then we explain how detectability and report likelihood are estimated.

Detectability

Detectability ($DET(b, t, s)$) is the probability that an observer detects (finds) a species s in block b in a time period t, when it is known to be present. It varies with species, space and time. There are many factors that influence detectability. Some of the important factors are species abundance, habits (roosting, flocking, singing), habitat preference, observer effort etc. But from Ω and LULC data, we can only consider the factors habitat preference and observer effort in estimating detectability because it is tough to infer species abundance and habits from Ω. Given the data limitations, we assume that detctability of a species varies with time and remains constant through space.

Report Likelihood

Report likelihood ($REL(o, t, s)$) is the probability that an observer o reports a species s in time period t when s/he detects it. $REL(o, t, s)$ depends on observer interests that vary with observer, species, and time. It is not a common practice for an observer to report all the individuals of a species or distinct species detected. From the section 3.2.6, it was understood that observers report rare species relatively more than common ones. For a simplified estimation of report likelihood we assume that report likelihood varies with species and time and remains constant through space and observers.

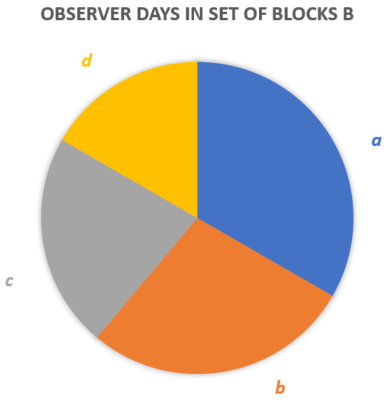
Attempt for separate estimation of detectability and report likelihood

We made an attempt to estimate *detectability* $DET(t, s)$ and *report likelihood* $REL(t, s)$ of species s, within time period t as two different quantities using the observations in the set of blocks **B**. However, this attempt did not succeed because of the data limitations discussed below.

In figure 4.3, a is the number of observer days in the set of blocks **B** during which species s was found and reported within time period t. Not all observers report all the distinct species they encounter, so b is the number of observer days during which s was found but not reported. c is the number of observer days during which s was present but was not found, and d is the number of observer days during which s was absent so it was not found.

Detectability of species s within time period t from observations in the set of blocks **B** can be estimated as the ratio of $a + b$ and $a + b + c$ as expressed in equation 4.3.

$$DET(t, s) = \frac{a + b}{a + b + c} \tag{4.3}$$



a : s was found and reported
b : s was found but not reported
c : s was present but not found
d : s was not present so not found

Figure 4.3: Observer days in the set of blocks $\mathbf{B}$ within time period t; a, b, c and d are number of observer days

Report likelihood of species s in time period t from observations in the set of blocks $\mathbf{B}$ can be estimated as the ratio of a and $a + b$ as expressed in equation 4.4.

$$REL(t, s) = \frac{a}{a + b} \tag{4.4}$$

In equations 4.3 and 4.4, a is known and b, c, d are unknown. It is difficult, if not impossible, to estimate the values of b, c, d from Ω. We have tried estimating $REL(s, t)$ from seasonal and annual population of species but only b can be derived from $REL(b, t)$, and we need c in addition to estimate $DET(t, s)$. However, it can be understood from definition 4.9 that $DET(t, s)$ and $REL(t, s)$ can be estimated together as one quantity, called $DETREL(t, s)$, and use it to estimate $ESP(b, t, s)$ as expressed in definition 4.10.

Estimation of detectability and report likelihood as one quantity

In this section, we discuss two estimates of combined values of detectability and report likelihood, seen as the product of the two, as $DETREL1$ and $DETREL2$. Both are estimated from our dataset Ω as explained below.

$DETREL1(t, s)$ and $DETREL2(t, s)$ of species s and time period t are two estimates for the combined probability for finding and reporting the species s in time period t. Both are estimated from the same dataset Λ, which we define as follows.

$$\Lambda(t, s) = \{\ (b, d) \in \mathbf{B} \times t \mid d \in t \text{ and } OD(b, \{d\}) \geq 2 \text{ and } ED(b, \{d\}, s) \geq 1\ \} \tag{4.5}$$

Our Λ is a set of block b, day d pairs, specific for a chosen species s and time period t. The block must have been visited by at least two observers on the given day, and the species must have been encountered in the block on that day. The idea here is that the stated restrictions allow us to focus on days where we know the species was present. If we then look at all observers visiting the block on the same date, we can determine which ratio of them reported the species as well. This ratio allows us to estimate detectability.

The metric $DETREL2(t,s)$ of species s in time period t is estimated as the ratio of sum of encounter days of b, d pairs in Λ reduced by 1, and sum of observer days of b, d pairs in Λ reduced by 1. The numerator, and denominator are reduced by 1 to remove the observation used to confirm the species presence.

$$DETREL2(t,s) = \frac{\sum_{(b,d)\in\Lambda(t,s)} ED(b,d,s) - 1}{\sum_{(b,d)\in\Lambda(t,s)} OD(b,d) - 1} \tag{4.6}$$

The second metric $DETREL1(t,s)$ of species s in time period t is estimated as the ratio of the sum of encounter days of b, d pairs in $\Lambda(t,s)$, and sum of observer days over the same set.

$$DETREL1(t,s) = \frac{\sum_{(b,d)\in\Lambda(t,s)} ED(b,d,s)}{\sum_{(b,d)\in\Lambda(t,s)} OD(b,d)} \tag{4.7}$$

The difference between the two $DETREL$ notions is that in the first we consider one observation of the species 'used up' as evidence for its presence, which is subsequently not accounted for in determining the ratio.

Either $DETREL1(t,s)$ or $DETREL2(t,s)$ can be used to account for variability in detectability, and report likelihood in estimating unbiased species presence as expressed in the definition 4.10. In this method we choose to work with $DETREL2(t,s)$.Out of 765 distinct species in Ω, we calculate $DETREL1(t,s)$ and $DETREL2(t,s)$ for 213 species. The species left out are either very rare or other varieties of the selected species.

4.5.2 Separating species presence from the three-step probabilistic process

An estimate for the probability of finding and reporting a species s in block b within time period t, $ESP(b,t,s)$, can be expressed as the ratio of observed hitrate $OHR(b,t,s)$ and $DETREL(t,s)$ as shown in the expression 4.9. We divide $OHR(b,t,s)$ with $DETREL(t,s)$ to separate the probability of species presence from the three-step probabilistic process mentioned above. $ESP(b,t,s)$ is estimated for each block in **B1(t)**. Then we determine the threshold for $ESP(b,t,s)$. All the blocks in **B1** with $ESP(b,t,s)$ value above the threshold are labelled as species-is-present and others blocks as species-is-absent. The classes species-is-present and species-is-absent will be imbalanced in most of the cases. SO, we balance them by random resampling and obtain the training classes **C0(t,s)** and **C1(t,s)** as explained in the section 4.6.

$$OHR(b,t,s) = \frac{ED(b,t,s)}{OD(b,t)} \tag{4.8}$$

$$ESP(b,t,s) = \frac{OHR(b,t,s)}{DETREL2(b,t,s)} \tag{4.9}$$

To determine the threshold for $ESP(b,t,s)$, we select eight (species, season) pairs that have high $DETREL2(t,s)$ values. High $DETREL2(t,s)$ means high probability of finding and reporting the species when it is known to be present. So, for the species selected during the respective seasons, the effect of detectability and report likelihood in the three-step probabilistic process is relatively low. We try different thresholds for $ESP(b,t,s)$ and obtain different training classes **C0(t,s)** and **C1(t,s)** for the selected (species, season) combinations. Then we train an RF classifier and generate range maps for the selected (species, season) pairs as explained in the section 4.7. Then the predicted range maps are compared with the range maps obtained from independent sources. The threshold for $ESP(b,t,s)$ that leads to better range maps of selected (species, season) combinations is 0.07. This threshold is applicable for any (species, season) combination. As all the selected (species, season) pairs have almost similar $DETREL2(t,s)$ values, we can use their average $DETREL2(t,s)$ value, here by called $cDETREL$, in expression 4.9 instead of using the $DETREL2(t,s)$ of the respective (species, season) combination as shown in expression 4.10. The value of $cDETREL$ is 0.3942.

$$ESP(b,t,s) = \frac{OHR(b,t,s)}{cDETREL} \tag{4.10}$$

As the selected (species, season) combinations have high and similar $DETREL2(t,s)$, we have used their average $DETREL2(b,t,s)$ in the expression 4.9 and obtained 4.10. But we cannot do the same for any given species because they might vary in terms of detectability and report likelihood compared to the selected combinations. So, $ESP(b,t,s)$ of (species, season) combinations other than the selected eight combinations can be defined as the ratio of product of $OHR(b,t,s)$ and $DETREL2(t,s)$; and $cDETREL$ as shown in the expression 4.11. Multiplying $OHR(b,t,s)$ with $DETREL2(t,s)$ adds the information on detectability and report likelihood of the species and dividing with $cDETREL$ removes the effect of detectability and report likelihood from the definition.

$$ESP(b,t,s) = \frac{OHR(b,t,s) \times DETREL2(t,s)}{cDETREL} \tag{4.11}$$

$ESP(b,t,s)$ of the selected eight (species, season) is defined as expressed in 4.10 and for other species it is defined as expressed in 4.11. The threshold for $ESP(b,t,s)$ 0.07 is applicable for all the (species, season) combinations.

4.6 SELECTION OF TRAINING DATA

$ESP(b,t,s)$ of species s during time period t is calculated for all the blocks in **B1**. Blocks with $ESP(b,t,s)$ value above 0.07 (the threshold for $ESP(b,t,s)$) are labelled as species-is-present and below as species-is-absent. The number of blocks in the classes species-is-absent and species-is-present might not be similar. If N is the number of blocks in the class species-is-present, and M is the number of blocks in the class species-is-absent, and $M > N$, then the classes are balanced by randomly picking N number of blocks in the class species-is-present. After the classes are balanced, the blocks in the class species-is-present become the members of **C1**, and the blocks in the class species-is-absent become the members of **C0**. These class memberships, **C0** representing absence of s during t, and **C1** representing presence of species s during t are used for training the RF classifier in the next stage.

4.7 TRAINING A RANDOM FOREST CLASSIFIER

A Random Forest (RF) classifier is trained to predict the occurrence of species s in time period t for all the blocks in the input data selected as explained in the section 3.4. The set of blocks **C0(t,s)** that represent the species absence and the set of blocks **C1(t,s)** that represent the species presence are used to train the classifier. For each block in **C0(t,s)** and **C1(t,s)**, the area size in hectares by 21 LULC classes, 30 water classes, and 306 soil classes discussed in section 3.3 are used as explanatory variables. As the model is generalised for multiple species in Ω, default parameters of the RF classifier in scikit-learn version 0.24.2 are used. Once the classifier is trained, it is used to predict the class values for all blocks, and the range map of species s and time period t is generated.

4.8 COMBINATION OF SEASONAL RANGE MAPS INTO AN ANNUAL RANGE MAP

All the steps described in sections 4.2 to 4.7 are repeated for the species for the three other seasons to obtain all seasonal range maps. The conditions used to combine these maps, classify a block into one of six classes in the annual range map using values from each seasonal range map are provided in table 4.1. The model first checks if the combination of seasonal occurrence of the species matches with the combination of resident class, if it matches the block will be labelled as the species-is-resident, otherwise it goes to the next condition in the table. For most of the species the conditions apply but there are some species that have non-breeding summer presence. The method maps such species under summer (breeding) presence though they do not breed but present in summer.

Table 4.1 Conditions to generate annual range map from seasonal range maps

Winter	Spring	Summer	Autumn	Class
1	0 or 1	1	0 or 1	Resident (**1**)
1	0 or 1	0	0 or 1	Non-breeding (winter) presence (**2**)
0	0 or 1	1	0 or 1	Summer (breeding) presence (**3**)
0	0 or 1	0	1	Post-breeding migratory presence (**4**)
0	1	0	0	Pre-breeding migratory presence (**5**)
0	0	0	0	Absent (**6**)

4.9 CALCULATION OF OBSERVER WEIGHTS

Not all observers are equal in terms of their contribution to the observations dataset. Observers who devote more effort and report more distinct species are considered as relatively superior to others. The skills of observers is quantified as $ObserverWeight(OW)$ which varies with observers and time. The observer weight $OW(o,t)$ of an observer o within time period t is calculated as the product of encounter days by observer o within time period t and the contributed observer days by o within t as expressed in definition 4.14. All values of $OW(o,t)$ within the same t are normalized to 0 and 1 using the formula 4.15. In this way, observers with a high observer effort and high number of distinct species reported per observer day are assigned a high weight value. Observer weights are added as an attribute $ObserverWeight(OW)$ in Ω, based on the observer and date of observation.

4.9.1 Definitional basis for observer weights

We count the distinct block visits of observer o on day d within time period t ($d \in t$) and express this as $ObserverDays(o, t)$:

$$\forall o \in Observers, t \in TimePeriods : \\ ObserverDays(o,t) = \\ count(unique(\mathbf{S}\ r.block, r.date\ \mathbf{F}\ r \in \Omega\ \mathbf{W}\ r.observer = o\ and\ r.date\ in\ t)) \quad (4.12)$$

Observe the analogy with definition equation 4.1, where we counted observer days per block.

We can also count the number of different species reported on separate days per observer in a time period. We express it as $EncounterDays(o, t)$:

$$\forall o \in Observers, t \in TimePeriods : \\ EncounterDays(o,t) = \\ count(unique(\mathbf{S}\ r.block, r.date, r.species\ \mathbf{F}\ r \in \Omega\ \mathbf{W}\ r.observer = o\ and\ r.date\ in\ t)) \quad (4.13)$$

Observe the analogy with equation 4.2.

The product of $ObserverDays(o, t)$ and $EncounterDays(o, t)$ is used as observer weight $OW(o, t)$ of observer o for time period t:

$$\forall o \in Observers, t \in TimePeriods : \\ OW(o,t) = ObserverDays(o,t) \times EncounterDays(o,t) \quad (4.14)$$

Observer weights obtained from definition 4.14 are normalized to a value between 0 and 1 using the equation 4.15 where each weight $OW(o, t)$ of an observer o per time period t is transformed. This process is repeated for all the observer weights in time period t obtained from definition 4.14.

$$NOW(o,t) = \frac{OW(o,t) - min_{p \in Observers} OW(p,t)}{max_{p \in Observers} OW(p,t) - min_{p \in Observers} OW(p,t)} \quad (4.15)$$

Here, $NOW(o, t)$ is the normalized weight of the observer o during tim period t, $max\ OW(p, t)$ is the highest weight among all observer (p) weights during time period t, and $min\ OW(p, t)$ is the lowest weight among all the observer weights during that time period.

4.9.2 Weighted observer days

The notion of weighted observer days is slightly different from observer days in section 4.2. The weighted observer days $wOD(b, t)$ of block b within time period t is quantified as the sum of observer weights of the observers who have visited the block b in day d over time period t (with $d \in t$) as expressed as:

$$\begin{aligned}
&\forall b \in Blocks, t \in TimePeriods: \\
&\qquad wOD(b,t) = \\
&sum(\mathbf{S}\ now\ \mathbf{F}\ unique(\ \mathbf{S}\ r.observer, NOW(r.observer,t)\ as\ now, r.date \\
&\qquad \mathbf{F}\ r \in \Omega \\
&\qquad \mathbf{W}\ r.block = b\ and\ r.date\ in\ t))
\end{aligned} \tag{4.16}$$

From the above definition, we get distinct $(observer, ow, date)$ triplets for block b within time period t. All the OW values obtained are added to determine the weighted observer days $wOD(b,t)$ of block b within time period t.

4.9.3 Weighted encounter days

Weighted encounter days are also calculated almost like $ED(b,t,s)$ as defined in section 4.4 but with a slight difference. Weighted encounter days $wED(b,t,s)$ of species s in block b within time period t is the sum of weights of distinct observers who have found and reported the species s in block b on day d over time period t (with $d \in t$) as expressed in the following definition.

$$\begin{aligned}
&\forall b \in Blocks, t \in TimePeriods, s \in Species: \\
&\qquad wED(b,t,s) = \\
&sum(\mathbf{S}\ now\ \mathbf{F}\ unique(\ \mathbf{S}\ r.observer, NOW(r.observer,t)\ as\ now, r.date, r.species \\
&\qquad \mathbf{F}\ r \in \Omega \\
&\qquad \mathbf{W}\ r.block = b\ and\ r.date\ in\ t\ and\ r.species = s))
\end{aligned} \tag{4.17}$$

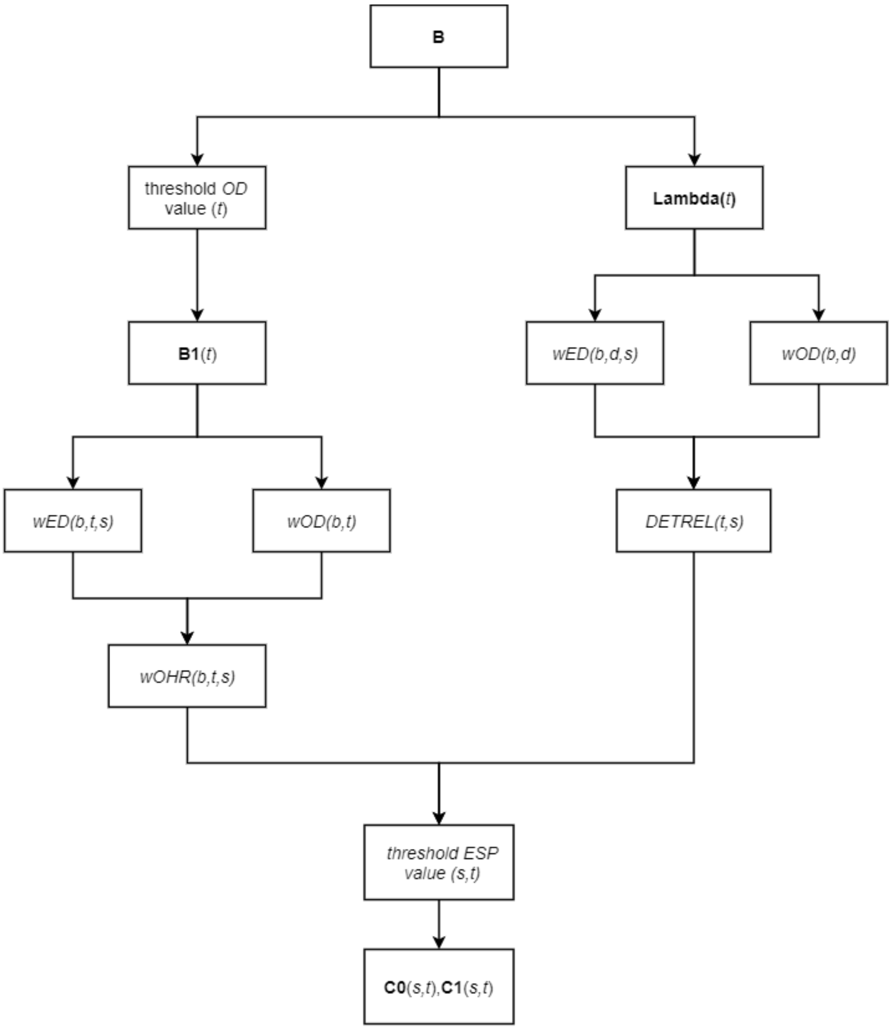
In the above definition, we obtain distinct $(observer, now(), date, species)$ quadruples of species s block b within time period t. All the NOW values obtained are added to determine $wED(b,t,s)$as a sum of $now()$ values for species s in block b within time period t.

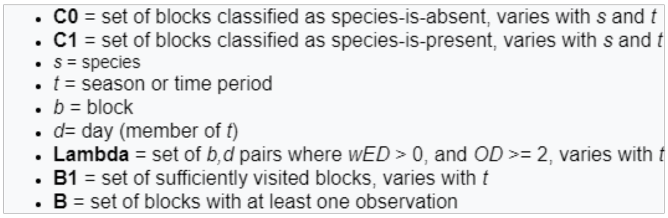
4.9.4 Weighted observed hitrate

Weighted observed hitrate $wOHR(b,s,t)$ is similar to $OHR(b,s,t)$ discussed in section 4.5 but it uses $wED(b,s,t)$ and $wOD(b,t)$ instead of $ED(b,t,s)$ and $OD(b,t)$. Weighted observed hitrate $wOHR(b,t,s)$ of species s in block b within time period t is the ratio of weighted encounter days $wED(b,t,s)$ of s in b within t and weighted observer days $wOD(b,t)$ of b within t as expressed in the definition 4.18.

$$wOHR(b,t,s) = \frac{wED(b,t,s)}{wOD(b,t)} \tag{4.18}$$

In order to account for variability in observer quality, we use weighted metrics and implement the steps from sections 4.2 to 4.7 to generate seasonal and annual range maps of species 423 and display them in chapter 5. The workflow to select **C0(t,s)**, **C1(t,s)** from **B1(t,s)**, and **B1(t,s)** from **B** using weighted metrics is illustrated in the figure 4.4.





- **C0** = set of blocks classified as species-is-absent, varies with *s* and *t*
- **C1** = set of blocks classified as species-is-present, varies with *s* and *t*
- *s* = species
- *t* = season or time period
- *b* = block
- *d*= day (member of *t*)
- **Lambda** = set of *b,d* pairs where *wED* > 0, and *OD* >= 2, varies with *t*
- **B1** = set of sufficiently visited blocks, varies with *t*
- **B** = set of blocks with at least one observation

Figure 4.4: Workflow to select **C0(t,s)**, **C1(t,s)** from **B1(t,s)**, and **B1(t,s)** from **B**

Chapter 5

Results

In this chapter, we zoom in on the species 423 and discuss the generation of seasonal maps and an annual range map using the method described in chapter 4. All the intermediate and final results obtained during the process to generate seasonal and annual range maps for species 423 are explained.

5.1 OVERVIEW OF INPUT DATA

Species 423 is a non-floating species. So, 7,250 blocks that are completely covered by water grouped under the set of blocks **A1** are by default classified as absent in all the seasonal range maps. Blocks that are not completely covered by water and have at least one observation are in the set of blocks **B**. The observations in 37,976 blocks in **B** is the input data of the model for generating seasonal and annual range maps of species 423.

5.2 SELECTION OF SUFFICIENTLY VISITED BLOCKS

As explained in section 4.9.2, we calculate weighted observer days $wOD(b, t)$ for each block in **B** and for all the four seasons. The seasonal distributions of $wOD(b, t)$ in the set of blocks **B** are visualised in figure 5.1. For each season, a threshold for wOD mentioned in table 5.1 is selected to label the blocks in **B** as sufficiently or insufficiently visited. The data in table 5.1 is independent of species so the seasonal thresholds for $wOD(b, t)$ apply for any given species.

Table 5.1 Species-independent threshold for wOD per season and number of blocks obtained

Season	Threshold	Blocks in **B**	Sufficiently visited blocks	Insufficiently visited blocks
Winter	17.1869	36,586	9,146	27,440
Spring	18.1109	37,725	9,432	28,293
Summer	11.5439	36,868	9,025	27,943
Autumn	11.3349	35,893	8,794	27,099

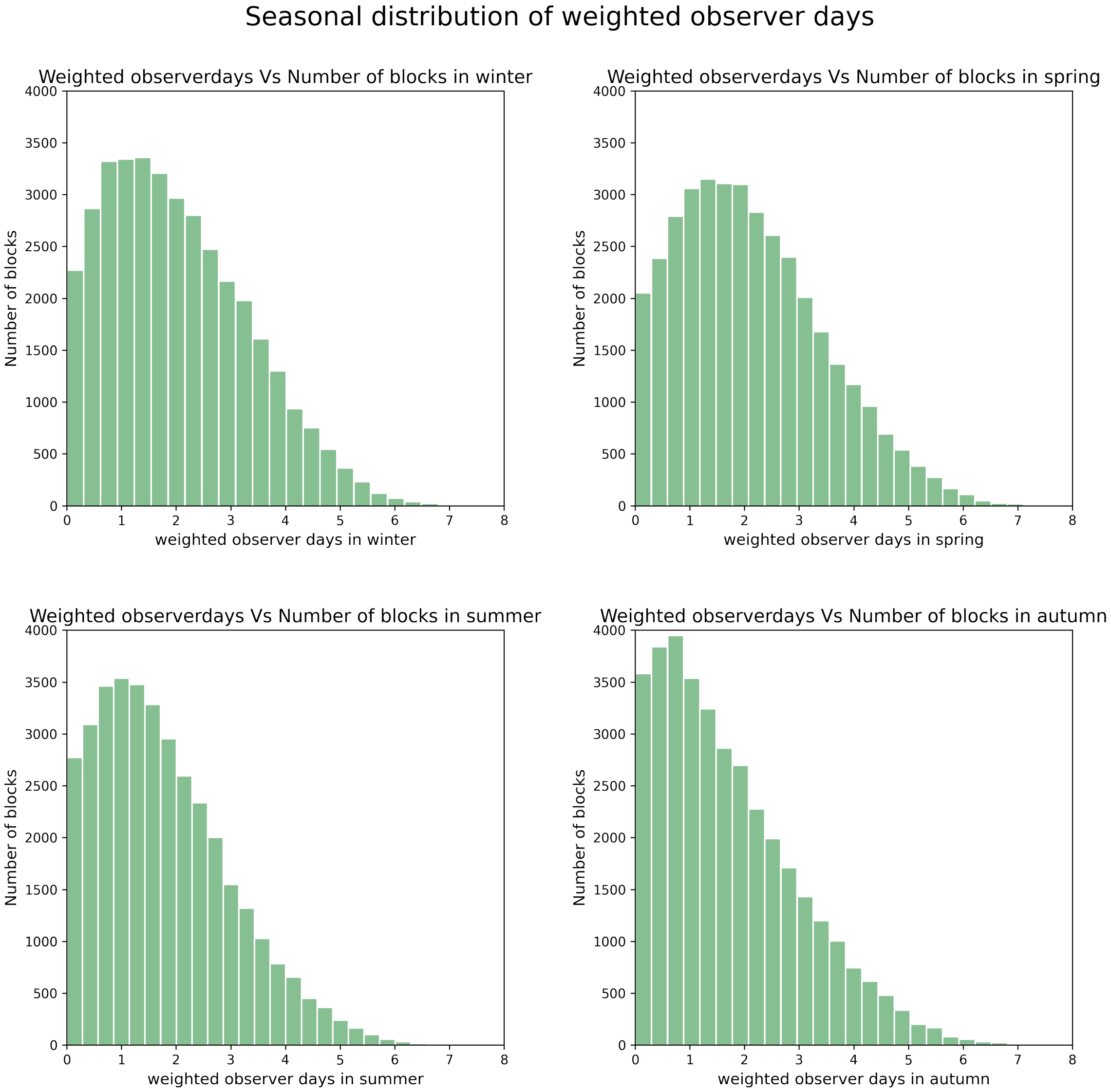


Figure 5.1: Distribution of weighted observer days in the set of blocks **B** for four different seasons. A bar with value 3 on X-axis represents the $wOD(b, t)$ value natural $log(3) = 1.098$ and the respective number on Y-axis represents number of blocks with $wOD(b, t)$=1.098

5.3 SELECTION OF THE CLASSES SPECIES-IS-PRESENT AND SPECIES-IS-ABSENT

For every block in **B1**, we calculate weighted encounter days $wED(b, t, s)$ of species 423 in all the seasons. This results in four values (one per season) for every block in **B**. The distribution of $wED(b, s, t)$ values of species 423 in all the four seasons is shown in figure 5.2. As the values of $wED(b, s, t)$ are skewed, blocks with zero $wED(b, s, t)$ are not included in the distributions in figure 5.2 for better visualization. Number of blocks with zero $wED(b, s, t)$ per season are mentioned in table 5.2. For every block in **B**, species 423, and four seasons, we calculate weighted observed hitrate $wOHR(b, s, t)$ as mentioned in the section 4.9.4. Distribution of $wOHR(b, t, s)$ of species 423 per season in **B1** is shown in figure 5.3. Similarly, blocks with zero $wOHR(b, t, s)$ values are not included in the histograms in figure 5.2. When $wOD(b, t, s)$ is zero, then $wOHR(b, t, s)$ will also be zero. So, the number of blocks with zero $wED(b, t, s)$ values is equal to number of blocks with zero $wOHR(b, t, s)$. The summary statistics of $oED(b, t, s)$ and $wOHR(b, t, s)$ per season are shown in tables 5.3 and 5.4.

Table 5.2 Number of blocks per season with $wED(b, t, s)$ and $wOHR(b, t, s)$ equal to 0

Season	$wED(b, t, s) = 0$ and $wOHR(b, t, s)$=0
Winter	3970
Spring	5206
Summer	7910
Autumn	4593

Table 5.3 Summary statistics of $wED(b, t, s)$ per season

Season	Mean	Std.	Min.	25%	50%	75%	Max.
Winter	2.775	5.866	9×10^{-5}	0.448	1.096	2.763	160.18
Spring	1.491	3.745	13×10^{-4}	0.292	0.635	1.458	108.34
Summer	0.982	2.551	4×10^{-5}	0.263	0.524	0.953	54.536
Autumn	2.639	8.868	4×10^{-5}	0.393	0.858	2.148	318.422

Table 5.4 Summary statistics of $wOHR(b, t, s)$ per season

Season	Std.	Min.	25%	50%	75%	Max.
Winter	5.576×10^{-2}	4.602×10^{-7}	9.236×10^{-3}	2.406×10^{-2}	5.721×10^{-2}	7.237×10^{-1}
Spring	2.791×10^{-2}	5.837×10^{-7}	4.398×10^{-3}	1.083×10^{-2}	2.416×10^{-2}	3.796×10^{-1}
Summer	2.88×10^{-2}	2.712×10^{-7}	5.77×10^{-3}	1.308×10^{-2}	2.598×10^{-2}	3.41×10^{-1}
Autumn	5.449×10^{-2}	9.369×10^{-8}	9.563×10^{-3}	2.359×10^{-2}	5.022×10^{-2}	5.846×10^{-1}

5.3.1 Selection of a threshold per season to obtain training classes

By multiplying $wOHR(b, t, s)$ with the ratio $DETREL1(t, s)$ and $cDETREL$ or with $DETREL2(t, s)$ and $cDETREL$ we obtain two $ESP(b, t, s)$ values per season for the species 423. However we chose to work with the latter ratio, $DETREL2(t, s)$ and $cDETREL$ because it gives comparatively better results . The value of $cDETREL$ (0.3942) is independent of species and season. Using the constant 0.07 as the threshold for $ESP(b, t, s)$ all the sufficiently visited blocks **B1** are labelled as species 423-is-present or species 423-is-absent. The two classes, 423-is-present and 423-is-absent, are balanced and $C0(t, s)$ and $C1(t, s)$ are obtained as mentioned in the section 4.6. The number of blocks in class 423-is-present and 423-is-absent, and number of blocks in **C0(t,s)** and **C1(t,s)** are mentioned in the table 5.5.

5.4 RESULTS OF RANDOM FOREST CLASSIFICATION

An RF classifier per season, species combination is trained using the class memberships **C0(t,s)**, **C1(t,s)**, and the explanatory variables mentioned in section 3.3. Each trained classifier is used

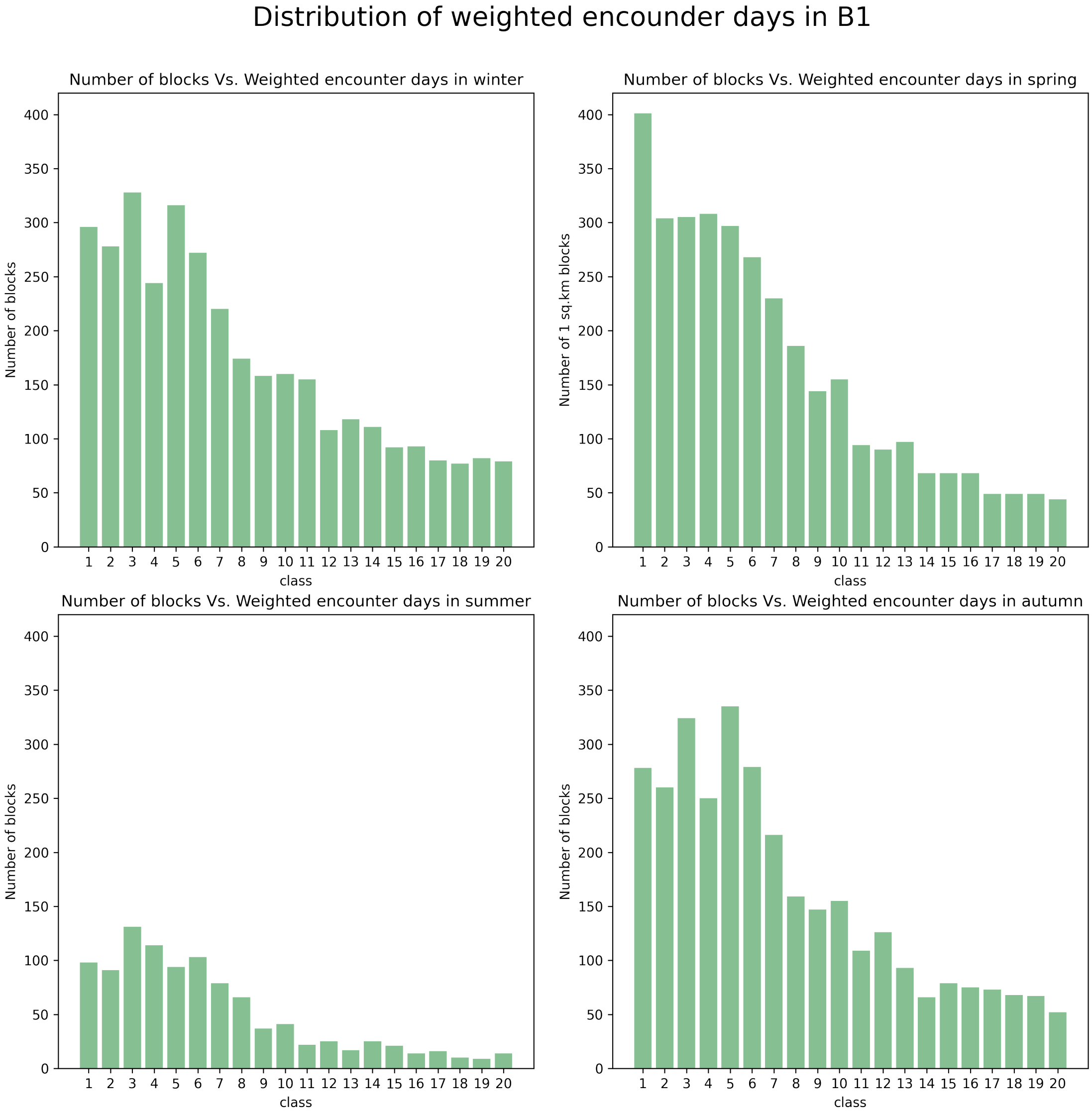


Figure 5.2: Distribution of weighted encounter days in the set of blocks **B1**. Class 1 on X-axis of all the histograms represent the values above 0 and below 0.1, class 2 represents values between 0.1 to 0.2, and so on, class 20 represents the values greater than 1.9. The range of each class from 1 to 20 is increased by a multiple of 0.1. Values on Y-axis represents the number of blocks. SO, each bar in the histogram represents the number of blocks on Y-axis with $wED(b, t, s)$ values within respective class and season on X-axis.

Table 5.5 Number of blocks per season in different classes with threshold for $ESP(b, t, s)$ at 0.07

Season	423-is-present	423-is-absent	**C0(t,s)** and **C1(t,s)**
Winter	981	8,165	981
Spring	158	9,031	158
Summer	74	8,960	74
Autumn	366	8,608	366

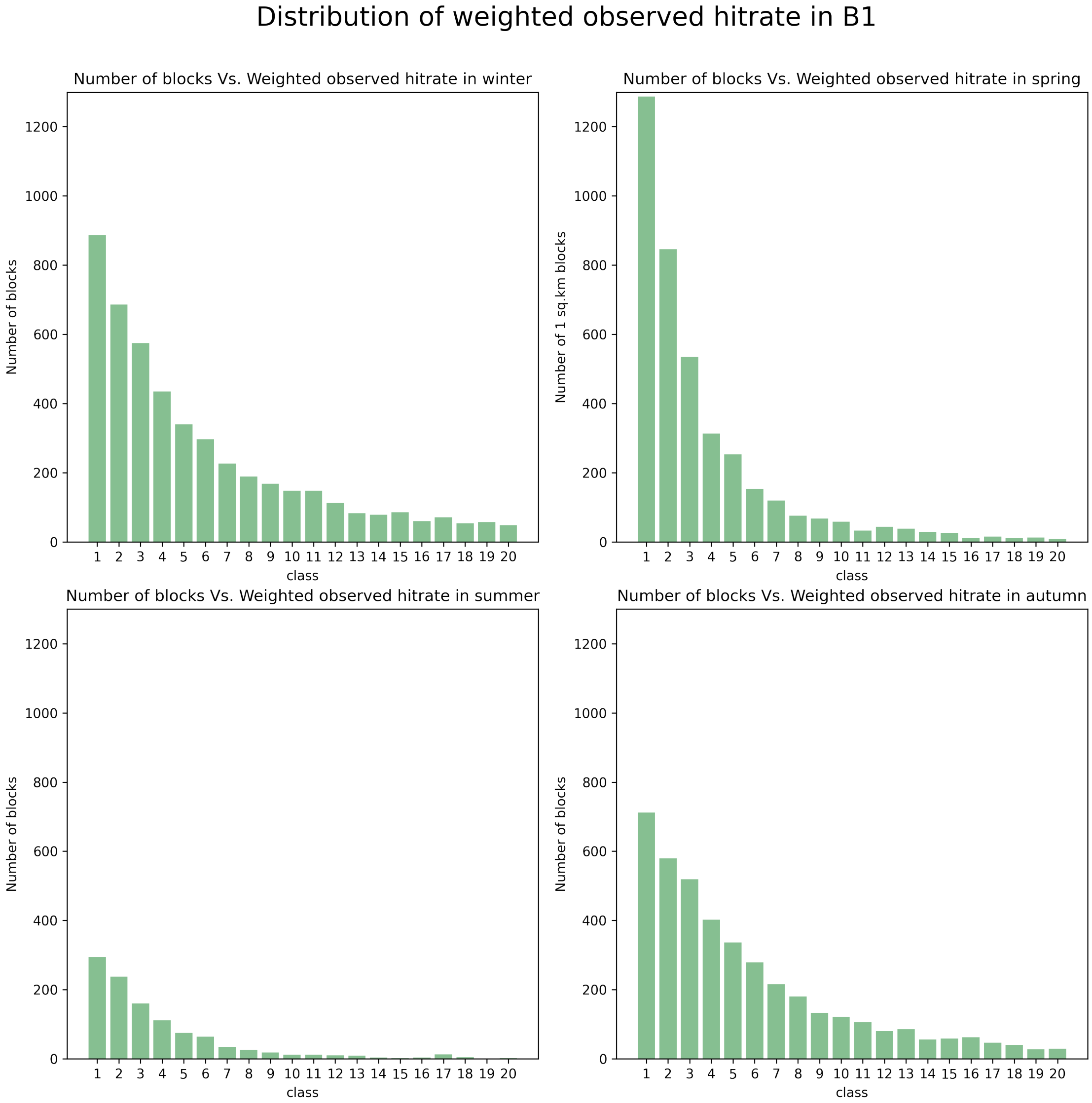


Figure 5.3: Distribution of $wOHR(b, t, s)$ in the set of blocks **B1**. Class 1 on X-axis represents the $wOHR(b, t, s)$ values above 0 and below 0.006, class 2 represents $wOHR(b, t, s)$ values between 0.006 to 0.012, class 3 represents $wOHR(b, t, s)$ values between 0.012 and 0.018, and so on, value 20 represents $wOHR(b, t, s)$ values above 0.114. The range of each class from 1 to 20 is increased by a multiple of 0.006. So, each bar in the histogram represents the number of blocks on Y-axis with $wOHR(b, t, s)$ values within respective class and season on X-axis.

to generate one of four seasonal range maps for species 423 shown in figure 5.4. Based on the conditions mentioned in section 4.2, these seasonal maps are combined to obtain the annual range map of species 423, shown in figure 5.5.

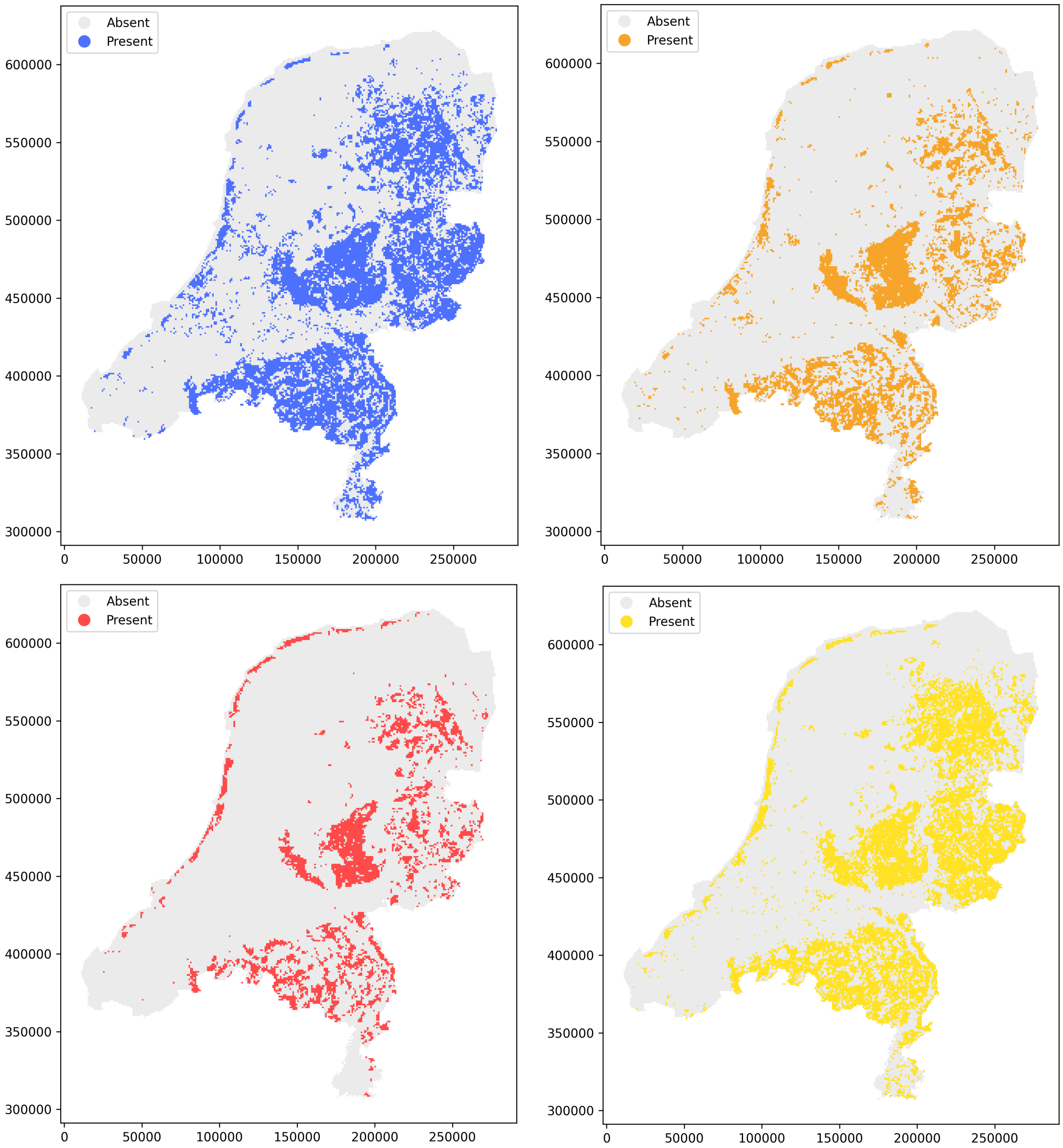


Figure 5.4: Seasonal range (presence) maps for species 423. The threshold for $ESP(b, t, s)$, the constant, is 0.07 and the constant $cDETREL$ is 0.3942. Top left: winter; top right: spring; bottom left: summer; bottom right: autumn. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841)

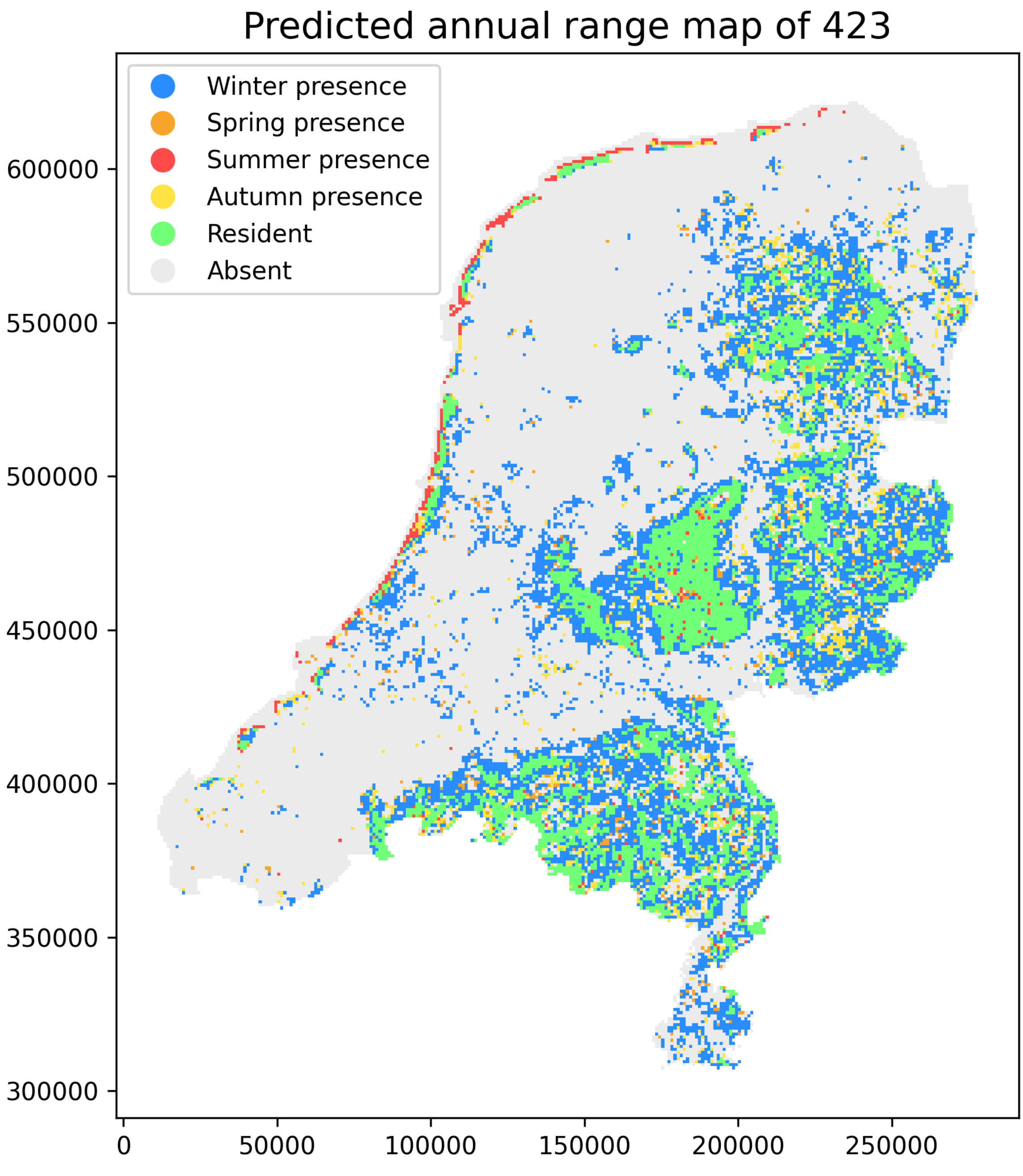


Figure 5.5: Predicted annual range map of species 423. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841)

Chapter 6

Validation of modelling outcomes

In this chapter, we explain how seasonal range map data for validating the predicted seasonal range maps was prepared from the seasonal abundance maps obtained from the Sovon Atlas (Sovon Vogelonderzoek Nederland, 2019). This atlas provides no data on spring, autumn and annual range, so all we can do is validate against summer and winter maps. Next, we explain the classification reports of winter and summer classifiers of species 423. Then the predicted range maps of eight (species, season) pairs selected to determine the constants $cDETREL$, that is 0.3942, and the threshold for $ESP(b, t, s)$, that is 0.07 are validated using the respective validation range map data obtained from Sovon Atlas. Weighted metrics($wOD(b, t)$, $wED(b, t, s)$, and $wOHR(b, t, s)$) are used instead of non-weighted metrics ($OD(b, t)$, $ED(b, t, s)$, $OHR(b, t, s)$)) to account for the variability in observer skills. In section 6.3, the range maps of the selected eight (species, season) pairs generated by models using weighted and non-weighted metrics are compared.

6.1 PREPARATION OF VALIDATION MAPS FROM THE SOVON ATLAS

Winter and summer abundance maps for the selected species, season pairs for validation are obtained from the Sovon Atlas (Sovon Vogelonderzoek Nederland, 2019). Range maps for spring and autumn are not available, so we do not validate spring and autumn range maps. Each map has eleven abundance classes as shown in the figureA.8, and A.9. In the abundance maps, all the blocks that are mudflats and blocks completely covered with open sea water are by default classified into class species-is-absent. But our model includes mudflats for all the species and considers water blocks for floating species during winter. These atlas abundance maps are the output of some model themselves, so these maps are not identical to true abundance maps, but they are the best possible data to compare against. However, we convert the abundance maps to range maps by labelling all the blocks into species-is-present and species-is-absent. When all the blocks with abundance values above zero were labelled into species-is-present, the resulting range map was more optimistic. So we have decided to work on a cut-off abundance class value where all the blocks with abundance class below the cut-off can be labelled as species-is-absent and above as species-is-present.

For each of the selected eight (species, season) pairs, four different seasonal range maps for validation were obtained by using different cut-off for the abundance to label the blocks in the respective seasonal abundance maps of the selected (species, season) pairs as species-present and species-is-absent. The four cut-off abundance class values used to obtain four validation maps are 0, 1, 2, 3 and 4. Then we have compared the predicted range map of each of the eight (species, season) pairs with the respective four validation maps and calculated the classification accuracies and F1 scores as shown in table A.1. It can be observed that the classification accuracy and F1 scores are high for majority of the species when the cut-off for abundance class is 3. So we have decided to use the cut-off for abundance class at 3 to obtain seasonal range map for validation for any species, season.

6.2 VALIDATION OF WINTER AND SUMMER CLASSIFIERS

To validate the generated winter and summer range maps of species 423 we compare the predicted range maps of 423 (by our RF model) with the range maps obtained from abundance maps of 423 (from the Atlas website). We use the metrics classification accuracy of the model; precision, recall, and F1 score of both present and absent classes to validate the modelling outcomes.

Classification accuracy of a classifier C for a species s and time period t is the ratio of number of blocks correctly classified by the classifier C and the total number of blocks classified by the classifier C. Precision of class c of species s and time period t is the ratio of number of blocks in the range map of s during t that are correctly classified into class c by the classifier C and the total number of blocks classified into class c by the classifier C. Recall of class c is the ratio of number of blocks classified correctly into class c and the total number blocks in class c. F1 score is calculated from precision and recall. F1 score of class c of species s and time period t can be calculated as expressed in the equation 6.1.

$$F1(s,t,c) = 2 \times \frac{precision(s,t,c) \times recall(s,t,c)}{precision(s,t,c) + recall(s,t,c)} \tag{6.1}$$

The higher the metrics, the better the performance of the classifier. Using only classification accuracy to assess the performance of a classifier might lead to incorrect conclusions. For example, a species s during time period t is present in 10 out of 100 blocks and if the classifier labels all the 100 blocks as species-is-absent, the classification accuracy would still be 90% which is an exaggerated score. So we calculate precision, recall, and F1 score for each class. We may observe from table 6.1 that the classification accuracy of summer classifier of species 423 is higher than its winter classifier. But the precision, recall, and F1 scores of both classes of winter classifier are higher than the summer classifier. From this, we conclude that the performance of the winter classifier of species 423 is better than that of the summer classifier of species 423.

By comparing the range maps in figure 6.1 and analysing the accuracy metrics in the table 6.1, we conclude that the predicted winter range map is reliable with decent classification accuracy and F1 scores. But the predicted summer range map is too generous and unreliable. Though the classification accuracy is 0.9, the F1 score of the class species-is-present is low at 0.49. The precision of class species-is-present is low at 0.37, that means, only 37% of the blocks classified as species-is-present are correct.

6.2.1 Validation of range maps of other species

Here we show the validation range maps obtained from abundance maps of Sovon Atlas, predicted range maps by our RF classifiers, and classification reports of the eight (species, season) pairs selected to determine the constants $cDETREL$ and threshold for $ESP(b, t, s)$. The seasonal

Table 6.1 Classification report for winter and summer classifiers of species 423. 0 represents absence class and 1 represents presence class

Season	Precision		Recall		F1 score		Classification accuracy
	0	1	0	1	0	1	
Winter	0.84	0.73	0.87	0.68	0.85	0.7	0.8
Summer	0.98	0.37	0.92	0.72	0.95	0.49	0.9

abundance maps of species 165, 224, 402, 639 and species 692, 706, 715, 753 are shown in figures A.8 and A.9 respectively. The validation maps from the Sovon Atlas and predicted seasonal range maps of species 165, 224; species 402, 639; species 692, 706; and species 715, 753 are shown in figures 6.2, 6.3, 6.4, and 6.5 respectively. Classification reports of the selected eight species are shown in table 6.2.

Table 6.2 Classification reports of selected species, season pairs used to determine the constants $cDETREL$ and threshold for $ESP(b, t, s)$

Species	Season	Precision		Recall		F1 score		Classification accuracy
		0	1	0	1	0	1	
165	Winter	0.85	0.69	0.82	0.73	0.83	0.71	0.79
224	Winter	0.68	0.53	0.79	0.39	0.73	0.45	0.64
402	Winter	0.85	0.48	0.80	0.58	0.82	0.52	0.74
639	Winter	0.89	0.37	0.70	0.68	0.78	0.48	0.70
692	Winter	0.77	0.56	0.90	0.32	0.83	0.41	0.74
706	Winter	0.91	0.84	0.93	0.79	0.92	0.81	0.89
715	Summer	0.77	0.75	0.79	0.73	0.78	0.74	0.76
753	Summer	0.75	0.83	0.90	0.61	0.82	0.70	0.78

From the seasonal range maps obtained from the Sovon Atlas displayed above, it can be understood that winter occurrence of species 224, 402, 639, and 692 is mostly along the coastal waters. Similarly, the winter occurrence of species 165, 706 and summer occurrence of species 715 and 753 is mostly inland. By comparing the classification accuracies and F1 scores of species that occur along coasts and species that occur inland from the table 6.2, it can be understood that the model performs better for inland species compared to species that occurs along coastal waters. This could be mainly because the validation maps from the Sovon Atlas, by default, labels blocks that are mud flats or completely covered with water as absent where as our model classifies mudflats for all species and water blocks for floating species during winter as species-is-present or species-is-absent. This could be one reason for classification accuracy and F1 scores.

6.3 COMPARISON OF RESULTS OBTAINED USING WEIGHTED AND NON-WEIGHTED METRICS

In the method used to obtain the results shown in chapter 5, we have selected eight species that have common and high $DETREL(t, s)$ to decide on optimal constant values of $cDETREL$ and threshold for $ESP(b, t, s)$. Then the threshold for $ESP(b, t, s)$ is used to obtain class memberships **C0(t,s)** and **C1(t,s)**. But to compare the results from weighted and non-weighted methods we do not account for the variability in detectability and report likelihood, that means, the constants

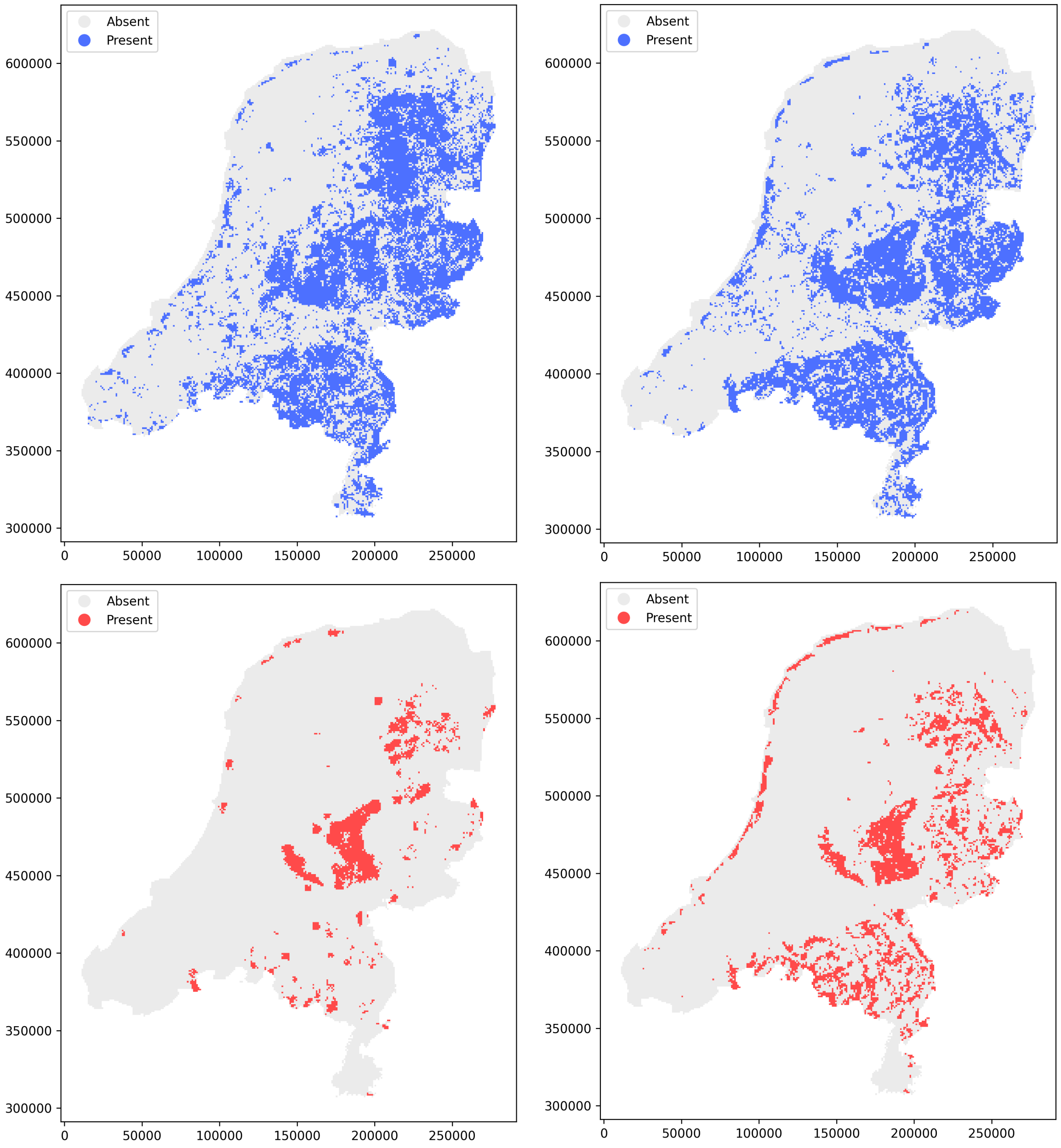


Figure 6.1: Seasonal range maps of species 423 obtained for validation from the Sovon Atlas and predicted seasonal range maps of species 423 from our RF classifiers. Top left: winter range of species 423 obtained for validation from the Sovon Atlas; top right: predicted winter range map of species 423 by our RF classifier; bottom left: summer range of species 423 obtained for validation from the Sovon Atlas; bottom right: predicted summer range map of species 423 by our RF classifier. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841)

$cDETREL(t, s)$ and threshold for $ESP(b, t, s)$ will not be used. Instead of using the threshold for $ESP(b, t, s)$ we use the value on 50th percentile of $OHR(b, t, s)$ to obtain $C0(t, s)$ and $C1(t, s)$ for non-weighted model and use the value on 50th percentile of $wOHR(b, t, s)$ to obtain $C0(t, s)$ and $C1(t, s)$ for weighted model. Using weighted and non-weighted metrics has given almost the

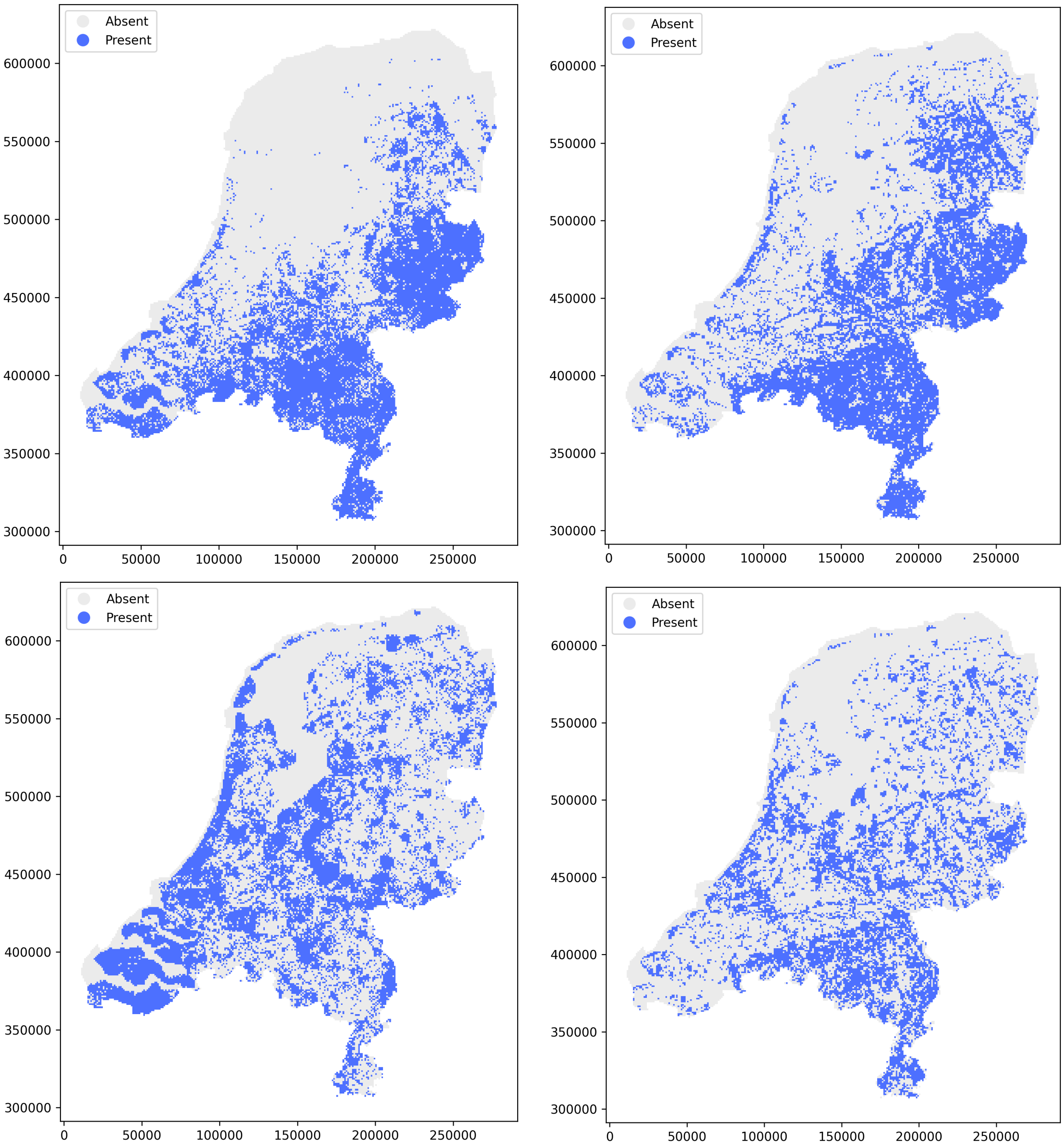


Figure 6.2: Seasonal validation and predicted range maps of species 165 and 224. Top left: validation map for species 165, winter; top right: predicted range map of species 165, winter; bottom left: validation map for species 402, winter; bottom right: predicted range map for species 224, winter. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841)

same results obtained using the threshold for $ESP(b, t, s)$. But for some (species, season) combinations the value on 50th percentile of $OHR(b, t, s)$ will be zero. This labels all the blocks in non-weighted model as species-is-present and there will be no species-is-present blocks that can be used for training. So, it is recommended to use weighted metrics to avoid such complication. The classification accuracies and F1 scores of the eight (species, season) combinations by using weighted and non-weighted model are shown in the table 6.3.

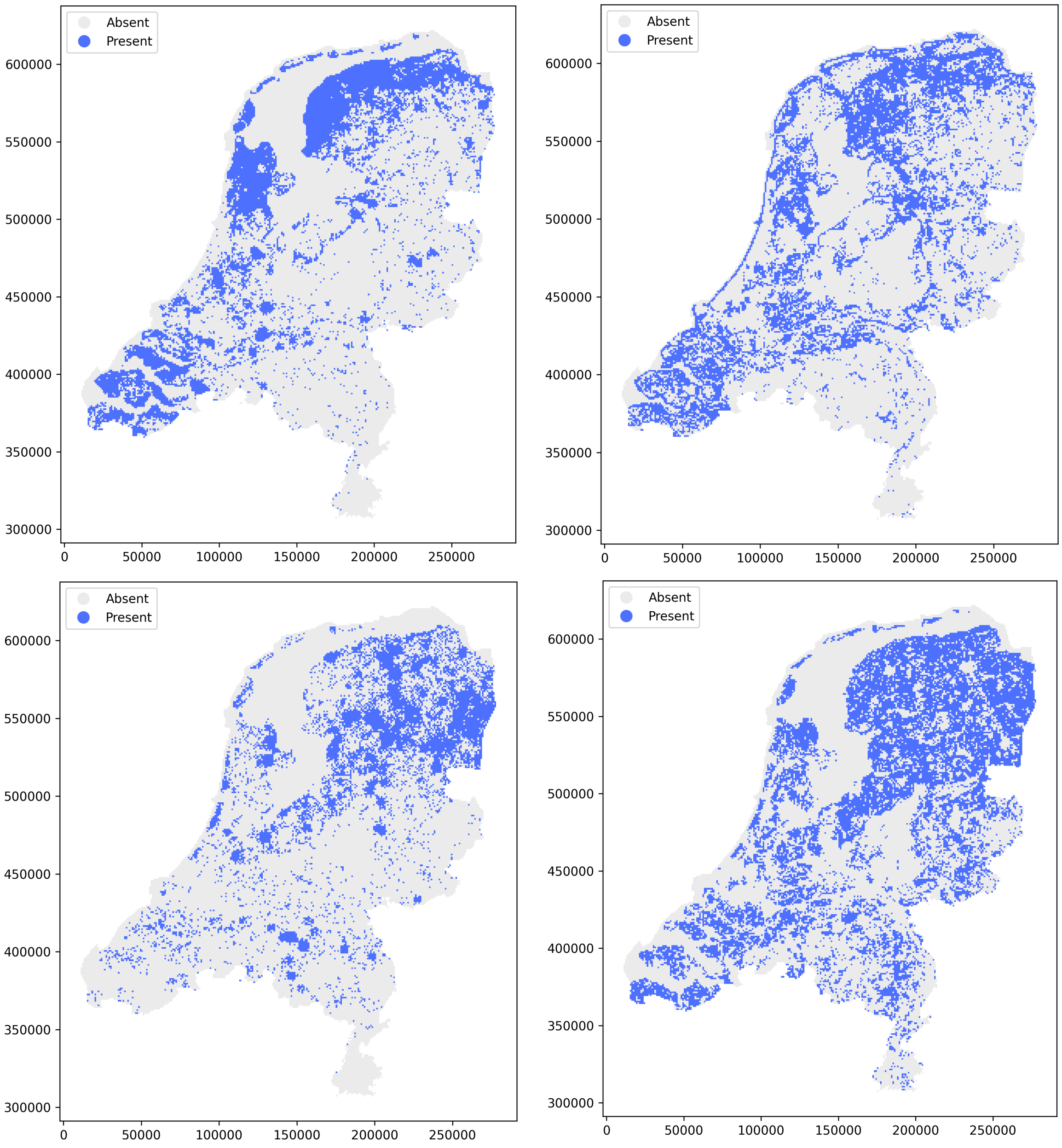


Figure 6.3: Seasonal validation and predicted range maps of species 402 and 639. Top left: validation map for species 402, winter; top right: predicted winter range map of species 402; bottom left: validation map for species 639, winter; bottom right: predicted range map for species 639, winter. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841)

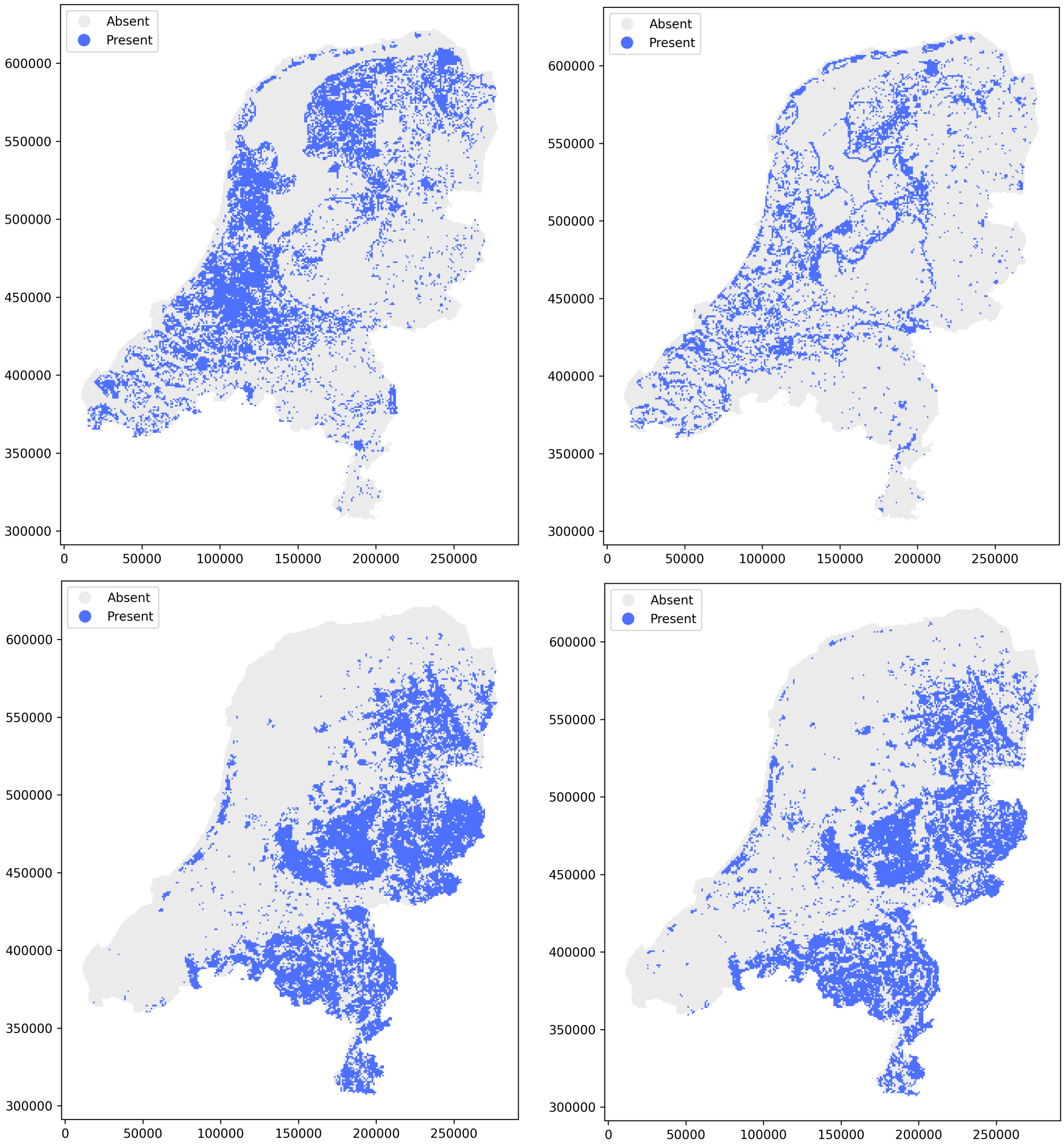


Figure 6.4: Seasonal validation and predicted range maps of species 692 and 706. Top left: validation map for species 692, winter; top right: predicted winter range map of species 692; bottom left: validation map for species 706, winter; bottom right: predicted range map for species 706, winter. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841)

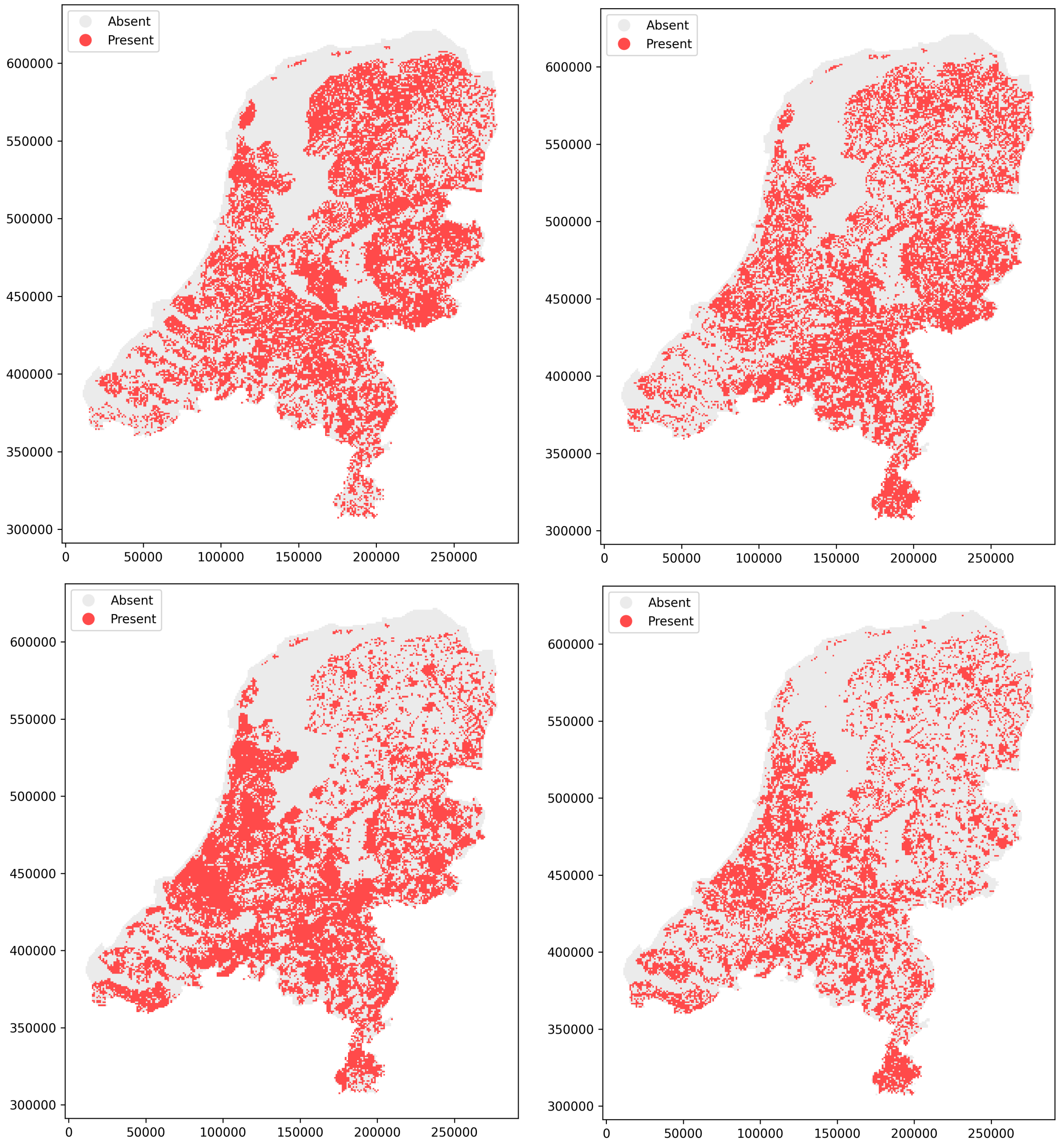


Figure 6.5: Seasonal validation and predicted range maps of species 715 and 753. Top left: validation map for species 715, summer; top right: prediction summer range map of species 715; bottom left: validation map for species 753, summer; bottom right: predicted range map for species 715, winter. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841)

Table 6.3 Classification reports of weighted and non-weighted models for the eight (species,season) pairs

	Weighted model			Non-weighted model		
Species/season	F1 score		Classification accuracy	F1 score		Classification accuracy
	0	1		0	1	
165/winter	0.83	0.71	0.79	0.83	0.71	0.78
224/winter	0.72	0.49	0.64	0.73	0.47	0.64
402/winter	0.82	0.53	0.74	0.82	0.55	0.75
639/winter	0.79	0.49	0.70	0.78	0.49	0.69
692/winter	0.84	0.44	0.74	0.83	0.46	0.75
706/winter	0.92	0.81	0.89	0.92	0.81	0.88
715/summer	0.78	0.74	0.76	0.77	0.70	0.74
753/summer	0.83	0.73	0.79	0.84	0.73	0.80

Chapter 7

Discussion of outcomes and Conclusion

7.1 INTRODUCTION

The challenge in this study was to account for different types of variability in the dataset Ω and automatically select training data for the given (species, season) pair. In data science and machine learning the modelling outcomes depends heavily on the quality of the input data. High voluminous data is not a solution if the data quality is poor. Though we have about 30 million observations collected over a period of ten years, we cannot directly generate reliable range maps from the data without accounting for the variability in the data. As described in the chapter 3 there is high variability in observer effort thorough space and time, observers differ greatly in terms of observer skills, and species are dynamic in terms of detectability and report likelihood. Accounting for each of the variability to get unbiased results is a challenging task.

An observation of species s in block b within time period t takes place only when an observer o is present in b during t detects and reports the species s. As the species s and observer o are living beings they will have certain habits and preferences that changes with space and time. These habits and preferences should be understood and quantified as much as possible in order to account for the variability and generate reliable results. Some of the observer habits are visiting the blocks, spending time in the blocks visited, detecting species, and reporting the detected species. Observer preferences include choosing which block to visit, when to visit, amount of time to be spent in the visited block, and whether to report the detected species or not. All the mentioned habits and preferences of the observer can be quantified from the dataset Ω except for the last preference i.e. to know the species that were detected but not reported by the observers. On the other hand, inferring species habits and preferences like roosting, flocking, singing, breeding etc., from Ω is limited. But we can infer habitat preferences of the species from the observations in Ω and LULC data.

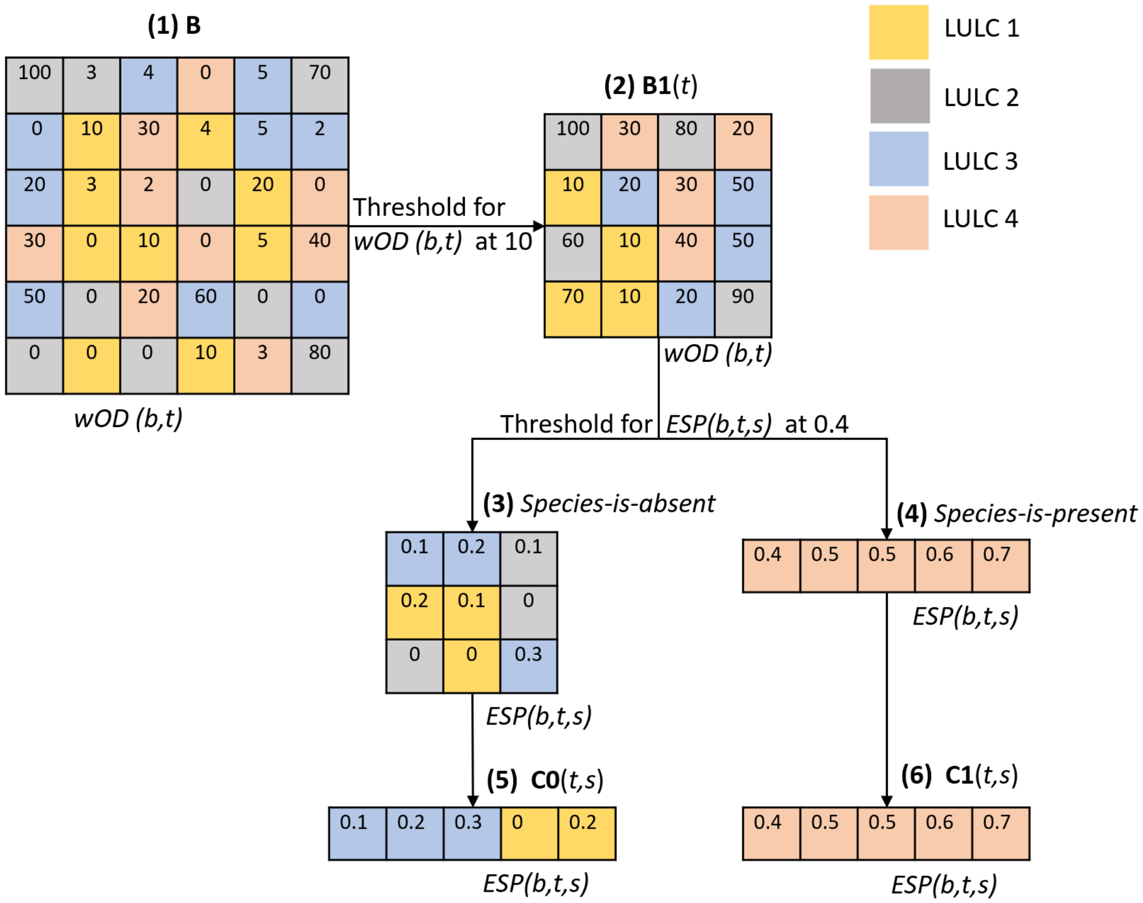
7.2 ACCOUNTING FOR VARIABILITY IN INPUT DATA SOURCES

To account for variability in observer effort through space, we have quantified observer effort as weighted observer days, $wOD(b, t)$ for each block in $\mathbf{B}$. Then we have determined a threshold for $wOD(b, t)$ to label each block in $\mathbf{B}$ as sufficiently or insufficiently visited. It means that all the blocks labelled as sufficiently visited have accumulated enough weighted observer days required to infer species presence. The threshold for $wOD(b, t)$ is similar for the species but changes with season. Selecting blocks with a decent weighted observer days value does not bring evenness in the observer effort but ensures that the selected blocks are sufficiently visited and the inferences made from them can be reliable. To account for variability in observer effort through time we have adjusted the calendar dates and season as mentioned in the section 3.2.1. Such adjustments brings same number of weekdays and weekends in each season removing the ‘weekend effect’. There is also variability in observer effort within the seasons but this is not accounted for. So, the generated annual range map can be biased. To account for variability in observer skills, we have calculated observer weights based on number of distinct block day visits made and sum of distinct species reported during each block day visit. To account for detectability and report likelihood of the species, we have estimated combined probability of species presence as $DETREL(t, s)$ and used it to separate the probability of species presence from the three-step probabilistic process.

7.2.1 Accounting for variability in observer effort through space

After accounting for variability in observer effort, the selected training classes **C0(t,s)** and **C1(t,s)** should be sufficiently visited and decently represent the study area. All the blocks in **B** are labelled as sufficiently and insufficiently visited based on the determined threshold for weighted observer days $wOD(b, t)$. This threshold is same for all the species but varies with season. In fact, the threshold should vary with species and season because the observer effort required to detect a species varies with species and time. Because of the data limitations, we could not define a decent estimate for detectability of the species and determine a species and time specific threshold for weighted observer days $wOD(b, t)$. In crowd-sourced data observers prioritize areas that are closer to where they live and areas that are easily accessible. For this reason, predicted occurrence patterns of species correlate more with the sampling patterns than the true occurrence patterns of the species (Kelling et al., 2018). In order to reduce such effect, the selected training classes **C0(t,s)** and **C1(t,s)** should decently represent the study area. It is difficult to check the representativiness of **C0(t,s)** and **C1(t,s)** because it keeps changing with species and time. So, we ensure the representativiness in $B1(t)$ from where **C0(t,s)** and **C1(t,s)** are selected. The representativiness of training classes is negatively effected when the classes species-is-present and species-is-absent are balanced to **C0(t,s)** and **C1(t,s)**. Detailed explanation of this limitation is explained as follows.

Assume that in figure 7.1, **1** is the set of blocks **B**, blocks with at least one observation; **2** is the set of sufficiently visited blocks **B1(t)** during time period *t*; **3** and **4** are the set of blocks labelled as species-is-absent and -present respectively using the threhold for $ESP(b, t, s)$ 0.07, **5** and **6** are the balanced classes obtained by random resampling as mentioned in the section 4.6. Values in **1** and **2** represents weighted observer days **wOD(b,t)** and values in **3**, **4**, **5**, **6** represents values of estimated species presence $ESP(b, t, s)$. Each colour represents a LULC class. There are 36 blocks equally distributed among the 4 LULC classes in **B** i.e, each LULC class covers 9 blocks. With 10 as the threshold for $wOD(b, t)$, we select **2** **B1(t)** from **1** **B**. Observer that **B1(t)** has all the 4 LULC classes and covers 4 blocks each. This shows that **B1(t)** completely represents **B**. Next, we calculate **ESP(b,t,s)** for each block in **B1(t)**. With 0.4 as the threshold for **ESP(b,t,s)** we obtain **3** species-is-absent and **4** species-is-present. Observe that **3** and **4** together represents **B** but the classes are imbalanced. As the number of blocks in class species-is-present is 5, we randomly select 5 blocks from the class species-is-absent. This random sampling results in the balanced classes **C0(t,s)** and **C1(t,s)**. Now the important issue is that the number of blocks have reduced from **3** to **C0(t,s)** and **C0(t,s)** do not represent all the LULC classes represented in **3**. So, it can be concluded that there is data loss and reduction in representativeness from **3** and **4** to **5** and **6**. A recommendation for these limitations is to generate synthetic data to balance the classes species-is-absent and species-is-present.



(1) B

100	3	4	0	5	70
0	10	30	4	5	2
20	3	2	0	20	0
30	0	10	0	5	40
50	0	20	60	0	0
0	0	0	10	3	80

(2) B1(t)

100	30	80	20
10	20	30	50
60	10	40	50
70	10	20	90

(3) Species-is-absent

0.1	0.2	0.1
0.2	0.1	0
0	0	0.3

(4) Species-is-present

0.4	0.5	0.5	0.6	0.7

(5) C0(t,s)

0.1	0.2	0.3	0	0.2

(6) C1(t,s)

0.4	0.5	0.5	0.6	0.7

Figure 7.1: Workflow explaining the data loss and reduction in representativeness from **3** and **4** to **5** and **6**. Assume that **1** is the set of blocks **B**, **2** is the set of sufficiently visited blocks **B1(t)**, **3** and **4** are the set of blocks labelled as species-is-absent and -present respectively, **5** and **6** are balanced training classes **C0(t,s)** and **C1(t,s)**. Spatial relation between the blocks from **2** to **6** are false.

7.2.2 Accounting for variability in observer effort through time

Courter et al., 2012 have found that observer activity in crowd-sourcing is high during weekends than weekdays (weekend effect). In the initial data exploration, from the figure 3.6, it was observed that the number of observations vary per week and season due to varying observer effort through time. The number of observations also increased from the year 2010 to 2019 due to increase in observers and observer activity. Following the suggestion of Courter et al., 2012, we have used day of the week and week number in defining the chosen temporal resolution. As the observer activity within a season does not vary much within a season, we chose to work with temporal resolution of one season that has 13 weeks. If we go by calendar dates there will be imbalance in the number of week days and weekends per season. So, as explained in section 3.2.1, the calendar dates of 10 years are aggregated to one year and adjusted in such a way that each season has 13 weeks with equal number of week days and weekends to account for the ‘weekend effect’. Such adjustments has a negative effect i.e, it can bring some observations of spring into winter, summer into spring, and autumn into summer. This increases the false positives in the training data that can impact the modelling outcomes. However, all the species do not follow the meteorological seasons to breed or migrate. So, this limitation might not be applicable to all species. For species that have decent number of observations through out the year, plotting its observations per week gives the temporal occurrence of the species as shown in the figure A.1. Aggregating ten years of data into one year assumes that the habitat preference have remained the same for all the ten years but that is not quite true. This assumption might negatively impact the range maps of some species whose habitat preferences have changed.

7.2.3 Accounting for variability in observer skills

Different observers have different habits and preferences. Some observers make frequent visits to the blocks and meticulously reports all the distinct species they detect while some report only the species that are of their interest. To account for the variability in observer skills, observer weights per observer and season were calculated. The observer weight of an observer during a season is calculated as the product of distinct block visits made by the observer during that season and number of distinct species reported per block visit by the observer during the same season as explained in the section 4.9. Two models, one using the weighted metrics and other using the non-weighted metrics are run to generate range maps for the selected eight (species, season) pairs. By comparing the classification accuracy and F1 scores of weighted and non-weighted models documented in table 6.3, we understood that calculating observer weights did not help in improving the classification accuracy and F1 scores. This could be because we might have missed some important factor that helps in understanding observer skills. One important factor that can be considered in estimating observer weights is the number of distinct blocks visited by the observer in the selected season. Observers who visit more blocks can be expected to be more experienced and skilled.

7.2.4 Accounting for variability in detectability and report likelihood of species

An observation reported can be considered as a result of a three-step probabilistic process. One, the species must be present. Two, the observer must detect the species present. Three, the observer must report the species detected. So, there is a need to estimate the probability of species presence, probability of finding the species present, and probability of reporting the species found. We have estimated detectability and report likelihood as one quantity. The factors that effect detectability of a species are abundance, observer effort invested, habitat preference, size etc. and the main factor effecting the report likelihood of the species is observer interest. Considering all the factors effecting in estimating detectability and report likelihood of species from Ω was challenging. Given the data limitations, we have assumed that detectability and report likelihood varies with species, time and remains constant through space. Out of all the factors effecting detectability and report likelihood, we have considered only the factor (observer effort invested) to estimate the combined probability of detecting and reporting the species as $DETREL(t, s)$. We do not expect $DETREL(t, s)$ to be accurate because many factors that effect detectability are not considered and the assumption that detectability is constant through space is not completely true.

7.3 SUMMARY OF CONCLUSION

First, We have accounted for varying observer effort through space but with two limitations. One, there is loss of data and representativeness in training classes as explained in section 7.2.1. Two, the threshold for weighted observer days $wOD(b,t)$ should be species and time specific but it changes with time and remains constant with species. Second, we have aggregated the observations of ten years to one year. This assumes that habitat preferences of species have remained constant through the ten year time period. This is not true for all the species. We adjusted the calendar days in such way that each season has same number of week days and week ends. This removes the 'weekend effect'. We have also observed variability in observer effort between the seasons but we did not account for it. As an effect the annual range map derived from the seasonal range maps can be biased. Third, to account for variability in observer skills we have calculated observer weights that vary with season. However, these weights did not help in improving the model performance. This could be because the observers in Ω do not differ much in their skills or we have missed some important factors in quantification of observer weights. Fourth, to account for varying detectability and report likelihood of the species, we have define the combined estimate of detectability and report likelihood as $DETREL(t,s)$. Given the data limitations, we have assumed that detcetability and report likelihood remains constant through space and varies with species and time. This is not quite true, they also vary through space. Out of all the factors effecting detectability and report likelihood, we consider only one factor in quantification of $DETREL(t,s)$ i.e, the observer effort invested to find and report the species that is known to be present. However, using $DETREL(t,s)$ did not help in improving the performance of the model because $DETREL(t,s)$ is not accurate enough. Fifth, to estimate the probability of species presence, we consider each observation reported as a result of three-step probabilistic process where the species must be present, the observer must detect the species, and report it. We have estimated detectability and report likelihood as $DETREL(t,s)$ and separated species presence, $ESP(b,t,s)$, from the three-step probabilistic process. Then we have selected eight (species, season) combinations to determine a threshold for $ESP(b,t,s)$ that can be used to label the selected sufficiently visited blocks as species-is-present and species-is-absent. Then we balance these classes to **C0(t,s)** and **C1(t,s)** and train an RF classifier to obtain seasonal range map. This process is repeated for three other seasons to obtain three other range maps. Finally, we combine the four seasonal range maps based on the conditions mentioned in 4.1. Overall, the performance of the classifiers for species that occur inland was better than for species that occur along the coastal waters. One reason could be that the validation maps masks the mudflats and water blocks but our model treats them differently.

7.4 OPTIONS FOR FUTURE WORK

There are five conceptual aspects that we believe can be tried in the future to improve the method. First, to quantify observer effort of a block b during time period t as amount of time spent in hours by distinct observers who have visited b during t. If the missing clock time stamps for the observations can be estimated, the amount of time spent by an observer can be estimated from his/her first and last observation in a day. If the observer has made only one observation in a day, then it is a complication. Second, to determine species and time specific threshold for weighted observer days to label the blocks in **B** as sufficiently or insufficiently visited. This can be done only after having a decent estimate for detectability and report likelihood. Third, to account for variability in observer skills, we have calculated observer weights based on only two factors. Using the third factor (number of distinct blocks visited by the observer in the given season) might help in calculating improved observer weights. Any other factors that explain the observer quality can also be included. Fourth, to further explore the possibilities of estimating detectability and report likelihood. We have explored multiple ways to estimate detectability and report likelihood from Ω and understood that the data in Ω is limited to decently estimate detectability and report likelihood. So, we suggest to try using some ancillary data like habitat suitability maps. It can be assumed that blocks with high suitability can have high abundance and high abundance leads to high detectability given the blocks were sufficiently visited. This provides opportunities to estimate detectability and report likelihood per block, species, and time period. Fifth, the seasons defined for this study are almost similar to the meteorological seasons. We assume that species winter in winter season, breeds in the summer season, and migrates in spring and autumn seasons. But in reality, different species migrate at different points of time in spring and autumn. Some species starts breeding earlier than others. Clearly, to obtain accurate range maps, there should be species specific winter, breeding, and migratory seasons.

Methodically there are three aspects that are worth studying to improve the performance of the model. One, the class memberships selected to train the model should be representative of the study area but they have some limitations that are explained in the section 7.2.1. These limitations can be addressed by generating synthetic data. There are many techniques to generate synthetic data and attain class balance in the training classes but the condition is that the selected technique should improve the representativeness of the training data. Two, the dimension of soil data used as explanatory variables is huge. If the dimensionality is reduced we expect the classifier to make better decisions. Three, the train and test accuracy of classifiers were different for different (species, season) combinations when default parameters of RF classifier were used. If the training and test accuracies of a classifier are low, it means the model performance will also be low. So a technique that can determine species and time specific classification parameters for the RF is expected to make huge difference in classification accuracy and F1 scores.

Appendix A

Other figures

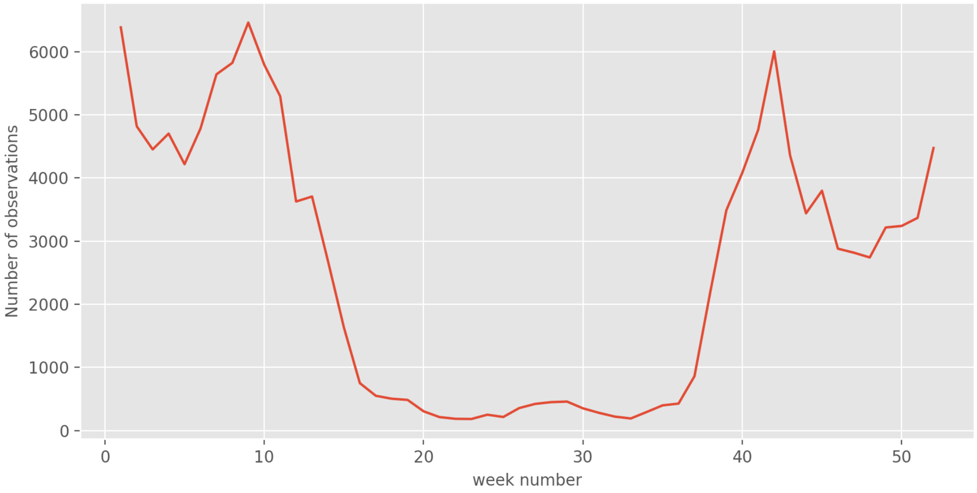


Figure A.1: Annual distribution of observations of species 423. Value on X-axis represents the week number and value on Y-axis represents the count of observations of species 423.

Table A.1 F1 scores and classification accuracies of the predicted range maps of eight (species, season) pairs when compared with validation maps obtained using different cut-offs for abundance

Species/season	Cut-off for abundance class	F1 score		Classification accuracy
		0	1	
165/winter	0	0.63	0.66	0.61
	1	0.78	0.74	0.76
	2	0.82	0.74	0.79
	3	0.83	0.71	0.79
	4	0.82	0.60	0.75
224/winter	0	0.35	0.50	0.43
	1	0.58	0.50	0.54
	2	0.65	0.49	0.58
	3	0.73	0.45	0.64
	4	0.78	0.38	0.67
402/winter	0	0.36	0.45	0.41
	1	0.76	0.53	0.68
	2	0.79	0.55	0.72
	3	0.84	0.55	0.76
	4	0.85	0.51	0.77
639/winter	0	0.51	0.61	0.57
	1	0.76	0.57	0.69
	2	0.78	0.52	0.70
	3	0.79	0.47	0.70
	4	0.86	0.37	0.77
692/winter	0	0.44	0.34	0.40
	1	0.75	0.41	0.65
	2	0.80	0.43	0.70
	3	0.83	0.43	0.74
	4	0.86	0.37	0.77
706/winter	0	0.49	0.55	0.52
	1	0.82	0.72	0.78
	2	0.87	0.77	0.84
	3	0.92	0.81	0.89
	4	0.92	0.77	0.88
715/summer	0	0.41	0.65	0.56
	1	0.62	0.71	0.67
	2	0.72	0.73	0.73
	3	0.78	0.71	0.75
	4	0.79	0.66	0.74
753/summer	0	0.36	0.53	0.46
	1	0.56	0.59	0.58
	2	0.69	0.64	0.66
	3	0.82	0.70	0.78
	4	0.86	0.70	0.81

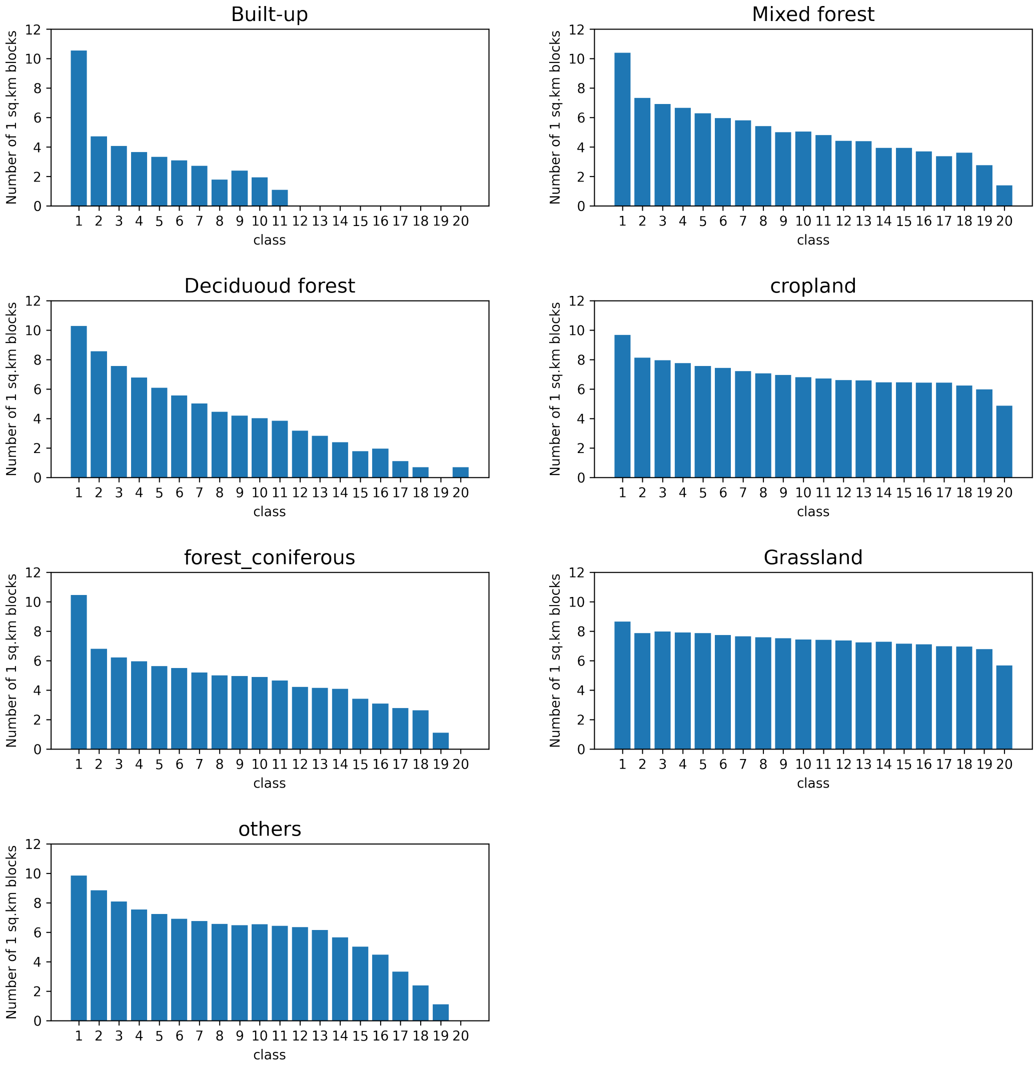


Figure A.2: Distribution of 7 LULC classes in NL. Each class on X-axis represents area in hectares within a particular range covered by the respective LULC class. class 1 represents values within range 0 to 5, class 2 represents values within range 5 to 10, and so on, class 20 represents the values above 95. The range of each class is increased by a multiple 5 from class 1 to class 20. The value on Y-axis represents the number of blocks transformed using natural log

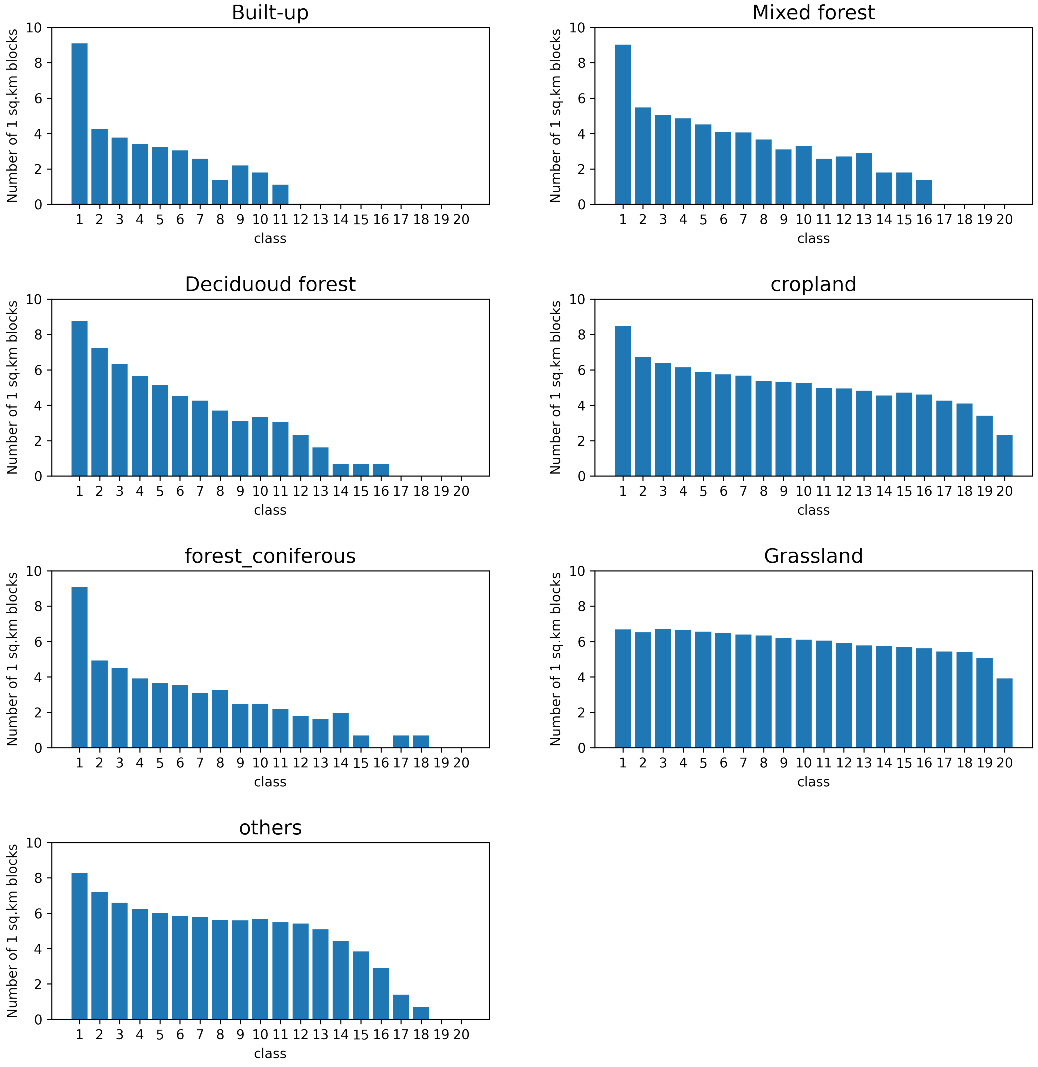


Figure A.3: Distribution of 7 LULC classes in sufficiently visited blocks in winter. Each class on X-axis represents area in hectares within a particular range covered by the respective LULC class. class 1 represents values within range 0 to 5, class 2 represents values within range 5 to 10, and so on, class 20 represents the values above 95. The range of each class is increased by a multiple 5 from class 1 to class 20. The value on Y-axis represents the number of blocks transformed using natural log

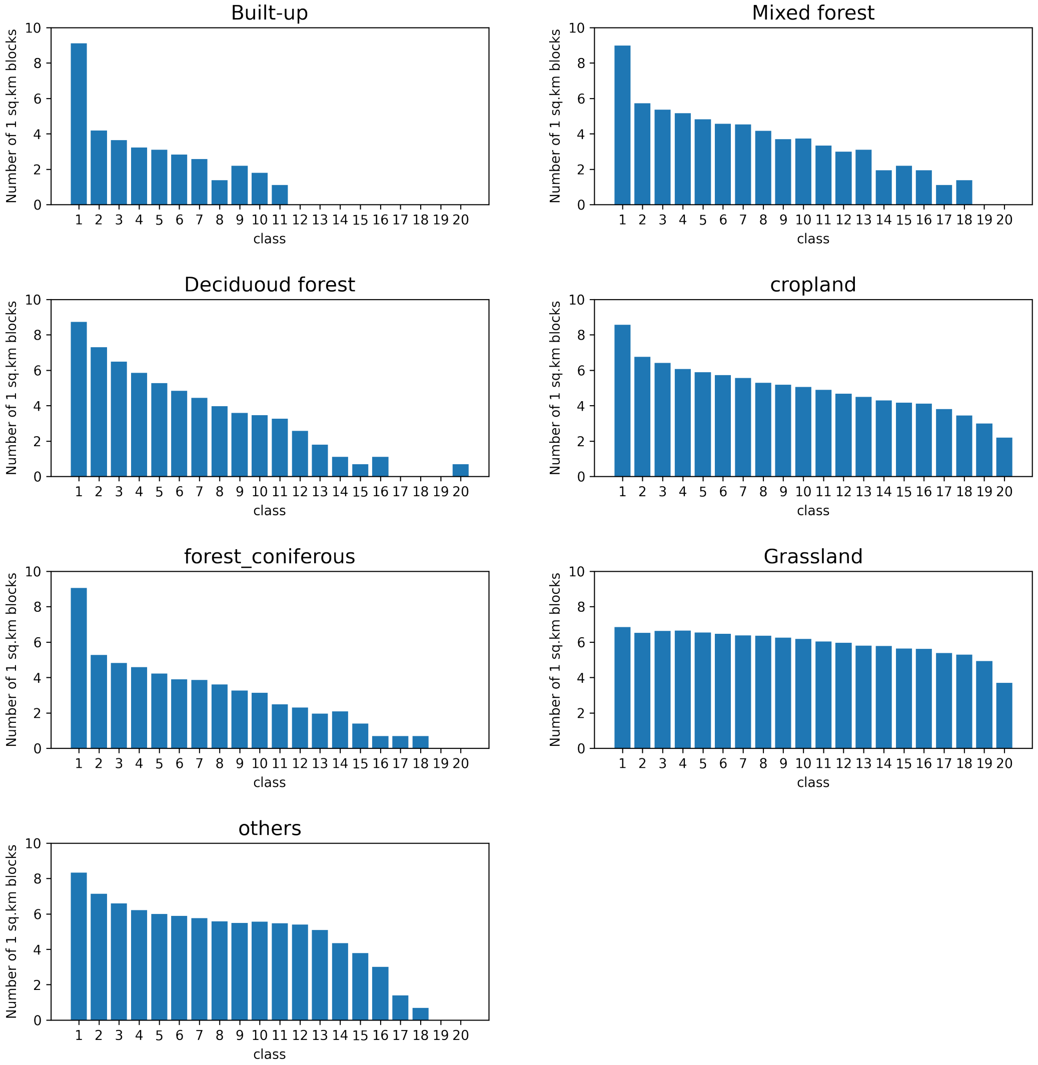


Figure A.4: Distribution of 7 LULC classes in sufficiently visited blocks in spring. Each class on X-axis represents area in hectares within a particular range covered by the respective LULC class. class 1 represents values within range 0 to 5, class 2 represents values within range 5 to 10, and so on, class 20 represents the values above 95. The range of each class is increased by a multiple 5 from class 1 to class 20. The value on Y-axis represents the number of blocks transformed using natural log

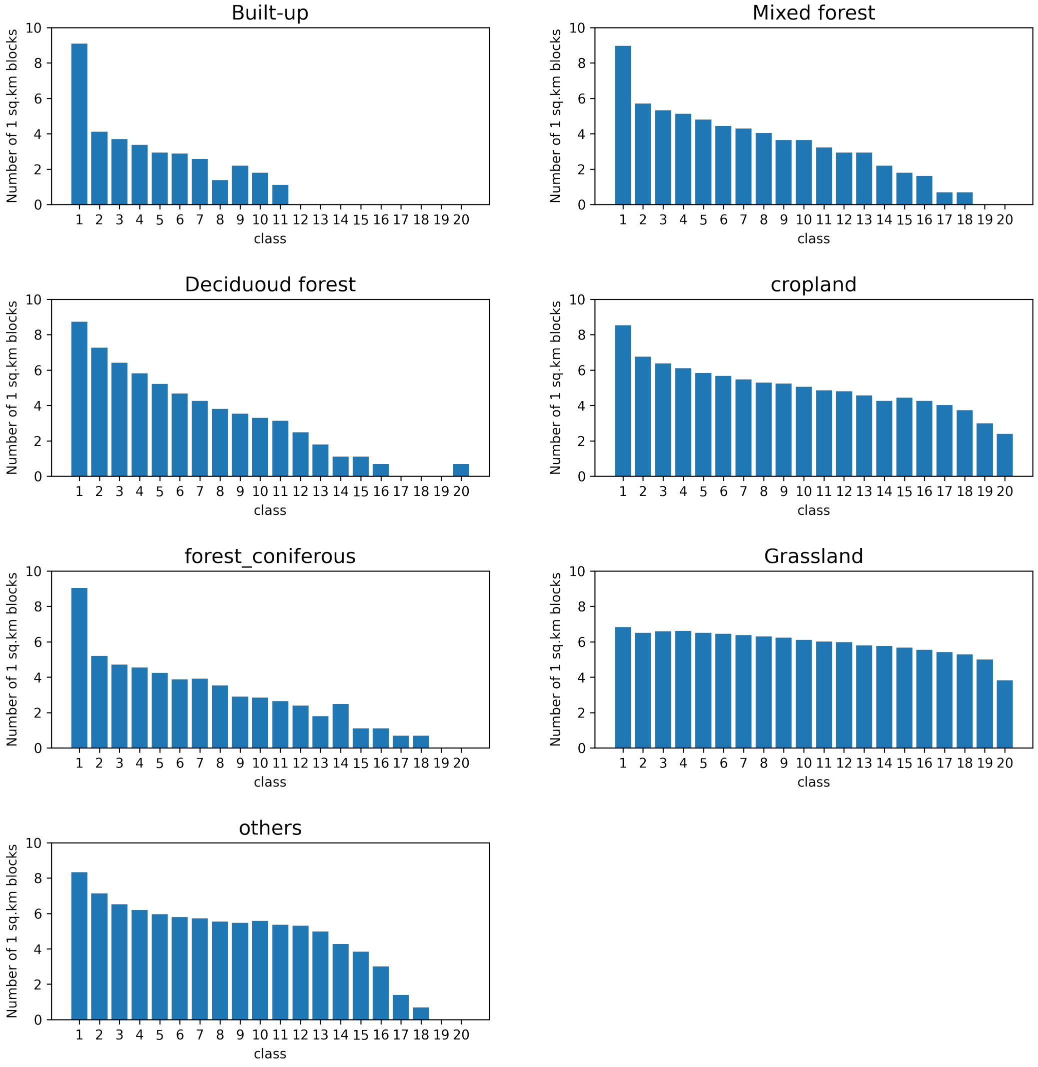


Figure A.5: Distribution of 7 LULC classes in sufficiently visited blocks in summer. Each class on X-axis represents area in hectares within a particular range covered by the respective LULC class. class 1 represents values within range 0 to 5, class 2 represents values within range 5 to 10, and so on, class 20 represents the values above 95. The range of each class is increased by a multiple 5 from class 1 to class 20. The value on Y-axis represents the number of blocks transformed using natural log

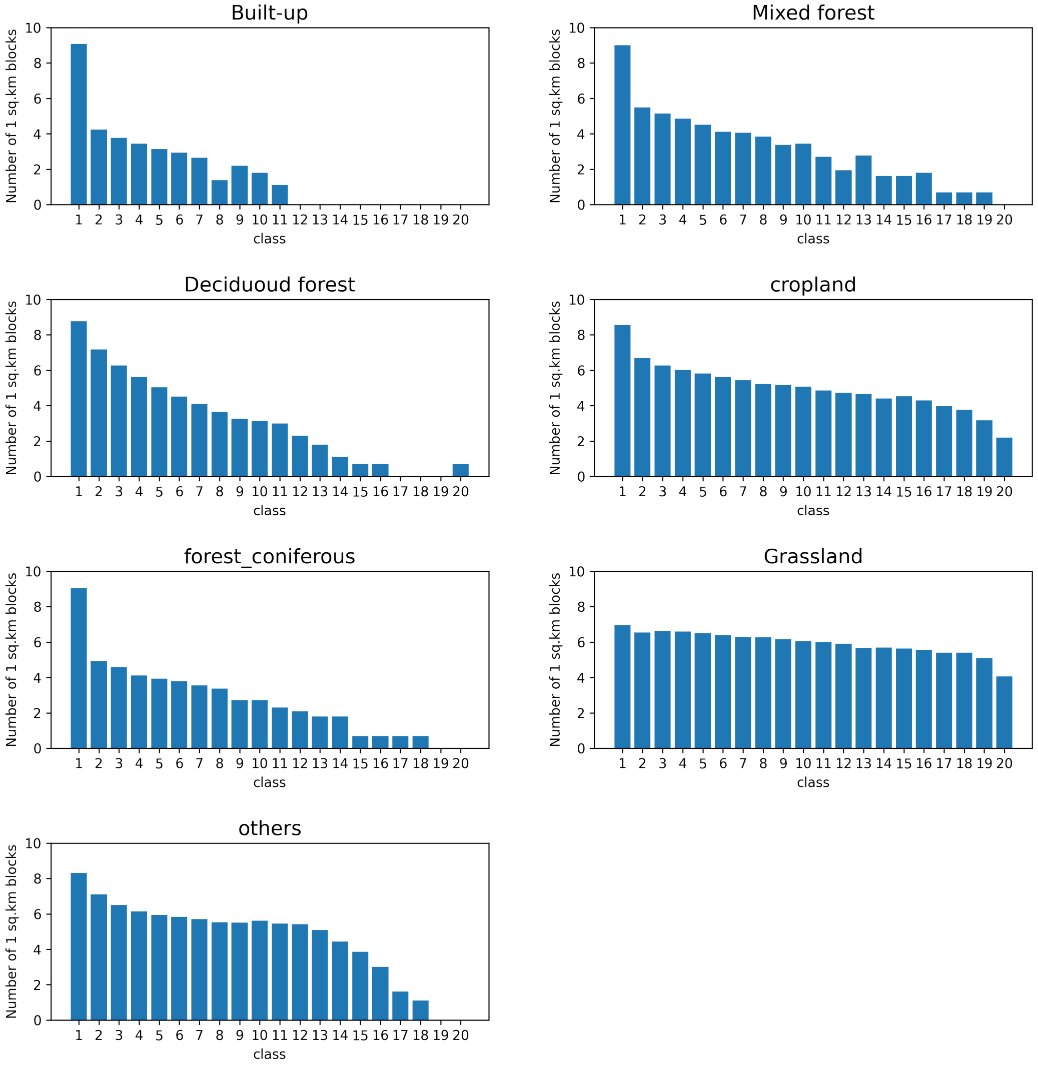


Figure A.6: Distribution of 7 LULC classes in sufficiently visited blocks in autumn. Each class on X-axis represents area in hectares within a particular range covered by the respective LULC class. class 1 represents values within range 0 to 5, class 2 represents values within range 5 to 10, and so on, class 20 represents the values above 95. The range of each class is increased by a multiple 5 from class 1 to class 20. The value on Y-axis represents the number of blocks transformed using natural log

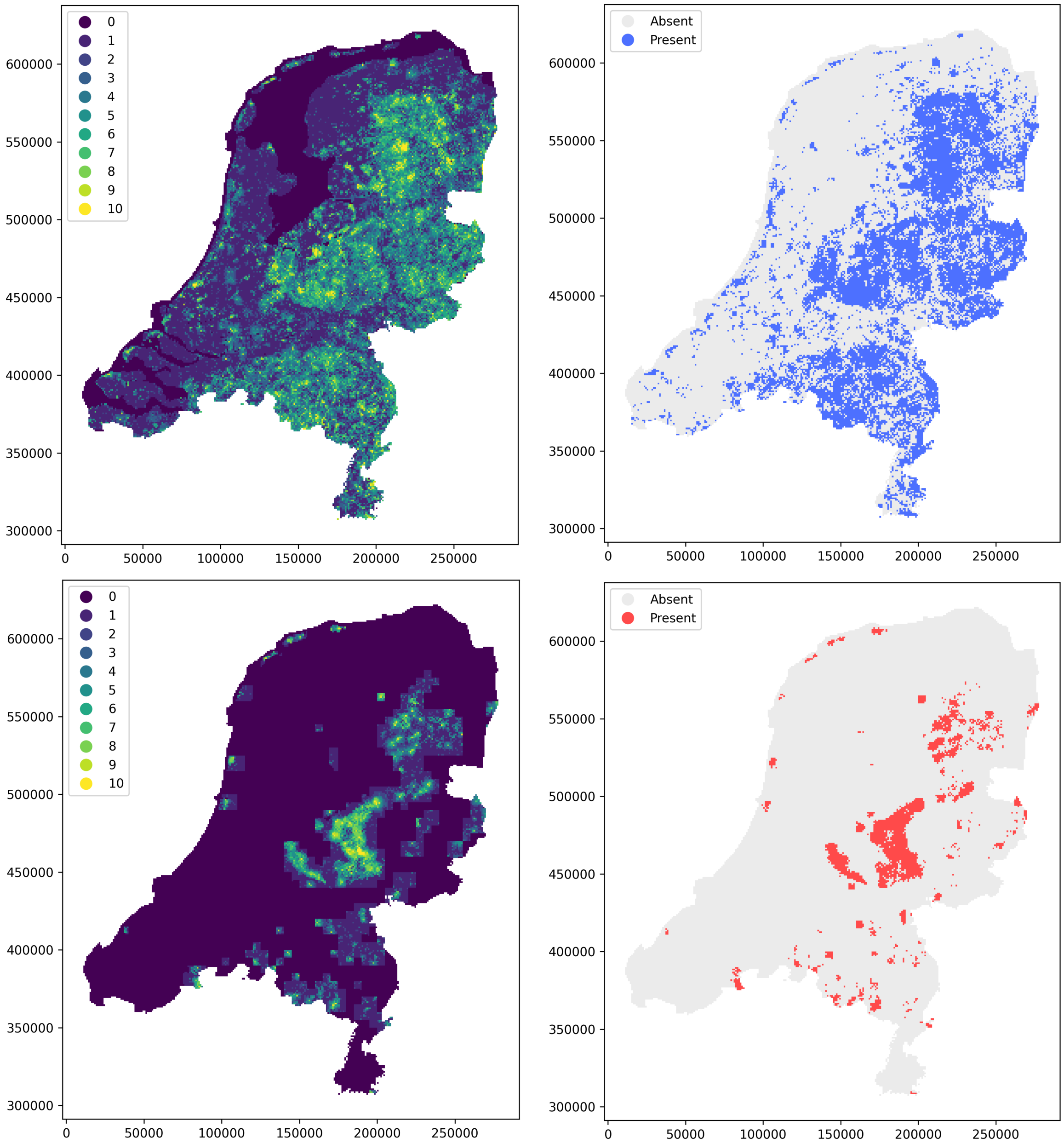


Figure A.7: Seasonal abundance and validation range maps of species 423 obtained from the Sovon Atlas. Blocks with abundance class above 3 are labelled under the class present and others under the class absent. Top left: winter abundance map of species 423; top right: winter validation map of species 423; bottom left: summer abundance map of species 423; bottom right: summer validation map of species 423. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841)

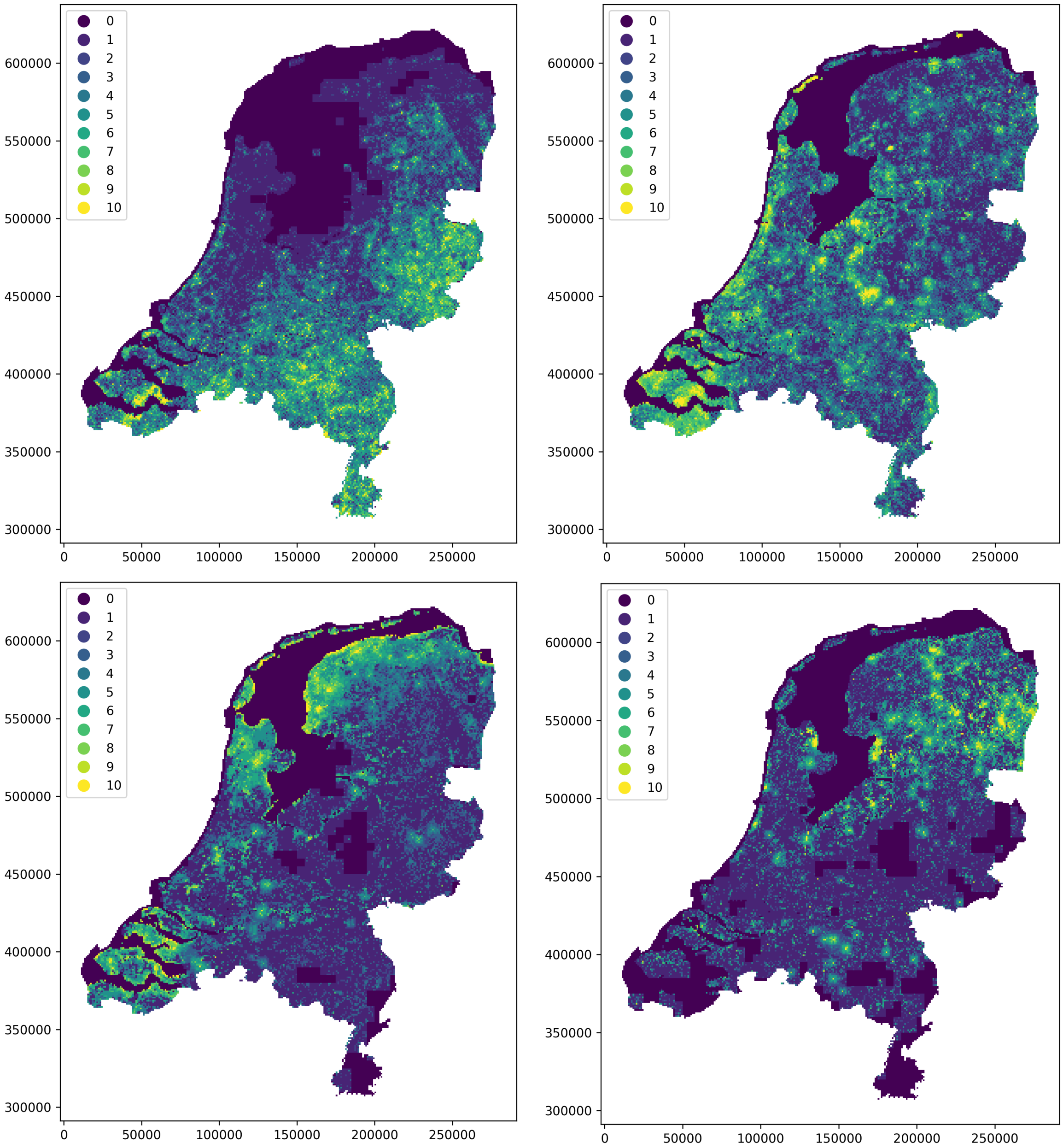


Figure A.8: Obtained seasonal abundance maps of selected species, season pairs. Top left: 165, winter; top right: 224, winter; bottom left: 402, winter; bottom right: 639, winter. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841)

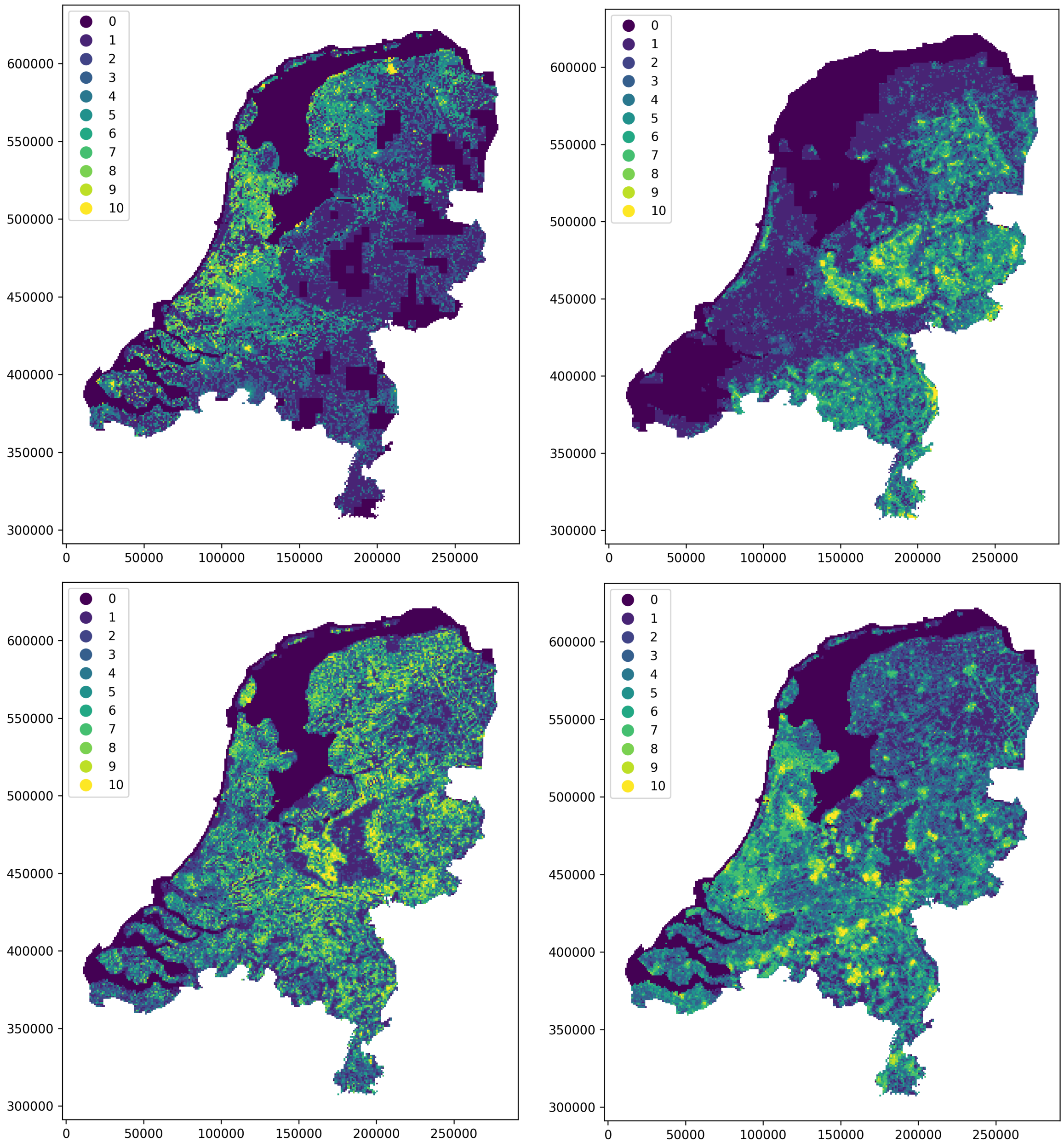


Figure A.9: Obtained seasonal abundance maps of selected species, season pairs. Top left: 692, winter; top right: 706, winter; bottom left: 715, summer; bottom right: 753, summer. X, Y values are coordinates in the Dutch grid 28992 (unit: meters, Geodetic CRS: Amersfoort, Datum: Amersfoort, Ellipsoid: Bessel 1841)

List of References

Bradter, U., Mair, L., Jönsson, M. T., Knape, J., Singer, A., & Snäll, T. (2018). Can opportunistically collected citizen science data fill a data gap for habitat suitability models of less common species. *Methods in Ecology and Evolution*, *9*, 1667–1678.

Buytaert, W., Zulkafli, Z., Grainger, S., Acosta, L., Alemie, T. C., Bastiaensen, J., Bièvre, B., Bhusal, J. K., Clark, J., Dewulf, A., Foggin, M., Hannah, D., Hergarten, C., Isaeva, A., Karpouzoglou, T., Pandeya, B., Paudel, D., Sharma, K., Steenhuis, T., … Zhumanova, M. (2014). Citizen science in hydrology and water resources: Opportunities for knowledge generation, ecosystem service management, and sustainable development. *Frontiers in Earth Science*, *2*, 26.

Callaghan, C. T., Nakagawa, S., & Cornwell, W. (2021). Global abundance estimates for 9,700 bird species. *Proceedings of the National Academy of Sciences*, *118*.

Conrad, C., & Hilchey, K. G. (2011). A review of citizen science and community-based environmental monitoring: Issues and opportunities. *Environmental Monitoring and Assessment*, *176*, 273–291.

Courter, J. R., Johnson, R. J., Stuyck, C. M., Lang, B. A., & Kaiser, E. (2012). Weekend bias in citizen science data reporting: Implications for phenology studies. *International Journal of Biometeorology*, *57*, 715–720.

Cutler, D. R., Edwards, T., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. (2007). Random forests for classification in ecology. *Ecology*, *88*, 2783–92.

Fink, D., Auer, T., Johnston, A., Ruiz-Gutierrez, V., Hochachka, W., & Kelling, S. (2019). Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications*, *30*.

Fink, D., Hochachka, W., Zuckerberg, B., Winkler, D., Shaby, B., Munson, M. A., Hooker, G., Riedewald, M., Sheldon, D., & Kelling, S. (2010). Spatiotemporal exploratory models for broad-scale survey data. *Ecological applications : a publication of the Ecological Society of America*, *20*, 2131–47.

Fox, A. D. (2004). Has Danish agriculture maintained farmland bird populations. *Journal of Applied Ecology*, *41*, 427–439.

Geldmann, J., Heilmann-Clausen, J., Holm, T., Levinsky, I., Markussen, B., Olsen, K., Rahbek, C., & Tøttrup, A. P. (2016). What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, *22*, 1139–1149.

Ghiraldi, L., & Aimassi, G. (2019). Extinct and endangered ('E&E') birds in the ornithological collection of the Museum of Zoology of Torino University, Italy. *Bulletin of the British Ornithologists' Club*, *139*, 28–45.

Gregory, R., van Strien, A. V., Voříšek, P., Meyling, A. W. G., Noble, D., Foppen, R., & Gibbons, D. (2005). Developing indicators for European birds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*, 269–288.

Isaac, N., & Pocock, M. (2015). Bias and information in biological records. *Biological Journal of The Linnean Society*, *115*, 522–531.

Kelling, S., Fink, D., Sorte, F. A. L., Johnston, A., Bruns, N. E., & Hochachka, W. (2015). Taking a 'big data' approach to data quality in a citizen science project. *Ambio*, *44*, 601–611.

Kelling, S., Gerbracht, J., Fink, D., Lagoze, C., Wong, W.-K., Yu, J., Damoulas, T., & Gomes, C. (2012). eBird: A human/computer learning network for biodiversity conservation and research. *IAAI*.

Kelling, S., Johnston, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Bonn, A., Fernandez, M., Hochachka, W., Julliard, R., Kraemer, R., & Guralnick, R. (2018). Finding the signal in the noise of citizen science observations. *bioRxiv*.

Klonner, C., Marx, S., Usón, T., Albuquerque, J. D., & Höfle, B. (2016). Volunteered geographic information in natural hazard analysis: A systematic literature review of current approaches with a focus on preparedness and mitigation. *ISPRS Int. J. Geo Inf.*, *5*, 103.

Munson, M. A., Webb, K., Sheldon, D., Fink, D., Hochachka, W. M., Iliff, M., Riedewald, M., Sorokina, D., Sullivan, B., Wood, C., & Kelling, S. (2010). The eBird Reference Dataset, Version 1 . 0. *Network*, 1–11. http://www.avianknowledge.net/content/features/archive/eBird_Ref

Phillips, S. J., Dudík, M., Elith, J., Graham, C., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological applications : a publication of the Ecological Society of America*, *19*, 181–97.

Robinson, O. J., Ruiz-Gutierrez, V., & Fink, D. (2018). Correcting for bias in distribution modelling for rare species using citizen science data. *Diversity and Distributions*, *24*, 460–472.

Sauer, J., & Butcher, G. (2014). The role of citizen science in bird conservation: The christmas bird count and breeding bird survey.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H., Milcinski, G., Niksic, M., Painho, M., Podör, A., Raimond, A., & Rutzinger, M. (2016). Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic information. *ISPRS Int. J. Geo Inf.*, *5*, 55.

Snäll, T., Kindvall, O., Nilsson, J., & Pärt, T. (2011). Evaluating citizen-based presence data for bird monitoring. *Biological Conservation*, *144*, 804–810.

Sovon Vogelonderzoek Nederland. (2019). *Vogelatlas van Nederland. Broedvogels, wintervogels en 40 jaar verandering*. Kosmos Uitgevers, Utrecht/Antwerpen.

Welvaert, M., & Caley, P. (2016). Citizen surveillance for environmental monitoring: Combining the efforts of citizen science and crowdsourcing in a quantitative data framework. *Springer-Plus*, *5*.