INTEGRATING OPPORTUNISTIC CITIZEN SCIENCE DATA AND SPECIES DISTRIBUTION MODELS FOR BIODIVERSITY CONSERVATION POLICY AND MANAGEMENT

Camille Van Eupen

Supervisors:
Prof. dr. Stijn Luca
Prof. dr. ir. Ben Somers
Prof. dr. Dirk Maes

Members of the Examination Committee:
Prof. dr. ir. Eddie Schrevens (chairman)
Prof. dr. ir. Pieter De Frenne
Dr. Marc Herremans
Prof. dr. Thomas Neyens
Prof. Tord Snäll
Prof. dr. ir. Karen Vancampenhout

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Bioscience Engineering (PhD)

June 2023

Doctoraatsproefschrift nr. 1859 aan de faculteit Bio-ingenieurswetenschappen van de KU Leuven

Uitgegeven in eigen beheer, Camille Van Eupen, Tremelo

ACKNOWLEDGEMENTS

Finishing this PhD would not have been possible without many friends, colleagues, mentors and most of all family. The last five years have given me so much more than a piece of paper and I could not have managed it without you. Most importantly, Seb, thank you for the immense support, pep talks, soiréekes and, above all, three beautiful daughters.

My deepest gratitude also to the tens of thousands of volunteers, the species experts and the technical and scientific support team that dedicate their time and effort to generate and ameliorate the citizen science database *waarnemingen.be*. This dissertation could not have existed without you. I would like to extend my sincere thanks to the Flemish Research Foundation for funding this research.

I am extremely grateful to have had three kind supervisors. Stijn, you have been a true support during the entire process. Thanks for always being there for questions and brainstorms, for explaining me complicated (and not so complicated) statistics whenever I needed it and for introducing me at the division in Leuven. Ben, thank you for the opportunity to work at FNL, for your critical questions during our meetings, your empathy and your generous feedback. Dirk, I am extremely grateful for your co-supervision. You were my go-to for anything species-related and got me started with distribution modelling and R. Together with our many "lunch meetings", those were definitely among the things I truly enjoyed during this PhD.

A special thank you to my assessors and jury members. Marc, thank you for your ecological insight, thorough feedback and for sharing your experience with the *waarnemingen.be* data. Karen, thanks for sharing your (below) bottom-up view on things. Your passion and commitment have been inspiring. Tord, Pieter and Thomas, I very much appreciated your critical review of the dissertation and your valuable feedback. I would like to extend my gratitude to Eddy, for successfully managing the whole process of the defence.

Last but not least, a big thank you to all my (ex) colleagues at FNL for the lab activities, coffee breaks and milestones that could take my head off work when I needed it. Also thanks to my thesis students, Ruben, Sarah and Joachim. Your work has contributed significantly to the quality and my own experience of this PhD. I truly enjoyed our collaboration. The same goes for the people at INBO and Natuurpunt Studie, who supported me with valuable expertise and technical support.

SUMMARY

Biodiversity is crucial to the well-being of our planet and its inhabitants and has become a priority in international treaties and policies. Protecting and restoring this natural capital requires collaboration between citizens, scientists and landowners, as well as evidence-based research.

The combination of citizen science data (CSD) and species distribution models (SDMs) can be a powerful resource for enhancing biodiversity conservation policy and management. Citizen science, which involves engaging citizens in scientific research, is an effective way to monitor species in an era of biodiversity informatics and big data. SDMs, on the other hand, are able to link species occurrence data and environmental variables to predict the distribution of species in space and time. Combining these two approaches creates opportunities to study species and areas that are not regularly surveyed.

Opportunistic CSD contain information on species presence and are collected by volunteers without them following specific guidelines. Despite their abundance, their quality is uncertain, which can lead to bias and error in SDM predictions. To improve data quality, data cleansing can be used as a first step to remove erroneous records from a dataset, for example, based on record attributes that provide information on the observation process or post-entry data validation. This is called data quality filtering (or also stringent filtering) and while it reduces uncertainty, it also reduces sample size, a trade-off that had remained relatively unexplored. However, this is an important consideration, as smaller sample sizes often have a negative effect on the performance of SDMs.

This dissertation addresses that knowledge gap, by exploring the combined impact of data quality and sample size on model performance (Chapter II). We applied data quality filters based on observer experience, record detail and record verification to opportunistic species records gathered from the citizen science platform *waarnemingen.be*. SDM performance was assessed before and after filtering while controlling for sample size. Results provided insight into the quantity-quality trade-off in data quality filtering but also revealed that species responded differently to filtering. A second study (Chapter III) consequently linked several species traits to the results of the first study to finetune filtering recommendations (BOX 2: Think before you shrink). The goal of the study, taxonomy and multiple species traits (especially proneness to misidentification, home range and familiarity) should be considered before choosing an appropriate filter. Caution was needed when filtering reduced sample size beyond a certain threshold (e.g. by more than half of the original sample size).

The research then focuses on a specific conservation case where opportunistic CSD and environmental data obtained through remote sensing were combined to support multi-scale habitat management in heathlands (Chapter IV). The study found that local vegetation structure, habitat heterogeneity and the landscape context impacted the habitat suitability of dry-heathland birds, butterflies and grasshoppers and crickets, with differences in small versus large heathlands. In large patches, vegetation structure and heathland heterogeneity generally benefitted habitat suitability while in small and fragmented patches, edge effects and species characteristics interacted more with the results.

The fifth chapter (Chapter V) provides a summary of the conducted research and discusses some important considerations when using the suggested methods for conservation applications. Finally, the dissertation elaborates on the application potential of this research in biodiversity conservation policy with a focus on Flanders (Chapter VI). Overall, these findings contribute to the existing literature on the use of opportunistic citizen science data for ecological research and species conservation. It does so by providing evidence-based recommendations for increasing data quality and illustrating the application potential in various conservation applications.

SAMENVATTING

Biodiversiteit is cruciaal voor het voortbestaan van de planeet en haar bewoners en is een prioriteit geworden in internationale verdragen en beleid. De bescherming en het herstel van biodiversiteit vereist samenwerking tussen burgers, wetenschappers en landeigenaren, evenals *evidence-based* onderzoek.

De combinatie van burgerwetenschap en soortverspreidingsmodellen (SDMs) kan een krachtig middel zijn om biodiversiteitsbeleid en -beheer te versterken. Burgerwetenschap, waarbij burgers worden betrokken bij wetenschappelijk onderzoek, is een effectieve manier om soorten te monitoren in tijden van biodiversiteitsinformatica en *big data*. SDMs daarentegen zijn in staat om verspreidingsgegevens van soorten te linken aan omgevingsvariabelen om zo de verspreiding van soorten in ruimte en tijd te voorspellen. Het combineren van deze twee methoden opent mogelijkheden voor het bestuderen van gebieden en soorten die niet systematisch onderzocht werden.

Opportunistische waarnemingen bevatten informatie over de aanwezigheid van soorten en werden verzameld door vrijwilligers zonder dat deze daarbij specifieke richtlijnen volgden. Ondanks hun grote aantallen is hun kwaliteit onzeker, wat kan leiden tot slechte voorspellingen uit SDMs. Om de kwaliteit van verspreidingsgegevens te verbeteren, worden onzekere waarnemingen vaak verwijderd. Datakwaliteitsfilters verwijderen bijvoorbeeld gegevens op basis van informatie over het observatie- of validatieproces. Hoewel dit de kwaliteit van de gegevens kan verbeteren, vermindert het ook hun aantal, een wisselwerking die relatief weinig onderzocht bleef. Dit is nochtans een belangrijke overweging, aangezien kleinere steekproefgroottes vaak een negatief effect hebben op de prestatie van SDMs.

Dit proefschrift onderzoekt daarom de simultane impact van kwaliteit en kwantiteit van opportunistische waarnemingen op SDM-prestaties (Hoofdstuk II). Verschillende datakwaliteitsfilters werden toegepast op gegevens van het burgerwetenschapsplatform *waarnemingen.be*. Ze waren gebaseerd op de ervaring van waarnemers, op het detail van ingevoerde waarnemingen en op hun verificatiestatus. Vervolgens werd het verschil in prestatie gemeten tussen SDMs mét en zonder gefilterde gegevens, waarbij gecontroleerd werd voor de steekproefgrootte. De resultaten gaven ons inzicht in de wisselwerking tussen kwantiteit en kwaliteit bij het gebruik van datakwaliteitsfilters, maar toonden ook verschillen aan tussen de onderzochte soorten. Een tweede studie (Hoofdstuk III) koppelde daarom verschillende soorteigenschappen aan de resultaten van de eerste studie, waardoor aanbevelingen voor het

toepassen van datakwaliteitsfilters werden verfijnd (*BOX 2: Think before you shrink*). Bij het kiezen van een geschikte filter, is er allereerst aandacht nodig voor de taxonomie, verschillende soorteigenschappen (voornamelijk de kans op een foutieve identificatie, de grootte van het soortverspreidingsgebied en de bekendheid van de soort) en het doel van de studie. Daarnaast dient het filteren te gebeuren mits de nodige voorzichtigheid wanneer de steekproefgrootte wordt gereduceerd (bv. met meer dan de helft van het aantal aanwezigheden).

Het proefschrift richt zich vervolgens op een specifiek geval van natuurbehoud, met name het beheer van heide op meerdere schaalniveaus (Hoofdstuk IV). Hiertoe werden opportunistische waarnemingen gecombineerd met omgevingsvariabelen verkregen uit remote sensing. Uit het onderzoek bleek dat lokale vegetatiestructuur, heidetype-heterogeniteit en landschappelijke context de habitatgeschiktheid voor heidevogels, -vlinders en -sprinkhanen en krekels beïnvloedden, met verschillen in kleine versus grote heidegebieden. In grote gebieden hadden vegetatiestructuur en heidetype-heterogeniteit een voornamelijk positieve impact op de habitatgeschiktheid, terwijl in kleine en gefragmenteerde gebieden, randeffecten en soorteigenschappen belangrijker werden om die impact te verklaren.

Het vijfde hoofdstuk (Hoofdstuk V) geeft een samenvatting van het onderzoek en bespreekt enkele belangrijke overwegingen bij het gebruik van de voorgestelde onderzoeksmethoden. Ten slotte gaat de thesis dieper in op mogelijke toepassingen in het biodiversiteitsbeleid in Vlaanderen (Hoofdstuk VI). Over het algemeen dragen deze bevindingen bij aan de bestaande kennis over het gebruik van opportunistische waarnemingen uit burgerwetenschap voor ecologisch onderzoek en het behoud van soorten. Dit doet het door aanbevelingen te formuleren om de kwaliteit van opportunistische waarnemingen te verbeteren en door verschillende mogelijke toepassingen te suggereren voor het gebruik van deze gegevens voor biodiversiteitsbeleid.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter I

Chapter II

Chapter III

Chapter IV

Chapter VI

LIST OF TABLES

Chapter III

Chapter IV

Chapter V

LIST OF ABBREVIATIONS

A	ACTIVITY filter
AE	Agri-Environment
AIC	Akaike's Information Criterion
AUC	Area Under the receiver-operating Curve
BD	Birds Directive
BDS	Biodiversity Scores
BVM	Biological Valuation Map
CAP	Common Agricultural Policy
CBD	Convention on Biological Diversity
CSD	Citizen Science Data
D	DETAIL filter
DCLF	Diggle-Cressie-Loosmore-Ford
DE	Deviance Explained
EBV	Essential Biodiversity Variables
EFAs	Ecosystem Functioning Attributes
EN	Endangered
EU	European Union
EVI	Enhanced Vegetation Index
FPS	Flemish Priority Species
GAM	Generalized Additive Model
GAMM	Generalized Additive Mixed Model
GBIF	Global Biodiversity Information Facility
GEO BON	Group on Earth Observations Biodiversity Observation Network
GLCM	Gray-Level Covariance Matrix
GLM	Generalized Linear Model
HSS	Habitat Specific Species
IPP	Inhomogeneous Poisson point Process
IUCN	International Union for Conservation of Nature
LC	Least Concern
LiDAR	Light Detection And Ranging
LST	land surface temperature
MA	Management Agreement
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NT	Near Threatened
ODMAP	Overview, Data, Model, Assessment and Prediction
PA	Presence-Absence
PCA	Principal Component Analysis
PO	Presence-Only
PPM	Point Process Model
SAC	Special Area of Conservation
SDM	Species Distribution Model
SPA	Special Protection Area
TNR	True Negative Rate
TPR	True Positive Rate
V	VALSTAT filter
VIF	Variance Inflation Factor
VLM	Flemish Land Agency

CHAPTER I. Introduction

1. Biodiversity conservation

1.1. A global biodiversity crisis

Biodiversity is the diversity within species, between species and of ecosystems (United Nations, 1992) and is essential for human survival and well-being. It provides various ecosystem services such as the provision of biomass for food and energy (productive), ecosystem resilience and pollination (regulating) and aesthetic value (cultural) (Mace et al., 2012). Over the past centuries, the pressures on terrestrial biodiversity have been escalating by land use conversion and intensification, climate change, pollution and invasive alien species (Early et al., 2016; IPBES, 2019; Leclère et al., 2020; Newbold et al., 2015; Urban et al., 2016). Many populations and species are either extinct or on the verge of extinction (Burns et al., 2021; IPBES, 2019). This sixth mass species extinction would take millions of years to recover from (Ceballos et al., 2020), a worrying prospect for citizens, scientists and governments worldwide. They have acknowledged the crucial role of biodiversity in sustaining life on this planet and recognized the accelerating degradation of biodiversity as a global crisis, which initiated many efforts for biodiversity conservation.

1.2. Biodiversity conservation policy: from global to regional

The global and European protection of biodiversity is the result of a few key policies and initiatives, supported by a growing public awareness of the value of biodiversity. These efforts started around the middle of the 20th century with the foundation of organisations, such as the International Union for Conservation of Nature (IUCN) in 1946 and the World Wide Fund for Nature (WWF) in 1961, the implementation of global treaties, such as the Ramsar Convention in 1971 (Navid, 1984) to safeguard waterfowl habitat, and legislations, such as the Birds Directive (79/409/EEC) in 1979, which protects wild bird species within the European territory.

In 1992, the Convention on Biological Diversity (CBD) was drawn up at the Earth Summit in Rio de Janeiro and aimed to preserve biodiversity and promote the sustainable use of its components (later defined as ecosystem services) (United Nations, 1992). In the same year, the

EU (here we mean the European Union and all former structures) adopted the Habitats Directive (92/43/EEC) that, together with the Birds Directive, obliged EU member states to delineate Special Protection Areas (SPAs) and Special Areas of Conservation (SACs). These areas were aimed to protect the most valuable and threatened habitat types, fauna and flora of Europe (De Knijf and Paelinckx, 2013) and formed the cornerstone of European biodiversity conservation policy: the Natura 2000 network (Decleer, 2007; Sundseth, 2008).

At the tenth meeting of the governing body of the CBD, the Conference of the Parties, in Nagoya in 2010 (UNEP/CBD/COP/10/27), the Aichi Biodiversity Targets were adopted as a strategic plan to address the growing biodiversity crisis. In line with these targets, the EU designed the European Biodiversity Strategy for 2020 which intended to halt and reverse the loss of biodiversity by 2020 and to improve the state of Europe's natural capital by 2050 (European Commission, 2011). Unfortunately, by 2020, the Aichi Biodiversity Targets were only partially achieved (IPBES, 2019) and also the EU Biodiversity Strategy had insufficiently protected and restored nature (European Commission, 2020). The new EU Biodiversity Strategy for 2030, adopted in 2020, therefore emphasized the expansion of conservation areas (Natura 2000) and stricter implementation of conservation regulations (European Commission, 2020). By 2030, at least 30% of the European land area must be protected, including 10% strictly protected areas. Additionally, the European Commission proposed a Nature Restoration Law that intends a full restoration of all degraded ecosystems by 2050 (European Commission, 2022). Member states are free to choose how to implement and monitor these regulations in their respective territories, but also have several international obligations, such as investing in green and blue infrastructure across borders (European Commission, 2019; Schneiders et al., 2016) and reporting on the conservation status of protected habitat types and species to the EU (European Commission, 2020).

1.3. Trends in biodiversity monitoring

1.3.1. Traditional methods

To evaluate and adapt biodiversity conservation policy, biodiversity needs proper monitoring of its actual state and trends (i.e. distributions, abundances, extinctions and genetic diversity) (Pereira et al., 2012) and evidence-based research on the drivers of its loss (Sutherland et al., 2004).

Biodiversity monitoring traditionally involved scientists conducting fieldwork or setting up small-scale experiments (Sagarin and Pauchard, 2010). However, with the increasing adoption of biodiversity conservation policies, the need arose for more systematic survey methods to study trends in species populations and the impact of management decisions. To ensure repeatability in space and time and international data sharing, cooperatives were established and monitoring schemes were designed. For instance, the Group on Earth Observations' Biodiversity Observation Network (GEO BON) (Scholes et al., 2012) proposed Essential Biodiversity Variables (EBVs) (Pereira et al., 2013), which can be implemented globally to align methods for, for example, developing biodiversity indicators (Pettorelli et al., 2016; Vihervaara et al., 2015) and Red Lists (IUCN Standards and Petitions Committee, 2022; Maes et al., 2019b). More recently, technological advancements have opened up new possibilities for biodiversity monitoring, such as satellite-based remote sensing, drones, acoustic recording devices, and environmental DNA (Stephenson, 2020).

While this variety of methods has an indisputable value for biodiversity monitoring, it also have drawbacks. Field surveys are labour-intensive and provide very localised information on species. Systematic surveys might upscale monitoring, but they remain labour-intensive and contain large gaps in their spatial and temporal extent (Bradter et al., 2018). Trained volunteers are often engaged by professionals to reduce the workload, but the coordination and implementation of these projects requires regular funding and it remains difficult to obtain a large spatial and temporal coverage (Maes et al., 2015b). Furthermore, modern methods still have questionable efficiency and cost-effectiveness and need better standardisation for them to be widely used (Stephenson, 2020). In response to these limitations, scientists have been looking for ways to effectively monitor biodiversity across large extents while reducing both effort and costs.

1.3.2. State-of-the-art

Three major trends that are particularly interesting for biodiversity monitoring and policy support are species distribution models (Guisan and Zimmermann, 2000), citizen science data (Dickinson et al., 2010; Dobson et al., 2020; Theobald et al., 2015) and remote sensing (He et al., 2015; Leitão and Santos, 2019; Nagendra, 2001; Randin et al., 2020). Their performance and applicability have advanced rapidly in the 21st century while often being readily available in large amounts and over large spatial and temporal extents. Moreover, they are usually supported by open-source software and numerous studies that constantly attempt to increase the quality of their output.

a) Species distribution models

Scientists have been looking for ways to fill the spatial and temporal gaps in information on species occurrence for decennia. Species distribution models (SDMs) (Guisan and Zimmermann, 2000) have now become a widely accepted method to study the mechanisms underlying biodiversity change, reflected in the publication of numerous studies and the uptake of SDMs into biodiversity conservation applications (Guisan et al., 2013; Guisan and Thuiller, 2005; Maes et al., 2019d; Urban et al., 2016). They usually predict the suitability of a location based on ecological principles supported by expert knowledge (i.e. mechanistic SDMs) (Kearney and Porter, 2009) or based on statistical techniques that correlate species occurrence data with environmental variables (i.e. correlative SDMs) (Elith and Leathwick, 2009). The main difference between mechanistic and correlative SDMs is that the first do not require actual information on species occurrence and are by consequence a discrete indication of potential habitat based on decision rules, while the latter can estimate a (relative) probability of occurrence based on statistic principles. In this dissertation, correlative SDMs were used (section 3.1), but it should be noted that they can also be combined with mechanistic models when feasible (Kearney and Porter, 2009; Maes et al., 2019d, 2016).

b) The era of citizen science

Species occurrence data used in correlative SDMs are ideally obtained through systematic surveys, performed by trained observers and with a clear description of both data collection and project objectives (Kosmala et al., 2016). Such highly structured data, however, is rarely available for extensive geographical areas or periods, nor for a wide range of species (Isaac et al., 2014; Urban et al., 2016). In response, bulky occurrence data with lower information content that are often collected by volunteers participating in citizen science initiatives have been explored (Guisan et al., 2013; Schmeller et al., 2008; Theobald et al., 2015). A major advantage of these citizen science data (CSD) is they contain a multiplicity of species records at broad spatial and recent temporal scales and they have, consequently, become an important source of information for ecologists worldwide.

The increasing availability of species data collected in online data platforms has opened up a wide range of possibilities for supporting biodiversity conservation. Citizens have been participating in ecological studies for centuries (Silvertown, 2009), but biodiversity informatics, i.e. new information technologies such as the internet, smartphones and image recognition, have changed how these data are collected and used (Anderson et al., 2020; Peterson et al., 2015).

Both historical data, such as records from museums and herbaria, and new data are now centralised in large online data platforms such as iRecord in the United Kingdom (https://www.brc.ac.uk/irecord/), Artportalen in Sweden (https://www.artportalen.se/), *waarnemingen.be* in Flanders (northern Belgium; https://www.waarnemingen.be) (BOX 1) or eBird (https://ebird.org/), iNaturalist (https://www.inaturalist.org/) and GBIF (https://www.gbif.org) worldwide. Some of these platforms are targeted at data aggregation (e.g. GBIF; but see Anderson et al. (2020) for planned efforts towards more interactive data use), while others are more interactive: from record verification and feedback by experts to organised collaborations (e.g. eBird, *waarnemingen.be*, Artportalen) (Dobson et al., 2020).

c) Remote sensing

Remote sensing involves the use of sensors to collect data about the Earth's surface and atmosphere from a distance and has been widely adopted as a method for biodiversity monitoring in international conservation strategies and initiatives such as the Aichi Biodiversity Targets and EBVs (Arenas-Castro et al., 2018; Pereira et al., 2013; Pettorelli et al., 2016). It can be used to map and monitor ecosystems and wildlife populations (Reif and Theel, 2017; Wachendorf et al., 2018), assess habitat quality (Schmidt et al., 2018) and support conservation planning and management (Besnard et al., 2015; Vila-Viçosa et al., 2020).

Different sensors provide different possibilities, depending on the data they collect and the spatial and temporal resolution they can deliver (He et al., 2015; Leitão and Santos, 2019). Sensors that provide information on natural land cover with the highest application potential for biodiversity monitoring are optical sensors (multispectral or hyperspectral depending on the number of spectral bands they can measure) and Light Detection And Ranging (LiDAR) sensors (Wehr and Lohr, 1999). Multispectral sensors (e.g. MODIS, Landsat, Sentinel-2) are often used for land cover classification or quantitative assessments of vegetation such as time series analysis for phenology, vegetation indices for vegetation health (Leitão and Santos, 2019), or image texture analysis for vegetation structure (Farwell et al., 2021, 2020). Hyperspectral sensors (e.g. Hyperion, PRISMA) can deliver additional information on vegetation chemistry such as canopy or leaf water content or species diversity (He et al., 2015). LiDAR sensors are commonly used to assess the 3D structure of land cover, such as topography, vegetation height or stem density (Bergen et al., 2009).

BOX 1: WAARNEMINGEN.BE

Waarnemingen.be is the largest citizen science platform for species records in Belgium[1] managed by *Natuurpunt Studie*, *Natagora* and *Stichting Natuurinformatie*. The platform was designed to collect, share and store biodiversity data in Belgium (Herremans et al., 2018; Swinnen et al., 2018). Records can be entered on the website or via the mobile applications ObsIdentify, ObsMapp (Android) and iObs (iOS). Observers must provide default information on the observation (date and time, species, number of individuals, location and geographical precision) and are free to add more detailed information such as life stage, sex, behaviour, notes, photographs or sound fragments. Incoming records are verified in a semi-automated validation system (Swinnen et al., 2018) (Figure 1).

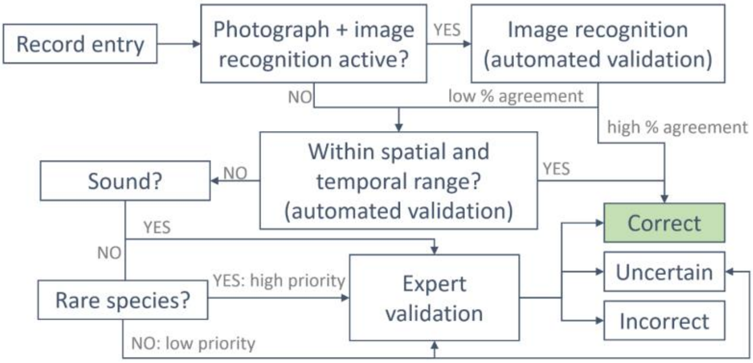


Figure 1: Semi-automated validation system in waarnemingen.be (adapted from Vanreusel et al., 2018).

The database currently holds one of the densest collections of species records in Europe: > 51 million species records of > 26,500 species by > 113,000 observers[2]. The majority of records are unstructured data (94%) but every year an increasing amount of semi-structured data[3] is submitted (Figure 2).

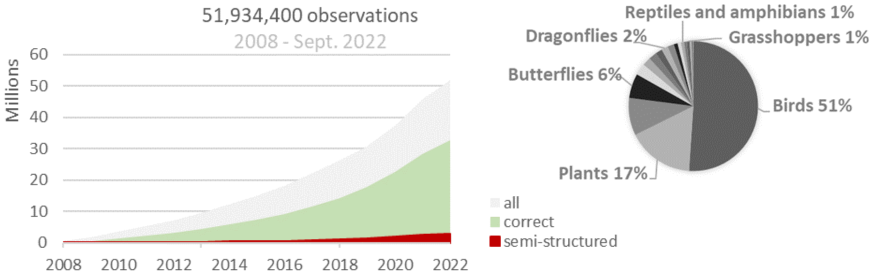


Figure 2: Cumulative number of observations and taxonomic representation in waarnemingen.be, with percentages of the taxonomic groups used in this study [2].

[1] https://www.waarnemingen.be in Flanders and https://www.observations.be in Wallonia

[2] Database query from situation on September 30th 2022

[3] i.e. checklists, point and transect counts, project observations such as moth traps and surveys from different research institutions and standardized monitoring schemes

Remotes sensing can provide direct measures of biodiversity, such as categorical land cover maps, species diversity and individual species. Land cover maps have been produced since the first remote sensing data became readily available in the 1970s. They are a valuable tool for assessing large-scale habitat loss and land use change, for example for restoration monitoring (Reif and Theel, 2017). Through spectral, temporal and textural features of multispectral images, broad classes of vegetation cover (e.g. CORINE land cover[4]) or vegetation communities (Thoonen et al., 2013) can be distinguished, but information on individual species or fine-scale habitat differences is often missed (Nagendra et al., 2013). Moreover, the aggregation of quantitative features into classes induces errors, is labour-intensive and might miss essential information on habitat requirements (Oeser et al., 2020).

Measures of species diversity are based on the habitat heterogeneity hypothesis which states that habitat heterogeneity benefits species diversity by increasing niche availability and diversifying environmental resources (MacArthur and Wilson, 1967). As habitat heterogeneity can be measured by the spatial variation in the remotely sensed signal, this is also known as the Spectral Variation Hypothesis (Palmer et al., 2002; Rocchini et al., 2007).

To detect individual species, high-resolution sensors such as hyperspectral and LiDAR sensors on drones or high-resolution satellite imagery combined with field observations are often necessary but are usually expensive and labour-intensive. Moreover, only a limited set of species can be monitored with remote sensing, often defined by their size (e.g. large vertebrates versus invertebrates or tree versus grass species) or spectral properties (e.g. invasive species detection (Sladonja and Damijanić, 2021) or crop monitoring (Wu et al., 2022)). As a result, information on individual species is typically obtained indirectly, such as by using remotely-sensed quantitative measures of habitat quality. The increasing spatial and temporal resolution, the availability of long-term time series, and the fine resolutions at which data can be obtained in an accessible and cost-effective manner (e.g. through Google Earth Engine) are promising advances for the use of remote sensing data as a source of environmental predictors in SDMs (Randin et al., 2020). Examples of remotely sensed predictors in SDMs are fractional land cover (Milanesi et al., 2017), landscape heterogeneity (Amici et al., 2015), spectral bands (Hubert-Moy et al., 2022), vegetation indices (Evens et al., 2021; Parviainen et al., 2013; Sheeren et al., 2014), ecosystem functioning attributes (EFAs) (Regos et al., 2020) and vegetation structure (Bellis et al., 2008; de Vries et al., 2021; Farwell et al., 2021; Wood et al., 2013).

[4] Retrievable from https://land.copernicus.eu/pan-european/corine-land-cover/clc2018

2. Quality of citizen science data

2.1. The trade-off between data quantity and data quality

Citizen science data (CSD) vary in quantity (i.e. the number of species records) and quality (i.e. the information content), depending on the underlying structure with which they were collected (Dobson et al., 2020). CSD with higher information content can produce more reliable scientific results but this is usually at the expense of limited data quantity (Bird et al., 2014; Figure 3). Unstructured data are voluntarily collected by individuals with different levels of expertise and in an unstandardized manner, resulting in large amounts of data with low information content hence uncertain data quality. In this dissertation, we consider all opportunistically collected data as unstructured, even when such data can also include some basic additional information on the observation (e.g. date, time, precision) or individual (e.g. sex, life stage) (cf. Kelling et al., 2019). Semi-structured CSD are collected by volunteers that follow some basic instructions and include checklist data, where users check observed species from a list that usually includes all the species from a particular taxonomic group. Users can indicate whether they looked for all listed species (complete checklists) or not (incomplete checklists). Structured CSD usually have the highest information content as they are collected in systematic surveys, where trained volunteers collect detailed information on species occurrence by following rigorous protocols designed by specialists, such as MEETNETTEN in Flanders (Westra et al., 2016) or the United Kingdom Butterfly Monitoring Scheme (Brereton et al., 2019).

The variation in data quality and quantity has different implications for using CSD in SDMs. Structured protocols are designed to reduce uncertainties and improve statistical inference. The collected data can give additional information on the observers (e.g. level of expertise), the observation process (e.g. checklist duration to quantify search effort) or the species (e.g. absences derived from transect surveys or complete checklists to estimate detectability). Moreover, more structured data are more likely to include information on both species presences and absences (presence-absence data) or species abundance (count data), while unstructured data are generally opportunistic records of species presences (presence-only data). Additionally, independent structured data is preferred for model evaluation (i.e. checking the performance of the SDM by comparing model predictions with real presence and absence locations; see section 3.3) given their high reliability.

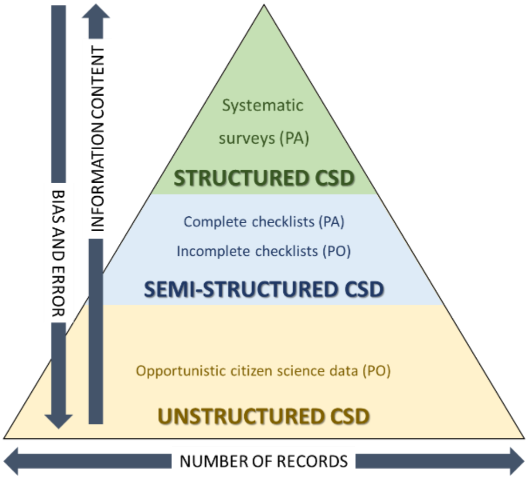


Figure 3: The trade-off between data quality and data quantity in citizen science data (CSD) (adapted from Isaac and Pocock (2015)). Unstructured CSD have lower information content but are usually available in large amounts. However, they are more likely to contain error and bias (e.g. imperfect detection or sampling bias – see section 2.2) which can cause misleading inferences on species distributions and diversity. Structured data are less available but contain more information. They include presences and absences (PA) and/or counts, while unstructured data contain information on presences only (PO). Semi-structured data can contain both PA and PO, depending on the protocol.

Despite their high quality, structured data have limited applicability as response data in large-scale SDM studies due to their restricted availability for a wide range of species and over extensive geographical areas (Urban et al., 2016). However, note for instance eBird, a citizen science initiative that collects bird observations, mostly in the form of checklist data. Their efforts to motivate users to provide semi-structured data have paid off and illustrate that long-term and large-scale citizen science projects can deliver high-quality data with relatively low effort (Johnston et al., 2018; Kelling et al., 2019, 2018). Also in other databases, the submission of checklists is promoted and semi-structured data are growing, for example in Artportalen (Henckel et al., 2020) or *waarnemingen.be* (BOX 1). Nevertheless, unstructured opportunistic CSD remain the largest data source and their high unparalleled spatial, temporal and taxonomic coverage makes them a promising tool for SDMs (Kosmala et al., 2016). To increase the application potential of opportunistic CSD in SDMs for biodiversity conservation, however, it is crucial to understand the various types of bias and error inherent to opportunistic CSD and to identify methods to reduce and mitigate them. Failure to address these issues may result in the misestimation of species distributions, leading to inadequate conservation measures (Guillera-Arroita et al., 2015; Vantieghem et al., 2017).

2.2. Quality issues in opportunistic citizen science data

Opportunistic CSD have high uncertainty about their completeness, correctness, and precision, leading to potentially biased and erroneous predictions (Isaac and Pocock, 2015). Literature on the different types of bias in CSD is abundant, and definitions are sometimes confused, yet two types are highly relevant for SDM studies, i.e. sampling bias and detection bias.

Sampling bias results from an uneven sampling of observations across space (spatial bias) or time (temporal of phenological bias), for example, due to differences in location accessibility (e.g. near hiking trails versus remote locations) or general observer activity (e.g. in summer versus winter). Oversampling of particular locations is more common in opportunistic CSD and might lead to residual spatial autocorrelation. Spatial autocorrelation is inherent to most species occurrence data and arises when the values of environmental variables at nearby locations are correlated. This is a natural process because species tend to occupy areas with similar environmental conditions and their distributions are constrained by biological factors like dispersal and competition (Dormann et al., 2007). Residual spatial autocorrelation occurs when the correlation between observations is not fully explained by the model, resulting in biased parameter estimates, higher risk of type I errors (i.e. assuming an effect when there is none) (Dormann et al., 2007) and inflated model accuracy (Segurado et al., 2006; Veloz, 2009).

Detection bias is either the result of imperfect detection (i.e. false negatives) or observation errors (i.e. false positives). Imperfect detection occurs when observers visit locations but fail or consciously choose not to record a species while it is present, for example, caused by environmental circumstances (e.g. low visibility), taxonomic differences (e.g. species phenology) or human differences (e.g. observer preferences or experience) (Kéry and Schmidt, 2008). This leads to underestimations of the true probability of occurrence, a commonly observed bias in opportunistic presence-only data (Lahoz-monfort et al., 2014). Imperfect detection generally does not impact the ranking of locations on habitat suitability, but attention should be paid to situations where detectability is negatively correlated with occupancy or with the predictors used to estimate occupancy (Guillera-Arroita et al., 2015; Lahoz-monfort et al., 2014). Observation errors, on the other hand, may be caused by low expertise or experience of observers (Fitzpatrick et al., 2009; Ratnieks et al., 2016) or phylogenetic relatedness of species (Vantieghem et al., 2017). False positives can cause both over-predictions and under-predictions of the probability of occurrence (Costa et al., 2015).

3. Opportunistic data in species distribution models: best practices

3.1. Presence-only species distribution models

Correlative SDMs have gained popularity with the emergence of online citizen science platforms. As these platforms largely provide opportunistic presence-only data, there has been an increase in research on methods that can effectively deal with them, such as logistic regression, machine-learning methods and point process models (Dorazio, 2014; Elith et al., 2006; Liu et al., 2013; Phillips et al., 2009, 2006; Renner et al., 2015; Valavi et al., 2022). When pseudo-absences can be inferred from the data (or absences are available), logistic regression is a natural choice for modelling species occurrence data. However, as information on absences is usually not available, presence-only SDMs typically contrast the environmental conditions at locations where species are present with the available environmental conditions in the study area (i.e. the background) (Elith et al., 2010), which is why presence-only data are also called presence-background data (Wang and Stone, 2019). Note that by presence-only SDMs, we do not mean methods that only consider presences such as climatic envelopes (e.g. BIOCLIM; Busby (1991)).

The selection of an appropriate SDM method should not only consider the type of data (e.g. presence-absence or presence-only) but also take into account the quality of the data and the requirements of its users (section 2.2; Guillera-Arroita et al., 2015). The challenge is to find a trade-off between the complexity and interpretability of the model and the bias reduction in its predictions (Elith and Leathwick, 2009). In our case, for example, semi-structured data were only poorly represented within the *waarnemingen.be* database at the start of this project in 2018. Meanwhile, there was a remarkable surge in the growth of extensive opportunistic data platforms, leading to an amplified demand for understanding how to effectively handle these vast amounts of largely unstructured records. Our objective was, therefore, to explore the potential applications of opportunistic CSD when structured data were unavailable instead of combining data with different information content (i.e. data-integration methods; see sections 3.2.3 and 23.1). Moreover, using relatively simple methods benefitted the application potential of our research, for example for users possessing basic statistical and programming skills, as well as for regions with limited IT infrastructure where it is preferable to avoid high computational demands and the associated costs.

The next sections briefly discuss potential SDM methods for presence-only data and motivate our choice for the methods used in this dissertation (i.e. Maxent in Chapters II, III and VI and Gibbs Point Process Models with a Geyer interaction process in Chapter IV).

3.1.1. Logistic regression

Logistic regression methods include Generalised Linear Models (GLMs), which parametrically fit linear, quadratic and/or cubic terms, Generalised Additive Models (GAMs), which non-parametrically fit non-linear terms (i.e. smoothers) (Guisan et al., 2002) and Multivariate Adaptive Regression Splines (MARS), which are similar to GAMs but use piecewise linear basis functions instead of smoothers (Elith and Leathwick, 2007). Using traditional approaches such as GLMs and GAMs in a presence-background setting is possible yet requires careful implementation of methods to avoid overfitting and minimize the impacts of class imbalance (i.e. a disproportionate number of presences versus background samples). Such methods, for example, include the selection of the locations and number of background points. Recommendations for that selection, however, can be confusing as they highly depend on external factors such as spatial scale and model algorithm (Barbet-Massin et al., 2012; Renner et al., 2015).

3.1.2. Machine-learning methods

Machine-learning methods can fit complex species-environment relationships and are extremely suited when high predictive performance is desired (Elith and Leathwick, 2009). In comparative studies on presence-only SDM performance, Boosted Regression Trees (Elith et al., 2008), down-sampled Random Forests (Chen et al., 2004) and Maxent (Elith et al., 2010; Phillips et al., 2006) outperformed most other methods (Elith et al., 2006; Valavi et al., 2022). Note, however, that tuning individual presence-only models in an ensemble approach might deliver even better results (Valavi et al., 2022).

Boosted Regression Trees (BRT) and Random Forests (RF) are ensembles of single non-linear regression trees which are selected in a stagewise (BRT) or bootstrap (RF) approach. Boosting increases model accuracy (Elith et al., 2008) and down-sampling deals with class imbalance in RF and is therefore preferred to regular RF in presence-background settings (Valavi et al., 2022). In general, the issues of class imbalance and class overlap (i.e. when background points are sampled at presence locations) for RF and computational time for BRT (Valavi et al., 2022) make these methods more challenging to implement as a presence-only method.

Maxent uses maximum entropy methods, which means that the algorithm will try to find the closest fit to a prior distribution (i.e. the predictor values at background locations) considering some restraints (i.e. the predictor values at presence locations) (Merow et al., 2013). It does so using features, i.e. mathematical transformations of the predictors, and regularization, i.e. optimizing model fit while avoiding over-fitting (Phillips et al., 2006). The algorithm predicts a relative occurrence rate under the assumption of spatial independence, i.e. no sampling bias (Phillips et al., 2017; also see section 3.1.3).

We decided to use Maxent, as it is still considered one of the best-performing presence-only methods and has relatively low computational power (Elith et al., 2006; Valavi et al., 2022). This was an advantage for the extensive (i.e. many repetitions) and large-scale assessment of the combined impact of data quality and sample size on model performance in Chapters II and III. For ecological studies at finer scales and when the goal is to study the impact of environmental covariates on species occurrence (e.g. Chapter IV), other methods such as point process models will be more suited.

3.1.3. Point process models and their equivalence to Maxent

Point Process Models (PPMs) have regained attention as presence-only SDMs due to their statistical agreement with Maxent (Renner et al., 2015; Warton and Shepherd, 2010). Maxent can be interpreted as an Inhomogeneous Poisson Point Process (IPP) (Phillips et al., 2017; Renner and Warton, 2013), i.e. a point process where the intensity depends on the underlying spatial environment.

Suppose a region *D* with surface *A* with *m* presence records at a set of locations $u_x = \{u_1, u_2, \ldots, u_m\}$. In the case of an IPP, the intensity $\lambda(u)$ of this point pattern (i.e. the expected number of presence records per unit area) is a log-linear function of a vector of real-valued covariates $Z(u)$:

$$\lambda(u) = exp(\alpha + \beta Z(u)) \qquad \textit{Equation 1}$$

where α is a normalizing constant to ensure that $\int_D \lambda(u)du$ equals the total number of occurrence records and β is the vector for covariate effects. The mean abundance in that region can then be estimated as follows:

$$\textit{Predicted mean abundance} = c_P\, A\, exp(\alpha + \beta Z(u)) \qquad \textit{Equation 2}$$

Note the unknown constant c_p, which basically implies that the estimated abundance will always be a relative estimate. In the absence of spatial dependence, the probability of presence can be derived by a link function, i.e. the complementary log-log (*cloglog*) link (Phillips et al., 2017), as follows:

$$\textit{Probability of presence} = 1 - \exp(-c_p\, A \exp(\alpha + \beta Z(u))) \qquad \textit{Equation 3}$$

This is the default output of Maxent and allows for an intuitive interpretation of model predictions. However, Maxent assumes spatial independence, an important consideration when choosing a presence-only SDM method (Renner and Warton, 2013; Yackulic et al., 2013). Opposed to Maxent, PPMs can incorporate spatial dependence by adding a random intensity function (i.e. Cox models) or a spatial interaction term (i.e. Gibbs models). Gibbs models explicitly postulate that spatial dependence is due to interactions between points (i.e. attraction or repulsion), while Cox models merely assume spatial dependence due to clustering defined by some random process (i.e. an unobserved external factor) (Renner et al., 2015). Both methods are feasible, yet we decided to use a Gibbs model following De Solan et al. (2019), where the spatial interaction term can be tuned and bias covariates are used to mitigate sampling bias (section 3.2.3).

In the case of Gibbs models, the intensity becomes a conditional intensity $\lambda(u|x)$ at a location u given a pattern of presences x and the maximum likelihood becomes a maximum pseudolikelihood, which is an approximation to reduce computational effort (Baddeley and Turner, 2000). The conditional intensity consists of a first-order term β (the trend or covariate effects) and a higher-order term γ (the interaction parameter). The main idea to incorporate spatial dependence is based on the following expression:

$$\lambda(u|x) = \beta(u)\gamma^{v(x,r,s)} \qquad \textit{Equation 4}$$

The higher-order term can take many forms, of which the Geyer saturation process is an interesting choice for modelling species occurrence data. Here, the estimated intensity (i.e. relative abundance) $\lambda(u|x)$ at a location will depend on the underlying environmental conditions ($\beta(u)$ in Equation 4) and the configuration of the surrounding points ($\gamma^{v(x,r,s)}$ in Equation 4). The exponent $v(x,r,s)$ can be interpreted as a weighting function that is defined by the configuration of the points x within a radius r from a point u. A saturation parameter s ensures that the conditional intensity cannot take arbitrarily large values in the case where $\gamma > 1$, which implies that points exhibit clustering (while $\gamma < 1$ suggests inhibition and $s = 0$ suggests a Poisson point process with no spatial interaction). Clustering increases $\lambda(u|x)$ and can be tuned by choosing r

as the distance at which spatial dependence occurs and *s* as an indicator of the strength of the spatial dependence. For a detailed statistical explanation of the Geyer process or Gibbs PPMs in general, see Chapter 13 in Baddeley et al. (2015).

3.2. Dealing with bias and error

Ideally, sources of bias and error are removed from the data before modelling. Quantifying bias can, for example, improve the sampling design or dirige sampling at locations where species occurrence data is under-represented (Araújo and Guisan, 2006; Ruete, 2015). However, to take full advantage of the readily available high quantity of opportunistic presence-only records, one of the priorities in SDM research has been finding ways to deal with bias and error, resulting in different methods and recommendations (Bird et al., 2014; Isaac et al., 2014). Removing erroneous observations and accounting for bias improves both model calibration and predictive performance (e.g. Boria et al., 2014; Johnston et al., 2018; Merow et al., 2017; Steen et al., 2019).

3.2.1. Data cleansing

Data are expected to be cleansed in preparation for an SDM study (Zurell et al., 2020), which includes the removal of spatial and temporal outliers, duplicates, and records with low precision (Serra-Diaz et al., 2017). When dealing with opportunistic data, cleansing usually also implies stringent filtering, where data are filtered based on record attributes that hold information on the observation process or post-entry data verification (Steen et al., 2019; Vantieghem et al., 2017). The main goal is to remove uncertainty, for example by only allowing observations verified by semi-automatic verification systems (Vantieghem et al., 2017) or observations from species experts (Steen et al., 2019).

3.2.2. Bias reduction

Pre-modelling methods for reducing sampling bias either subsample the data or manipulate the background. Probably the most implemented method for dealing with sampling bias (and spatial autocorrelation) is the subsampling or spatial (and temporal) filtering of occurrence records (Boria et al., 2014). Spatial filtering or spatial thinning aggregates records within a predefined distance, but this method must be implemented with care as the distance used for spatial filtering automatically also defines the spatial resolution (or grain) at which the model can be interpreted and reduces sample size. This might lead to losses of information on species occurrence (El-Gabbas and Dormann, 2017), biological processes such as dispersal mechanisms (McPherson and Jetz, 2007) and impacts of fine-scaled environmental conditions (Connor et al., 2017). To account for imperfect detection, temporal aggregation of opportunistic records can be useful, as a species is more likely to be detected in repeated visits (MacKenzie et al., 2006).

Environmental filtering is another subsampling technique to reduce sampling bias, where records with similar environmental conditions are aggregated to reduce the impact of an oversampled environmental situation (Varela et al., 2014). However, this method should be implemented with consideration of species prevalence and its response to environmental gradients and with similar considerations regarding sample size reduction as with spatial filtering (Gábor et al., 2020). Another approach to reducing sampling bias is background manipulation, such as target-group background selection (Phillips et al., 2009) or background thickening (Vollering et al., 2019), which can be used to avoid coarse models or extremely low sample sizes.

3.2.3. Bias mitigation

Apart from pre-model bias reduction, there are several in-model techniques for bias mitigation. All SDMs are designed to fill gaps in space and time, yet different statistical techniques have been explored to better tackle different types of bias, such as bias covariates, spatial interaction terms (see section 3.1.3) and data-integration methods. The use of bias covariates is a relatively easy in-model method to mitigate bias with a large application potential in presence-only SDMs. Here, known sources of bias can be used to account for their impact on model predictions. Bias covariates are used in presence-only SDMs to train the model and are consequently kept constant for model predictions (Warton et al., 2013). Examples include measures (or combinations) of search effort (e.g. relative intensity of species observations in its taxonomic group or the number of unique sampling dates), accessibility (e.g. distance to roads or urban areas and road density) or detection probability (e.g. date or weather conditions) (e.g. De Solan et al., 2019; El-Gabbas and Dormann, 2018; Fletcher et al., 2019; Simmonds et al., 2020; Warton et al., 2013).

Data-integration methods mitigate bias by combining information from one or more datasets (Fletcher et al., 2019; Isaac et al., 2020). Occupancy-detection models (MacKenzie et al., 2006), for example, retain information on occurrence and detectability by modelling data from repeated visits in a hierarchical structure. This method allows us to account for different sources of imperfect detection, such as differences in observer expertise (Johnston et al., 2018; Yu et al., 2010). When both opportunistic presence-only data and structured survey data are available, integrated SDMs can be used to mitigate bias (Miller et al., 2019). The idea behind the method is that bias can be quantified by assessing the difference between unstructured and structured data while keeping the strengths of both data types (Isaac et al., 2020). We chose to focus on methods that deal with opportunistic presence-only data only (this was motivated in section 3.1).

3.3. Model calibration, prediction and evaluation

This section focuses on the modelling step (Figure 4, page 22, presents an overview of the different steps and methods in this study). Maxent was chosen to model species occurrence at coarse resolutions and when the objective was to assess the predictive performance of the model based on an independent dataset (Chapters II and III). PPMs with a Geyer saturation process were chosen to assess the impact of fine-scaled environmental variables on species occurrence (Chapter IV).

In an SDM study, models are usually first calibrated and, depending on the goal of the study, then used for predictions. Model calibration is improving the agreement between the species occurrence data and the model predictors. This includes the selection of a modelling method and predictors (Guisan and Zimmermann, 2000). Model prediction is using the fitted model to predict a probability of occurrence, for example in regions where no data on species occurrence was collected. It is important to know that presence-only SDMs have a couple of limitations because species prevalence is unknown due to imperfect detection. First, they are unable to estimate the intercept (i.e. the average prevalence of the species) due to the lack of information on species absences (Fithian and Hastie, 2013; Warton and Shepherd, 2010). Predictions from presence-only SDMs will thus always be a measure of relative abundance (or intensity) or relative probability of presence (Phillips et al., 2017). Second, when imperfect detection is not accounted for, only a ranking of the relative occurrence probability (or habitat suitability) can be obtained. This might not be proportional to the true probability of occurrence (Guillera-Arroita et al., 2015). Finally, presence-only SDMs cannot be used to study populations or trends (Guillera-Arroita et al., 2015; Kamp et al., 2016; Lee-Yaw et al., 2022).

The performance of an SDM can be assessed for calibration and prediction. In regression modelling, calibration performance is assessed by “goodness-of-fit” measures based on the deviance, i.e. the difference between observed and fitted values. Common examples are the pseudo R^2, which measures the proportion of variance explained by the predictors, and the AIC (Akaike’s Information Criterion), which does the same but penalizes for the number of parameters (Burnham et al., 2011). However, those common maximum-likelihood-based methods do not apply to Gibbs point process models because the assumption of spatial independence is violated (Baddeley et al., 2015). Instead, one can rely on simulation envelopes of summary functions and related tests, such as the Diggle-Cressie-Loosmore-Ford test (Baddeley et al., 2014). For Maxent, although model calibration is incorporated in the algorithm (section 3.1.2), additional information on calibration performance can be obtained by evaluating the model on the training data (Phillips, 2017). Note that when sampling bias or imperfect detection are not accounted for, calibration performance is limited (Guillera-Arroita et al., 2015; Pearce and Ferrier, 2000) and covariate effects might be affected (Lahoz-monfort et al., 2014).

Assessing a model's predictive performance (i.e. model evaluation) is preferably done by measuring the ability of the model to predict independent data (Fielding and Bell, 1997). As such independent data are not always available, a common approach in SDM studies is to perform a repeated cross-validation on subsets of the training data which were set aside for model calibration (Guisan and Zimmermann, 2000). The results of such 'internal' cross-validation should always be interpreted with care as they are susceptible to different sources of bias (Roberts et al., 2017) and cannot be compared between species (Lobo et al., 2008). Additional bias mitigation methods (to the ones in section 3.2.3) are spatial block cross-validation or checkerboard cross-validation for model selection, which can reduce the impact of spatial autocorrelation and consequent inflated model predictive performance (Roberts et al., 2017).

4. Towards evidence-based biodiversity conservation policy

Biodiversity monitoring has seen significant advancements in recent years, particularly in the form of species distribution models (SDMs), citizen science data (CSD), and remote sensing (as introduced in section 1.3.2). These trends have been exciting governing bodies and conservation practitioners as they allow them to strengthen biodiversity policy measures and management decisions with evidence-based research (Sutherland et al., 2020). Evidence-based research on species distributions and the drivers of their change is important to assess the ecological feasibility of conservation measures (Dicks et al., 2014; Downey et al., 2021), allocate funding (Parks et al., 2022) and ease acceptance by policymakers and land owners (Sutherland and Worldley, 2018).

Opportunistic CSD can be used in multiple ways to support biodiversity conservation policy. They can give direct information on the current distributions or status of species, or be used as input for SDMs. SDMs can be implemented for individual or multiple species, for example by stacking their predictions to indicate species richness or biodiversity hotspots (Demolder et al., 2014; Dubuis et al., 2011; Vila-Viçosa et al., 2020) or by modelling biodiversity directly (Dorazio et al., 2006; Pollock et al., 2014).

4.1. Opportunistic citizen science data as a complementary information source

Opportunistic CSD can support biodiversity monitoring and conservation and might even deliver the same information as structured surveys when enough data is available (Callaghan et al., 2020). However, different biases make this unlikely for most taxonomic groups, as these data usually lack the necessary information to provide direct measures of species distributions or population trends. They can, however, complement structured survey data for IUCN Red List assessments (Maes et al., 2015, 2018) and estimates of species richness (Soroye et al., 2018). Additionally, opportunistic CSD can support the design of monitoring protocols (Westra et al., 2016) or provide information on current distributions to refine maps of potential habitat suitability created with mechanistic models (Maes et al., 2016).

4.2. Opportunistic citizen science data as input for species distribution models

A quick literature query[5] indicated that 7176 studies mentioned presence-only SDMs (Maxent or PPMs specifically) with possible management applications. Of those, only 164 studies (2%) mentioned opportunistic or citizen science data, of which only 36 studies were published at the start of this research in January 2018. These numbers illustrate the multitude of studies that use presence-only SDMs for conservation applications but the limited adoption of opportunistic CSD to support them, although numbers have been increasing over the past four years.

Management recommendations are often still based on SDMs built with systematic survey data (Demolder et al., 2014; Evens et al., 2021; Seavy et al., 2009; van den Berg et al., 2001), which is the better option when such data are sufficiently available (Simmonds et al., 2020; Suhaimi et al., 2021). When structured data are limited, which is usually the case, they are ideally combined with unstructured data (also see sections 3.2.3 and 23.1). Unfortunately, structured

[5] A query was conducted on the 21st of March 2023 on https://kuleuven.limo.libis.be/ with the following search terms: species distribution model, ecological niche model or habitat model; presence-only or presence-background; Maxent or PPM; application or management; (opportunistic and citizen science).

data are unavailable for a majority of species and regions. Other studies have compared the effectiveness of using structured versus opportunistic data in SDMs. Results have been contradictory, depending largely on the adopted methods. Not accounting for sampling bias in opportunistic CSD gave misleading insights into species-environment relationships (Broman et al., 2014). For estimating relative habitat suitability, logistic regression methods based on opportunistic data with inferred absences outperformed presence-only methods like Maxent (Bradter et al., 2018) and even systematic survey data for rare species (Henckel et al., 2020). By further reducing the uncertainty surrounding the use of opportunistic CSD in SDMs, through quality improvement and case studies, our research will make valuable contributions to biodiversity conservation.

We distinguish three main conservation applications of SDMs that can also integrate opportunistic citizen science data:

(i) the delineation and prioritization of areas for biodiversity conservation and monitoring,
(ii) risk assessments under future long-term scenarios of land conversion and climate change,
(iii) and habitat management.

SDMs can produce habitat suitability maps that aid in the delineation and prioritization of areas for biodiversity conservation and monitoring. These maps indicate areas with suitable environmental conditions, whether they are currently occupied or could be in the future, and provide guidance for conservation efforts. For example, SDMs can prioritize areas with high conservation value or so-called 'hotspots' (Prendergast et al., 1993), identify ecological corridors to facilitate spontaneous migration (Vanden Broeck et al., 2017) or guide the reintroduction (Cianfrani et al., 2013; Maes et al., 2019c; Miranda et al., 2019) or translocation of species (Eyre et al., 2022). SDMs have also proven useful in guiding field surveys for species monitoring (Carvalho et al., 2016), even for critically endangered species (Eyre et al., 2022). In addition, model-based stratifications can help target areas where additional sampling effort is needed to build better SDMs (Araújo and Guisan, 2006; Ferrier et al., 2004; Guisan et al., 2006).

SDM predictions can also be used for risk assessments under future long-term scenarios of land conversion and climate change (Elith and Leathwick, 2009; Maes et al., 2010; Urban et al., 2016). Such assessments can, for example, include studies on the impact of habitat fragmentation on species distributions in human-dominated landscapes (Rutten et al., 2019), the impact of climate change on species with low dispersal ability (Sanczuk et al., 2022), the colonisation risk of invasive species (Bradley and Mustard, 2006; Truong et al., 2017) or human-wildlife conflicts (Rutten et al., 2019; Swinnen et al., 2017). Furthermore, extrapolating SDM predictions or running SDMs in different climatic regions can predict species responses to extreme weather conditions and support assisted migration (Sanczuk et al., 2022; Van Daele et al., 2021). While climate SDMs are beyond the scope of this dissertation, improving the quality of CSD and illustrating their application potential for addressing other ecological and conservation challenges will also support their use for climate-related research.

Finally, SDMs can support habitat management, by providing new insights or scientific evidence that underpins the ecological principles of species-environment relationships. For example, they can assess the impact of habitat heterogeneity at large extents and different scales (de Vries et al., 2021 (butterflies); Seavy et al., 2009; Van den Berg et al., 2001 (birds); Sillero and Gonçalves-Seco, 2014 (reptiles)) or the impact of nitrogen pollution on invertebrates (Nijssen et al., 2017; Vantieghem et al., 2017). SDMs can also support habitat management outside designated conservation areas, such as for designing agri-environment schemes (Sullivan et al., 2017).

5. Research focus

The main objective of this research is to reduce the uncertainty surrounding opportunistic citizen science data (CSD) and to promote their uptake in biodiversity conservation management and policy. Opportunistic CSD might be a valuable source of information given its high quantities, yet its value for conservation applications has not been fully explored due to its uncertain quality (Dobson et al., 2020; Guillera-Arroita et al., 2015). Stakeholders that might profit from the results and methods presented in this dissertation include policymakers, conservation practitioners, educators, environmental organisations, database managers and citizen scientists. Although our research has various stakeholders, the applications in this dissertation will largely focus on conservation policy as there is a strong need for evidence-based action plans for conservation management (IPBES, 2019; Kadykalo et al., 2021; Louette et al., 2015; Maes et al., 2017a; Sutherland et al., 2004; Wood et al., 2018). These needs will be tackled in two main parts, with part 1 focussing on data quality and part 2 focussing on using novel predictors for conservation applications (Figure 4).

Exploring cheap, transparent and accessible ways to scientifically support biodiversity conservation is extremely important as it will enable the new generation of conservation practitioners and policymakers to make informed decisions (Downey et al., 2021; Parker et al., 2016; Sutherland et al., 2020). However, this dissertation does not wish to diminish the value of field experience nor does it claim that biodiversity conservation can be supported merely by interpreting the results of statistical models that integrate state-of-the-art methods such as big data analysis and remote sensing. Best practices for conservation management depend on many additional factors, such as local socioeconomic or ecological restrictions.

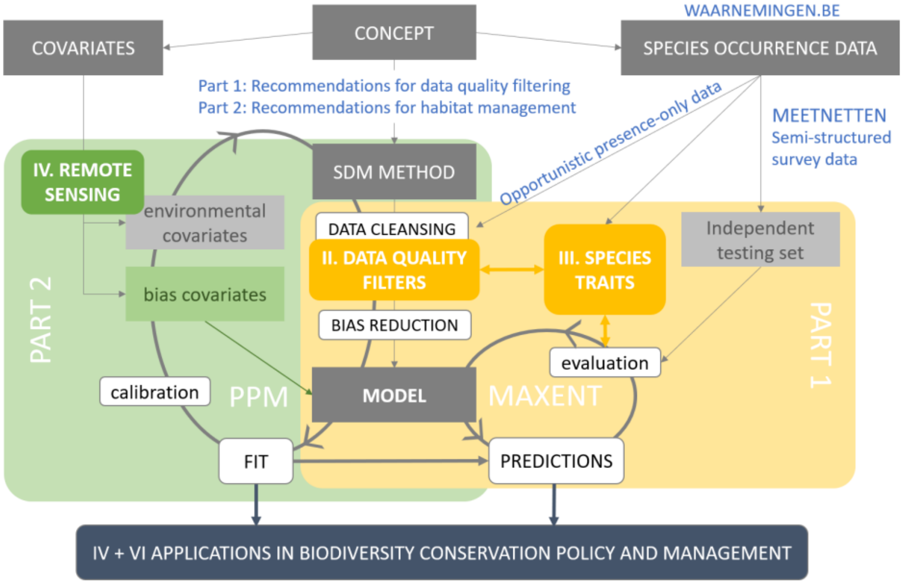


Figure 4: Overview. The first part of the dissertation will focus on formulating recommendations for data quality filtering of opportunistic presence-only citizen science data (CSD) retrieved from the waarnemingen.be (BOX 1) database. More specifically, Chapter II will explore the data quantity-quality trade-off by applying different data quality filters to opportunistic records while controlling for sample size. Maxent predictions will be evaluated to assess the impact of data quality filtering on species distribution model (SDM) performance. Chapter III will assess how species traits impact the results from Chapter II. The second part of the dissertation (Chapter IV) will integrate CSD and remote sensing data to formulate recommendations for habitat management. Point Process Models (PPMs) with a spatial interaction term and bias covariates will be used to mitigate sampling bias (section 3.2.3). The results of both parts are discussed in Chapter V, including important considerations and suggestions for future research. Chapter VI will elaborate on potential applications in biodiversity conservation policy and management. The figure is based on figures in Guisan and Zimmermann (2000) and Zurell et al. (2020), adapted for the methodology in this research.

5.1. (part 1) Chapter II: The data quality-quantity trade-off in stringent filtering

Knowledge gap **–** While much is known about the separate impact of sample size and data quality on the performance of SDMs, the simultaneous impact of both increasing data quality and reducing sample size on model performance has remained relatively unexplored.

Objective **–** The first specific objective of the dissertation was to test the impact of stringent filters on the data quality of opportunistic CSD when these are used as input for a presence-only SDMs. In addition, the study aimed to determine the threshold of sample size at which the trade-off between data quality and data quantity becomes unfavourable. Based on these results, recommendations for data quality filtering will be formulated and feedback on data quality management can be given to data collectors (i.e. database managers and citizen scientists).

RQ *– How can SDM performance be increased by quality filtering of opportunistic CSD when both data quality and data quantity are taken into account?*

5.2. (part 1) Chapter III: Insight into the drivers of data quality with species profiles

Knowledge gap **–** While stringent filtering is common practice when using opportunistic CSD in SDMs, filtering recommendations have remained relatively general and associations between species traits and filtering recommendations are sparse.

Objective **–** The second specific objective of the dissertation was to group species into species profiles based on their traits and their response to data quality filtering. The recommendations for data quality filtering from the first study will be improved with novel insights.

RQ *– How can recommendations for data quality filtering be fine-tuned using species characteristics?*

5.3. (part 2) Chapter IV: Integrating citizen science and remote sensing data for habitat management

Knowledge gap **–** Biodiversity conservation often requires habitat management, especially in landscapes that have been impacted by human activities. However, developing effective habitat management strategies for fragmented landscapes under anthropogenic pressures can be challenging, particularly at multiple scales. Traditional approaches have been relying on small-scale experiments and personal experiences, making it difficult to generalize best practices across different regions and ecosystems. In addition, habitat management efforts often focus primarily on vegetation, and animal species are sometimes neglected.

Objective – The third specific objective of the dissertation was to illustrate how integrating citizen science and multispectral satellite data can support habitat management at multiple scales. Based on these results, recommendations will be formulated to support management decisions in heathlands in fragmented and anthropogenic regions.

RQ – *Can opportunistic CSD and remote sensing data be integrated for supporting biodiversity conservation practices at multiple spatial scales?*

6. Overview

Chapter II explores the quality-quantity trade-off in stringent filtering. Three data quality filters were applied to opportunistically collected species records, by labelling them according to attributes that were derived from the *waarnemingen.be* database (BOX 1). For a selection of species from four well-studied taxonomic groups (birds, butterflies, plants and dragonflies), Maxent SDM performance was measured by evaluating model predictions on an independent high-quality presence-absence testing set. The difference in model performance between models based on filtered and unfiltered data of different sample sizes was assessed to disentangle the impact of data quality filtering on model performance.

Chapter III aims to gain more insight into the drivers of quality. We hypothesized that data quality-related species traits could influence the recording of species observations by volunteers. Therefore, we first assessed the impact of these traits from 91 species from three taxonomic groups (birds, butterflies and dragonflies) on the results of the first study, i.e. the difference in model performance caused by filtering for different degrees of sample size reduction. Second, a principal component and clustering analysis were performed to define five species profiles for which filtering recommendations could be formulated.

Chapter IV combines citizen science data with remote sensing data for a conservation application. We used a second-order texture metric (Haralick, 1979) derived from multispectral Sentinel-2 data to quantify small-scaled vegetation structure in heathlands in the Campine (northeastern) region of Belgium. Point process models were run to assess the impact of vegetation structure, together with heathland size and heathland heterogeneity, on the habitat suitability of typical dry-heathland fauna. Heathlands were divided into areas surrounded by open, closed or anthropogenic features to assess the impact of the landscape context on the relationship between vegetation structure, heathland heterogeneity, heathland size and species occurrence.

Chapter V discusses the findings in both parts of the dissertation in an integrated manner and highlights some important considerations when interpreting our results. It also gives future perspectives on the use of opportunistic CSD in presence-only SDMs to support biodiversity conservation policy by suggesting new research.

Chapter VI discusses the general application potential of our results specifically and elaborately illustrates the application potential in Flemish conservation policy. Applications in both nature conservation areas (e.g. Natura 2000) and areas of intensified land use (e.g. agri-environment schemes) will be considered and two case studies with results from a preliminary analysis will be presented.

CHAPTER II. Is There a Data Quality-Quantity Trade-off in Stringent Filtering?

Adapted from

Van Eupen, C., Maes, D., Herremans, M., Swinnen, K.R.R., Somers, B., Luca, S., 2021. The impact of data quality filtering of opportunistic citizen science data on species distribution model performance. Ecol. Modell. 444, 109453. https://doi.org/10.1016/j.ecolmodel.2021.109453

Author contributions

Camille Van Eupen: conceptualization, methodology, software, validation, formal analysis, writing – original draft, visualization; **Dirk Maes:** conceptualization, writing – review & editing, supervision; **Marc Herremans:** conceptualization, data curation, writing – review & editing; **Kristijn Swinnen:** data curation, writing – review & editing; **Ben Somers:** conceptualization, writing – review & editing, supervision; **Stijn Luca:** conceptualization, writing – review & editing, supervision

ABSTRACT

Opportunistic citizen science data are often used for species distribution models (SDMs) when high-quality data collected through standardized recording protocols are unavailable. While opportunistic data are abundant, uncertainty is usually high, e.g. due to observer effects or a lack of metadata. To increase data quality and improve model performance, we filtered species records based on record attributes that provide information on the observation process or post-entry data validation. Data filtering does not only increase the quality of species records, it simultaneously reduces sample size, a trade-off that remains relatively unexplored. By controlling for sample size in a dataset of 255 species, we were able to assess the combined impact of data quality and sample size on model performance. We applied three data quality filters based on observers' activity, the validation status of a record in the database and the detail of a submitted record, and analysed changes in AUC, sensitivity and specificity using Maxent with and without filtering. The impact of stringent filtering on model performance depended on (1) the quality of the filtered data: records validated as correct and more detailed records lead to higher model performance, (2) the proportional reduction in sample size caused by filtering and the remaining absolute sample size: filters causing small reductions that lead to sample sizes of more than 100 presences generally benefitted model performance and (3) the taxonomic group: plant and dragonfly models benefitted more from data quality filtering compared to bird and butterfly models. Our results also indicate that recommendations for quality filtering depend on the goal of the study, e.g. increasing sensitivity and/or specificity. Further research must identify what drives species' sensitivity to data quality. Nonetheless, our study confirms that large quantities of volunteer-generated and opportunistically collected data can make a valuable contribution to ecological research and species conservation.

7. Introduction

Appropriate conservation measures must mitigate the alarming declines in biodiversity caused by global pressures such as climate change (Urban et al., 2016), invasive species (Early et al., 2016) and intensifying land use (Newbold et al., 2015). Choosing proper conservation measures requires evidence of the state of biodiversity and species distributions. Ideally, such evidence is gathered through standardised protocols, performed by trained observers and with a clear description of both data collection and project objectives (Kosmala et al., 2016). Such highly structured data, however, is rarely available for a wide range of species, nor for extensive periods or geographical areas (Urban et al., 2016).

In response, less structured but bulky occurrence data with varying information content, often collected by volunteers participating in citizen science initiatives (Theobald et al., 2015), are being explored for biodiversity conservation purposes (Guisan et al., 2013). The value of data with information on detectability or information on absences is indisputable and their applications are abundant, e.g. for species distribution models (SDMs) (Guisan and Zimmermann, 2000; Van Strien et al., 2013; Wood et al., 2018) or Red List compilations (e.g. Maes et al., 2015). In contrast, the value of data with little information on the observation process is uncertain and conservation applications are limited (Dobson et al., 2020; Guillera-Arroita et al., 2015). When such unstructured occurrence data consist of occasional observations of species presences, they are termed opportunistic presence-only data (Giraud et al., 2016) or presence-background data (Wang and Stone, 2019). They are generally used in SDMs (e.g. Maxent (Phillips et al., 2006) or point process models (Renner et al., 2015)) that contrast available environmental conditions in the study area (the background), with the conditions at locations where the species was observed (Elith et al., 2010).

Using opportunistic presence-only data for SDMs has both advantages and disadvantages. The main advantage is the abundance of available data, because easy data collection leads to the coverage of a large number of species over large geographical areas, at a fine scale and over potentially long periods (Kosmala et al., 2016). Online platforms and smartphone applications facilitate an easy recording of species for a volunteer observer, and the number of active observers on data platforms such as iRecord in the United Kingdom (https://www.brc.ac.uk/irecord/), *waarnemingen.be* in Flanders (northern Belgium; https://waarnemingen.be/) or iNaturalist worldwide (https://www.inaturalist.org/) is indeed growing by the hundreds (e.g. *waarnemingen.be*) or even thousands (e.g. iNaturalist) every year. Since the quantity and extent of this data can never be reached by standardised monitoring schemes, opportunistic data can make a valuable contribution to science if processed correctly (Giraud et al., 2016; Soroye et al., 2018). Two major disadvantages of opportunistic presence-only data limiting their application potential (Dobson et al., 2020; Guillera-Arroita et al., 2015), however, are the incapability of delivering probabilistic model outputs (Guillera-Arroita et al., 2015) and a high risk of bias and error (Bird et al., 2014; Isaac and Pocock, 2015). The awareness of these uncertainties reflects in the scepticism towards data quality of opportunistic observations or citizen science data in general (Burgess et al., 2017), because when disregarded in the modelling or decision-making process, these disadvantages can lead to misguided conservation measures (Isaac et al., 2014).

Different strategies are applied to increase the quality of opportunistic datasets. A first strategy is rather bottom-up, where the underlying protocol of a citizen science project is changed (Kosmala et al., 2016). This requires a regime shift and takes time, but can be fruitful (e.g. eBird; Sullivan et al., 2014). A second and promising strategy is data integration (Miller et al., 2019), where multiple sources of opportunistic presence-only data are combined (Lin et al., 2017) or presence-only data is treated as complementary to structured presence-absence data (Robinson et al., 2019). A third strategy, integrated into many national citizen science databases, is data validation, where the identification of the species is verified, often together with the spatial and temporal plausibility of a record. It is common practice in, for example, eBird (Sullivan et al., 2009), *waarnemingen.be* (Swinnen et al., 2018) and iRecord (https://www.brc.ac.uk/irecord/records-verified). However, even with the best experts and state-of-the-art methods (e.g. image recognition), it is challenging to verify thousands of records entering data repositories every day, particularly those without corroborating picture evidence. As a result, many researchers apply a fourth strategy, where data reliability is maximised by data cleansing. This can be done by error detection (e.g. Serra-Diaz et al., 2017), outlier removal (e.g. Kallimanis et al., 2017), filtering in geographic or environmental space (e.g. Varela et al., 2014), or deleting species records based on data attributes (e.g. Rutten et al., 2019), so-called "stringent filtering" (Steen et al., 2019).

The desired effect of stringent filtering is an increase in quality, by reducing bias and error (Steen et al., 2019). However, sample size is inevitably reduced by filtering, which impacts model performance (Gábor et al., 2020; Wisz et al., 2008) and leads to a trade-off between data quality and sample size. To our knowledge, the combined impact of data quality and sample size in stringent filtering on the performance of SDMs remains underexplored. Studies that explored the impact of stringent data filters found a negligible effect on bird occurrence predictions when retaining only structured survey data (Kamp et al., 2016) or data from observers with higher expertise (Steen et al., 2019). On the other hand, predictions were more accurate when using only records validated as correct for a butterfly genus prone to misidentification (Vantieghem et al., 2017), or by using only eBird checklists of observers who travelled larger distances to make their observations (Steen et al., 2019).

In this chapter, we will expand on previous findings by applying different quality filters on a regional species occurrence database *waarnemingen.be* (BOX 1). The database consists of both structured and unstructured recordings in Flanders since 2008 and currently holds more than 51 million species records[6] and one of the densest collections of species records in Europe (Herremans et al., 2018). We aim to identify which quality filters increase the discrimination accuracy of Maxent and to formulate recommendations based on taxonomic group and data characteristics. Every citizen science database is unique and while the considered taxonomic groups in *waarnemingen.be* are blessed with a relatively high proportion of quality data, this might not be the case in all data repositories. The properties of *waarnemingen.be* allowed us to evaluate the impact on model performance of different changes in data quality, for a wide range of changes in sample size. This not only provides more insight into the trade-off between data quality and sample size in stringent filtering but also ensures the transferability of our results to datasets of lower quality and/or record density.

8. Materials and methods

8.1. Dataset and quality filters

We assessed the impact of data quality filtering on opportunistic citizen science data gathered in the Flemish species occurrence database *waarnemingen.be* (BOX 1). The dataset contained both “structured data” or observations supported by guidelines or a protocol (varying from standardized monitoring schemes to small project observations), and “unstructured data” or incidental observations. For a detailed description of the data selection and model testing procedure in this chapter, see section 8.2 and Appendix A. Structured records were separated for model testing (n = 161,782) to measure the performance of the species distribution models (see sections 8.3 and 8.4) and unstructured records were used for model training (n =

[6] Database query from situation on September 30th 2022

5,547,750). We adopted the ODMAP protocol (v1.0, Zurell et al., 2020) and describe the different steps (Overview, Data, Model, Assessment and Prediction) in Appendix B.

We selected three dichotomous filters as a measure for data quality, based on available metadata (Table A.1). The first filter "ACTIVITY" refers to the annual average number of active recording days of an observer, in the study period. We calculated the individual activity rate of observers, including the observers with the highest number of records first and stopped when we reached the observers that cumulatively collected 80% of the data. The threshold for a high activity rate was set to the first quartile of the activity rate of this group, i.e. 92 recording days in one year. We considered this a proxy for observer experience, presumably leading to lower rates of both false-negative and false-positive errors (Farmer et al., 2012; Kallimanis et al., 2017; Kelling et al., 2015). The second filter "DETAIL" reflects whether observers provide information beyond the default date, location and species name, such as species behaviour, photographs or additional comments. Records submitted with more effort are of higher quality when effort is defined by the 'distance travelled for a checklist' (Steen et al., 2019). Because we applied filters to unstructured data only, we used 'record detail' as a measure of effort instead. The third filter "VALSTAT" is based on the status of a record in the internal validation system of the database, indicating if it was evaluated as correct or as uncertain. Records marked as correct are meant to contain no misidentification errors (e.g. Vantieghem et al., 2017), even though an occasional human or software error might occur. Records marked as uncertain have either not been validated or were hard to judge correctly, due to a lack of additional information (Swinnen et al., 2018).

8.2. Data selection

All species records from four well-studied taxonomic groups in Flanders, i.e. birds, butterflies, dragonflies and plants were subjected to some initial data restrictions: (1) records were limited to our study area, the Flemish region of Belgium, (2) observations dated from January 2014 to September 2019, (3) we included only records with sufficient precise geographical location (≤ 500 metres), (4) for birds, only birds that breed in Flanders were used (Vermeersch et al., 2020), and (5) we removed absences (zero-counts) and entries validated as incorrect.

After the initial selection, we divided the data into records for model training and model testing (also see Appendix A and Figure A.1). Structured data were used solely for model testing and never for model training, and were further reduced to high-quality testing records. This was done by selecting only structured records that were validated as correct and from observers with a high activity rate. The model training records consisted of unstructured presence-only data, a data type found in many large-scale datasets of opportunistic species records (e.g. GBIF; https://www.gbif.org). Model training records were subjected to the three quality filters and their combinations, resulting in seven filtered datasets (Figure A.2).

Per species, records from each set (training or testing) were aggregated in a 1x1 km grid, a frequently used resolution in Flemish biodiversity research (e.g. Demolder et al., 2014; Rutten et al., 2019; Vantieghem et al., 2017), resulting in one presence per grid cell per species. This aggregation of records is also known as ‘spatial thinning’ or ‘spatial filtering’, a common technique to reduce spatial bias (Kramer-Schadt et al., 2013) and improve model performance (Boria et al., 2014). The high-quality presences of the model testing set were complemented with absences derived from grid cells with high search effort for the associated taxonomic group, but where the target species was not observed. We kept only species with at least 50 presences in the testing set, and at least one filtered training set with at least 100 presences. This resulted in a dataset of 255 species in four taxonomic groups (full list in Table C.1).

8.3. Species distribution model

We evaluated the impact of stringent filtering on the performance of Maxent (software version 3.4.1, implemented in the R package ´dismo` v1.1-4 (Hijmans et al., 2017)). Maxent is a commonly used presence-only algorithm (Elith et al., 2010; Phillips et al., 2006), which models a relative probability of occurrence based on a species’ presence records and background points. Background points are used to define the contrast between what is available in the environment and what is used by the species (Elith et al., 2010). We included all of the 13552 cells in our study area as background and did not adjust the background selection to correct for sampling bias (e.g. Phillips et al., 2009; Vollering et al., 2019) to ensure comparability of our models (Merow et al., 2013). Comparability was further supported by allowing only linear, quadratic and product features for every model, by setting a minimum sample size of 100, ensuring that the regularization coefficient was kept to 0.05, and by using identical predictors in all Maxent models.

The predictor set represented a range of environmental conditions in our study area and comprised twelve continuous predictors and two factor variables (see Table C.2 for a summary). We aggregated the land use in Flanders into eleven classes: agriculture, forest, semi-natural grassland, scrub, heathland, saltmarshes, wetlands, dunes, urban areas, water and other green areas (i.e. green areas outside the urban area that are not mapped as agricultural or natural land use) (Poelmans and Van Daele, 2014). The area of these classes in each 1x1 km cell was calculated and cells were removed if the cumulative area of land use was less than 50% of the total area (i.e. cells close to regional borders). We removed one class “agriculture” from the set because of the relatively high collinearity with other classes and because of the problem with perfect multicollinearity in compositional data (Aichison, 2003). The ten other land use classes were used to describe the variation in the extremely fragmented landscape in Flanders (Antrop, 2004). Two additional continuous predictors were the mean annual temperature and mean annual precipitation, BIO1 and BIO12 from WorldClim2 respectively (Fick and Hijmans, 2017). The first factor variable was a grid cell’s dominant soil texture class (Maréchal and Tavernier, 1974), a direct or indirect influencer of a species’ microclimate (Titeux et al., 2009). The second was ‘Ecoregion’ (Couvreur et al., 2004), which is a region with similar biotic and abiotic conditions. Since Flanders has limited geographical and environmental gradients (e.g. 240 km across, 0 to 288 m elevation and relatively uniform climatic conditions) and species use similar biotopes throughout the region, we assumed that the environmental response of a species was similar across the entire study area (Chen et al., 2020).

8.4. Model evaluation

Model calibration is incorporated in the Maxent algorithm (section 3.1.2). For evaluating model predictive performance, we chose three metrics: the Area Under the Receiver Operating Curve (AUC), sensitivity (i.e. true positive rate) and specificity (i.e. true negative rate) (Fielding and Bell, 1997), based on three rationales. First, using AUC alone as a summary metric of the ROC curve would lead to a loss of information about model performance (Jiménez-Valverde, 2012). Second, these metrics are measures of model discrimination and are independent of species prevalence which is unknown in presence-background situations (Lawson et al., 2014). Third, we evaluated our models on an external testing test that contained both presences and absences, enabling a reasonable calculation of the two threshold-dependent metrics (sensitivity and specificity) and justifying the use of these metrics for model evaluation (Jiménez-Valverde, 2012; Jiménez and Soberón, 2020). Sensitivity and specificity were calculated by transforming the continuous model predictions into a binary response. The threshold was set to the value that maximized the sum of sensitivity and specificity calculated on the species’ testing set, thereby minimizing misclassification errors (Kaivanto, 2008). The difference in model performance (Δ AUC, Δ sensitivity and Δ specificity) was used to evaluate the impact of data quality filtering. Four choices facilitated the comparison of evaluation metrics within one species (Elith et al., 2010; Lobo et al., 2008; Merow et al., 2013): (1) an identical testing set, (2) identical Maxent settings (features and regularization coefficient), (3) identical background selection and (4) identical predictors.

8.5. The impact of data quality on model performance

We repeatedly (20 times) selected a random sample from the unfiltered and filtered training sets, at six predefined levels of 100, 250, 500, 1000, 2000 and 4000 presences (also see Figure A.3). Model evaluation metrics were compared between training sets of constant fixed sample size but with different quality, resulting from the application of the different filters.

For the evaluation of data quality, species were divided into one of the six sample size levels. Species were classified at the highest level possible, based on the number of presences in the training set formed by the 3-filter combination ACTIVITY-DETAIL-VALSTAT (ADV). This way, all filters could be compared per species and sample size was kept as close as possible to the number of recorded presences in the database. This way we prevented large differences between the original and the filtered (fixed) sample size impacted model performance (Hanberry et al., 2012).

8.6. The impact of absolute sample size on model performance

For the evaluation of absolute sample size, we included models from different fixed sample sizes per species. We kept data quality constant by comparing results per filter and not between filters. Per filter, species were grouped in one out of six intervals of sample size that indicate the sample size of the original training sets: [100, 250[or [250, 500[or [500, 1000[or [2000, 4000[or ≥ 4000. Species were thus constant across absolute sample sizes but not across filters or intervals.

8.7. The combined impact of data quality and sample size on model performance

The impact on model performance of a change in data quality and a change in sample size will occur simultaneously. To evaluate this combined impact, we analysed 30,724 combinations of unfiltered and filtered training sets, with different changes in quality and sample size. We used all training sets of fixed sample size (at the six predefined levels) that we could obtain for each species, together with the original training sets, with sample size equal to the number of aggregated presences from the dataset.

Model performance was averaged across the 20 repetitions for the fixed sample sizes (i.e. per species, filter type and sample size level) and we looked at the mean differences in model performance (Δ AUC, Δ sensitivity, Δ specificity) between models of an unfiltered training set and the filtered training sets. To fully capture the impact of the change in sample size, we assessed two ‘sample size variables’: the remaining sample size after filtering and the proportional reduction in sample size. The latter is defined as the proportion of presences removed from an unfiltered training set by applying a single filter or a combination of filters. See Figure A.3 for an example of how many different datasets we could extract for one species and filter.

The combined impact of data quality and sample size on the difference in model performance was assessed using Generalized Additive Mixed Models (GAMMs) with species as a random effect, implemented in the ´mgcv` R package v1.8-31 (Wood, 2017). To account for the doubly-bounded character of our response variable, we rescaled Δ AUC, Δ sensitivity and Δ specificity to fall between 0 and 1 and used the ‘betareg’ family with logit-link. Smoothing functions were used to fit both sample size variables, with cubic spline method and k = 5 to reduce overfitting. We included interactions by allowing different smoothers per filter and by including the product of the remaining sample size and proportional reduction in the equation. Per taxonomic group, the model which best explained the difference in model performance while keeping model complexity low was selected, by comparing the Akaike’s Information Criterion (AIC) (Burnham et al., 2011) of multiple a priori GAMMs (full list in Appendix F) in the R package ´MuMIn` v1.43.17 (Barton, 2019). The relative importance of data quality (filter type) and sample size (sample size after filtering and proportional reduction) was assessed by comparing the proportion of explained deviance of those variables in the best model identified by our model selection.

We performed all analyses for the three evaluation metrics (AUC, sensitivity and specificity) across all species and within species groups and show the main results for AUC in the main text. All other results can be found in Appendices D through H. Models and statistical analyses were run in R v4.0.1 (R Core Team, 2021).

9. Results

Throughout the results section, the filters will be referred to as ACTIVITY (A): retaining records collected by observers with a high activity rate, DETAIL (D): retaining records that were submitted with information beyond the default date, location and species name, and VALSTAT (V): retaining records marked as ‘correct’ in the data platform’s validation system.

9.1. The impact of data quality on model performance

Figure 5 shows that for all species, filtered data could deliver higher AUCs than unfiltered data, but with differences among sample size levels. Smaller sample sizes of filtered data were more likely to result in higher AUCs compared to large sample sizes of filtered data. At 100 presences, all filters could result in a higher AUC, while at 250 and 500 presences VALSTAT and DETAIL could deliver positive results. For larger sample sizes, VALSTAT and its combinations (at 1000 presences) or no filters at all (at 2000 and 4000 presences) benefitted model performance.

Plants were most sensitive to data quality, where DETAIL and VALSTAT, and also ACTIVITY at 100 presences, resulted in higher AUCs throughout. Birds were sensitive to data quality at the low and intermediate sample sizes, where the best option was VALSTAT. At 500 and 1000 presences, VALSTAT alone already increased AUC. At 100 and 250 presences, VALSTAT had to be combined with at least one other filter. For butterflies, AUCs increased when using ACTIVITY: alone or in combination with one or two other filters at 4000 presences, or in combination with VALSTAT at 1000 presences. For dragonflies, single filters were not powerful enough to increase AUC. Combining DETAIL with VALSTAT at 500 presences or with ACTIVITY at 1000 presences did deliver higher AUCs.

Similar results to AUC were found for specificity, but mostly for plants at small sample sizes of 100 presences (all filters increased specificity) and 250 presences (DETAIL, VALSTAT, A+D and A+V increased specificity). At 500 presences, we noted increases in specificity for dragonflies (A+D and A+D+V) and decreases in specificity for plants (DETAIL, A+D and D+V). At larger sample sizes of 1000 presences or more, a higher specificity was found only for birds (filter combinations). Data quality did not impact specificity for butterflies (Figure D.2).

Results for sensitivity showed more negative impacts of using filtered data compared to AUC and specificity, yet also increases in sensitivity were noted for plants at 250 presences (DETAIL and its combinations) and 500 presences (all filters except ACTIVITY and A+V), and for butterflies at 4000 presences (ACTIVITY and its combinations). A lower sensitivity was found for plants at 100 presences (VALSTAT and its combinations), for dragonflies at 500 presences A+D and A+D+V) and for birds at 100 presences (A+D), 2000 presences (ACTIVITY and combinations with VALSTAT) and 4000 presences (DETAIL and its combinations) (Figure D.1).

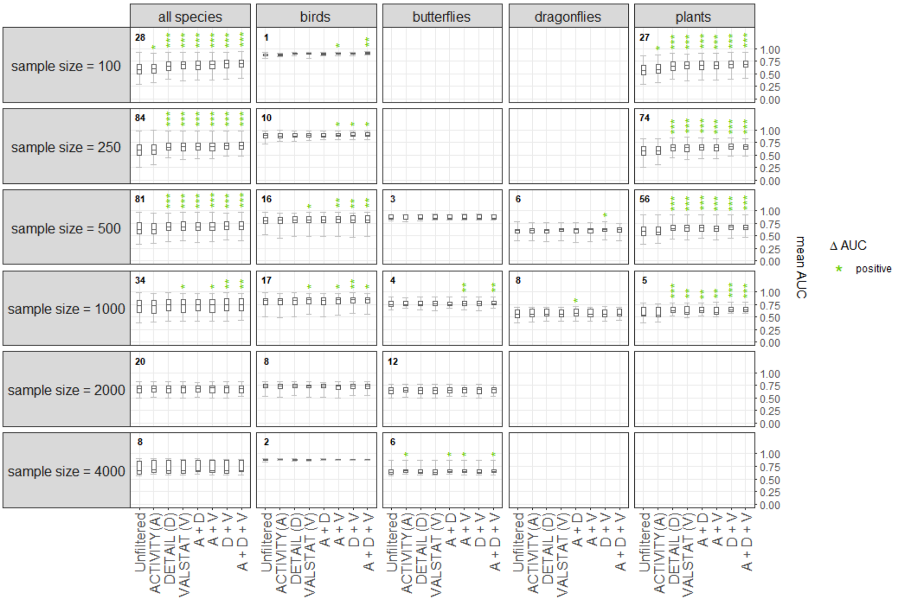


*Figure 5: The impact of data quality on AUC for all species and per taxonomic group, when absolute sample size is constant at six levels: 100, 250, 500, 1000, 2000 and 4000 presences. Per level, species were limited to those that could be modelled with all filters at the considered level, including the 3-filter combination ACTIVITY-DETAIL-VALSTAT. Species were subsequently classified at the highest level possible, meaning that AUC results cannot be compared between sample size levels, because species are different. The number of species in each comparison is presented in the top left corner of the graphic areas. Not all levels could be assessed for all taxonomic groups, because for example for butterflies there were no species with less than 500 presences in our dataset, so all species were classified at level 500 or higher. Boxplots represent medians, upper and lower quartiles with whiskers extending to the minimum and maximum values. Asterisks show significant differences in AUC compared to the unfiltered data, tested by a multiple comparison test with Benjamini & Hochberg (1995) correction (*** p<0.001, ** p<0.01, * p<0.05). Colours indicate only positive changes (green) for AUC. Results for the impact of data quality on sensitivity and specificity are found in Figure D.1 and D.2 respectively.*

9.2. The impact of absolute sample size on model performance

Figure 6 shows that reducing absolute sample size beyond a certain level always impacted AUCs negatively. This level depended more on the original sample size than on the applied filter. At lower original sample sizes (< 2000 presences), reducing sample size by 50% did not cause significant decreases in AUC for most filters, with exceptions for DETAIL, VALSTAT, A+D and A+V at 500 to 1000 presences. At larger original sample sizes (> 2000 presences), sample size could be reduced by 75% for most filters, with exceptions for VALSTAT and D+V at 2000 to 4000 presences. Reducing sample size to 100 presences, no matter what the original sample size was, always resulted in lower model performance. For birds and butterflies, the impact of sample size on AUC was similar to that of all species (Figures E.3 and E.4). Dragonfly and plant models appeared less sensitive to sample size (Figures E.5 and E.6).

Similar to AUC, the impact of smaller sample sizes on specificity was generally negative across all species with a higher tolerance for larger reductions when original sample sizes were high, yet with more variation among filters (Figure E.2). Specificity of butterfly and plant models (Figures E.12 and E.14) appeared more sensitive to smaller sample sizes compared to bird and dragonfly models (Figures E.11 and E.13).

In contrast with results for AUC and specificity, the impact of smaller sample sizes on sensitivity is generally positive. Significant increases in sensitivity were more likely to occur for higher quality data (filter combinations) at lower original sample sizes and for lower quality data (unfiltered data and single filters) at higher original sample sizes (Figure E.1). For butterflies, dragonflies and plants, sensitivity generally increased (Figures E.8, E.9 and E.10) when specificity decreased (Figures E.12, E.13 and E.14). For birds, this contrast was less pronounced and we even noted more decreases in sensitivity than increases when sample size was reduced (Figure E.7).

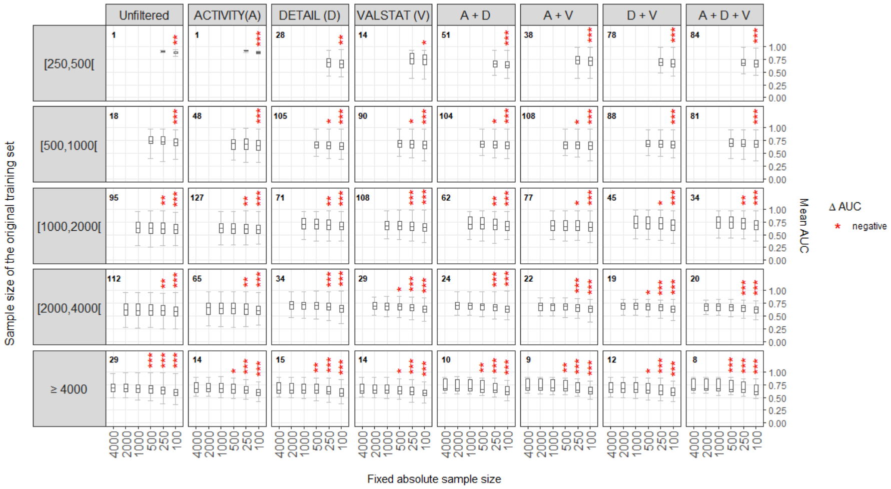


*Figure 6: The impact of absolute sample size on AUC for all species when data quality is constant. Per filter, species were grouped in one of the six specified intervals of sample size (left) that indicate the available sample sizes of the original training sets. AUCs were compared between models resulting from a repeated and random selection of different fixed sample sizes. Because species differ, results can only be compared within the graphic areas, i.e. between fixed sample sizes, but not between filters (horizontal) or intervals (vertical). The number of species in each comparison is presented in the top left corner of the graphic areas. Boxplots represent medians, upper and lower quartiles with whiskers extending to the minimum and maximum values. Asterisks show significant differences in AUC compared to the highest sample size, tested by a multiple comparison test with Benjamini & Hochberg (1995) correction (*** p<0.001, ** p<0.01, * p<0.05). Colours indicate only negative changes (red) for AUC (Δ AUC < 0). Results for the impact of absolute sample size on sensitivity and specificity are found in Figures E.1 and E.2 respectively.*

9.3. The combined impact of data quality and sample size on model performance

Up to this point, the absolute sample size of unfiltered and filtered data remained identical. In reality, however, sample size usually decreases when applying quality filters. Therefore, the impact of sample size was quantified with two variables in this section: the ‘proportional reduction in sample size’ and the ‘sample size after filtering’ (also called ‘remaining sample size’). A detailed summary per species of all the filters and their impact on model performance showed that model performance mostly increased after filtering (depending on the applied filter, for 55 to 80% of the species for AUC, 49 to 55% for sensitivity and 51 to 58% for specificity), but that various filter-species combinations also show a negative impact on model performance (Supplementary Information 2[7]).

Per taxonomic group, we selected the 'best' GAMM (Appendix F), i.e. the model with the least parameters and a small difference in AIC (Δ AIC < 1) compared to the top model, to evaluate the combined impact of data quality and sample size on the change in model performance caused by filtering. Figure 7 shows the relative importance of the variables in the GAMM for Δ AUC.

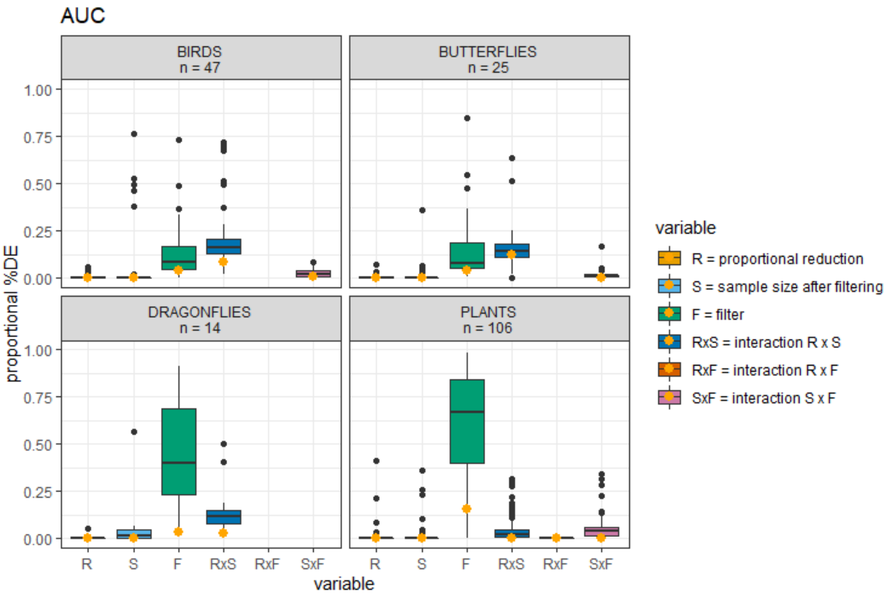


Figure 7: The relative variable importance for the impact of data quality and sample size on Δ AUC, based on the proportion of the percentage of deviance explained (%DE) by the different explanatory variables in the best GAMM (Generalized Additive Mixed Model) per taxonomic group (orange dots), and the relative variable importance across species, in the GAMs (Generalized Additive Models) where the random species effect was excluded (boxplots). The proportional %DE is the decrease in %DE between the full model and the model where the variable was excluded (but with identical smoothing parameters), relative to the %DE of the full model to summarize effects across n species. Species of which the full model could not be estimated due to convergence issues were excluded from the summary. The relative variable importance for the impact on Δ sensitivity and Δ specificity are found in Figures G.1 and G.2 respectively.

[7] File available at: https://www.sciencedirect.com/science/article/pii/S0304380021000260

Considering the averages across species (boxplots), the change in quality (the filter type) explained most of the variation in Δ AUC for plants and dragonflies, yet with high variability in percentage deviance explained (%DE) among species. The interaction between proportional reduction and sample size after filtering explained the most variation in Δ AUC for bird and butterfly models and is also important for dragonfly models. For plants, however, more variation in Δ AUC was explained by the interaction between quality and sample size after filtering. This interaction was also more important when considering the variation in Δ sensitivity and Δ specificity, and the differences between the proportional %DE for the variables 'filter', 'interaction RxS' and 'interaction SxF' became smaller. The filter type remained the most important variable for plants for predicting both Δ sensitivity and Δ specificity yet with less variability among species compared to AUC (Figures G.1 and G.2).

The predictions for Δ AUC of the best GAMM are presented in Figure 8, along a continuous scale of proportional reduction and for three sample sizes after filtering, that we chose based on data availability: 100, 500 and 1000 presences. Predictions for Δ Sensitivity and Δ Specificity are found in Appendix H. The combined impact of filtering varies among taxonomic groups and we find the highest impacts for plant models (AUC and Sensitivity) and dragonfly models (Sensitivity), with the largest differences in model performance among filters. The predictions for birds and plants in Figure 8 show that the best filters (i.e. the filters leading to increases in AUC) can differ between remaining sample sizes, confirmed by the relatively higher importance of the interaction between filter and sample size after filtering (Figure 7). For plants, for example, the best filter was A+D+V at small, but D+V at large remaining sample sizes. Similar patterns were detected for Sensitivity (birds, dragonflies and plants in Figures G.1 and H.1) and for Specificity (all groups in Figures G.2 and H.2). In general, filters that resulted in high-quality data usually increased model performance (Figures 1, D.1 and D.2). The proportional reduction in sample size could also be higher for those filters before a negative impact on model performance was detected.

Overall, filtering increased AUCs and sensitivity for plants (i.e. $\Delta > 0$) and decreased sensitivity for birds (i.e. $\Delta < 0$), while in other cases, both increases and decreases in model performance were noted. Different trends described the impact of proportional reduction on model performance. The shape of the trend depended on the remaining sample size, with different trend slopes for all taxonomic groups and even different trend directions for birds (sensitivity), butterflies (AUC), dragonflies (sensitivity and specificity) and plants (sensitivity).

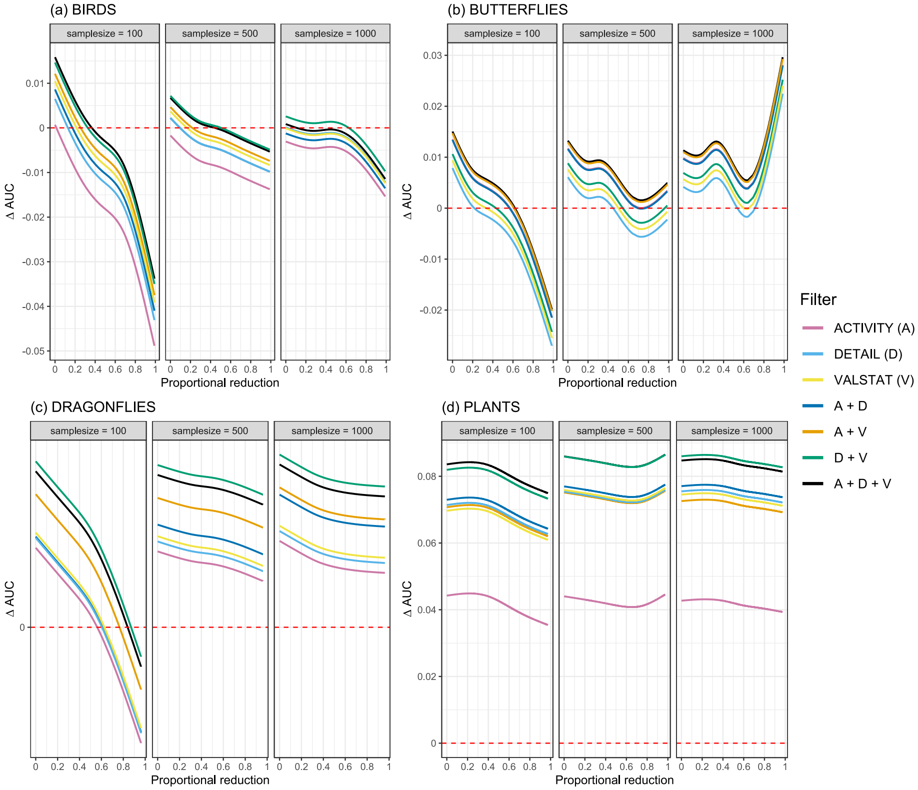


Figure 8: The combined impact of data quality and sample size on Δ AUC per taxonomic group. The full lines are the predictions for Δ AUC (AUC$_{\text{filtered data}}$ – AUC$_{\text{unfiltered data}}$) from the 'best' GAMM (Generalized Additive Mixed Model) along a continuous scale of proportional reduction in sample size and for three sample sizes after filtering that we chose based on data availability: 100, 500 and 1000 presences. Colours represent the different filters (data quality). The red dotted line equals a Δ AUC of 0, i.e. filtering did not impact model performance. We used the REML-method (restricted maximum likelihood) in the 'gam' function of the ´mgcv' R package v 1.8-31 (Wood, 2017) to model our data. Filter type was modelled as factor variable and species as random effect. Smoothing functions were used to fit both sample size variables (proportional reduction and sample size after filtering), with cubic spline method and k = 5. Δ AUC was rescaled to fall between 0 and 1, so that we could use the 'betareg' family with logit-link, because of the double-bounded character of the response variable (Δ AUC). The combined impact of data quality and sample size on Δ sensitivity and Δ specificity are shown in Figures H.1 and H.2 respectively.

For AUC and specificity, trends at small remaining sample sizes of 100 presences were negative, and filtering decreased model performance (i.e. Δ < 0) beyond a certain maximal threshold of proportional reduction. Depending on the filter, maximum reductions in sample size could range from 0-35% (AUC) for birds, 20-60% (AUC) or 10-30% (specificity) for butterflies, 55-85% (AUC) or 35-65% (specificity) for dragonflies and 5-85% (specificity) for plants. For sensitivity, trends at a remaining sample size of 100 presences were positive, except for birds. Depending on the filter, reductions had to be at least 0-10% for butterflies and 35-70% for dragonflies before an increase in model performance was noted.

For larger remaining sample sizes of 500 and 1000 presences, trends in the impact of proportional reduction on Δ AUC and Δ specificity remained negative for birds. For butterflies, trends for Δ AUC flattened with increasing sample size after filtering and Δ AUCs became largely positive, except for DETAIL, VALSTAT and D+V at reductions above 45%. We even saw a positive trend when reductions above 70% resulted in larger sample sizes of 1000 presences. For dragonflies, trends were flattened for AUC and specificity at larger remaining sample sizes and, except in the case of specificity and VALSTAT, model performance generally increased after filtering. Trends even became positive for specificity at larger remaining sample sizes of 1000 presences and reductions above 20%. For sensitivity, however, trends became more negative for dragonflies at higher remaining sample sizes and only VALSTAT, at 500 presences and reductions below 70%, led to increases in model performance.

10. Discussion

We applied three dichotomous filters to opportunistic species records of citizen scientists as single filters and in combinations to test their impact on species distribution model performance. We retained records from more active observers (ACTIVITY), detailed records, i.e. submitted with information beyond the default date, location and species name (DETAIL) and validated records, i.e. marked as ‘correct’ in the data platform’s validation system (VALSTAT). Results indicated that the impact of stringent filtering on model performance (measured by changes in AUC, sensitivity and specificity) depended on the quality of the filtered data, both the proportional reduction in sample size caused by filtering and the remaining absolute sample size, and the taxonomic group. To illustrate how filtering can impact relative occurrence maps, Appendix H shows model predictions based on the unfiltered data and three situations of reduced sample size when using the best filter (i.e. the filter that caused the largest positive difference in AUC). We did this for one species per taxonomic group, i.e. the species where the highest positive change in AUC was observed (Figures H.3 to H. 6).

A recurring pattern was that specificity results (true negative rates) generally agreed more with AUC results than sensitivity results (true positive rates). Moreover, specificity usually increased when sensitivity decreased and vice versa, which happens when evaluating model predictions on an external data set (Jiménez-Valverde, 2012). In the discussion that follows, we will focus on AUC results and we refer to the different results for specificity and sensitivity in the results section and Supplementary Information (Appendices D to H). The reader must keep in mind that the choice of an optimal threshold for threshold-dependent metrics depends on the characteristics of the SDM study (e.g. the goal of the study or the availability of information on species prevalence) (Jiménez-Valverde and Lobo, 2007) and that this choice might influence the recommendations for the most suited approach for quality filtering.

The quality of validated and detailed records was generally higher than the quality of records from more active observers. Luckily, validation of occurrence data entering large repositories, by synergies between human experts and computer intelligence, has been common practice (e.g. in eBird; Kelling et al., 2013). The main benefits for data quality of such an internal validation system are (i) the quick and relatively easy identification and correction of false-positive errors, as they can impact model performance negatively (Costa et al., 2015), and (ii) an increased observer skill by the interaction between data managers and users (Sullivan et al., 2009).

Metadata cannot only hold important information to improve SDMs by overcoming problems with imperfect detection (e.g. Kéry et al., 2009) or other types of systematic bias (e.g. Johnston et al., 2017), but our results also indicate that the very act of supplying additional information can benefit data quality. We, therefore, agree that observer dedication and effort (linked to DETAIL) are more fit measures of data quality than observer experience and recording rates (linked to ACTIVITY) (Henckel et al., 2020; Steen et al., 2019). Like in several other studies on data quality, it remains tough to detect changes in model performance due to observer-related measures of quality (e.g. observer skill and reporting consistency in Henckel et al. (2020) or observer expertise in Steen et al. (2019)). Combining multiple observer characteristics in observer profiles (Boakes et al., 2016; Isaac and Pocock, 2015) might be of added value here. Nonetheless, selecting data from active observers did significantly increase data quality for eight butterfly species that were among the most observed species in our dataset. We hypothesize that these common species are susceptible to misidentification by the inexperienced observer (Farmer et al., 2012), because of their highly familiar names in Dutch (*Aglais io L.*, *Gonepteryx rhamni L. and Vanessa atalanta L.*) or because they are hard to distinguish from congeners (*Pieris rapae L., Maniola jurtina L.* and *Pararge aegeria L.*) (Vantieghem et al., 2017).

When deciding whether or not to filter, it is not only important to consider the obtained data quality, but also both the proportional reduction in sample size and the remaining absolute sample size after filtering. Large reductions or small remaining sample sizes do not always cause lower model performance, and while we agree that small sample sizes generally lead to worse models (Jiménez-Valverde et al., 2009; Liu et al., 2019), the relative change in sample size must not be ignored (Hanberry et al., 2012). Both measures of sample size co-define which filters are suited for model performance improvement. They have a limited impact on the selection of the best or worst filters based on AUC results, as the relative impact on AUC of the different filters remained largely constant across different changes in sample size. However, here we must mention that when the goal is to increase sensitivity or specificity, the remaining sample size after filtering does need to be considered (Appendix G).

The different drivers of model performance make the interpretation complex but also highlight the importance of analysing multiple aspects of data manipulation together (Gábor et al., 2020). We add data quality to the list of drivers that can notably impact model performance, such as species characteristics, modelling technique and sample size (Gábor et al., 2020; Tessarolo et al., 2014). Compared to these factors, previous studies found marginal importance of the impact of sampling bias (Gábor et al., 2020; Tessarolo et al., 2014) and we have no reason to contest this finding based on our results (but note that we partially controlled for sampling bias by presence thinning (Kramer-Schadt et al., 2013)). Disentangling the different drivers of model performance in stringent filtering could be more feasible in a virtual species setting (Hirzel et al., 2001; Meynard et al., 2019), however, we argue that the simulation of filtered data of different quality is not trivial. This would require a more profound understanding of how data quality is impacted by data and species characteristics.

We can recommend stringent filtering for taxonomic groups where model performance is more impacted by data quality and less by sample size, such as the plants and dragonflies in this chapter. For plant models, we even observed that an increase in quality can mitigate the negative impact on AUC of reducing sample size to 100 presences (Figures 8 and E.6). For the other taxonomic groups, this is only true below certain proportional reductions. Models from species with specific habitat conditions, such as dragonflies, are less sensitive to sample size and also profit from data quality increase. Such species have a more distinct link with their habitat and are easier to model compared to species with a broader niche (Hernandez et al., 2006). On the other hand, mobile species that have large home ranges, such as most studied birds and butterflies, are more difficult to model. It might be that this issue cannot be resolved by data quality filtering alone. Nevertheless, caution is needed, because the impact of data quality on model performance shows large variation among plant and dragonfly species (Figure 7) and is different when considering other evaluation metrics (Appendix E).

For taxonomic groups where model performance is more impacted by sample size and less by data quality, such as the birds and butterflies in this chapter, we advise being more careful. We observed that filtering is less beneficial for these groups, probably because their abundant data already leads to relatively high model performance. Especially for birds, unfiltered data appeared very suited for modelling and filtering did not improve AUCs, certainly when less than 50% of the sample size remained. For these groups, even filters that do not cause large reductions nor lead to a small sample size could cause model performance to decrease. Nonetheless, choosing the right filter can mitigate the negative impact of sample size if the obtained quality is high enough (e.g. extracting data from active observers for butterflies or combining validated and detailed records for birds).

In this chapter, we focussed on the combined impact of data quality and sample size in stringent filtering, but we acknowledge that other factors, such as environmental filtering (Gabor et al., 2019), scale (Connor et al., 2017; Gottschalk et al., 2011), species traits (Hernandez et al., 2006; McPherson and Jetz, 2007) and SDM technique (Liu et al., 2019) will probably impact the sensitivity of a dataset to stringent filtering as well. For example, the proportion of high-quality data in a model training set is scale-dependent, because a coarse resolution gives a higher chance that at least one high-quality observation falls in a grid cell. Presence thinning is therefore not only a way to remove spatial bias (Boria et al., 2014) but also to reduce other sources of uncertainty (Kramer-Schadt et al., 2013), such as the presence of data with uncertain quality. We also detected variation among species, and as taxonomic groups still show plenty of variation in species traits (Maes et al., 2019a), it might be more efficient to formulate recommendations for stringent filtering based on species traits rather than on taxonomy. Species prone to misidentification, for example, can benefit from retaining only records validated as correct based on photos supplied by the observer (Vantieghem et al., 2017) and we have indications that, for example, habitat-specificity, mobility and popularity impact the sensitivity of a species to data quality filtering as well. That aspect of data quality filtering will be further investigated in the next chapter. Our recommendations are limited to the discrimination accuracy of Maxent. As Maxent usually comes out as a relatively more robust SDM technique (Thibaud et al., 2014), our conclusions are likely to be conservative. We, therefore, expect at least a similar, if not a larger, impact of data quality filtering for other SDM techniques.

11. Conclusions

We conclude that data quality filtering has the potential to improve predictions of species distributions, especially for species where SDMs are less sensitive to decreases in sample sizes. However, data quality should not be pursued at any cost, because filtering can also impact model performance negatively, e.g. for species with abundant data or when filtering leads to low sample sizes or causes high sample size reductions. We encourage the further development and adoption of techniques that can increase the availability of high-quality data, to be able to fully profit from the benefits of opportunistic citizen science data. The value of a database-integrated validation system demonstrates the potential of bulky datasets from platforms and applications where the focus is on the identification and validation of species observations, such as iNaturalist (https://www.inaturalist.org/), Pl@ntnet (https://www.plantnet.org) or ObsIdentify (Hogeweg et al., 2019). We advise to always ‘Think before you shrink’ because volunteer-generated data can make valuable contributions to science if processed correctly.

CHAPTER III. Insights Into the Drivers of Quality with Species Profiles

Adapted from

Van Eupen, C., Maes, D., Herremans, M., Swinnen, K.R.R., Somers, B., Luca, S., 2022. Species profiles support recommendations for quality filtering of opportunistic citizen science data. Ecol. Modell. 467. https://doi.org/10.1016/J.ECOLMODEL.2022.109910

Author contributions

Camille Van Eupen: conceptualization, methodology, software, validation, formal analysis, writing – original draft, visualization; **Dirk Maes:** conceptualization, writing – review & editing, supervision; **Marc Herremans:** conceptualization, data curation, writing – review & editing; **Kristijn Swinnen:** data curation, writing – review & editing; **Ben Somers:** conceptualization, writing – review & editing, supervision; **Stijn Luca:** conceptualization, writing – review & editing, supervision

ABSTRACT

Opportunistic citizen science data are commonly filtered in an attempt to improve their applicability for relating species occurrences with environmental variables. Recommendations on when and how to filter, however, have remained relatively general and associations between species traits and filtering recommendations are sparse. We collected six traits (body size, detectability, classification error rate, familiarity, reporting probability and range size) of 52 birds, 25 butterflies and 14 dragonflies. Both absolute (values not rescaled) and relative traits (values rescaled per taxonomic group) were linked to filter effects, i.e. the impact on three different measures of species distribution model performance caused by applying three different quality filters, for different degrees of sample size reduction. First, we applied multiple regressions that predicted the filter effects by either absolute (including taxonomic group) or relative traits. Second, a principal component and clustering analysis was performed to define five species profiles based on species traits that were retained after a multiple regression model selection. The analysis of the profiles indicated the relative importance of species traits and revealed new insights into the association of species traits with changes in model performance after data quality filtering. Both taxonomic group (more than absolute traits) and relative species traits (mainly classification error rate, range size and familiarity) defined the impact of data quality filtering on model performance and we discourage the selection of a quality filtering strategy based on one single species trait. Results further confirmed the importance of considering the goal of the study (i.e. increasing model discrimination capacity, sensitivity or specificity) as well as the change in sample size caused by stringent filtering. The general species knowledge among citizen scientists (importance of observer experience), together with the mechanism of record verification in an opportunistic data platform (importance of verifiable metadata) have the largest potential for enhancing the quality of opportunistic records.

12. Introduction

Chapter II highlighted the importance of considering both the type of filter and the resulting change in sample size, yet variation among species in their response to data quality filtering remained large. Grouping species according to their taxonomy revealed that filtering benefitted some groups (i.e. plants and dragonflies) more than others (i.e. butterflies and birds). In this chapter, we aim to verify whether grouping species according to a-priori-selected life history and/or ecological traits could better substantiate recommendations for data quality filtering.

Species traits have been linked extensively to SDM performance and those that cause the most variation can usually (but not exhaustively) be compiled into the following three: (1) traits that define the species-environment relationship (e.g. range size, niche breadth (Brotons et al., 2007; Stockwell and Peterson, 2002) and habitat association (Chefaoui et al., 2011)), (2) traits that impact the detectability of the species in space and time (e.g. conspicuousness (Seoane et al., 2005), migratory behaviour (Carrascal et al., 2006) and lifespan (Hanspach et al., 2010)), and (3) traits that influence the proneness to misidentification (e.g. phylogenetic relatedness (Vantieghem et al., 2017)).

Notwithstanding the vast amount of proof of the link between species traits and absolute SDM performance, few studies have successfully linked species traits to the change in SDM performance caused by stringent filtering of species occurrence records (but see e.g. Steen et al., 2019, where models of more restricted species performed better when using data collected with lower effort). This could be due to the higher quality of the unfiltered data in most of these studies (e.g. semi-structured data in Steen et al. (2019)) or due to the conflicting character of the simultaneous impact of data quality filtering, i.e. an increase in data quality and a decrease in sample size (Chapter II). By assessing this twofold effect on an extensive dataset of opportunistic records, *waarnemingen.be,* we will aid the optimisation of the data cleansing process that is essential for high-quality SDMs (Zurell et al., 2020).

13. Materials and methods

13.1. Species data and impact of quality filtering

We used the results of the analysis in Chapter II (see sections 8.1 and 8.2 and Appendix A for a description of the data quality filters and data selection). The three filters were: '**ACTIVITY**', based on an observer's average annual activity rate, where the filter consists in removing records from less active observers; '**DETAIL**', based on the presence of metadata beyond default requirements (i.e. species name, location, date and observer id), where the filter consists in removing records that were submitted without any additional information (e.g. sex, count, behaviour); and '**VALSTAT**', based on the validation status of a record in the data platform, where the filter consists in removing doubtful and unevaluated records (Table 1). These are all records that could not be verified by species experts because key information was missing or because the record was not assessed yet by an expert at the moment the dataset was extracted.

For the analysis in this study, we extracted the change in model performance (i.e. Δ AUC, Δ sensitivity and Δ specificity) (Table 1), after using the three single filters (ACTIVITY, DETAIL and VALSTAT) for 52 birds, 25 butterflies and 14 dragonflies. Plant observations were not used in the present analysis because their traits are not directly comparable to animal species traits. For a summary per species of the data used for model testing and model training (unfiltered and filtered data) and of the impact on model performance, we refer to Table C.1 and Supplementary Information 2[8] respectively.

13.1. Species traits

We used six species traits that can be related to data quality in opportunistic citizen science data based on literature review and expert opinion: body size, detectability, classification error rate, familiarity, reporting probability and range size (Table 1). Abundance was not considered because the largely unstructured *waarnemingen.be* database contains unreliable count data that are mostly without a clear reference to time and space. All trait values can be found in Table I.1.

Body size equals the wing length for birds (Storchová and Hořák, 2018) and butterflies (Bink, 1992) and head-to-tail length for dragonflies (https://www.vlinderstichting.nl/libellen/).

The **classification error rate** reflects how likely it is for an average observer to wrongly identify a species. This was quantified by the number of erroneous photo records (i.e. observations accompanied by a photograph) of a species in the *waarnemingen.be* data portal, relative to its total number of photo records. The portal keeps track of changes in the identification of a species, and we considered only the changes at the species level as erroneous (and for example not the changes from family or genus to species level). Auto-corrections made by the observer were excluded.

[8] File available at: https://www.sciencedirect.com/science/article/pii/S0304380021000260

Table 1: Overview and definitions of the used variables.

Data quality filters		
	description:	*based on:*
ACTIVITY	removes records from less active observers	an observer's average annual activity rate
DETAIL	removes records that were submitted without any additional information	the presence of metadata beyond default requirements
VALSTAT	removes doubtful and unevaluated records	the validation status of a record in the data platform
Species traits		
	description:	*source:*
Body size	wing length (birds and butterflies) or head-to-tail length (dragonflies)	Bink (1992); Storchová and Hořák (2018), https://www.vlinderstichting.nl/libellen/
Classification error rate	the number of erroneous photo records (i.e. observations accompanied by a photograph) relative to the total number of photo records.	the *waarnemingen.be* data portal during the study period
Detectability	the probability of detecting a species on the condition that it is present	quantified by applying site occupancy models to complete checklist data, retrieved from the *waarnemingen.be* data portal
Familiarity	reflects how well-known a species is by the average observer	the number of Belgian websites (searched on Google) with the Dutch name of the species in the title (Żmihorski et al., 2013)
Reporting probability	the likelihood that a species is reported by an average observer, on the condition that it is present and that the taxonomic group it belongs to is surveyed	a species' relative (per taxonomic group) average reporting rate divided by its detectability, retrieved from the *waarnemingen.be* data portal
Range size	the distribution range size	the total number of grid cells (km²) in which a species has been recorded during the study period, retrieved from the *waarnemingen.be* data portal
	description:	
Absolute traits	unscaled trait values as retrieved by the different methods described	
Relative traits	scaled trait values; using the following transformation per taxonomic group: $y = \frac{\mathrm{x} - \min(\mathrm{x})}{\max(x) - \min(x)}$	
Impact on model performance		
Δ AUC	change in the area under the receiver operating characteristic	
Δ sensitivity	change in the true positive rate (TPR) after data quality filtering $TPR = \frac{true\ positives}{true\ positives + false\ negatives}$	
Δ specificity	change in the true negative rate (TNR) after data quality filtering $TNR = \frac{true\ negatives}{true\ negatives + false\ positives}$	
Filter effects		
All combinations of data quality filters and impact on model performance	the impact of data quality filtering (Δ AUC, Δ sensitivity and Δ specificity) by the three filters ACTIVITY, DETAIL and VALSTAT	
Sample size situations		
actual reduction	the actual reduction in the number of presences after data quality filtering	
50% reduction	a relative reduction in the number of presences after data quality filtering of more than 50%	
reduction to 100 presences	a reduction to 100 presences after data quality filtering	

Detectability is the probability of detecting a species on the condition that it is present (MacKenzie et al., 2017). Species detectability was retrieved from applying site occupancy models to complete semi-structured checklist data extracted from *waarnemingen.be,* following Johnston et al. (2021). Detection histories consisted of five to ten repeated visits to a specific site (a 1 km grid cell) by the same observer in a period of closure (i.e. a period with no supposed changes in occupancy). A period of closure was defined as 20 consecutive days in the peak active season of a species. The peak active season was defined as every 10 days with an observation count above the average count of all observations in a year, excluding egg, larva, pupa and caterpillar observations. Covariates used to describe the detection process were: checklist duration (in minutes), starting time of the checklist, search effort (i.e. the number of species recorded at a specific location, based on the principle of species accumulation curves (Colwell et al., 2004)), and open habitat (grasslands, wetland, marshes and water) versus closed habitat (forest and woodland), because of an increased detectability (visually and, for birds, also auditory) in open habitat types (Johnston et al., 2014; Morton, 1975). Detection probabilities were predicted for all grids with covariate values and averaged to attain one value per species.

Familiarity refers to how well-known a species is by the average observer and was quantified by the number of Belgian websites with the Dutch name of the species in the title, retrieved from the Google search engine (Żmihorski et al., 2013). We added two extra search terms that specified the taxonomic group (in Dutch) and excluded the *waarnemingen.be* website to avoid counting individual observations on the used data platform, e.g. *"Bruinrode Heidelibel" site:.be libel -waarnemingen.be.* An Incognito window was used to unlink search results from the used Google account.

Range size is the distribution range size of the species during the entire study period 2014-2019 in the study area and was quantified as the total number of grid cells (km²) in which a species has been recorded (McPherson et al., 2004).

Reporting probability is the likelihood that a species is reported by an average observer, on the condition that it is present and that the taxonomic group it belongs to is surveyed. To meet these requirements, we looked at the peak of the active season and calculated the relative number of species observations to the number of observations for a taxonomic group. This was averaged across locations and observers. We subsequently divided this number by the average detectability across locations where the species was present to correct for the impact of detectability on reporting rate.

13.2. The impact of data quality filtering

To build recommendations for data quality filtering based on species traits, we first analysed the multivariate relationship between species traits and the filter effects. Consequently, species were grouped in species profiles characterised by the most highly associated traits to assess if such groups presented a similar response to data quality filtering. By filter effect, we mean the impact of data quality filtering by the three filters ACTIVITY (only observations from active observers), DETAIL (only detailed observations) and VALSTAT (only approved observations) on three evaluation metrics: AUC, sensitivity and specificity. All analyses were conducted in R (R Core Team, 2021).

13.2.1. Multi-trait analysis

Relationships between species traits and filter effects were examined using multiple (multi-trait) regressions. The data were modelled in beta-regressions (*betareg* package v3.1-4, Cribari-Neto & Zeileis, 2010), because of the bound character of the response variable (Δ AUC, Δ sensitivity and Δ specificity theoretically range from -1 to 1). Filter effect values were rescaled to fall between 0 and 1 with the following transformation: $y = \frac{\mathrm{x-min(x)}}{\mathrm{max}(x) - \mathrm{min}\ (x)}$. To reduce the impact of outliers, data points with a cook's distance of more than four times the mean cook's distance of all data points were removed (Ferrari and Cribari-Neto, 2004). As trait values showed taxonomic differences (Figure 9), continuous values (absolute traits) were rescaled per taxonomic group (relative traits), using the aforementioned transformation (Table 1). The relative values can be informative for patterns across taxonomic groups that would go unnoticed otherwise (e.g. birds are always larger than butterflies, but similar impacts from filtering might be observed for large birds as well as large butterflies).

First, multi-trait regressions were performed using the log-transformed absolute trait values as continuous variables and the taxonomic group as a factor variable. Second, relative traits were regressed against the filter effects. Trait values were standardised and multicollinearity was reduced by retaining only those variables with a Variance Inflation Factor (VIF) below 5 (Menard, 2001). We modelled the absolute and relative traits separately because of high pairwise correlations among most of these variables (Figure I.1). We also quantified variable importance by leaving out each trait one by one and calculating the decrease in pseudo-R^2 compared to the full model. Finally, we performed a model selection based on three conditions to obtain parsimonious models for each filter effect: (1) the increase in the Akaike's Information Criterion (AIC) had to be smaller than (a conservative) five (Burnham et al., 2011), (2) the model should at least contain the most important variable and (3) the simplest model was selected (i.e. the model with the least parameters).

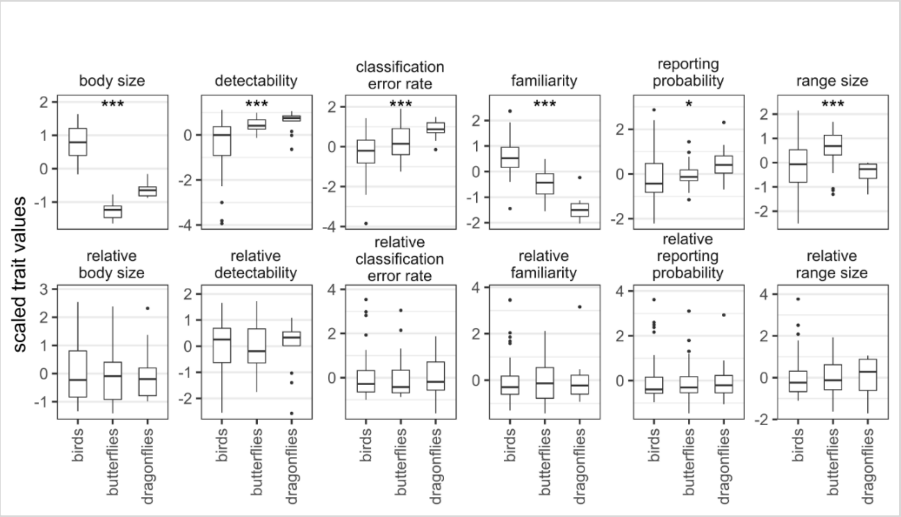


*Figure 9: Summary of the species traits per taxonomic group after value transformation and standardisation. Absolute traits (top row) were rescaled to relative traits (bottom row) per taxonomic group to assess patterns across taxonomic groups. Stars indicate differences in the medians of the trait values between taxonomic groups (*** = p < 0.001, * = p < 0.05).*

13.2.2. Species profiles

To test whether species with similar traits can improve recommendations for data quality filtering, we clustered species into groups with similar traits using the ‘FactoMiner’ package v1.34 (Le et al., 2008). The package allows using a principal component analysis (PCA) as a pre-processing method for hierarchical clustering (Husson et al., 2010). The PCA was performed on the active variables, i.e. the species traits that contributed most to the change in model performance across filters, resulting from the multi-trait regression model selection. All species can be described by the resulting principal components (PCs). All PCs, or a selection of PCs that explain the most variance in the active variables, can consequently be used in an agglomerative hierarchical clustering. The resulting clusters were defined as species profiles. Supplementary variables were added to characterise the clusters further, without impacting the clustering itself, and comprised: the remaining traits and the impact of filtering on model performance per filter (quantitative), and the taxonomic group (qualitative).

We delineated the profiles based on four conditions. First, traits used were those remaining after model selection (see section 13.3.1) in at least one multi-trait regression. Second, the number of clusters was chosen based on the increase of inertia between two consecutive aggregation steps in the hierarchical tree (Husson et al., 2010). Third, profiles were ideally associated with one or more distinctive filter effects: (1) an increase in AUC, (2) a decrease in AUC, (3) an increase in sensitivity and/or decrease in specificity or (4) an increase in specificity and/or decrease in sensitivity. Fourth, profiles should be ecologically meaningful, where we relied on species experts to evaluate the profiles' species composition. To this end, we built the final profiles by experimentally excluding the PCs that explained the lowest percentages of the total variance and by choosing different heights in the hierarchical tree to change the number of clusters.

13.2.3. Impact of sample size

The role of sample size in the relationship between species traits and filter effects was assessed by adding two sample size situations based on previous recommendations in Chapter II, where filtering was not advised when sample size was reduced by more than 50% or when the resulting sample size was 100 presences. We looked at (1) the actual reduction, i.e. the filter effects when sample size after filtering was not altered (sample size was reduced by an amount that depended on the applied filter), (2) the 50 % reduction, i.e. the filter effects when sample size was reduced by 50% or more and (3) the reduction to 100 presences, i.e. the filter effects when sample size was reduced to 100 presences. Adding these two situations could aid interpretation and simulate situations occurring in datasets of lower quality (i.e. where fewer presences are kept after stringent filtering).

14. Results

14.1. Multi-trait analysis

Figure 10 shows that the relative importance of the different traits in their association with the filter effects varies among filters and model evaluation metrics. Considering the absolute traits (Figure 10a), the taxonomic group was the most important variable in five out of nine cases. When the goal was to increase AUC, it was best to use data from active observers or approved observations for butterflies and dragonflies, or detailed observations for dragonflies. Specificity could be increased for dragonflies by using detailed observations. No other significant differences ($p < 0.05$) between taxonomic groups were detected in the multiple regressions. AUC of models from large-bodied species could best be increased by using data from active observers, AUC of models from species with a low error rate by using detailed observations and AUC of models from species with a restricted range size by using approved observations. Sensitivity of models from unfamiliar species benefitted from using detailed observations, as did specificity of models from species with restricted range sizes. Specificity of models from small-bodied species benefitted from using approved observations.

Considering the relative traits (Figure 10b), using observations from active observers worked best for large-bodied species (to increase AUC), for species with low reporting probability (to increase sensitivity) or for species with high reporting probability (to increase specificity). Using detailed observations most benefitted familiar species (to increase AUC), species with high reporting probability (to increase sensitivity) or species with a restricted range size (to increase specificity). Using approved observations was most valuable for species with a restricted range size (to increase AUC), for species with a low classification error rate (to increase sensitivity) or for species with a high classification error rate (to increase specificity).

Multicollinearity among absolute and relative traits was negligible (VIF < 5), so all traits were included in the multi-trait regressions. Neither absolute nor relative detectability was retained after model selection as these traits explained less variation compared to others. We did observe that detectability was negatively correlated with familiarity ($r = -0.38$, $p < 0.001$), which can be explained by the taxonomic differences found in both traits (Figure 9 and Figure I.1). Detectability was also negatively correlated with reporting probability ($r = -0.62$ and $r = -0.65$ for absolute and relative traits respectively, $p < 0.001$), which can be explained by their inverse dependence (Figure I.1 and section 13.2).

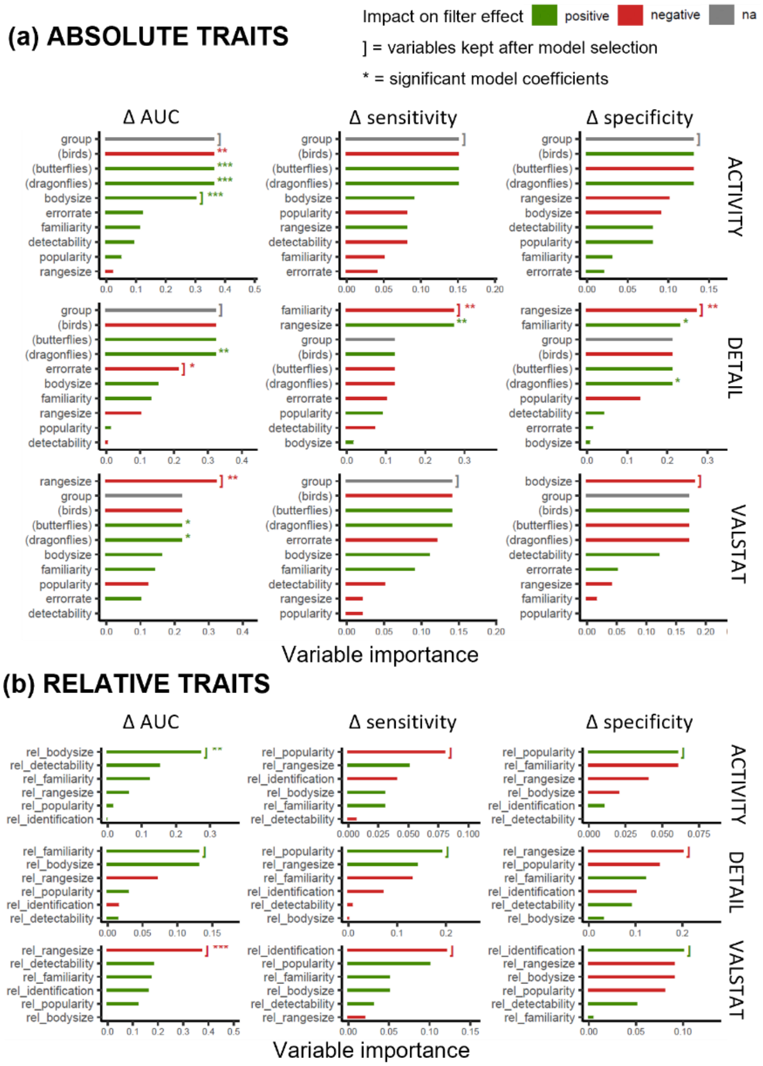


*Figure 10: Variable importance in the multi-trait regressions for absolute (a) and relative (b) species traits per filter (ACTIVITY, DETAIL and VALSTAT) and change in model evaluation metric (Δ AUC, Δ sensitivity, Δ specificity). Variable importance is expressed as the square root of the change in pseudo-R^2 when leaving out one variable at a time from the full model. Colours indicate a positive (green) or negative (red) impact of the trait on the filter effect, factor variables have a grey (n/a) colour. Square brackets indicate the variables kept after model selection (i.e. the simplest model with an increase in the Akaike's Information Criterion (AIC) of less than 5 compared to the best model where at least the most important variable was included). Asterisks indicate significant model coefficients (*** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$).*

14.2. Species profiles

The active variables that we used in the PCA and clustering analysis were the continuous variables kept after model selection in the multi-trait analysis: i.e. body size, relative body size, classification error rate, relative classification error rate, familiarity, relative familiarity, relative reporting probability, range size and relative range size. In the experimental phase, three clusters (i.e. profiles) based on six PCA dimensions provided the best separation of species and captured 96 % of the variation in species traits. However, the impact on model performance still showed large variations within profiles. Ecologically, these three profiles also separated species into quite general groups and we cross-checked the inclusion of more dimensions and the clustering of the species into four or more profiles with species experts. Eventually, we selected all PCA dimensions and five clusters appeared as the best outcome while keeping cluster size at a reasonable level (minimum cluster size equalled 7 species) (Table 2). Note that using all dimensions for clustering is similar to performing a clustering without the pre-processing PCA step (Husson et al., 2010). However, the PCA and clustering method prosed by Husson et al. (2010) enriches the descriptive analysis and gives a framework for data visualization. We found that the package was extremely useful to characterize groups of individuals using multiple variables in the case where some variables (i.e. the active variables) are more important than others (i.e. the supplementary variables). The results of the PCA and clustering analysis with the resulting five clusters are presented in Appendix J (Table J.1 and Figure J.1).

*Table 2: Recommendations for data quality filtering for the five species profiles, described by five relative traits (body size, classification error rate, familiarity, reporting probability and range size) and four absolute traits (body size, classification error rate, familiarity and range size). Recommendations are positive (green – all values in the 90% confidence interval are positive), cautious (blue - the average filter effect is positive but the 90% confidence interval also includes negative values), alarming (orange - the average filter effect is negative but the 90% confidence interval also includes positive values) or negative (red - all values in the 90% confidence interval are negative). The taxonomic distribution of the species is given, as well as the most characterising species per profile (in bold are the species closest to the cluster centre and in italic are the species furthest away from the other cluster centres). The asterisks indicate the significance level at which traits, filter effects or taxonomic groups are associated with a profile (*** = .001, ** = 0.01, * = 0.05). For taxonomic groups, (+) and (-) indicate whether the group is significantly more or less represented in a profile.*

	PROFILE 1	**PROFILE 2**	**PROFILE 3**	**PROFILE 4**	**PROFILE 5**
Relative Traits	High error rate *** Widespread ***	Small body size *** Restricted range size ** Low error rate * Unfamiliar *	Large body size *** Restricted range size ** Unfamiliar *	Familiar *** Widespread *** Large body size ** Low error rate *	Familiar *** High reporting probability *** Low error rate *
AUC recommendations	**ACTIVITY ** > VALSTAT > DETAIL**	**VALSTAT > DETAIL** **ACTIVITY ** **	**VALSTAT** **DETAIL** **ACTIVITY**	**ACTIVITY ** > DETAIL** **VALSTAT**	**VALSTAT > DETAIL** **ACTIVITY * **
sensitivity recommendations	**ACTIVITY < VALSTAT < DETAIL**	**VALSTAT** **ACTIVITY < DETAIL**	**DETAIL > VALSTAT > ACTIVITY**	**ACTIVITY > DETAIL > VALSTAT**	**DETAIL > VALSTAT** **ACTIVITY**
specificity recommendations	**ACTIVITY > VALSTAT > DETAIL**	**ACTIVITY > DETAIL** **VALSTAT**	**DETAIL < VALSTAT < ACTIVITY**	**DETAIL** **VALSTAT < ACTIVITY**	**ACTIVITY** **DETAIL < VALSTAT**
Taxonomic group	20 species 4 birds *** (-) 6 butterflies 10 dragonflies *** (+)	35 species 20 birds 11 butterfly 4 dragonflies	17 species 17 birds *** (+)	12 species 4 birds 8 butterflies ** (+)	7 species 7 birds * (+)
Absolute traits	High error rate *** Unfamiliar **	Restricted range size *** Small body size *** Low error rate *	Large body size *** Low error rate *	Widespread *** Low error rate *	Familiar *** Low error rate *
Characterising species	**Pieris napi** ***Sympetrum striolatum*** **Sympetrum sanguineum** **Aeshna cyanea** **Maniola jurtina** *Pieris brassicae* *Pieris rapae* *Enallagma cyathigerum* *Larus canus*	**Oenanthe oenanthe** **Turdus pilaris** **Tachybaptus ruficollis** **Delichon urbicum** **Rallus aquaticus** *Platycnemis pennipes* *Colias crocea* *Calopteryx splendens* *Pyrrhosoma nymphula* *Motacilla alba*	**Tadorna tadorna** **Circus aeruginosus** **Numenius arquata** **Egretta garzetta** **Branta leucopsis** *Cygnus olor* *Branta canadensis* *Anser anser* *Ardea alba* *Corvus frugilegus*	**Vanessa cardui** **Polygonia c.album** ***Gonepteryx rhamni*** ***Vanessa atalanta*** **Falco tinnunculus** *Buteo buteo* *Aglais io* *Papilio machaon*	***Cuculus canorus*** ***Alcedo atthis*** **Perdix perdix** **Carduelis carduelis** ***Athene noctua*** *Ciconia ciconia* *Alopochen aegyptiaca*

Positive (green) or negative (red) recommendations were only noted when the goal was to increase AUC. When the goal was to increase sensitivity or specificity, recommendations were either cautious (blue) or alarming (orange) and in most cases, filter recommendations for increasing sensitivity and specificity were opposite to each other. Similar impacts between profiles on one evaluation metric might have a different impact on other metrics (e.g. similar impact on AUC but a different impact on sensitivity and specificity in profiles 1 and 4).

Absolute traits appeared highly associated with the most represented taxonomic group for four out of five profiles (Table 2 and Figure 9). Profile 1 contained more dragonflies, indeed species with a higher classification error rate that are less familiar. Profile 4 contained more butterflies, species with a large range size, yet not necessarily a lower error rate. Profile 3 contained birds only, which have larger body sizes and lower error rates. There is a difference with profile 5 though, also containing only birds, where higher familiarity and lower error rates are characterising traits. Profile 2 is not associated with one of the three taxonomic groups, but species in this profile are mostly small, with a restricted range size and a lower error rate, which are also relative traits that characterise this profile.

Recommendations based on relative traits were mostly similar to the results in the multi-trait analysis, with a few exceptions (Table 2 and Figure 10). Model AUC for large species increased when using observations from active observers, confirmed by negative and positive recommendations in profiles 2 and 4 respectively. In profile 3, however, body size seemed subordinate to the taxonomic group. Higher reporting probability was associated with a higher Δ sensitivity when using detailed observations and with a lower Δ sensitivity when using observations from active observers, confirmed in profile 5. Familiarity had a positive impact on Δ AUC when using detailed observations (DETAIL), confirmed by positive and cautious recommendations in profiles 4 and 5, yet only as the second-best option. Using DETAIL did not necessarily worsen model AUC for unfamiliar species (profiles 2 and 3), but not all species in these profiles were unfamiliar (indicated by the weak significance level). For using only approved observations, range size was a good indicator of a change in AUC (profiles 2, 3 and 4), except for the widespread species in profile 1 where range size seemed to be subordinate to error rate. For using only detailed observations, however, range size did not seem to drive filter recommendations when the goal was to increase model specificity, except for the cautious recommendation for species with a restricted range size in profile 2. Finally, the association between error rate and model sensitivity and specificity supported filter recommendations when using only approved observations (profiles 1, 2, 4 and 5).

14.3. Impact of sample size

Reducing sample size further (beyond the already occurring decrease in sample size after data quality filtering) impacted filter effects both positively and negatively (Figure 11). The impact on AUC was generally negative, but the impact on sensitivity and specificity could in a few cases also be positive. It, therefore, depends on the goal of the study whether reducing sample size further might have a desirable effect. When predicting suitable presence locations is the main interest, like for the delineation of conservation areas, a high sensitivity is desired (few omission errors, i.e. predicting absences when the species is actually present) (Lobo et al., 2008; Thomaes et al., 2008). However, for assisted monitoring, like for invasive species, a high specificity is desired (few commission errors, i.e. predicting a presence when the species is actually absent) (Guillera-Arroita et al., 2015). Note that there is usually a trade-off between sensitivity and specificity (when sensitivity increases, specificity usually decreases and the other way around) (Jiménez-Valverde, 2012). Variability in the impact on model performance also increased with decreasing sample size, except for species with a restricted range size (profiles 2 and 3).

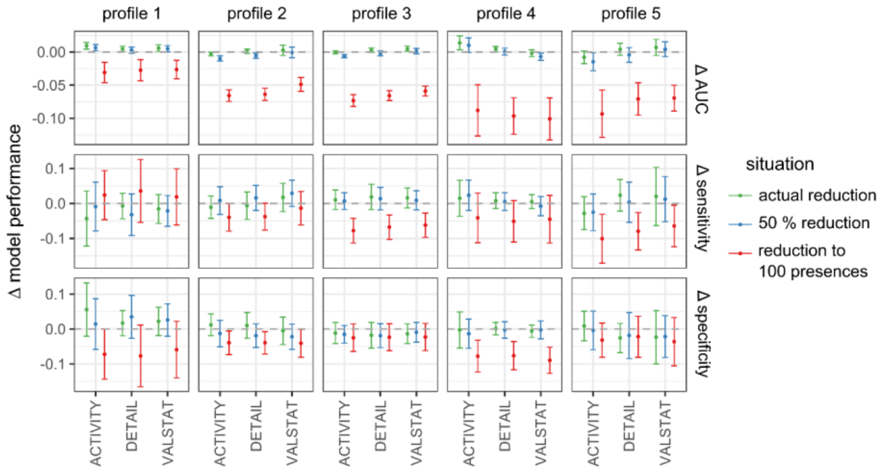


Figure 11: Recommendations for data quality filtering for the five species profiles in the three situations of reduced sample size. Dots and error bars are the means and 90% confidence intervals for the filter effects.

Reducing sample size further mostly worsened model performance, especially when sample size was reduced to 100 presences, where recommendations became alarming or negative in most cases. In our dataset, this meant that sample size was reduced by at least 77% (the lowest unfiltered sample size equalled 432 presences). There were a few exceptions where reducing sample size further did have a positive impact on model performance. For example, recommendations for increasing sensitivity could change from alarming to cautious when reducing sample size by over 50% for profile 2 (using observations from active observers or detailed observations) and up till sample size reached 100 presences for profile 1 (all filters). Results also showed that model sensitivity was more (and specificity was less) impacted by sample size reduction for profiles with birds only (profiles 3 and 5) compared to profiles with more dragonflies and butterflies (profiles 1 and 4).

14.4. Recommendations for data quality filtering

Recommendations for data quality filtering were built on the various results presented in this article. In general, users of opportunistic records should always pay attention when filtering reduces sample size by more than half of its original size, leading to small sample sizes, and we generally advise against filtering when sample size is reduced by more than 75%. We further interpreted the filtering recommendations of the PCA and clustering analysis (Table 2) together with the results of the multi-trait analysis (Figure 10). In the following paragraphs, recommendations are formulated with the aim to increase AUC unless specified otherwise.

Results showed that taxonomic group (more than absolute traits) and relative traits formed the best basis for filtering recommendations and when we discuss traits in the following paragraphs, we mean the relative values unless specified otherwise. We recommend using only data from active observers when filtering opportunistic records of large or widespread butterfly and dragonfly species (profiles 1 and 4) and approved observations when filtering bird records unless they are very familiar and widespread (profile 4). In the cases where absolute traits were retained after model selection (Figure 10), it was the relative rather than the absolute trait that was causing the filter effect. For example, dragonflies and butterflies benefitted more from using observations from active observers (ACTIVITY) compared to birds, yet a higher absolute body size also impacted this effect positively. This meant that dragonflies and butterflies with a higher (relative) body size benefitted most from using the ACTIVITY filter. Keeping bird observations from more active observers only was generally not recommended, except for widespread species with a high classification error rate (profile 1).

Recommendations based on the taxonomic group seemed to overrule the impact of body size (profile 3) and we advise against using body size as a motivator for filtering bird species data. Recommendations based on the taxonomic group were also superior to the impact of familiarity (profiles 4 and 5), yet we still recommend using more detailed observations (DETAIL) for familiar species, especially when they have high reporting probability and an increase in sensitivity is desired. It must be said that recommendations for using the filter DETAIL showed more inconsistencies compared to the other filters and this filter effect could less clearly be linked to species traits.

We recommend using approved observations for species with a restricted range size, especially for large birds. One noted exception was for the widespread species with a high classification error rate (profile 1), where approved observations did impact model AUC positively.

When model AUC increased after filtering, sensitivity mostly increased and specificity decreased, with two exceptions noted. First, species in profiles 1 and 2 were generally more difficult to identify, reflected by either a high classification error rate (profile 1) or because they were small-bodied and unfamiliar to an average observer (profile 2). For these profiles, we see that an increase in data quality by using either filter could reduce the impact of false positives on model performance (i.e. increase specificity), except for using approved observations for widespread species in profile 2. A side-effect was that sensitivity had a greater potential to increase when sample size was reduced beyond the actual reduction, even at high reductions (Figure 11). A second exception, where specificity increased after filtering, was noted for familiar species when using more detailed observations (profile 4) or data from active observers (profile 5). While the positive impact of using only data from active observers on Δ specificity could be linked to higher reporting probability (profile 5), the positive impact of using only detailed observations for familiar and widespread species contradicted the negative association of Δ specificity with range size (Figure 10).

15. Discussion

In this study, we built recommendations for data quality filtering of opportunistic citizen science data when used as input in species distribution models (SDMs), based on a set of a priori-defined species traits. Traits associated with a change in model performance after filtering were: body size, classification error rate, familiarity, reporting probability and range size. Based on these traits, it was possible to generate ecologically meaningful species profiles and make filtering recommendations (section 14.4). The analysis of the species profiles (section 14.2) mostly agreed with the results of a regression analysis (section 14.1) but also gave new insights on the relative importance of the different traits and trait combinations that lead to specific filtering recommendations.

One of the main results was that, when choosing a quality filter, the taxonomic group a species belongs to should be considered. This confirms the findings in Chapter II and makes sense as taxonomic groups by default present differences in most of the considered species traits due to differences in appearance, appeal, distribution etc. In an attempt to simplify the results presented in this chapter, we have tested different approaches to generate the species profiles: considering relative traits only, clustering of species for each taxonomic group separately and including filter effects as active variables. Unfortunately, none of these approaches lead to profiles that were ecologically more meaningful compared to the profiles suggested here (Table 2 and Table I.1; evaluated by species experts). Moreover, they lead to less consistent results (sections 14.1 and 14.2) or less explicit filtering recommendations (i.e. larger confidence intervals in Figure 11). This confirms previous expectations that filtering recommendations can differ between taxonomic groups, but that there might also be common traits among these groups that can refine them (Chapter II). The selected approach indeed revealed that it is possible to formulate recommendations based on taxonomic group and relative traits only (Table 2 and section 14.4). Absolute traits did not directly support recommendations but aided the formation and interpretation of the species profiles as they either characterized the most represented taxonomic group(s) or confirmed a profile's association with relative traits.

The taxonomic bias towards bird species in citizen science data could explain some results, as it indicates greater knowledge by the general public of this species group versus other groups such as butterflies and dragonflies (Troudet et al., 2017; https://waarnemingen.be/stats/). As increased observer activity can lead to higher experience and expertise (Johnston et al., 2017), this can explain why observer activity mattered more for the less-known taxonomic groups in this study (i.e. butterflies and dragonflies). For example, experienced observers were better at detecting individuals of low-density insect populations (Fitzpatrick et al., 2009) and increasing volunteer performance through training could reduce false positive observations for pollinating insects (Ratnieks et al., 2016). These results can also be generalised to other well-known taxonomic groups such as plants. Observer experience, for example, did not increase volunteer performance in identifying an invasive plant species (Crall et al., 2011). Here, observers' self-identified comfort level was a better predictor of volunteer success.

The positive impact of using approved observations for birds, and especially for species with a restricted range size, can be linked to the mechanism of record verification in the database (Swinnen et al., 2018), whereby records that can be verified by photograph or sound play an important role. The verification procedure consists of two main steps: (1) automated record validation by either image recognition or both spatial and temporal proximity of new records to existing approved records and (2) manual expert verification (when there is uncertainty in step 1) (BOX 1, Figure 1). A decent photograph or sound record can thus easily lead to multiple approved records and, as consequence, high photo or sound rates have more chance of leading to approved (filtered) datasets of higher quality. Photo rates were, for example, generally higher for bird and butterfly species with restricted range sizes (Table I.1), which can explain why they benefitted from using approved records. High photo or sound rates can also reduce the negative impact of locational errors on model performance, especially for small sample sizes (Mitchell et al., 2017). Photographs are often made from a closer distance, especially with the available easy-to-use identification apps (e.g. *ObsIdentify*), leading to observations with lower locational uncertainty. When they are made from larger distances, mostly for larger species (i.e. birds in this study), smartphone cameras will not suffice and an observer needs a stronger camera lens. We believe that this is a pastime largely practised by more experienced birders that are more likely to correctly register an individual's exact location compared to an inexperienced observer. As for the importance of sound fragments, bird song usually indicates territorial behaviour (Catchpole and Slater, 2008), hence observations made by sound are usually made in birds' respective habitats. Additionally, the prevalence of locational errors in opportunistic bird data will be larger compared to invertebrate species because of their high mobility (Maes et al., 2019a), even at a scale of 1 km², which was the resolution used in this study.

Large range size is associated with lower model performance because wide-ranged species usually occupy a broad environmental niche and have less distinctive links with their habitat compared to species with a restricted range size that usually have a narrow niche (e.g. Hernandez et al., 2006; Stockwell & Peterson, 2002). While increasing model performance for more widespread species through statistical methods or survey design has been observed to be difficult (Brotons et al., 2007; Tessarolo et al., 2014), we observed that using filtered data, especially from more active observers, had a positive impact on model performance for widespread species. We argue, however, that range size in those cases is subordinate to either classification error rate or the taxonomic group. Firstly, improving data quality is always important for any species with a high misidentification risk (Table 2, profile 1). Misidentification errors can distort estimates of species distributions (Costa et al., 2015; Cruickshank et al., 2019; Miller et al., 2011), even though such errors were reduced by spatial aggregation of records (section 3.2.3; Kramer-Schadt et al., 2013). Misidentification risk has been found higher for species with similar physical appearance, for example, because they are genetically related (Vantieghem et al., 2017) or have mimicking congeners (Ratnieks et al., 2016). Secondly, widespread species in profile 4 are mostly large butterflies and, as previously discussed, this taxonomic group might benefit more from using data from active observers. Moreover, based on the relative traits only (i.e. not considering the taxonomic group) one would intuitively assume that data quality filtering does not have such a pronounced positive impact in profile 4 because these widespread species were also more familiar and had lower error rates.

While retaining observations from active observers or approved observations showed clear associations with taxonomic groups or relative species traits, retaining detailed observations showed more inconsistencies, except for the positive impact on model performance for familiar species. Familiarity might reflect the level of detail at which a species' ecology is known, hence data quality can be increased by retaining more detailed observations for species that are familiar to an average citizen scientist. Because retaining only detailed observations on average had the largest impact on sample size (Figure A.2), the impact of sample size may be overruling the effect of the increase in data quality. Reducing the sample size of presences generally impacts presence-only SDMs negatively as model performance decreases, especially at low sample sizes, and performance variability increases (Hernandez et al., 2006; Liu et al., 2019; van Proosdij et al., 2016). An increase in variability was mostly noted for widespread species, as these species are more sensitive to small sample sizes (Liu et al., 2019). While large reductions in sample size require attention, it remains important to realise that filtering simultaneously increases data quality and thus model performance can also increase, especially when less than half of the presences in a dataset are removed (sections 9.3 and 10).

Detectability did not appear to be an important trait in this study, while it has repeatedly been proven to impact model performance positively (e.g. Pöyry et al., 2008; Seoane et al., 2005), and variation in detectability is directly linked to the problem of imperfect detection in opportunistic presence-only data (Dorazio, 2014; section 2.2). Species traits that are associated with increased detectability are, for example, high abundance (Mccarthy et al., 2013), high singing rates (Sólymos et al., 2018), large body size (Johnston et al., 2014; Pöyry et al., 2008), long lifespan and migratory behaviour (Carrascal et al., 2006). However, we did not find proof that any of these traits were confounded with detectability in our analysis. One trait that could have influenced the outcome for detectability was reporting probability because the way we calculated reporting probability caused a moderate negative correlation between relative reporting probability and relative detectability (Figure I.1). However, reporting probability characterised only one profile and thus implications for filtering recommendations would remain marginal.

While the highly fragmented (Antrop, 2004) and easily accessible landscape in our study region, Flanders, has many benefits for studying species distributions, it was also one of the limitations. The largest benefit was the consequent high spatial and temporal density of records in the *waarnemingen.be* database (Herremans et al., 2018). On the other hand, because of the high density, the low importance of detectability in our study could be an underestimation when studying regions with less fragmented and larger conservation areas.

Another limitation was the insufficient availability of structured data for external model validation in the original dataset (Figure 2; Figure A.2), leading to two restrictive features. First, data consisted of relatively common species (minimum sample size was 432 presences). Rare habitat specialists from habitats with restricted distribution ranges in Flanders (e.g. heathlands) were thereby excluded from this analysis. Since these are often targeted species in national and international biodiversity policy (De Ro et al., 2021; Vanden Broeck et al., 2017), it would be useful to adjust the model validation strategy used in the previous chapter (section 8.4; Appendix A) for those species to be able to formulate generic recommendations. Based on the findings in this chapter, building SDMs with validated data (for species with a restricted range size) or with data from more active observers (for conspicuous invertebrates) could deliver the best results. Second, the data showed sub-optimal representativeness of the taxonomic groups by the studied species. We argue, however, that this imbalance in species representation is often inherent to opportunistic datasets (e.g. over-representation of large birds in Callaghan et al. (2021)).

Finally, some filter effects might have been impacted by the temporal and spatial aggregation of records over the period 2014-2019 and in grid cells of 1x1 km. While a 1 km² resolution is a standard resolution in Flemish biodiversity studies (e.g. Demolder et al., 2014; Rutten et al., 2019; Vantieghem et al., 2017), performing the analysis at different scales might reveal higher or lower impacts of some traits.

16. Conclusions

Many have attempted to disentangle the relationships between species ecology and model performance, and this chapter adds to that knowledge with recommendations for data quality filtering for three commonly studied taxonomic groups. Clustering species into species profiles based on traits that resulted from a multiple regression analysis both highlighted the relative importance of species traits and revealed new insights, and it is important to realise that one single trait does not necessarily predict a species' response to filtering. We found that both the taxonomic group (more than absolute traits) and relative species traits (rescaled values that can be compared among taxonomic groups) defined the impact of data quality filtering on model performance. Our findings were largely based on (1) the general species knowledge among citizen scientists, with a high importance of data quality for widespread and familiar species in general and, more specifically, a high importance of observer experience for less known taxonomic groups, and (2) the mechanism of record verification in an opportunistic data platform, with a high importance of submitting observations that can easily be verified, especially for species with restricted range sizes. We encourage the further improvement of general species knowledge and optimisation of record verification protocols in large citizen science projects. While adopting these recommendations, it is always important to keep the goal of the study in mind (i.e. increasing AUC, sensitivity and/or specificity) and to keep an eye on the change in sample size caused by stringent filtering.

CHAPTER IV. Integrating Citizen Science and Remote Sensing Data for Habitat Management

Adapted from

Van Eupen, C., Maes, D., Heremans, S., Swinnen, K.R.R., Somers, B., Luca, S., (in prep.) Integrating Citizen Science and Multispectral Satellite Data for Multiscale Heathland Management. [under review at Biodiversity and Conservation].

Author contributions

Camille Van Eupen: conceptualization, methodology, software, validation, formal analysis, writing – original draft, visualization; **Dirk Maes:** conceptualization, writing – review & editing, supervision; **Stien Heremans:** conceptualization, writing – review & editing; **Kristijn Swinnen:** data curation, writing – review & editing; **Ben Somers:** conceptualization, writing – review & editing, supervision; **Stijn Luca:** conceptualization, writing – review & editing, supervision

ABSTRACT

Habitat management is necessary for the conservation of threatened species, yet best practices in fragmented human-dominated landscapes have remained difficult to generalise. We show that multi-scale vegetation management decisions in heathlands can be supported by integrating opportunistic citizen science data and multispectral satellite data.

Opportunistic observations were gathered from ten typical, mostly threatened animal species of dry heathlands in Flanders as point records with specified precision. We considered vegetation structure at the local scale, quantified by image texture within 0.25 hectares derived from multispectral satellite data, and heathland heterogeneity at the habitat scale, quantified by the diversity in heathland vegetation communities within 50 hectares. Additionally, locations inside heathlands were attributed to an open, closed or anthropogenic landscape context. Point process models were used to test the impact of heathland size, vegetation structure and heathland heterogeneity on the habitat suitability of the studied species.

We found that (1) heathland vegetation management can benefit habitat suitability in fragmented heathlands, but with a different approach for local management of vegetation structure in small versus large heathlands (e.g. due to micro-fragmentation effects), (2) the landscape induces positive and negative edge effects (e.g. due to a high versus low resource availability), especially in small heathlands and (3) habitat suitability is driven by both vegetation structure and heathland heterogeneity but with a different relative importance for birds, butterflies and grasshoppers (e.g. due to differences in mobility).

17. Introduction

Dry heathlands are human-shaped habitats prioritised in Annex I of the European Habitats Directive (92/43/EEC). They can provide a variety of ecosystem services, such as food and water supply, landscape and biodiversity conservation, carbon sequestration and aesthetic/recreational value. They are, however, threatened by land conversion and privatisation, recreation and soil eutrophication from intensive agriculture that causes moss, grass and tree encroachment (Fagúndez, 2013; Webb, 1998). These pressures have led to the fragmentation and reduced habitat quality of heathlands, and an ever-increasing proportion of heathland fauna appearing on national red lists (Maes et al., 2019b). Conservation of species that rely on habitats under anthropogenic pressures remains challenging and is in strong need of evidence-based action plans (Maes et al., 2022; Olmeda et al., 2020). In European dry heathlands, conservation management is traditionally designed from a flora perspective with a focus on preserving typical successional heathland vegetation (De Blust, 2022; Webb, 1998). Typical management schemes are designed to prevent nutrient accumulation and natural succession to forest, for example by sod-cutting, burning or grazing (De Blust, 2022; Fagúndez, 2013). It has become generally accepted, however, that heathland fauna profits from management that includes exposure of bare soil, diversifies vegetation communities and increases structure complexity (Byriel et al., 2023; De Blust, 2022; de Vries et al., 2021; Schirmel et al., 2011; van den Berg et al., 2001), yet evidence-based action plans in conservation policy remain scarce.

Habitat quality can be increased for animal species of conservation interest by providing a broad range of environmental resources (for example for shelter, nesting space and foraging) through vegetation management that increases habitat heterogeneity (MacArthur and Wilson, 1967; Tews et al., 2004). Habitat quality is also impacted by the landscape context, which can provide opportunities for habitat connectivity (Gibson et al., 2004; Haddad and Baum, 1999) or induce positive or negative edge effects (Dupont and Overgaard Nielsen, 2006; Fagúndez, 2013; Neilan et al., 2019; Pfeifer et al., 2017). While the impact of edge effects on heathland vegetation is spatially confined (for example, eutrophication effects on heathland vegetation and soil were detected up to ca. 8 metres into the patch according to Piessens et al. (2006)), the impact on heathland fauna might reach further (e.g. Pfeifer et al., 2017). Additionally, the potential habitat quality of small and isolated patches should not be neglected (Wintle et al., 2019).

Heathland vegetation management can be implemented at different spatial scales, of which we distinguish two in this chapter: heathland heterogeneity at the larger habitat-type scale and vegetation structure at the smaller local scale. Heathland heterogeneity is the horizontal variation of vegetation communities (i.e. habitat subtypes such as wet and dry heathlands, peatlands or Nardus grasslands), with high heterogeneity leading to higher habitat suitability for species that need complementary resources and higher species richness (see reviews of Stein et al., 2014; Tews et al., 2004). Vegetation structure is the structural complexity of vegetation within a habitat (Bergen et al., 2009) and a crucial determinant of species' habitats (Bergen et al., 2009; Randin et al., 2020). It has, for example, been positively evaluated as a predictor for the habitat suitability of birds in forests (Farrell et al., 2013; Goetz et al., 2010; Graf et al., 2009; Huber et al., 2016; Seavy et al., 2009) and grasslands (Bellis et al., 2008), butterflies in grasslands and woodlands (de Vries et al., 2021) and lizards in a river valley (Sillero and Gonçalves-Seco, 2014).

Heathland heterogeneity and the landscape context can easily be quantified by a landscape analysis, for example by using landscape metrics (Gustafson, 1998; Hesselbarth et al., 2019). Vegetation structure is less trivial to capture because of its different components and the most appropriate method depends on the intended use. For modelling habitat suitability, Sentinel-2-derived image texture (Haralick, 1979) has recently been proposed as a proxy for vegetation structure (Farwell et al., 2021) and has several advantages over image texture derived from other satellite sensors and over more traditional methods, such as field campaigns and Light Detection and Ranging (LiDAR) data (Wehr and Lohr, 1999). First, Sentinel-2 data deliver finer resolution data (10 metres) compared to other satellite sensors such as Landsat (30 metres) or MODIS (100 metres) and are thus especially suited to investigate fine-scaled drivers of species distributions. Second, Sentinel-2 data have higher temporal coverage than LiDAR data. Surely, LiDAR data deliver fine-resolution data and in some regions have high spatial coverage, yet they usually have limited temporal coverage (Moudrý et al., 2022) that can cause an undesired temporal mismatch between environmental data and species occurrence data (Randin et al., 2020). Moreover, when LiDAR images are taken in the leaf-off season, they fail to capture several structural habitat characteristics in low-stature habitats (e.g. in grasslands; de Vries et al., 2021). Third, texture measures can capture both vertical and horizontal components of vegetation structure (Farwell et al., 2021) and Sentinel-2 outperformed other satellite sensors and field measurements in quantifying both LiDAR-derived metrics as well as with field-based metrics of vegetation structure (Farwell et al., 2021; Wood et al., 2012). Finally, Sentinel-2 data are freely accessible, with many regions having had free access to 10-metre resolution images every 5 to 10 days since April 2017, while LiDAR and field campaigns are often expensive.

Both the availability of multispectral Sentinel-2 data for over five years now and the increasing quantity and density of species occurrence data through large citizen science initiatives, such as *waarnemingen.be* in Flanders (https://www.waarnemingen.be) and observation.org (https://observation.org/) or iNaturalist (https://www.inaturalist.org/) worldwide, facilitate the use of fine-grained habitat suitability models**.** The availability of fine-resolution continuous measures of vegetation cover over large spatial extents is a promising advance in the field of species distribution modelling (Milanesi et al., 2017; Randin et al., 2020) as categorical land cover maps contain errors, are labour-intensive to develop and might miss essential information on habitat requirements (Oeser et al., 2020). Popular measures are, for example, vegetation indices that quantify vegetation health or greenness, such as the Normalized Difference Vegetation Index (NDVI) (Pettorelli et al., 2005) and the Enhanced Vegetation Index (EVI) (Huete et al., 2002), and image texture that quantifies spatial heterogeneity in vegetation cover (Wood et al., 2012). Fine-grained remote sensing products are especially suited when using point process models (Renner et al., 2015), where presence-only data are treated as point events and the intensity of species occurrence is modelled. Here, environmental data can be extracted for each event at various and small spatial extents, which is the strength of this method.

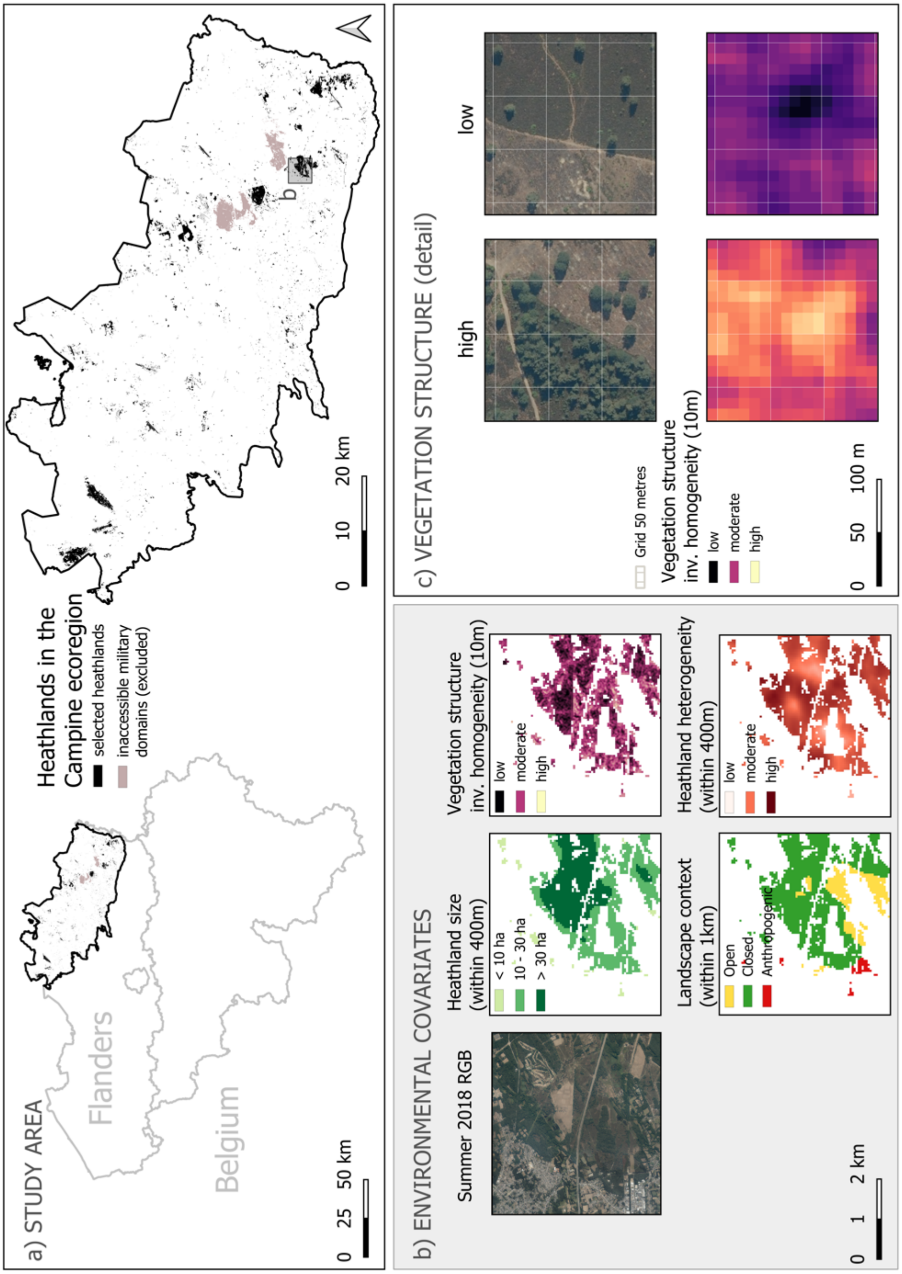
This chapter will test the possibility of integrating opportunistic citizen science data and multispectral satellite data to support multiscale management decisions for the conservation of animal species in anthropogenic regions (Maes et al., 2022). More specifically, we will analyse whether the habitat suitability of dry heathland specialists across different taxonomic groups is driven by vegetation structure and/or heathland heterogeneity and whether this relationship depends on the heathland size and landscape context. We hypothesize that heathland management can benefit habitat suitability for species of conservation interest, even in small heathlands (Gábor et al., 2022; Wintle et al., 2019), that it should consider the landscape matrix due to positive and negative edge effects (Fahrig, 2003) and that it requires an integrated multispecies approach (Bonari et al., 2017; Maes and Van Dyck, 2005). We will use Gibbs point process models models with a Geyer saturation process (Baddeley et al., 2015) to account for spatial dependence in the species occurrence data and add bias covariates to account for known sources of sampling bias (Renner et al., 2015; Warton et al., 2013). Vegetation structure will be quantified by a second-order texture measure (Haralick et al., 1973), using a 10-metre Sentinel-2 pixel image of the Enhanced Vegetation Index (EVI; Liu and Huete, 1995) as this has been evaluated as a good proxy for vegetation structure by Farwell et al. (2021). Heathland size and heterogeneity and the landscape context will be quantified based on the Biological Valuation map, a detailed inventory of the land cover in Flanders (De Saeger et al., 2017)**.**

18. Materials and methods

18.1. Study area

The study region was the Campine region in Flanders in the northeast of Belgium (Figure 12a), holding about 13,000 hectares of heathland (De Saeger et al., 2020) and characterised by sandy soils (Couvreur et al., 2004). We limited our study area to heathland patches with more than 40 per cent classified heathland on the 2020 Biological Valuation Map (BVM) (De Saeger et al., 2020) which is a database for land cover in Flanders that includes a map of habitat classes (De Saeger et al., 2017). We omitted three military domains (Figure 12a), because of a strong negative observation bias due to their inaccessibility, and patches with urban elements.

Figure 12: a) Study area and studied heathlands (inaccessible military domains were excluded); b) Environmental covariates used to predict the relative habitat suitability of dry-heathland species of conservation interest at an example location. The landscape context was the dominant surrounding land cover class in a one-kilometre radius around points on a regular grid of 50 metres. Heathland size and heathland heterogeneity were calculated by calculating the mean heathland size and the Shannon diversity in heathland subtypes (such as dry and wet heathlands and heathlands with and without trees), respectively, within a 400-metre radius around points on a regular grid of 50 metres. Vegetation structure is the inverse of the homogeneity (a gray-level co-occurrence matrix (GLCM) second-order texture metric) calculated at a resolution of 10 metres, supplemented with the average homogeneity in a 50-metre radius around points on a regular grid of 50 metres in the patch edges with missing values; c) An example of low and high vegetation structure. The location with a high vegetation structure is characterised (from left to right) by plantings of Scots pine (Pinus sylvestris L.) with undergrowth of shrubs and trees, a woody edge of broom thicket (Cytisus scoparius L.) and dry heather vegetation (Calluna vulgaris L.) with shrub or tree stands. The location with low vegetation structure is characterised by dry heather vegetation communities (Calluna – Genista) with an occasional tree or shrub (De Saeger et al., 2020).


a) STUDY AREA
Heathlands in the Campine ecoregion
selected heathlands
inaccessible military domains (excluded)
Flanders
Belgium
b
0
25
50 km
0
10
20 km
b) ENVIRONMENTAL COVARIATES
Summer 2018 RGB
Heathland size (within 400m)
< 10 ha
10 - 30 ha
> 30 ha
Vegetation structure inv. homogeneity (10m)
low
moderate
high
Landscape context (within 1km)
Open
Closed
Anthropogenic
Heathland heterogeneity (within 400m)
low
moderate
high
0
1
2 km
c) VEGETATION STRUCTURE (detail)
high
low
Grid 50 metres
Vegetation structure inv. homogeneity (10m)
low
moderate
high
0
50
100 m

18.1. Species observations

We considered dry-heathland fauna of conservation interest in Flanders, meaning that they are either species of regional conservation interest (Annex II or IV of the Habitats Directive (92/43/EEC) or Annex I of the Birds Directive (79/409/EEC)) (Paelinckx et al., 2009), Flemish Priority Species (De Knijf et al., 2014; Herremans et al., 2014) or habitat-specific species (Habitats Directive habitat types 2310, 2330, 4030) (De Knijf and Paelinckx, 2013) (Tables 4 and K.1). Critically endangered species were excluded (e.g. Northern wheatear is almost extinct in Flanders (Vermeersch et al., 2020)). Observations from eighteen species from four taxonomic groups (i.e. four birds, five butterflies, seven grasshoppers and two reptiles) were extracted from the data portal *waarnemingen.be* (Herremans et al., 2018; https://www.waarnemingen.be). They were point observations with specified geographical precision for the study region and study period 2017-2021. Only the months from April to August were considered as this period provided a good overlap between the growing season in Flanders and the reproductive seasons for the species under study. The data was cleansed, checking for wrong coordinates, removing incorrect observations and keeping only observations with a precision of fewer than 50 metres.

To construct the model training sets, we extracted opportunistic/unstructured records and first applied data quality filtering according to previous recommendations made by in Chapter III. Data verified as correct were retained based on the taxonomic group, range size and relative body size. Second, we applied spatial thinning at 50 metres per observation date to reduce the impact of duplicates (i.e. observations from an individual at a similar location on the same date).

18.2. Model predictors

We used existing maps, satellite imagery and species occurrence data from *waarnemingen.be* to construct the model predictors, i.e. heathland heterogeneity, vegetation structure, heathland size, the landscape context and two sampling bias covariates (accessibility and search effort) (Table 3; sections 18.3.1 to 18.3.5). Most model predictors (all except vegetation structure) were rasterised at a resolution of 50 metres by applying calculations (i.e. summary statistics, landscape metrics, vector lengths) in a buffer area with varying radii around each point at a regular grid of 50 x 50 metres, further called ‘dummy points’. Per species, all predictors were tested for multicollinearity by extracting their values at all training presence locations and calculating variance inflation factors (VIFs) and Pearson correlations in the R package ‘fuzzySim’ version 4.3 (Barbosa, 2015). All calculations were performed in R version 4.2.1 (R Core Team, 2022) and QGIS version 3.16.9.

Table 3: Methods applied to obtain the model predictors that were used to predict the relative habitat suitability of dry-heathland species of conservation interest. Dummy points are points at a regular grid of 50 x 50 metres throughout the study area. BVM = Biological Valuation Map, rasterised at 5 metres; EVI = Enhanced Vegetation Index; GLCM = gray-level co-occurrence matrix; [1] De Saeger et al. (2020); [2] retrieved from Google Earth Engine; [3] retrieved from https://land.copernicus.eu/pan-european/corine-land-cover/clc2018; [4] retrieved from https://www.geopunt.be/; [5] Herremans et al. (2018). All calculations were performed in R version 4.2.1 (R Core Team, 2022) and QGIS version 3.16.9.

Predictor	***Source***	***Calculation (per pixel)***	***Scale***	***Res.***	***min-max***
ENVIRONMENTAL COVARIATES					
Heathland heterogeneity	BVM version 2020 [1]	4 heathland subtypes (each with/without trees or shrubs) - dry heathlands - wet heathlands - peatlands - Nardus grasslands	Shannon diversity index within 400 metres (≈ 50 ha) around each dummy point	50 m	0.00 - 1.71
Vegetation structure	Sentinel-2A images [2] April to August 2017-2021 Heathlands (> 40% heathland) and semi-natural edges (> 10% heathland)	i. Masked clouds, snow/ice and unreliable pixels ii. Calculated EVI (kept values between 0.1 and 1) iii. Average of the annual median composites iv. GLCM 2nd order texture: homogeneity (inverse)	5 x 5 moving window (≈ 0.25 ha) in steps of 10 metres Pixels with missing values in moving window: average homogeneity (inverse) within 50 metres	10 m	0.06 - 0.97
Heathland size (hectares)	BVM version 2020 [1]	Percentage of heathland (converted to hectares)	Mean within 400 metres (≈ 50 ha) around each dummy point	50 m	0.22 - 49.85
Landscape context	*Flanders*: BVM version 2020 [1] *Outside Flanders:* CORINE version 2018 [3]	Formed 3 land cover classes: - closed (forest) - open (other semi-natural) - anthropogenic (urban and agricultural)	Dominant class within one kilometre around each dummy point	50 m	NA (factor)
SAMPLING BIAS COVARIATES					
Accessibility (km road/km^2)	Wegenregister version 2.0 [4]	Length of road segments	within 100 metres (≈ 3.14 ha) around each dummy point	50 m	0.00 - 0.05
Search effort (n° species)	*waarnemingen.be* [5] April to August 2017-2021	The annual average number of species observed within the considered taxonomic group	within 100 metres (≈ 3.14 ha) around each dummy point	50 m	0.0 - 37.4

We chose to include only measures of vegetation structure and habitat composition, although we acknowledge that including measures of soil water, such as the topographic wetness index (Besnard et al., 2015; Moore et al., 1993), or soil biochemistry, such as nitrogen (N) and phosphorus (P) content (Vogels et al., 2017), might have led to additional insights. We motivate the choice of our predictors by the objective of the study (i.e. to illustrate how integrating citizen science and multispectral satellite data can support multiscale heathland vegetation management) and the absence of multicollinearity (Table K.2; also see section 22.1). Soil water, for example, might be correlated with both heathland heterogeneity and vegetation structure, as soil moisture impacts the composition of vegetation communities and the presence and growth of certain plant species (Schellenberg and Bergmeier, 2020).

18.2.1. Heathland heterogeneity

Heathland heterogeneity, heathland size (section 18.3.3) and the landscape context (section 18.3.4) were calculated based on the Biological Valuation Map (BVM) as it includes a detailed classification of habitat types and a classification of land cover in Flanders (De Saeger et al., 2017). Heathland heterogeneity was quantified by the Shannon Diversity Index (shdi) in the R package ‘landscapemetrics’ version 1.5.4 (Hesselbarth et al., 2019), applied to four sub-types of heathland as classified in the BVM version 2020 (De Saeger et al., 2020): dry heathland, wet heathland, peat and Nardus grasslands. We also distinguished subtypes with and without trees or shrubs. The BVM was rasterised at 5 metres and the shdi was calculated in a 400-metre radius (≈ 50 hectares) around each dummy point (Figure 12b). The BVM is a vector but was rasterized because the ‘landscapemetrics’ package takes rasters as input. We chose 400 metres as the maximum radius for all species to facilitate comparability among results, although an alternative would be to tune the radius according to a species’ maximum dispersal ability.

18.2.2. Vegetation structure

We used Sentinel-2A imagery to quantify vegetation structure as this satellite has been delivering multispectral data across large spatial extents since April 2017 at a high spatial and temporal resolution (10 x 10 metres every 5 to 10 days for Flanders). Vegetation structure was quantified by calculating the homogeneity, a second-order texture measure for image smoothness (Haralick, 1979; Haralick et al., 1973), of a Sentinel-2 EVI (Enhanced Vegetation Index) composite (Liu and Huete, 1995) (Figure 12b). When calculating second-order texture measures, the spatial configuration of pixel values are taken into account by first constructing a gray-level co-occurrence matrix (GLCM; Haralick et al., 1973). Second-order homogeneity characterizes mainly vertical complexity with ancillary information on horizontal plant diversity and was suggested to sufficiently capture vegetation structure relevant to species' habitat suitability (Farwell et al., 2021). Figure 12c shows example locations with low structure (i.e. a mostly uniform vegetation cover) and high structure (i.e. spatial variation in vegetation communities and height).

For each 10-metre pixel in the Campine region, annual median EVI composites from April to August in the study period 2017-2021 were obtained from the near-infrared, blue and red band of the image collection “Sentinel-2 MSI: MultiSpectral Instrument, Level 2A” in Google Earth Engine. Before calculating the EVI, pixels with scene classification labels 1 to 3 and 8 to 11 were omitted (i.e. unreliable pixels, clouds and snow/ice). The annual EVI values were averaged, excluding values below 0.1 and above 1 as they mostly indicated buildings, paved soils or solar panels. Homogeneity was calculated using the R package ‘glcm’ version 1.6.5 (https://cran.r-project.org/web/packages/glcm/) with a kernel size of 5 (i.e. a moving window of 5 x 5 pixels or 50 x 50 metres). Vegetation structure was calculated in steps of 10 metres and the inverse of homogeneity was taken as low values indicated a high vegetation structure and vice versa.

We adapted our approach to increase the availability of pixels available for modelling despite the large number of edges in our study area. Since the study area was not a spatially continuous patch, edges were abundantly present inducing one or more missing EVI values in the moving windows used to calculate homogeneity. To reduce the impact of these edge effects and hence increase the number of raster pixels with predictor values for vegetation structure, we took three actions. First, for texture calculations, we included the EVI values from semi-natural edges (i.e. connected patches of semi-natural habitats of which at least 10% was identified as heathland). Second, we chose a small kernel size to reduce the chance of missing values for texture calculations. Third, we calculated the average homogeneity in a 50-metre radius around each dummy point in the patch edges with missing values and added this information to the raster layer for vegetation structure.

18.2.3. Heathland size

To quantify heathland size, we attributed the percentage of heathland associated with each pixel in the rasterized BVM (see section 18.3.1) following the distribution key of the different habitat units per patch (De Saeger et al., 2020). Consequently, we calculated the mean percentage of heathland in a 400-metre radius (≈ 50 hectares) around each dummy point (Figure 12b). Models were run with continuous heathland size as predictor to assess its impact on habitat suitability. However, we also categorised heathland size into three classes in the results section for the dual purpose of simplifying the presentation of the results and formulating tangible recommendations. We distinguish between (1) small patches (≤ 10 hectares), i.e. mostly small and isolated patches with an occasional heathland patch edge largely surrounded by different land cover, (2) intermediate patches/patch edges (10-30 hectares), i.e. mostly edges of large heathland patches with an occasional medium-sized patch, and (3) large patches (> 30 hectares), i.e. core areas of large heathland patches (Figure 12b).

18.2.4. Landscape context

To describe the landscape context, we categorised the land cover into three classes: closed (i.e. forest), open (i.e. all other semi-natural land covers) and anthropogenic (i.e. urban and agriculture) land cover. The dominant class in a one-kilometre buffer around each dummy point was taken as the landscape context (Figure 12b). Land cover within the Campine region was taken from the BVM, while at the borders of Flanders, we used the CORINE land cover classification[9].

18.2.5. Sampling bias covariates

In a point process setting, it is common to include covariates that can accommodate sampling bias instead of modifying the background (i.e. the quadrature scheme) (Renner et al., 2015), opposed to, for example, a target group background selection in Maxent (Phillips et al., 2009). We added one accessibility covariate: road density (km road per square km, calculated based on the *Wegenregister* version 2.0[10]); and one search effort covariate: the annual average number of species observed within the considered taxonomic group in the study period (extracted from *waarnemingen.be*). Both were calculated in a 100-metre radius around each dummy point. Accessibility accounted for the impact of high observation density around roads (both paved and unpaved) while search effort accounted for the impact of observer activity.

18.3. Habitat suitability

18.3.1. Gibbs Point Process Model

Gibbs point process models (Baddeley et al., 2015) were used to study the impact of heathland size, vegetation structure and heathland heterogeneity on the habitat suitability of dry-heathland fauna in different landscape contexts. We ran models per species and landscape context (i.e. open, closed and anthropogenic) with two-way interactions between heathland size *hsize* and vegetation structure *VS* and heathland heterogeneity *HH*. The conditional intensity $\lambda(u|x)$ at a

[9] Retrieved from https://land.copernicus.eu/pan-european/corine-land-cover/clc2018

[10] Retrieved from https://www.geopunt.be/

location *u* given a pattern of presences *x* consists of a first-order term (the trend or covariate effects) and a higher-order term γ (the interaction parameter). The model formula can be simplified as follows:

$$\lambda(u|x) = (hsize + VS + HH + hsize{:}VS + hsize{:}HH + accessibility + search\ effort) * \gamma$$

We used a binary window to delineate the study area and the translate border correction as it is the recommended method in a binary window setting (Baddeley and Turner, 2005). A Geyer saturation process was implemented to model spatial interaction as it deals well with clustered data (Baddeley et al., 2015) and most of the species occurrence data records were identified as inhomogeneous point processes with spatial interaction (mostly clustering at small distances) (Figures K.2 to K.11). Geyer radius and saturation values were defined for each species using the *profilepl* function in the R package ‘spatstat’ version 2.3-4 (Baddeley and Turner, 2005) on a range of 50 to 500-metre radii, with 50-metre intervals, and saturation values of either 1 or 2. We chose dummy points at a regular grid of 50 metres for model fitting (eps = 50), which was an adequate resolution for estimating the maximum pseudolikelihood considering covariate resolution (Baddeley and Turner, 2000; Renner et al., 2015). These are combined with the presence points to generate quadrature points and quadrature weights before model fitting.

Only dummy points (i.e. background points) within five kilometres from a presence point were included. Goodness-of-fit was evaluated with a Diggle-Cressie-Loosmore-Ford (DCLF) test (Baddeley et al., 2014) and predictive performance was assessed in a spatial block cross-validation using the R package ‘blockCV’ version 3.1-1 (Valavi et al., 2019). We encountered some model fitting problems in an exploratory analysis and set a threshold of 60 presences to avoid poorly fitted or invalid models. Eight species and four models in the anthropogenic landscape context were omitted for further analysis (see Table K.1). We finally kept ten species with valid models in at least two landscape contexts (Table 4).

18.3.1. Model predictions

We are interested in the impact of the environmental variables, and their interactions, on species' habitat suitability. Therefore, we fitted the first-order trend of the model, which can be seen as the conditional intensity without spatial interaction (i.e. the conditional intensity of an empty point pattern) (Baddeley et al., 2015), and kept sampling bias covariates constant (Warton et al., 2013).

Table 4: List of selected species with their Red List Status in Flanders (LC = Least Concern, NT = Near Threatened, EN = Endangered) (Devos et al., 2016; Maes et al., 2021, 2017b), *Conservation Interest (BD = Birds Directive, FPS = Flemish Priority Species, HSS = Habitat Specific Species with Habitats Directive Annex I habitat types)* (De Knijf et al., 2014; De Knijf and Paelinckx, 2013; Herremans et al., 2014; Paelinckx et al., 2009)*, and the number of observations and average intensities of the point processes.*

Species	**English name**	**Red List status in Flanders**	**Conservation Interest**	**Number of observations** *Intensity of the point process (.10^{-5})*		
				Open	**Closed**	**Anthr.**
BIRDS						
Anthus trivialis	Tree Pipit	NT	HSS ^{2310}	907 *4.35*	2638 *6.37*	137 *1.00*
Caprimulgus europaeus	European Nightjar	NT	BD Annex I HSS ^{4030}	151 *0.72*	462 *1.10*	-
Lullula arborea	Woodlark	NT	BD Annex I HSS $^{2310, 4030}$	492 *2.36*	1213 *2.88*	80 *0.59*
Saxicola rubicola	European Stonechat	LC	HSS $^{2310, 4030}$	935 *4.48*	1542 *3.66*	130 *0.95*
BUTTERFLIES						
Callophrys rubi	Green Hairstreak	EN	HSS $^{2310, 4030}$	265 *1.27*	321 *0.76*	-
Hipparchia semele	Grayling	EN	FPS HSS $^{2310, 2330, 4030}$	302 *1.45*	330 *0.78*	485 *3.55*
Plebejus argus	Silver-studded Blue	EN	HSS ^{4030}	621 *2.98*	483 *1.15*	-
GRASSHOPPERS						
Gryllus campestris	Field Cricket	EN	HSS $^{2310, 2330}$	118 *0.57*	324 *0.77*	-
Myrmeleotettix maculatus	Mottled Grasshopper	LC	HSS $^{2310, 2330}$	68 *0.33*	243 *0.58*	82 *0.60*
Oedipoda caerulescens	Blue Winged Grasshopper	LC	HSS $^{2310, 4030}$	112 *0.54*	296 *0.70*	189 *1.38*

19. Results

Species occurrence sets showed spatial interaction at radii of 50 to 250 metres (Table K.1 and Figures K.2 to K.11). Predictors showed no multicollinearity (Table K.2; VIF < 3 and average Pearson correlations $r = -0.007 \pm 0.193$, $r = 0.023 \pm 0.188$ and $r = -0.060 \pm 0.261$ in open, closed and anthropogenic landscape contexts respectively). Models fitted the data reasonably well, with no test rejecting the null hypothesis at a 0.01 significance level (Table K.1). Bird models performed better than most butterfly and grasshopper models, with the latter also presenting more variation in model performance (Table L.1).

We will discuss the aggregated results for all species and per taxonomic group in the main text and refer to Appendix M for the results per species. We found that larger heathlands were more suitable than intermediate or small heathland patches for all three taxonomic groups, especially in an anthropogenic landscape context (Figures 13 and 14). Habitat suitability was impacted at different spatial scales (vegetation structure versus heathland heterogeneity) and results varied with heathland size and the landscape context (Figures 13 and 14). Results in semi-natural (i.e. open or closed) contexts were generally different from those in an anthropogenic context.

The impact of vegetation structure and/or heathland heterogeneity on habitat suitability depended on the surrounding heathland size for most species in at least one landscape context (Appendices L to N). A high vegetation structure became more important at larger heathland sizes, but in small patches in a closed or open landscape context also a low vegetation structure could benefit habitat suitability, especially for birds and butterflies (Figure 14). A high heathland heterogeneity mostly impacted habitat suitability positively (Figures 13 and 14) and became more important when heathland size increased in an open landscape context, while being equally important across heathland sizes in a closed or anthropogenic context (Figure 14).

Pooling all species (boxplots in Figure 14), habitat suitability in an open landscape context was highest in large patches with high heathland heterogeneity or small patches with a low vegetation structure. In a closed landscape context, high heathland heterogeneity is beneficial, as are small heathland patches with a low vegetation structure and large patches with a high vegetation structure. In an anthropogenic landscape context, habitat suitability was highest in patches with a high heathland heterogeneity and a high vegetation structure, especially in large patches.

Habitat suitability of species in all three taxonomic groups was impacted by both vegetation structure and heathland heterogeneity, depending on heathland size and the landscape context (Figures 13 and 14). Bird habitat suitability was positively impacted by heathland heterogeneity, especially in large patches in a semi-natural landscape context. Habitat suitability further increased in small and intermediate patches/patch edges with a low vegetation structure in a semi-natural context, and in intermediate patches/patch edges and large patches with a high vegetation structure in an anthropogenic context. Butterfly habitat suitability was positively impacted by both a high vegetation structure and a high heathland heterogeneity in large patches in an open or anthropogenic context. In a closed context, a high vegetation structure increased habitat suitability in large patches and a high heathland heterogeneity did so in intermediate patches/patch edges. In small patches, habitat suitability for butterflies increased with a low vegetation structure, combined with low heathland heterogeneity in an open landscape context and high heathland heterogeneity in a closed context. Grasshopper habitat suitability was higher at a high vegetation structure, especially in intermediate patches/patch edges, and at a high heathland heterogeneity in a closed or anthropogenic landscape context.

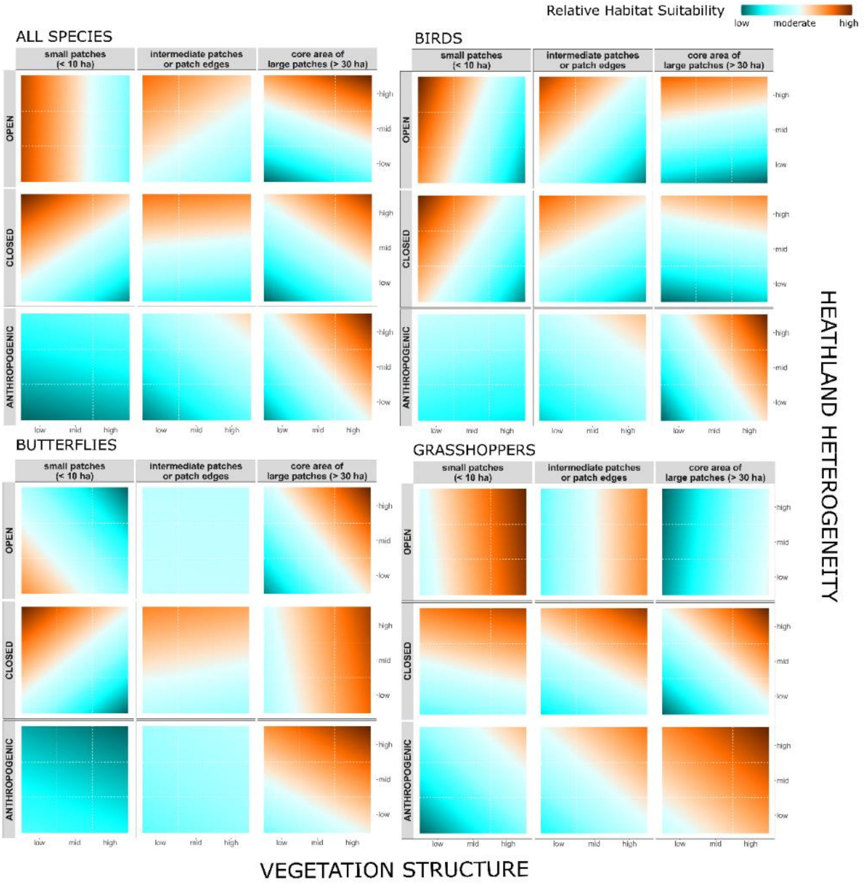


Figure 13: Relative habitat suitability - the impact of vegetation structure (x-axis) and heathland heterogeneity (y-axis) on habitat suitability in different landscape contexts, summarised in three classes of heathland sizes. Predicted intensities were first log-transformed to generate a linear output and then scaled and averaged across all considered dry-heathland species and according to taxonomy in different landscape contexts (blue = low relative suitability, orange = high relative suitability). These values are the results of different Gibbs point process models with Geyer saturation process per landscape context, including two-way interactions between heathland size and vegetation structure/heathland heterogeneity. For four species (i.e. Caprimulgus europaeus, Callophrys rubi, Plebejus argus and Gryllus campestris), the model in the anthropogenic landscape context was omitted (see section 18.4.1).

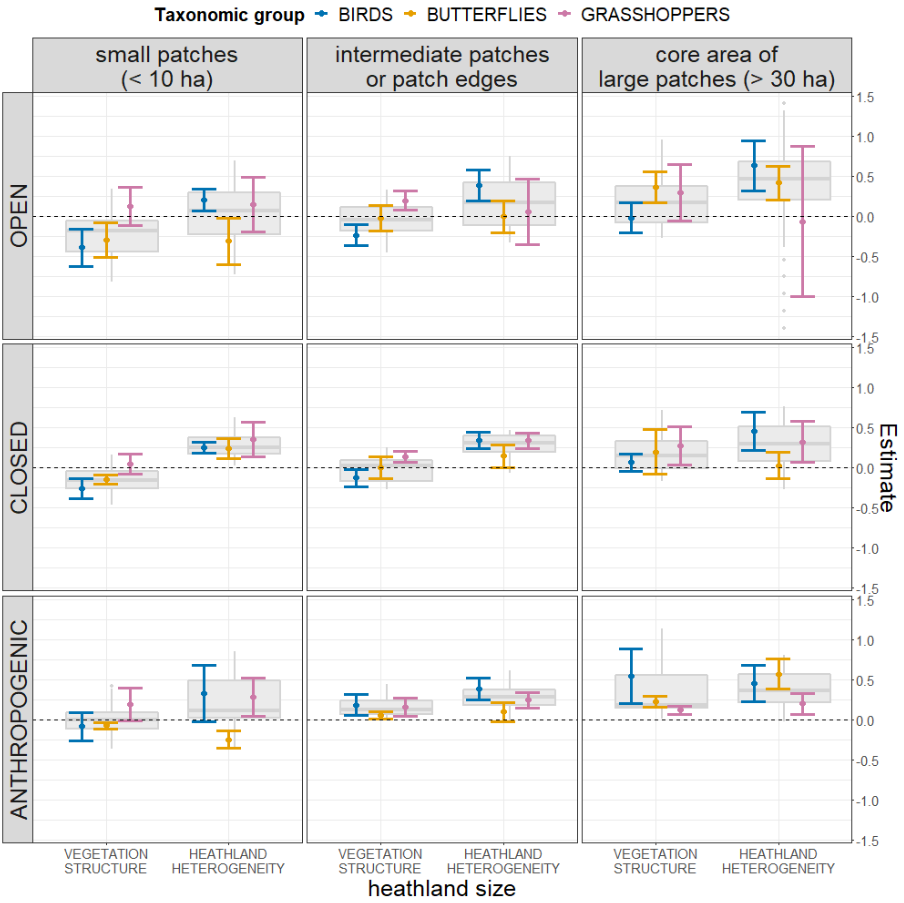


Figure 14: Model coefficients - the impact of vegetation structure and heathland heterogeneity on habitat suitability in different landscape contexts, summarised in three classes of heathland sizes. The distribution of model trend coefficients is shown for all species (boxplots) and grouped according to taxonomy, with dots and error bars representing mean estimate values and standard deviations. These values are the results of different Gibbs point process models with Geyer saturation process per landscape context, including two-way interactions between heathland size and vegetation structure/heathland heterogeneity. For four species (i.e. Caprimulgus europaeus, Callophrys rubi, Plebejus argus and Gryllus campestris), the model in the anthropogenic landscape context was omitted (see section 18.4.1).

20. Discussion

By integrating opportunistic citizen science data and multispectral satellite data in point process models, we have substantiated the importance of managing vegetation structure for heathland fauna (Byriel et al., 2023; Maes et al., 2017c) and highlighted some important considerations when working in human-dominated and fragmented landscapes, such as the impact of edge effects from the surrounding land cover and the characteristics of the considered taxonomic group. Quantifying vegetation structure and heathland heterogeneity in a standardized and spatially contiguous way enabled us to produce generalisable results, beyond local studies, an important asset for biodiversity policy and conservation, for example, for designing essential biodiversity variables (EBVs) (Valbuena et al., 2020; Vihervaara et al., 2015). Heathland vegetation management could increase the habitat suitability of ten species from different taxonomic groups at two spatial scales: local-scale vegetation structure (0.25 hectares) and habitat-scale heathland heterogeneity (50 hectares).

According to the habitat heterogeneity hypothesis (e.g. MacArthur & Wilson, 1967), structural complexity and habitat heterogeneity increase niche availability and diversify environmental resources. Although animal species distributions and diversity are mostly affected positively by increased habitat heterogeneity (Ampoorter et al., 2020), this relationship can also remain undetected or even be negative, largely depending on the spatial scale, the type of heterogeneity measure and the taxonomic group considered (Stein et al., 2014; Tews et al., 2004). Results generally demonstrated positive impacts of heathland heterogeneity, and also of vegetation structure in core areas of large heathland patches. In small and fragmented patches, however, local-scale vegetation structure was often associated negatively with habitat suitability for the studied birds and butterflies. This might be explained by the habitat preference of the studied birds and by the effect of micro-fragmentation (Laanisto et al., 2013). All four birds prefer open to semi-open heathlands with occasional trees or shrubs as a viewing point for foraging. These relatively large species will need relatively large areas with low-structure vegetation cover, especially in small and fragmented patches. Micro-fragmentation, on the other hand, implies that small-scale heterogeneity can cause niche isolation for less mobile species. While intuitively birds should be less affected by micro-fragmentation at the considered scale (i.e. 0.25 hectares), lower food availability of species that are negatively affected by micro-fragmentation, such as invertebrates and plants (Laanisto et al., 2013; Tamme et al., 2010), could also explain the negative relationship with vegetation structure for birds. Grasshoppers were not affected, although this might have been the result of a mismatch between predictor

(0.25 hectares) and home ranges (Guisan and Thuiller, 2005; Oliveira et al., 2021) and grasshopper models also showed more variability (Table L.1 and Figures M.3 and M.6).

Habitat edges induce edge effects that become stronger when habitats become more fragmented (Ewers et al., 2007; Fahrig, 2003), which is probably why we found the largest differences between landscape contexts in small patches and patch edges. Edges in a semi-natural landscape context can provide resources for the inhabiting species such as shelter, nesting or foraging opportunities (Dupont and Overgaard Nielsen, 2006; Evens et al., 2018) and deliver specific habitat conditions such as forest edges (Moquet et al., 2018; Pfeifer et al., 2017). Small and isolated patches can thus have high habitat quality (Wintle et al., 2019) if located in a resourceful environment. The surrounding semi-natural land cover might even enhance the structural complexity to a point where maintaining characteristic heathland vegetation (i.e. dwarf shrubs, quantified by a low vegetation structure; Figure 12c) will become relatively more important, especially for species that rely on them for food and reproduction (Byriel et al., 2023) such as Grayling, European Stonechat and Silver-studded Blue. Similarly, the diversity in vegetation communities can be enhanced by an open landscape context, which consists of all semi-natural land cover except forest. In a closed landscape context that consists of forest only, however, maintaining heathland heterogeneity remains essential.

Butterflies are considered an umbrella taxon for insect conservation (e.g. van Swaay et al., 2006) and birds are often used as indicators of general habitat quality (De Bruyn et al., 2009). Yet, results among taxonomic groups, even among invertebrates, showed dissimilarities (Figures 13 and 14). Taxonomic groups respond to different components of 3D vegetation structure at different spatial scales (Atauri and De Lucio, 2001; Davies and Asner, 2014; de Vries et al., 2021; Tews et al., 2004) and we stress the importance of targeting multiple taxa at multiple scales for proper heathland management, especially in small patches and around patch edges. The impact of local vegetation structure on bird habitat suitability, for example, would not have been detected by large-scale measures of habitat heterogeneity and certainly not by those derived from coarse categorical land cover maps (Coops and Wulder, 2019). Additionally, possible benefits of edges can be higher for larger (birds) or more mobile (birds and butterflies) taxonomic groups (Pfeifer et al., 2017) as opposed to small and less mobile taxonomic groups (grasshoppers). For the latter, we noted an overall positive impact of vegetation structure which was also noted in an earlier study for Blue Winged Grasshopper and Mottled Grasshopper (Schirmel et al., 2011). Additionally, the “enemy-free space hypothesis” states that prey species prefer dense vegetation with a high structure to escape from predators (Price et al., 1980). This was found to be true for large carabid beetles (Brose, 2003) and can also count for the grasshoppers in this chapter.

Although pooling species into taxonomic groups revealed some patterns regarding the impact of the predictors on habitat suitability, individual species might respond differently to multiscale vegetation management. Conservation planners must, therefore, consider additional knowledge on habitat requirements of dry-heathland species, especially those of conservation interest. For example, bird habitat suitability was generally impacted positively by a high heathland heterogeneity, although this was less pronounced for European Nightjar (Caprimulgus europaeus) (Figure M.4). This bird species requires complementary habitats for foraging (extensive grasslands) and breeding (heathlands) which may be separated up to several kilometres (Evens et al., 2018). The size of those habitats and the landscape configuration and heterogeneity will likely be more important than the heterogeneity of habitat subtypes (Evens et al., 2021). Another example is the overall preference for a low vegetation structure in a closed landscape context for Silver-studded Blue. This preference was also detected in intermediate and large patches, as opposed to the other two butterfly species which preferred a high vegetation structure in larger heathlands (Figure M.5; Table N.1). Figure 12c showed that a low vegetation structure can indicate a location with characteristic dry heather shrubs. Considering that Silver-studded Blue uses Calluna vulgaris as a host plant (Diemont et al., 2015) and has relatively low mobility, this can explain the importance of low vegetation structure for this species.

Our results support that restoring and maintaining large and structurally complex habitats with patchy vegetation is a good approach for fauna conservation in heathlands (Byriel et al., 2023; De Blust, 2022; de Vries et al., 2021; Schirmel et al., 2011; van den Berg et al., 2001). The positive impact of an increased heathland size for most species is expected as habitat loss threatens biodiversity (Newbold et al., 2015) and positive relationships between quantitative measures of a species associated land cover or habitat type and occurrence are common, especially for habitat specialists (Fahrig, 2003; Milanesi et al., 2017; Rutten et al., 2019; van den Berg et al., 2001). Heathland enlargement becomes especially important in an anthropogenic landscape context (i.e. urban land use and agriculture), due to negative edge effects and low quality of the surrounding land cover for species of conservation interest (Fletcher et al., 2018; Newbold et al., 2015; Olivier et al., 2016). When large patches are located in an anthropogenic landscape context, however, increasing vegetation structure and heathland heterogeneity becomes even more important in comparison to a semi-natural context, as habitat heterogeneity must also be present in the heathland itself.

Although we could not assess the impact of the landscape context on habitat suitability parametrically due to correlations with all other predictors (Figure K.1), our results highlighted that heathland management needs to consider the landscape matrix in which fragmented heathlands are located. Traditionally, habitat quality was negatively associated with fragmentation (Hanski, 1998; MacArthur and Wilson, 1967), but the surrounding land cover might also increase habitat suitability for species that can benefit from edge effects (Dupont and Overgaard Nielsen, 2006; Evens et al., 2018; Pfeifer et al., 2017), although direct anthropogenic influences, such as nitrogen deposition from agriculture or industry should be avoided (Vogels et al., 2017). We further emphasize the importance of using multiple species from different taxa as conservation umbrella, which has become especially feasible in light of the unprecedented quantity of species occurrence data collected on citizen science data platforms, even from lesser-known taxonomic groups (Maes and Van Dyck, 2005). Comparing different management practices in depth was beyond the scope of this study, so we kept our recommendations general and mostly focused on translating our findings into suitable heathland management.

The analyses in this chapter revealed that large heathland patches had higher habitat suitability for all three studied taxonomic groups (birds, butterflies and grasshoppers), especially in an anthropogenic landscape context. Enlarging and connecting heathland patches (Piessens et al., 2005; Worboys et al., 2010) is, therefore, urgently needed. In regions with highly fragmented and isolated patches facing anthropogenic pressures, however, this can be challenging due to policy restrictions, budgetary limitations or land ownership (Diemont et al., 2015). In this light, it is essential to understand that even small patches can have adequate habitat quality for typical (threatened) heathland species when habitat heterogeneity and/or vegetation structure are sufficiently high. Nevertheless, if the landscape matrix allows it, increasing heathland area can be achieved by restoring heathland habitat, for example by cutting down (non-native) coniferous forests (Diemont et al., 2015).

Increasing heterogeneity in nitrogen-polluted heathlands is often realised by large-scale removal of above-ground vegetation (e.g. by clearcutting, machine cutting or burning) or of both vegetation and soil top layers (i.e. sod-cutting or choppering) (De Blust, 2022). Those large-scale and intensive management practices homogenise the vegetation cover and deplete nutrients from the soil, which is beneficial for restoring typical heathland vegetation (Jones et al., 2017; Schellenberg and Bergmeier, 2020), but can also have a detrimental effect on invertebrates and larger predators, such as birds, that feed on them (Maes et al., 2017c; Vogels et al., 2021, 2017). Therefore, intensified large-scale management practices should be avoided when possible, especially in and around areas where species of conservation interest are known to be present, however always with consideration of present endangered flora or habitat.

The proxy that was used to quantify vegetation structure characterizes mainly vertical complexity with ancillary information on horizontal plant diversity (Farwell et al., 2021), yet both components are inextricably linked. Increasing the vertical complexity of vegetation cover at smaller scales will automatically allow for more plant diversity and can be achieved relatively fast, for example by removing above-ground vegetation and preventing grass encroachment of bare soil by mosaic mowing, cutting trees or low-intensity grazing, while allowing patches to reach older successional stages (Byriel et al., 2023). Note that, even when mostly evaluated as a good management practice for maintaining high vegetation structure, grazing can take many forms, such as variation in herbivore species or grazing intensity, and thus cause various responses of heathland flora and fauna (Diemont et al., 2015; Fagúndez, 2013). Furthermore, we recognize that additional changes in soil biochemistry (e.g. by liming) might be needed for nutrient-polluted soils (Vogels et al., 2017).

While using a multivariate structural measure has been shown to outperform single components of vegetation structure for estimating species distributions and diversity (e.g. Brose, 2003; Farwell et al., 2021), it also complicated the interpretation of which aspect of vegetation structure (vertical complexity or horizontal diversity) impacted habitat suitability at small scales. Combining a continuous measure of structure derived from multispectral remote sensing data with LiDAR, for example, might help to disentangle the individual impact of the components of 3D vegetation structure (Bergen et al., 2009; de Vries et al., 2021; Moudrý et al., 2022). Future research can also include microclimate data at fine scales obtained from remote sensing (Zellweger et al., 2019). This can, for example, shed further light on the importance of vegetation structure for invertebrates in heathlands as regulator under climatic extremes (Maes et al., 2019c; Mantilla-Contreras et al., 2012; Schirmel et al., 2011; Schirmel and Fartmann, 2014).

We remain careful to generalise our definition of multiscale management to a ‘small versus large-scale approach’. We did find important indications that heathland size, the landscape context and taxonomy affect the scale at which heathlands are best managed, yet additional findings from a sensitivity analysis (where vegetation structure and heathland heterogeneity are quantified at different spatial scales) could further support management recommendations and might highlight some keystone structures (Tews et al., 2004) in heathland ecosystems. For example, a maximum radius of 250 metres (≈ 20 hectares) might have been more appropriate, as correlations between landscape metrics and species diversity at this scale were highest for both birds and invertebrates (Morelli et al., 2013; Schiegg, 2000; Schindler et al., 2013). A kernel size of 750 metres to quantify vegetation structure was also believed to be too coarse to fully capture habitat requirements of birds (Farwell et al., 2021). On the other hand, an exploratory analysis proceeding this study showed that heathland size and heathland heterogeneity quantified at different spatial scales were highly correlated, and different kernel sizes for quantifying vegetation structure also did not impact estimations of bird density in previous studies (Wood et al., 2013).

CHAPTER V. General Discussion

21. Summary of the results

21.1. Recommendations for data quality filtering of opportunistic citizen science data

Chapters II and III formulated recommendations for data quality filtering (also called stringent filtering) of opportunistic citizen science data (CSD). This allowed us to use CSD optimally for biodiversity conservation supported by species distribution models (SDMs). To our knowledge, this was the first extensive study on the quantity-quality trade-off in stringent filtering. Both the questionable data quality in CSD (Burgess et al., 2017) and the negative impact of low sample size on SDM performance, especially when dealing with presence-only data (Liu et al., 2019), are widely recognized issues. Several studies have assessed the impact of sample size using data with constant quality (Chefaoui et al., 2011; Stockwell and Peterson, 2002; van Proosdij et al., 2016), while other studies have assessed the impact of stringent filtering on model performance without controlling for sample size (Kamp et al., 2016; Steen et al., 2019). However, recommendations for filtering CSD remained relatively general.

BOX 2 shows clear and specific evidence-based recommendations for stringent filtering of opportunistic CSD based on an analysis of five and a half million opportunistic records from 255 species across four taxonomic groups in Chapter II (Figure A.2; Table C.1) and 91 species across three taxonomic groups in Chapter III (Table I.1). The results in Chapter II indicated that the impact of stringent filtering on model performance depended on the quality of the filtered data (i.e. the filter type used) and both the proportional reduction in sample size caused by filtering and the remaining absolute sample size. Additionally, results showed that plant and dragonfly models benefitted more from stringent filtering than bird and butterfly models yet with variation in the impact on model performance among species. Chapter III confirmed that taxonomy can guide filtering recommendations, but that species traits should also be taken into account.

The value of our research was further increased by using an external evaluation set for model testing. This reduced the risk of inflated model evaluation metrics (Elith et al., 2006), facilitated the comparison of model performance within one species (Elith et al., 2010) and enabled the assessment of three metrics for model evaluation (i.e. AUC, sensitivity and specificity).

BOX 2: THINK BEFORE YOU SHRINK

Recommendations for data quality filtering based on the results of an extensive analysis of the data quantity-quality trade-off (Chapter II) and the impact of species traits (Chapter III) in stringent filtering. Maxent performance (AUC, sensitivity, specificity) was compared before and after filtering opportunistic data of well-surveyed bird, butterfly, dragonfly (and plant[11]*) species (resolution = 1 km²; study area = 13,552 km²).*

ACTIVITY *removes records from less experienced observers based on an observer's average annual activity rate*

Use for widespread species that are **difficult to identify,**
for **familiar and widespread** butterflies and
to minimize commission errors, e.g. for monitoring invasive species.

DETAIL *removes records that were submitted without any additional information based on the presence of metadata beyond default requirements*

Use for **plants** and
for **familiar** species with a **high reporting probability.**

VALSTAT *removes doubtful and unevaluated records based on the verification status of a record in the data platform*

Use is **generally recommended**, **except for** familiar and widespread butterflies.
Use to minimize omission errors, e.g. for prioritizing conservation areas.

CAUTION is needed when data quality filtering **reduces sample size**
to **less than 100** presences and
by **more than 50 %** (for widespread species)
or 75% (for species with restricted home ranges)

It was recommended that the goal of the study should be kept in mind when applying stringent filters, yet this was not elaborately discussed in the previous chapters. When predicting suitable presence locations is the main interest, like for the delineation of conservation areas, false negatives (omission errors, i.e. predicting absences when the species is actually present) should be avoided (Lobo et al., 2008; Thomaes et al., 2008). We, therefore, concur with the common practice of using verified records for increasing the quality of opportunistic CSD (also see Table 2). However, for assisted monitoring, like for invasive species, false positives (commission

[11] Plant species were only considered in chapter II

errors, i.e. predicting a presence when the species is actually absent) should be avoided (Guillera-Arroita et al., 2015). Using observations from more experienced observers might then be more effective (Table 2). Note that in relatively small regions, such as Flanders, an occasional false positive observation of an invasive species might be less problematic (because it can be validated relatively easily) compared to larger regions and countries.

We encourage the further development and implementation of semi-automated verification systems and the collection of metadata on observer experience in large citizen science platforms (Table 5). Verification systems can exist of (combinations of) image recognition, automated validation within known spatial and temporal ranges and expert or user validation (Figure 1; Swinnen et al., 2018). Metadata on observers can be collected in the form of observer classifications (such as the gold stars in iRecord https://irecord.org.uk/how-do-i) or observer-specific information on their number of entries, reported species, misidentifications etc.. The latter can, for example, support the method for data quality filtering based on observer activity described in Chapter II (section 8.1 and Table A.1).

Table 5: Examples of citizen science data platforms that collect species occurrence data. The second column indicates which methods or systems are used to provide information on data quality and increase it (Image = automated validation by image recognition; Range = automated validation within a reasonable spatial and temporal range of a verified record; Expert = manual record validation by experts; User = community consensus by user validation). The third column indicates how metadata on observers are provided.

Platform	*Semi-automated validation system*	*Metadata on observers' experience*
Waarnemingen.be	Image, Range, Expert (BOX 1)	On request
iRecord	Partially integrated into the NBN Record Cleaner[12] (data cleansing, range) Image, User ("Research grade" records imported from iNaturalist)[13] Expert	Gold stars = observers' level of certainty
Artportalen	Expert, Image (under development)	On request
eBird	Image, Range, Expert	Metadata from checklist data (Kelling et al., 2019)
iNaturalist	Image, User Data quality assessment protocol (e.g. Aristeidou et al., 2021)	Metadata in user profiles (non-downloadable)
GBIF	Data quality requirements for data publishers Automated data cleansing with Issues and Flags system[14]	Only at the level of the data publisher

[12] https://nbn.org.uk/tools-and-resources/nbn-toolbox/nbn-record-cleaner/
[13] https://irecord.org.uk/linking-inaturalist
[14] https://data-blog.gbif.org/post/issues-and-flags/

21.2. Recommendations for multiscale vegetation management in dry heathlands

Recent trends in biodiversity monitoring have made it possible to model species occurrence data at fine scales. In Chapter IV, opportunistic CSD and multispectral satellite data were integrated into SDMs to study the impact of heathland size, vegetation structure and heathland heterogeneity on the habitat suitability of dry-heathland fauna in three different landscape contexts (i.e. closed semi-natural, open semi-natural and anthropogenic context). Using a Point Process Model (PPM) with a Geyer saturation process (Baddeley et al., 2015; Renner et al., 2015) allowed us to make use of the fine resolutions at which species occurrence data (i.e. point observations with a precision of 50 metres or less) and remote sensing data (i.e. 10 x 10 metres) were available. This led to additional evidence of the importance of vegetation structure for heathland fauna at multiple scales. The importance of vegetation structure is widely recognized but often based on professional field experience or the habitat heterogeneity hypothesis (e.g. MacArthur and Wilson, 1967). Most studies link habitat heterogeneity to species diversity and have been tested at larger resolutions (Tews et al., 2004), although studies have been using LiDAR data for small-scale assessments as well (e.g. de Vries et al. (2021)). At larger scales, we confirmed previous findings on the importance of high structural complexity (e.g. Graf et al., 2009; Huber et al., 2016; Seavy et al., 2009), however, at small scales in isolated patches, we found negative species-environment relationships (Figures 13 and 14; section 20) (Stein et al., 2014). These results highlighted the importance of multiscale assessments of habitat heterogeneity over large spatial extents.

Integrating opportunistic CSD and remote sensing data is a promising advance in biodiversity conservation monitoring. By including heathlands of different sizes and the landscape context, management recommendations for fauna conservation could be formulated in highly fragmented landscapes. In summary, we recommend restoring and maintaining large and structurally complex heathlands with patchy vegetation. Conservation should also include action plans to connect fragmented heathlands (e.g. by cutting down pine plantations). When sufficient natural resources are available in the direct (semi-natural) environment of smaller heathlands (e.g. for foraging or nesting), heathland management plans should also emphasize the importance of maintaining characteristic dry-heathland shrub vegetation. In anthropogenic landscapes, on the other hand, simply increasing the structural complexity of patches without increasing their surface area might not be enough to avoid the local extinction of species of conservation interest. This is probably the largest challenge for conservation management, as enlarging heathlands in human-dominated landscapes will need the field experience of local managers, the sensitisation of citizens and additional actions and funding at the policy level.

22. Important considerations

In addition to the limitations of working with opportunistic presence-only data in SDMs introduced in Chapter I (sections 2.2 and 3), this section highlights some important considerations for the implementation of this research.

22.1. Correlation and causality

The goal of Chapter IV was to support management recommendations in habitats under anthropogenic pressures using fine-scaled opportunistic and quantitative remote sensing data. Inferring causal relationships from correlative SDMs (like in Chapter IV) is controversial as correlation does not necessarily imply causation (Box, 1966). This can cause problems when environmental covariates are correlated, which is called multicollinearity (Dormann et al., 2013). For example, Dormann et al. (2012) pointed out the potential danger of multicollinearity in correlative models for predictions under future scenarios when variables are correlated at present but not necessarily so in the future. Arif and MacNeil (2022) highlighted the inappropriateness of model selection techniques based on Akaike's Information Criterion (AIC; Burnham et al., 2011) when the aim is to select the covariates that best explain the data. Furthermore, Kühn (2007) warned of flips in covariate signs when covariates are ill-specified or correlated with an unspecified environmental gradient that impacts the observed pattern of species occurrence.

It is a misconception, however, that functional relationships among species occurrence and environmental parameters cannot be derived from SDMs (Arif and MacNeil, 2022). It is, however, essential to choose ecologically plausible parameters, check for multicollinearity, formulate appropriate hypotheses and include methods to account for spatial autocorrelation (Arif and MacNeil, 2022; Dormann et al., 2012; Kühn, 2007). Additionally, it is important to distinguish between ecological patterns in species occurrence data and patterns related to sampling bias. In Chapter IV, the selected parameters (i.e. heathland size, vegetation structure, heathland heterogeneity and the landscape context) exhibited no multicollinearity. Furthermore, these parameters were all proven or known (through professional expertise) to affect habitat quality (e.g. de Vries et al., 2021; Dupont and Nielsen, 2006; Piessens et al., 2005) so that relevant hypotheses could be formulated (section 17). The methodology also accounted for sampling bias and spatial autocorrelation by (1) applying spatial filtering at 50 metres for records on the same date, (2) using a Geyer saturation process to model spatial dependence (Baddeley et al., 2015), (3) choosing two bias covariates for bias mitigation (Warton et al., 2013) and (4) performing cross-validation in spatial-block design (Valavi et al., 2019). We argue that the predictions of the trend of the selected models were sufficiently validated by both their ecological plausibility and a reasonable model fit (sections 19 and 20). As residual spatial autocorrelation can increase type I errors (i.e. falsely rejecting the null hypothesis) but seldomly reverse the sign of a coefficient (Dormann et al., 2007; but see, e.g., Kühn, 2007), we are also confident that, by accounting for sampling bias, the interpretation of the relative impact of the covariates on relative habitat suitability was valid.

We acknowledge, however, that true causality between species occurrence and environmental covariates cannot be proven (Arif and MacNeil, 2022) and that our variables might be correlated with other unspecified environmental drivers such as soil water or microclimate (see section 23.1.3). Nevertheless, we maintain that we used the best available methods and refer to previous studies that employed similar methodologies to support our findings (e.g. De Solan et al., 2019). A related issue, and the reason why we did not model plant species in Chapter IV, is the challenge of explaining plant species occurrence using quantitative remote sensing predictors, as such measurements are influenced by both habitat quality and the spectral characteristics of the vegetation (Bradley et al., 2012).

22.2. Transferability of the results

In the two parts of the dissertation (Chapters II and II versus Chapter IV; Figure 4), we used two different presence-only SDMs for different species at different spatial scales and with different methods for bias correction. This was motivated by the different objectives of the study, i.e. performing a large-scale assessment of the change in model performance of a presence-only SDM before and after filtering (Chapters II and III) and performing a multiscale assessment of the impact of several predictors on the habitat suitability of species of conservation interest (Chapter IV). However, this raises two questions regarding the transferability of this research:

i. How generic are the filtering recommendations from Chapters II and III and, more specifically, do they apply to the methods in Chapter IV?
ii. Could we not have used the same methods for all three Chapters (II, II and IV)?

We will discuss these questions through three topics: (1) methods for modelling and bias correction, (2) species and sample size and (3) scale and study area.

22.2.1. Methods for modelling and bias correction

Chapters II, III and IV all supported the main objective of the research, i.e. to reduce the uncertainty associated with opportunistic citizen science data in biodiversity conservation applications (section 5). However, we argue that the specific objectives (improving predictive performance vs. causal inference) and potential applications (e.g. delineation of conservation areas vs. management recommendations) imposed limitations that caused a choice of different methods for modelling and bias correction. Choosing the same methods might have led to suboptimal recommendations for the intended conservation applications. Additionally, we refer to the interpretation of Maxent as an inhomogeneous PPM and argue that the implemented modelling methods were similar apart from the (important) assumption of spatial independence in Maxent (Renner and Warton, 2013).

Maxent (Phillips et al., 2006) was used in Chapter II (and III), as the objective was to assess the relative predictive model performance and Maxent is, to this date, still evaluated as among the best-performing and most-used SDM methods for presence-only data (Elith et al., 2006; Valavi et al., 2022). This choice for this method, however, imposed some limitations regarding spatial scale and bias correction methods. Maxent assumes spatial independence (section 3.1; Renner and Warton, 2013), so sampling bias had to be reduced before modelling to avoid problems associated with spatial autocorrelation (section 3.2.2; Kramer-Schadt et al., 2013). We chose a resolution of one kilometre for spatial filtering which was a good trade-off between reducing sampling bias and an appropriate resolution for conservation policy applications on a regional scale such as Flanders (e.g. Demolder et al., 2014; Rutten et al., 2019; Vantieghem et al., 2017). We acknowledge that other bias correction methods could have been tested, such as background manipulation (Phillips et al., 2009; Vollering et al., 2019) or bias covariates (El-Gabbas and Dormann, 2017; Warton et al., 2013) (section 3.2.3). However, this would have jeopardized the comparability of our models (section 8.3; Merow et al., 2013) and we chose to use the same background for all models instead. Another important limitation of Maxent, and presence-only SDMs based on opportunistic CSD in general, is that they can only provide a ranking of the relative habitat suitability of locations when information on detectability is unknown (section 3.3; Guillera-Arroita et al., 2015).

A Gibbs Point Process Model (PPM) (Baddeley et al., 2015; Renner et al., 2015) was used in Chapter IV for two reasons. First, the objective was to support habitat management by analysing the impact of specific environmental predictors on relative species occurrence (Fithian and Hastie, 2013). The interpretation of coefficients and their interactions in PPMs is more straightforward compared to Maxent because the former are fitted by linear regression methods and the latter by maximum entropy methods, using features and tuning parameters to optimize model fit (Merow et al., 2013; Phillips et al., 2017). Second, the objective included an assessment of the impact of a fine-scaled habitat feature (i.e. vegetation structure) on relative habitat suitability. Therefore, we had to increase the resolution and could not apply spatial filtering like in Chapter II. The Geyer process combined with bias covariates offered a good solution as spatial interaction between points (Baddeley et al., 2015) and spatial variation in sampling (Warton et al., 2013) were accounted for.

22.2.2. Species and sample size

The selected methods and species had implications for the minimum sample size considered in the different chapters of the dissertation. In Chapter II, species were selected by setting a minimum sample size of 100 presences to ensure comparability among results, i.e. by setting identical Maxent settings and ensuring that all three filters (i.e. ACTIVITY, DETAIL, VALSTAT; BOX 2) and their combinations could be compared per species for at least one level of sample size (section 8.4; Figure A.3). In Chapter IV, we selected dry-heathland fauna of conservation interest as these are target species for biodiversity conservation policy. The sample size was set to a minimum of 60 presences because of model fitting problems below this threshold (section 18.4.1). A consideration when using small sample sizes is that they tend to generate models with low statistical power (Chefaoui et al., 2011), hence significant covariate effects can imply overestimations of that effect (Yang et al., 2022). We highlight, however, that we did not base our conclusions on the statistical significance of the covariate effects in the PPMs but rather on the sign of the effect and their relative magnitude.

Due to the different SDM methods (Maxent versus Gibbs Point Process Models), recommendations regarding data quality filtering in Chapter IV could not rely on sample size. However, as Maxent and PPMs respond similarly to sampling and detection bias (Guillera-Arroita et al., 2015), we could expect a similar impact of data quality filtering when spatial dependence was accounted for. Therefore, we adopted the recommendations based on species traits (i.e. a restricted home range, relatively small body size and taxonomic group) and decided to use only verified records. Additional analysis on the impact of stringent filtering at a smaller scale and targeted at species of conservation interest could be valuable. In retrospect, as we only compared the single filters in Chapter III, the list of species used for the species profiles could have been expanded to species with more restricted home ranges. However, we do not expect changes in filtering recommendations because the positive effect of using correctly verified data for species with restricted home ranges was relatively important (Figure 10) and their models have higher transferability (Wogan, 2016).

22.2.3. Scale and study area

A different scale (i.e. the extent of the study area and the grain of the model) was used in Chapters II and III versus Chapter IV. In Chapters II and III, models were run for Flanders at a resolution of 1 kilometre, while in Chapter IV, models were run in the Campine region (the northeastern part of Flanders) at a resolution of 50 metres. The choice for scale was motivated by the objectives of the respective studies and the best options to meet them, as explained in section 22.2. While running the PPMs at larger scales would have resulted in a loss of information, running Maxent at smaller resolutions would have been feasible. We argue, however, that this could violate the assumption of spatial independence (Renner and Warton, 2013).

We repeatedly chose the same scales for all species to facilitate comparability among results. However, we acknowledge that information on the dispersal ability and/or mobility of the species could be used to further improve model predictions (Chapter II) or to gain additional insight into species-environment relationships (Chapter IV). Supposed that spatial dependence was not an issue, increasing the resolution (especially for butterflies, dragonflies and plants) towards one hectare rather than one squared kilometre (100 hectares) might have led to different recommendations for data quality filtering (Chapters II and III). Remember that the proportion of high-quality data in a model training set is scale-dependent because a coarse resolution gives a higher chance that at least one high-quality observation falls in a grid cell (section 10). Data quality filtering of the same dataset, but aggregated at a finer resolution, might hence cause a larger reduction in sample size. Whether this will impact the thresholds at which the data quality-quantity trade-off becomes unfavourable, has yet to be explored.

Section 20 in Chapter IV already mentioned that additional findings from a sensitivity analysis (where vegetation structure and heathland heterogeneity are quantified at different spatial scales) could further support management recommendations and might highlight some keystone structures (Tews et al., 2004) in heathland ecosystems. Although we do not expect major changes in the impact of heathland size and heathland heterogeneity, gathering more fine-scaled data on vegetation structure with, for example, LiDAR and/or microclimate sensors could definitely lead to new insights (see section 23.4.1).

We assumed that a species responded uniformly to the environmental gradients throughout the study regions as Flanders has limited geographical and environmental gradients (e.g. 240 km across, 0 to 288 m elevation and relatively uniform climatic conditions) and the Campine region is a region with similar biotic and abiotic conditions (Klijn and de Haes, 1994). We acknowledge, however, that on larger scales, the impact of climatic variables on habitat suitability becomes more prominent and species populations might respond differently to similar local environmental conditions (Chen et al., 2020). For example, heathland butterflies responded similarly to environmental conditions within the Campine region (Vanreusel et al., 2007) but might respond differently to spatial structure in other regions (De Ro et al., 2021; Schirmel and Fartmann, 2014).

The applied methods in Chapter IV might not be transferable to every habitat type, as quantifying habitat heterogeneity and vegetation structure possibly need different approaches or considerations. For example, habitat heterogeneity in farmland is not only impacted by variability in habitat subtypes (such as arable land, cultural grassland and orchards) but also by crop configuration and composition (Fahrig et al., 2011) and by the presence of small landscape features such as hedgerows or flower strips (Dochy, 2014).

23. Future research

Throughout the dissertation, different suggestions have been made for expanding this research. To improve filtering recommendations of opportunistic CSD, we encourage the implementation of the methods used in Chapters II and III at different scales (taking into account the limitations of sampling bias in opportunistic CSD), in other presence-only (or integrated) SDMs and for more citizen science databases. This will help to formulate both generic and specific filtering recommendations and increase the uptake and value of opportunistic CSD in biodiversity conservation applications.

We suggest the expansion of our analysis on the impact of multi-scale vegetation heterogeneity on habitat suitability to other habitat types and associated species and encourage the integration of habitat area size and the landscape context as environmental variables. With the increasing availability of fine-resolution data on species occurrences (through CSD) and quantitative measures of habitat quality (through remote sensing), we also believe that our methods can be used at even finer resolutions across large extents. The next sections formulate some additional suggestions regarding modelling methods and additional environmental predictors.

23.1. Integrated species distribution models

Integrated SDMs were introduced in section 3.2.3 as a method to mitigate bias. In the current section, we use the term "integrated SDM" for a model that incorporates submodels for different types of data. These submodels estimate shared parameters by utilizing a joint likelihood or correlation structure. Integrated SDMs typically employ a point process framework (Renner et al., 2015) in either a single-species (e.g. Dorazio, 2014) or multi-species (e.g. Botella et al., 2021; Fithian et al., 2015) approach. This framework facilitates the integration of various data types and allows for more robust modelling. We concur with the notion that data integration methods hold immense potential for fully harnessing the benefits of opportunistic presence-only CSD (Johnston et al., 2023). By employing data integration methods, model performance can be improved, and bias can be reduced in comparison to using single presence-only SDMs. For a comprehensive overview of data integration methods, we recommend referring to the reviews conducted by Miller et al. (2019) and Isaac et al. (2020).

In recent years, the number of checklist observations in *waarnemingen.be* has been growing exponentially and recently reached the milestone of three million records (Figure 2). Therefore, we particularly want to highlight methods that integrate semi-structured checklist data (presence-absence data) with opportunistic presence-only data (e.g. Dorazio, 2014; Fithian et al., 2015; Pacifici et al., 2017). Using integrated SDMs becomes especially interesting when few presence-absence data are available, for example at a ratio of structured to unstructured data of less than 5 % in a study by Simmonds et al. (2020). However, simply integrating unbiased presence-absence data will not be sufficient to improve model predictions and also the presence of unknown biases can significantly impact the performance of integrated SDMs (Simmonds et al., 2020; Suhaimi et al., 2021). Possible solutions to overcome these limitations include weighted joint likelihoods to account for sample size differences (Fletcher et al., 2019), adding terms that quantify sampling bias, such as bias covariates (Bradter et al., 2018) or a flexible spatial term (Simmonds et al., 2020), or using correlation methods (Suhaimi et al., 2021).

23.2. Joint species distribution models

Limitations or opportunities imposed by co-occurring species in the studied regions were not considered. Stacking model predictions without consideration of these effects might lead to over- or underestimations of species richness (Clark et al., 2014). Techniques such as joint SDMs can be explored to predict the suitability of locations for multiple species more accurately (Ovaskainen et al., 2010; Ovaskainen and Soininen, 2011; Pollock et al., 2014). Joint SDMs often rely on a hierarchical framework, such as the flexible Hierarchical Modelling of Species Communities (HMSC) proposed by Ovaskainen et al. (2017). However, they only recently started to include presence-only data (Escamilla Molgora et al., 2022), which is a promising advance for community modelling when no or little structured data are available.

23.3. Include temporal aspect

Further research on the impact of data quality filtering on model performance could, for example, assess whether data quality changed over the years and between seasons. Data quality might have increased over time due to increased participation and experience, but at the same time, mobile applications have made citizen science platforms more accessible to a broad public. This leads to more observers with low expertise and experience (i.e. lower ACTIVITY; BOX 2) and fewer observations submitted with accompanying metadata (i.e. lower DETAIL; BOX 2), which might both affect data quality negatively (Chapters II and III).

We assumed that a species' response to environmental variables was constant across the studied period (section 8.3; section 22.2.3) and temporal aggregation of records was performed to reduce bias (section 3.2.2). However, temporal patterns might be missed by our methods as species respond to annual and seasonal changes in landscape and climate (Zurell et al., 2009) and different acquisition dates of remote sensing images can impact model performance (Bonthoux et al., 2018; Sheeren et al., 2014). Methods that capitalize on the high temporal and spatial resolution of remotely-sensed variables include phenological predictors (see section 23.1.3.d), seasonal SDMs (Oeser et al., 2020), multi-state SDMs (Frans et al., 2018) and dynamic SDMs (Milanesi et al., 2020). Seasonal and multi-state SDMs account for the fact that a species occupies different habitats at different lifecycle stages, such as breeding, nursing, and overwintering. However, the resolution of the data must be much smaller than the species' migration area, so multi-state SDMs (and to a lesser extent, seasonal models) are primarily suitable for species with high data availability, long behavioural periods, and large dispersal areas, such as GPS-tracked large mammals (Frans et al., 2018; Oeser et al., 2020) or birds

(Vanermen et al., 2020). Dynamic SDMs either average individual models from different periods or link species occurrences to the environmental situation at a specific time range (Milanesi et al., 2020). Nonetheless, few attempts have fully exploited the time series and temporal dynamics of remotely-sensed variables (Randin et al., 2020). Although these methods seem promising, their applicability to opportunistic CSD may be restricted by the higher incidence of bias in temporally disaggregated data.

23.4. Additional predictors

The selection of model predictors was motivated by the objectives in the different chapters, the characteristics of the study area and the need for comparability in Chapters II and III. However, the following predictors could be explored in further research.

23.4.1. Microclimate

Microclimates are an important driver of local species occurrence, as differences in local land cover and topography can create ecological conditions that differ from the average macroclimate conditions measured by common weather stations (Lembrechts et al., 2019). Microclimates have gained renewed attention for species distribution modelling with the increased accessibility of high-resolution remote sensing data, such as LiDAR and hyperspectral data (Zellweger et al., 2019). They can, for example, be used to interpolate data obtained from microclimate sensors (Lembrechts et al., 2020; Zellweger et al., 2019).

Vegetation structure impacts microclimate in heathlands (section 20; Barclay-Estrup, 1971; Mantilla-Contreras et al., 2012; Schirmel et al., 2011; Schirmel and Fartmann, 2014). However, capturing fine-scaled variations in near-surface temperature and humidity for low-stature habitats across large extents remains a challenge (Maclean et al., 2021; Zellweger et al., 2019) and research is currently ongoing. When available, microclimate data can be adopted in SDMs to further analyse the relationship between vegetation structure and habitat suitability for heathland species (Chapter IV; Maes et al., 2019b; Mantilla-Contreras et al., 2012; Schirmel et al., 2011; Schirmel and Fartmann, 2014). Moreover, using fine-scaled species occurrence data in Gibbs point process models is a promising strategy to adopt microclimate data in SDMs for predicting relative habitat suitability across large spatial extents. This can be important for delineating conservation areas, especially in light of climate change (Lenoir et al., 2017).

23.4.2. Soil parameters

We included soil texture in the predictor set in Chapter II as combining land cover and soil data is known to improve predictions of species distributions (Titeux et al., 2009). While at these coarse scales, we believe that large differences in soil type are reflected by the difference in land cover, we do admit that a measure of soil water could have improved predictions. It is unlikely, however, that the assessment of relative model performance would have been impacted by adding one predictor.

In Chapter IV (section 18.3), it was noted that soil predictors could be tested, but with consideration of possible collinearity with other predictors. We acknowledged that including measures of soil water or soil biochemistry, such as nitrogen (N) and phosphorus (P) content (Vogels et al., 2017), might have led to additional insights. Those insights can support integrated soil-vegetation management, such as the combination of sod-cutting and P addition. Sod-cutting removes both vegetation and soil top layers, which depletes nutrients from the soil and homogenises the vegetation cover (De Blust, 2022). While this is beneficial for restoring typical heathland vegetation (Jones et al., 2017; Schellenberg and Bergmeier, 2020), the induced P limitation affects the nutritional quality of plants for invertebrates and by consequence also for larger predators, such as birds, that feed on them (Vogels et al., 2017, 2021). The methods presented in Chapter IV offer a way to investigate this effect for multiple species at fine scales and large extents.

23.4.3. Landscape metrics

Among the most used landscape metrics in SDMs are metrics related to landscape heterogeneity (i.e. the spatial variation in habitat types or land cover classes) due to its positive relationship with biodiversity, especially at large scales (Stein et al., 2014; Tews et al., 2004). We warn, however, that a high landscape heterogeneity might be a measure of high fragmentation in anthropogenic regions such as Flanders (Maes et al., 2022) and thus might have impacted species occurrence both positively and negatively in Chapter II.

Habitats contain various resources that appeal more to some species than others (Stein et al., 2014; Tews et al., 2004), hence integrating within-habitat heterogeneity in Chapter II might have improved predictions. We argue, though, that using one measure of within-habitat heterogeneity for all habitat types is not appropriate in a coarse-scale analysis. Heterogeneity in forests (De Frenne et al., 2021), for example, will not have the same effect on species occurrence as heterogeneity in heathlands or grasslands (Bar-Massada and Wood, 2014; de

Vries et al., 2021). Moreover, abiotic differences such as soil type can impact which habitat sub-types occur or dominate (Thoonen et al., 2013).

For studies at smaller scales, however, including more landscape metrics could have helped the interpretation of our results. For example, many of the findings in Chapter IV related to edge effects, hence edge metrics such as edge density (Lustig et al., 2017) could be used in future studies.

23.4.4. Remotely-sensed ecosystem functioning attributes

Using remotely-sensed ecosystem functioning attributes (EFAs) as a predictor in SDMs has been gaining attention as they can capture responses to changes in habitat much earlier than climatic or landscape variables (Mouillot et al., 2013) and can be used in the framework of the essential biodiversity variables (EBVs) (Alcaraz-Segura et al., 2017; Pereira et al., 2013). EFAs have been shown to successfully predict annual range shifts (Alcaraz-Segura et al., 2017), habitat suitability and abundance for protected plant species (Arenas-Castro et al., 2019, 2018; Vila-Viçosa et al., 2020) and bird distributions (Regos et al., 2019), yet the latter with low temporal transferability (Regos et al., 2020).

EFAs include, for example, land surface temperature (LST) (Arenas-Castro et al., 2018), Albedo (Regos et al., 2020) or indicators of seasonal dynamics quantified by summary statistics of the Enhanced Vegetation Index (EVI) (Alcaraz-Segura et al., 2017; Arenas-Castro et al., 2019, 2018) and Normalized Difference Water Index (NDWI) (Vila-Viçosa et al., 2020). In their definition as EFAs, their use is relatively new, but note that many of these variables have been used before in SDMs (Cord and Rödder, 2011). We encourage further research that includes EFAs obtained from high-resolution sensors (e.g. Sentinel-2) and that validates their use as model predictors for multiple species and species groups. We concur with their esteemed potential for biodiversity conservation applications (Arenas-Castro et al., 2019, 2018) due to their high spatial and temporal coverage and cheap collection.

23.4.5. Binary predictors of land cover

A recent study demonstrated that the probability of occurrence does not necessarily increase with increasing habitat size because it might be enough to have a certain amount of habitat to support a vital population (Gábor et al., 2022). In Maxent, this could easily be integrated by allowing threshold features (Merow et al., 2013). It would be interesting to see (i) whether such responses are detected by a presence-only algorithm (as Gábor et al. (2022) used presence-absence data) and (ii) how they impact model performance. For example, jumps in the response curve were detected when using non-linear features in Maxent for an endemic plant species in South Africa (Merow et al., 2013), while in a study on rare squirrels in Florida, conclusions did not change (Tye et al., 2017). Additionally, threshold features are preferably used based on existing knowledge of a species' ecology as they quickly overcomplicate model interpretation (Merow et al., 2013).

CHAPTER VI. Application Potential

24. General application potential

This research has diverse application potential in biodiversity conservation policy and can guide end-users of large citizen science platforms that collect biodiversity data. SDMs are increasingly used to support conservation policy in different domains (section 4.2). When the goal is to predict the potential habitat suitability of species, Chapters II and III offered a set of recommendations for data quality filtering. When the goal is to study relationships between species occurrence and environmental data at fine resolutions, Chapter IV provided an example study, where fine-scaled data on species occurrence and vegetation characteristics were used in a point process setting to support evidence-based habitat management. While the first application is more interesting for conservation strategies at coarse scales (i.e. delineation and prioritization of areas for biodiversity conservation and monitoring), the second is also relevant to small-scale conservation practices (i.e. supporting habitat management) (also see section 22.2). We stress that the potential habitat suitability maps and management recommendations are mostly targeted at short-term biodiversity conservation efforts. Long-term future risk assessments, for example by extrapolating model predictions under the Representative Concentration Pathway (RCP) scenarios for climate change (e.g. Van Daele et al., 2021), were beyond the focus of this research.

There is a growing consensus that opportunistic CSD can make valuable contributions to biodiversity conservation, if processed correctly (Chapters II, II and IV; Henckel et al., 2020; Soroye et al., 2018; Van Strien et al., 2013). Ideally, the strengths of both opportunistic and structured survey data should be combined (Fletcher et al., 2019; Henckel et al., 2020; Simmonds et al., 2020). While that was not the focus of our research, we contributed to a study on allergenic tree species in which structured data and relative habitat suitability maps based on opportunistic data were combined to improve abundance estimations in urban areas of Wallonia (the southern region of Belgium) (Dujardin et al., 2022).

Furthermore, the value of this research, as with much ecological research, is only realised when science, policy and practice meet in a transparent way (Parker et al., 2016; Wood et al., 2018). Still too often, research does not find its way into policy and policy measures are evaluated as ineffective after significant investments (Downey et al., 2021) or even international policy reforms such as the Greening of the Common Agricultural Policy (Lakner et al., 2019; Pe'er et al., 2017). We have shown that with relatively simple, accessible and cheap methods, evidence-based recommendations for biodiversity conservation policy and management can be supported. These findings can reach practitioners in the field, for example, by training them to interpret scientific results (Downey et al., 2021), by using so-called 'evidence bridges' (Kadykalo et al., 2021) who can also provide feedback to scientists on possible caveats, or by publishing scientific studies into local nature journals such as *Natuurfocus* in Flanders. Nevertheless, the expertise of all parties must be treasured (Molnár and Babai, 2021). A scientist, for example, has a deeper understanding of statistical methods and can make well-founded choices among different methods to reach a specific goal. A practitioner, on the other hand, has priceless field experience and will be better at estimating the feasibility of certain applications, while, lastly, policymakers have more insight into the socio-economic implications.

Considering the potential applications for presence-only SDMs in conservation which were introduced in Chapter I (section 4.2) and the known limitations of opportunistic CSD and presence-only SDMs (sections 3.3 and 22), this chapter will elaborate on possible applications in Flanders.

25. Application potential in Flanders

25.1. The current state of biodiversity (policy) in Flanders

Flanders (the northern region of Belgium) is an area of 13,625 km² characterized by a high population density (492/km²), intensive anthropogenic land use (46% agriculture and 29% urban areas) and a high fragmentation of the remaining semi-natural areas (https://www.statbel.fgov.be; Maes et al., 2022). Biodiversity has been suffering tremendously under these pressures which, in combination with climate change, have led to serious declines in species and populations and brought almost a third of the species in Flanders on a IUCN Red List (Schneiders et al., 2020).

In Flanders, nature conservation policy is primarily built around the Natura 2000 network (Figure 15). The protected area currently covers 1663 km² (12% of the total area) (Schneiders et al., 2020) of which 940 km² (7% of the total area) is under conservation management (Vught et al., 2020). Additionally, *Natuurpunt*, one of the largest NGOs working in the nature sector in Flanders, also happens to be one of the largest private landowners, contributing another 200 km² of protected areas.

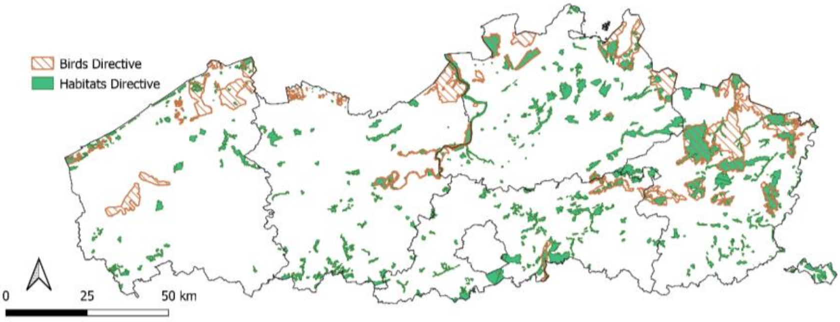


Figure 15: The Natura 2000 network in Flanders, including the special areas of conservation (SACs) under the Birds Directive (2009/147/EG) and the special protected areas (SPAs) under the Habitats Directive (92/43/EEG) (situation on 22/07/2005 and 18/01/2013 respectively)[15]. *The SACs and SPAs respectively cover 7% and 8% of the total area (with an overlap in 3% of the total area).*

Following the EU Biodiversity Strategy for 2020, the Flemish Natura 2000 program should realize a set of conservation goals before 2050 (Agentschap voor Natuur en Bos, 2017). This program included management plans for special protection areas (SPAs) and special areas of conservation (SACs), nature development plans, plans to reduce negative pressures on biodiversity such as the programmatic approach to nitrogen, species protection plans and EU LIFE projects. Every six years, intermediate goals are assessed and reported to the EU. Unfortunately, the most recent report showed that none of the six targets of the EU Biodiversity Strategy for 2020 were achieved (Schneiders et al., 2020). At the start of the new EU Biodiversity Strategy for 2030, governing bodies in Flanders are challenged to find ways to expand and connect the existing Natura 2000 network and enhance the protection and restoration of biodiversity directed by the anticipated Nature Restoration Law (European Commission, 2022). To reach these goals, it is now, more than ever, time to join forces with citizens, scientists and land owners.

Historically, the designation of nature conservation areas in Flanders (for example, the Natura 2000 areas in 2001 or the Flemish Ecological Network in 2003) was largely biotope-driven and highly defined by socio-economic factors and politics. It was only in 2005, that the first study

[15] Retrieved from https://www.geopunt.be/ on the 14th of March 2023

highlighted the importance of considering species distributions for biodiversity conservation policy (Maes et al., 2005). The uptake of data on species occurrence in policy research has since evolved gradually and developments such as, for example, the GeoDynamiX toolbox (https://vito.be/en/product/geodynamix-spatial-modelling-tools) and the standardisation of Red List criteria (Maes et al., 2019b) have supported many technical reports (e.g. Maes et al., 2015, 2019b). New initiatives, such as the design of the structured survey protocols for priority species "MEETNETTEN" (Westra et al., 2016), will further expand policy applications of species occurrence data.

Structured data collected in systematic surveys might outperform opportunistic CSD as input for SDMs, yet such high-quality data is often not available (Wood et al., 2018). Today, CSD platforms can deliver millions of species records with high geographic precision, which can be combined with remote sensing data to facilitate biodiversity conservation and monitoring on a much larger scale and with greater efficiency (Leitão and Santos, 2019). While some may be sceptical about using uninformed citizen scientists for data collection, there are additional benefits to this approach. For example, these individuals collect data without prior knowledge of the species distribution or ecology, which can be useful for detecting early range expansions or behavioural changes (Broman et al., 2014).

The role of opportunistic CSD in Flemish nature policy is acknowledged, yet their full potential remains untapped. Currently, CSD are primarily used to support more established methods and uncertainty is generally reduced by using general practices such as the use of verified data or spatial filtering of records to reduce sampling bias (Chapter II). The Research Institute for Nature and Forest (INBO) developed mechanistic models to map potentially suitable habitats for 245 species of conservation priority (Maes et al., 2017a). The maps indicate where species can occur at a 20-metre resolution (1 = present, 0 = absent), without guarantee of actual species presence. To finetune these potential habitat suitability maps, opportunistic CSD have been used to delineate areas that are relevant for the modelled species by taking into account (opportunistic) presence records and dispersal distances (Maes et al., 2016). Our research aims to facilitate the integration of opportunistic CSD and SDMs in biodiversity conservation policy, particularly in two domains that were earlier identified in Chapter I (section 4.2): the delineation and prioritization of areas for conservation and monitoring; and habitat management.

25.2. Examples of potential applications in Flanders

To illustrate the application potential with examples, we ran SDMs for 33 species (Table O.1) that were either Flemish priority species or bound to farmland. We used opportunistic presence-only data to run Maxent models at a 500-metre resolution in Flanders with 15,000 randomly selected background points. Opportunistic presence-only records were retrieved from the *waarnemingen.be* database for the period 2018-2022. Presences were cleansed (removing bad coordinates, wrong observations and observations with a precision of more than 250 metres), spatially thinned and quality filtered (Chapters II and III). The predictions from the resulting models were stacked to generate maps with biodiversity scores, i.e. weighted sums of model predictions from different species (in this example, weighted by their red list status (Demolder et al., 2014)) (Figures 16 and 17; Table O.1).

25.2.1. CASE STUDY 1: Biodiversity 2030

The most recent EU Biodiversity Strategy states that by 2030, at least 30% of the European land area must be protected, including 10% strictly protected areas (European Commission, 2020). Member states are free to choose how to implement and monitor these guidelines in their respective territories and are faced with the challenge of enlarging and connecting the existing Natura 2000 network. This challenge is especially prominent in fragmented and anthropogenic regions such as Flanders (Maes et al., 2022).

Correlative SDMs have some advantages to the mechanistic approach (GeoDynamiX toolbox) described earlier. First, these SDMs are based on actual species occurrences that reflect a species' realized niche rather than its fundamental niche (Lobo et al., 2010). This implies that, when model predictors have a longitudinal/latitudinal gradient, dispersal will automatically be taken into account (Sillero, 2011). Second, locations will be ranked according to the predicted habitat suitability so that policy-defined thresholds for conservation can easily be met. As a first case study, we have delineated protected and strictly protected areas in Flanders by indicating respectively 30% and 10% of the total land area with the highest biodiversity scores (Figure 16). In this example, biodiversity scores were derived from the stacked Maxent predictions of 14 Flemish priority species (De Knijf et al., 2014; Herremans et al., 2014). The map with biodiversity scores was overlayed with the existing SPAs under the Habitats Directive (92/43/EEG) to check for overlap and differences.

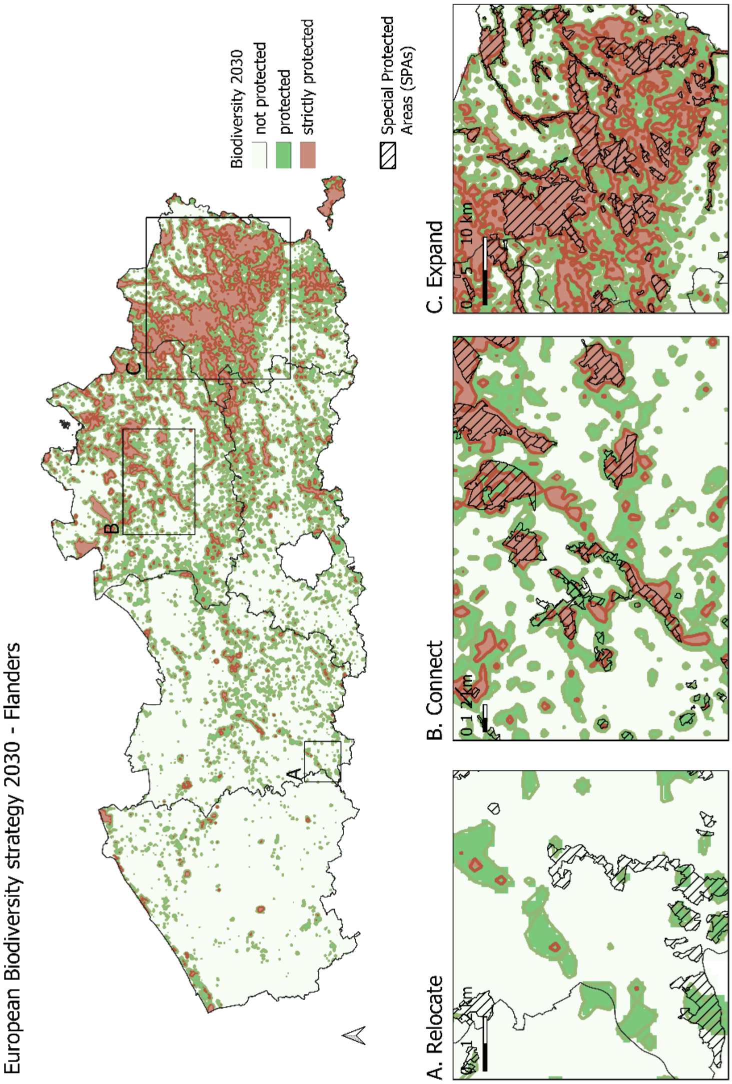


Figure 16: Prioritization of the protected (30%) and strictly protected (10%) areas in Flanders for the EU Biodiversity Strategy 2030, based on biodiversity scores (i.e. stacked relative occurrence rates from 14 Flemish priority species, weighted by Red list status). The map can be used to prioritize current Special Protected Areas (SPAs) under the Habitats Directive (92/43/EEG) in the Natura 2000 network for monitoring and possible relocation (A) and to indicate areas that are suitable for connecting (B) and expanding (C) the existing Natura 2000 habitat network.

In general, predictions overlapped well with the SPAs, with 85% of the current areas indicated as protected and 54% as strictly protected. This means that 15% of the current SPAs could be prioritized for additional monitoring and might be considered for relocation. Results on the strictly protected area also agreed well with previous studies as we noted an overlap of 42% with Natura 2000 SPAs (cf. 43% of the grids with the highest biodiversity scores overlapped with Natura 2000 area (Maes et al., 2005)). Only 22% of the protected area is now designated as SPA, yet this is not surprising as the threshold for protected area was 30% of the total land area and SPAs are only 8% of the total land area. This leaves 78% of the protected areas and 58% of the strictly protected areas for potential reserve expansion and connection. Note that SACs, i.e. areas designated under the Birds Directive, were not yet considered in this example.

25.2.2. Biological valuation Map

The current biological valuation map (BVM) is based on a detailed inventory of the biological environment and land cover in Flanders between 1998 and 2010 (De Saeger et al., 2010). It has been widely adopted as a spatial base layer by governments, scientists and NGOs to support landscape planning and nature conservation (De Saeger et al., 2017). After 2013, the map has been mostly updated inside protected areas as this requires labour-intensive fieldwork, leading to long mapping cycles (18 years for forests and 12 years for other semi-natural habitats). Occasionally, adjustments are also made based on the topographic reference layer for Flanders and the agricultural land declaration, but these areas are usually mapped with less detail. With the increasing availability of free high-resolution data from remote sensing, this method is also being explored to assist in monitoring and reduce fieldwork (Dumortier et al., 2022; Vanden Borre et al., 2011).

We see two applications of our research for the BVM, on the condition that the layer is not used to construct environmental predictors (like in Chapter IV, for example)[16]. First, areas with high biodiversity scores could be prioritized for monitoring. This can, for example, be harmonized with remote-sensing-based change detection (Tarantino et al., 2015; Vanden Borre et al., 2011; Williams et al., 2006). For practical reasons, the indicated areas for such guided field visits should be (much) smaller than 25 ha (cf. the 500x500m grids used for the biodiversity scores in the case studies), hence additional methods for dealing with sampling bias should be considered. Although the valuation in the BVM is largely plant-based, we highlight the value

[16] Other layers that might be used are small landscape features (Dochy, 2014), vegetation height (e.g. the Flanders *Groenkaart*; available on https://www.geopunt.be), remotely-sensed EFAs (section 23.4.4), the soil map (*Bodemkaart*; available on https://www.dov.vlaanderen.be/), water bodies (Packet et al., 2018).

of animal species as they have intrinsic value, are biodiversity indicators and deliver ecosystem services (Deliège and Neuteleers, 2015; IPBES, 2019; Peters et al., 2016). Moreover, many priority species for conservation in Flanders are animals (De Knijf et al., 2014; Herremans et al., 2014). As remote sensing is being explored as a new technique for habitat quality assessment, we encourage the adoption of, for example, measures of vegetation structure as this mostly impacted habitat suitability of animal species positively in large heathlands (Figure 14).

Second, a quality assessment could evaluate whether the current mapping cycles are sufficient to capture habitat changes by contrasting biodiversity scores with the biological valuation classes in the BVM (e.g. Figure 1 in De Saeger et al. (2017)). One could compare areas that have been inventoried regularly with those that have not in different habitats[17]. We expect to confirm that cycles can be longer in forests than in heathlands, for example, as these latter are characterized by fast natural succession (De Saeger et al., 2017; Fagúndez, 2013).

25.2.3. MEETNETTEN

The "MEETNETTEN" are networks for monitoring plant and animal species of conservation interest in Flanders (i.e. breeding birds, wintering waterfowl and 77 species of other taxonomic groups[18]) (Westra et al., 2016). Each network consists of predefined locations where citizen scientists can count individuals by following structured survey protocols. The resulting data are extremely valuable to monitor population trends and are a major success in their completeness since the project started in 2016[19]. Given their success, new networks will be operational soon, for example for the monitoring of farmland biodiversity (Dumortier et al., 2022) or invasive alien species[20]. Locations with high relative habitat suitability can be prioritised for more structured monitoring, both to design new networks and to optimise and expand existing networks.

25.2.4. Common agricultural policy

The Common Agricultural Policy (CAP) is the European legal framework that provides funding for eligible farmers in all EU member states. It was designed in 1962 as an answer to the growing need for food while ensuring fair incomes for farmers. Over the years, the CAP became more 'green' with increasing attention towards ecosystem services and biodiversity. Besides

[17] Inventory dates are saved as metadata in the Biological Valuation Map (available on https://www.geopunt.be).

[18] The MEETNETTEN currently include birds, amphibians, plants, mammals, molluscs, butterflies, dragonflies and other invertebrates https://meetnetten.be/

[19] Personal notes from the Biodiversity Spring Market. 22 March 2023. INBO, Brussels

[20] Personal notes from the Biodiversity Spring Market. 22 March 2023. INBO, Brussels

the delivery of positive services (e.g. food production, habitat creation and some cultural services), farming also has a detrimental effect on nature through eutrophication, acidification and drainage and also farmland biodiversity has seen major declines over the past decades (Tscharntke et al., 2005). On the other hand, farmers can benefit from ecosystem services such as pollination and pest control (Van Eupen, 2017).

The CAP reform of 2014-2020 failed on its so-called 'Greening' measures that were targeted at safeguarding and restoring biodiversity on farmland (Lakner et al., 2019). In return for funding, Flemish farmers mostly selected catch crops and nitrogen-fixing crops, which do not necessarily benefit biodiversity (Dicks et al., 2014; Pe'er et al., 2014) and agri-environment (AE) schemes, where farmers preferred schemes with easy implementation, although they do consider the environmental impact (Ghyselinck, 2021). The newest reform of the CAP 2023-2027[21] presented a new green architecture, with the largest changes being the abandonment of the ecological focus areas (Van Eupen, 2017) and the introduction of eco-schemes such as biological pest control, buffer strips and flower strips (Runge et al., 2022).

Biodiversity monitoring and conservation in and around farmland comes with challenges regarding the use of opportunistic data and the current positions of agricultural vs. nature organisations and policy in Flanders. First, lower accessibility of farmland can lead to sampling bias and imperfect detection of farmland species. Additionally, farmers may not report rare species to avoid unwanted visitors on their land, further complicating monitoring efforts (Stubbe, 2021). Second, transparent communication between farmers and (non-) governmental organisations is crucial for effective measures but mostly lacking (Stubbe, 2021; Verdonckt, 2018). Farmers need more evidence-based information on the current and potential value of their land for biodiversity, including species presence and how to protect them (Ghyselinck, 2021; Stubbe, 2021). Nature organisations and scientists also need to engage with farmers and allow them to participate in the design and monitoring of conservation measures.

Our methods have the potential to support biodiversity conservation in the Flemish implementation of the CAP in various ways. For instance, they can help prioritize areas for effective implementation of agri-environment schemes, as we will illustrate in a second case study in the next section. Additionally, our methods can improve the valuation of farmland by assisting in the biological valuation of habitats in the BVM (as described in section 25.2.2), which is used for the biodiversity component of the high nature value farmland indicator in

[21] https://agriculture.ec.europa.eu/common-agricultural-policy/cap-overview/cap-2023-27/key-reforms-new-cap_en

Flanders (Andersen et al., 2004; Danckaert et al., 2009a; Paracchini et al., 2008). Finally, the methods presented in Chapter IV can also support biodiversity-friendly farmland management or provide additional evidence of negative impacts from intensified land use, such as in the context of the nitrogen policy (Overloop et al., 2001; Vantieghem et al., 2017; Vogels et al., 2017). Implementing more efficient management practices can enhance positive outcomes and promote intrinsic motivation among farmers (Ghyselinck, 2021; Wilson and Hart, 2001).

25.2.5. CASE STUDY 2: Agri-environment schemes

Agri-environment (AE) schemes are a CAP funding tool to support environmentally friendly practices by landowners, supervised by the Flemish Land Agency (VLM) in Flanders. Currently, most of them are situated in specific regions (southwest of West-Flanders and southeast of Flemish-Brabant, south Limburg). The VLM employs professional landscape managers to advise farmers and assess the feasibility of AE schemes, which can be concluded in management agreements (MAs) with the VLM in terms of five years. Currently, farmers can choose out of twenty-two options and variations[22].

Figure 17 illustrates that farmland biodiversity scores can promote equitable participation of farmers throughout Flanders while taking into account budgets and recommendations from existing projects. In the PARTRIDGE project (Ghyselinck, 2021), for example, land managers observed that in the 500-hectare demonstration sites, roughly 10% should be assigned to an AE scheme (Stubbe, 2021). The figure indicates which areas of 625 ha should be considered according to this principle following a few assumptions. Farmers receive on average 1700 euros per hectare per year for implementing fauna-related AE schemes [7], meaning that 1176 hectares (or almost 9% of the total land area) can be declared as AE schemes with a budget of 10 million euros (cf. 8 million euros were paid out to farmers in Flanders in 2016 [23]). For this exercise, we first identified the 500-metre grids (i.e. 25 hectares) with the 10% highest farmland biodiversity scores per province and consequently counted the eligible grids per area of 625 hectares. Areas with no eligible grids are indicated as not suitable for AE schemes. Areas with one or two eligible grids (4 to 8%) are indicated in orange, three to ten eligible grids (12 to 40%) in light green and more than ten eligible grids (> 40%) in dark green.

[22] Retrieved from: https://www.vlm.be/nl/themas/beheerovereenkomsten/Paginas/default.aspx on March 27th 2023

[23] Retrieved from: https://www.vlm.be/nl/themas/beheerovereenkomsten/Wouter_Rombouts/Paginas/default.aspx on March 27th 2023

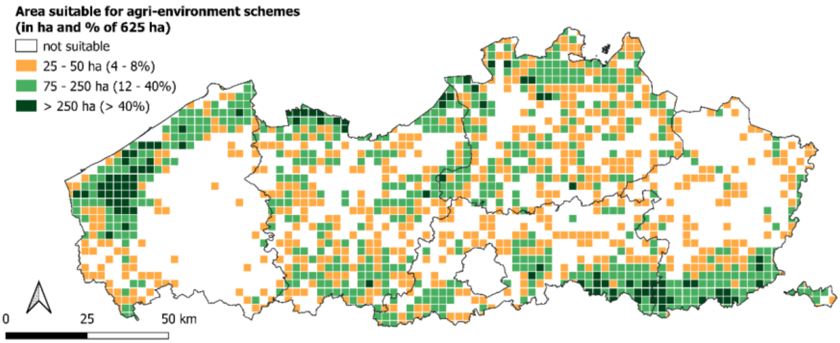


Figure 17: The area suitable for agri-environment (AE) schemes in Flanders. Colours indicate the number of 25-hectare areas with the 10% highest biodiversity scores for farmland fauna in areas of 625 hectares.

The coloured areas in Figure 17 are all suitable for AE schemes, as they contain at least one area of 25 hectares with high farmland biodiversity scores at a regional level. However, different approaches might be considered for farms located in the light green, orange and dark green grids. The most optimal choice in terms of budget, biodiversity and equitability would be to visit farms in the light-green grids across Flanders. Farms situated in orange areas can also be considered, and might especially be interesting for biodiversity restoration initiatives. Farms in the dark green can be encouraged to diversify AE schemes on their farms, for example by forming collectives. The white areas are mostly characterized by high urbanisation or intensive farming (Danckaert et al., 2009b), which is why they have less potential for biodiversity conservation (also see Figure 16). These maps could be further improved by considering additional predictors for farmland biodiversity, such as soil texture, soil water-related parameters, habitat heterogeneity, historical land use, and small landscape features.

A related study was performed in 2019, which generated relevant potential habitat maps for a set of indicator species for which management agreements can contribute to the protection of biodiversity in and around farmland (De Bruyn et al., 2019a, 2019b). The maps were based on mechanistic models (GeoDynamiX toolbox) and species occurrence data and were stacked to quantify potential species richness. A drawback of this method is that area prioritization is not possible because the produced maps are binary (presence or absence) and thus give no ranking of habitat suitability (opposed to correlative SDMs).

25.3. Discussion of the case studies

Preliminary results in the two case studies (sections 25.2.1 and 25.2.5) confirmed that stacking models for species of conservation priority is good practice (Guillera-Arroita et al., 2015). However, the presented maps (Figures 16 and 17) did not include all socioeconomic restrictions or thresholds such as the minimum percentage of SACs that should fall in Natura 2000 areas (Decleer, 2007). For example, while biodiversity scores can be a good first indication of the most suitable areas for protection and connection (Figure 16), the method should be expanded to include also cost-effectiveness (Wätzold et al., 2010) and stakeholder interests. Reaching cost-effectiveness based on opportunistic presence-only SDMs is challenging, as they can only provide a ranking of the relative habitat suitability of locations (section 3.3; Guillera-Arroita et al., 2015). We highlight that even without such restrictions it will take many years, even decades, to reach the Biodiversity targets set by the EU (European Commission, 2020) due to the high eutrophication levels and anthropogenic fragmentation in Flanders.

We further propose to validate our method for delineating conservation areas by comparing their output to that of well-established mechanistic methods such as the GeoDynamix toolbox. A master thesis in the context of this research has demonstrated that mechanistic models usually predict a higher share of the study area as suitable compared to Maxent (Deschuytter and Somers, 2022). However, this study compared the binary output of both methods to indicate spatial differences between predictions of potential habitat suitability. We propose a Spearman rank correlation test to indicate whether model predictions differ between methods and an evaluation of both methods on an independent evaluation set if such data are available (Norberg et al., 2019). The method for constructing an external validation set in Chapter II (Appendix A) can be used, supplemented with MEETNETTEN data (see section 25.2.3).

Recently, the European project 'BirdWatch'[24] started in Flanders, aiming to monitor and increase habitat suitability for farmland birds while considering various stakeholders. The project will use remote sensing data and species distribution models to model habitat suitability, and we recommend that they adopt the methods presented in this research. These can facilitate the use of opportunistic CSD by increasing their quality (Chapters II and III) and improve SDMs by incorporating measures of heterogeneity at different scales (Chapter IV). We also encourage similar initiatives for other taxonomic groups.

[24] https://www.vlaanderen.be/inbo/en-GB/projects/effectiviteit-van-beheerovereenkomsten-voor-akker-en-weidevogels

CHAPTER VII. General conclusions

This dissertation provides further evidence that opportunistic citizen science data (CSD) can be a valuable resource for biodiversity conservation policy and management if processed correctly.

Chapters II and III zoomed in on data cleansing through stringent filtering of opportunistic CSD collected in large online data platforms and how this can be fine-tuned by taking into account the traits of the species under study. The findings in Chapter II showed that there is a quantity-quality trade-off in stringent filtering and that filtering should not be performed blindly. In general, we recommend using verified records or records from more experienced observers for animal species and records submitted with extra detail for plant species, although with caution when the sample size is reduced beyond a certain threshold (BOX 2) and with consideration of the goal of the study. We encourage the integration of semi-automated verification systems and the provision of metadata on observer experience, such as observer ratings or the number of observations per taxonomic group.

Chapter IV used opportunistic CSD in combination with environmental data obtained through remote sensing to support multi-scale habitat management in heathlands. Since heathland size and the landscape context are static environmental variables, especially in anthropogenic landscapes, heathland management should include heathland vegetation management at multiple scales, with consideration of the species of conservation interest. The integration of opportunistic CSD and remote sensing data is a promising advancement in biodiversity conservation monitoring, which should be tested in other habitat types and regions.

While the results in this dissertation can contribute to the uptake of opportunistic CSD, species distribution models (SDMs) and remotely-sensed predictors in biodiversity conservation applications, the possible drawbacks of our methods needed to be recognized. Presence-only SDMs predict relative habitat suitability only and the objectives of the studies largely defined the applied methods (i.e. methods for modelling and bias correction, species, minimum sample size, scale and study area). For assessing the impact of specific environmental drivers on habitat suitability, it is important to choose ecologically plausible parameters, check for multicollinearity, formulate appropriate hypotheses, and account for spatial autocorrelation. In retrospect, the best option for reconciling methods in the dissertation would have been to use point process models, yet this may have led to suboptimal recommendations for the intended conservation applications.

Some initial suggestions for future research were made throughout the dissertation, where we encouraged additional studies to enhance the transferability of our results to other methods, scales, habitats and species. Our research focussed on the use of opportunistic CSD as the only data source, yet encourages applications where these data can be integrated with (semi-) structured survey data (e.g. through integrated SDMs). Additional suggestions made in Chapter V were the consideration of species co-occurrence (e.g. through joint-SDMs), the integration of a temporal aspect (e.g. through dynamic SDMs) and the uptake of other predictors related to microclimate (e.g. near-surface temperature), soil (e.g. soil water), landscape metrics (e.g. edge density), ecosystem functioning (e.g. remotely-sensed ecosystem functioning attributes) and environmental thresholds (e.g. the presence or absence of a habitat type).

Providing land owners, governing bodies and policymakers with evidence-based research on the state of biodiversity and the drivers of its change is a conservation priority. This study can guide end-users of large citizen science platforms by reducing the uncertainty associated with opportunistic biodiversity data. In Chapter VI, several examples were presented to illustrate the application potential in different domains of biodiversity conservation (i.e. prioritization of areas for monitoring and conservation and assisting habitat management). Species distribution models that provide relative estimates of habitat suitability were proposed as an instrument for policy based on thresholds such as minimum area (e.g. EU Biodiversity Strategy 2030) and maximum budget (e.g. agri-environment schemes). Moreover, stacking techniques such as biodiversity scores are ideal to support biodiversity conservation in different policy fields in a flexible way, i.e. by adjusting species and weighting methods. While opportunistic presence-only data cannot replace systematic survey data for the monitoring of population trends, they can direct surveys to prioritize locations for monitoring. We recognize that the conservation potential in the two case studies may have been overestimated or misplaced due to socio-economic restrictions.

To conclude, this dissertation illustrates the value of funded research with a focus on application, yet its value can only be realized when science, policy, and practice meet in a transparent way, and expertise from all parties is treasured. We should all take advantage of the current momentum for change and we hope these results inspire conservation practitioners and governing bodies to enhance biodiversity conservation policy and management with citizen-science-based research.

REFERENCES

Agentschap voor Natuur en Bos, 2017. Ontwerp Vlaams Natura 2000-Programma. Eerste cyclus 2016-2020. VR 2017 1407 DOC.0775/2BIS.

Aichison, J., 2003. The Statistical Analysis of Compositional Data. Blackburn Press, Caldwell.

Alcaraz-Segura, D., Lomba, A., Sousa-Silva, R., Nieto-Lugilde, D., Alves, P., Georges, D., Vicente, J.R., Honrado, J.P., 2017. Potential of satellite-derived ecosystem functional attributes to anticipate species range shifts. Int. J. Appl. Earth Obs. Geoinf. 57, 86–92. https://doi.org/10.1016/j.jag.2016.12.009

Amici, V., Eggers, B., Geri, F., Battisti, C., 2015. Habitat Suitability and Landscape Structure: A Maximum Entropy Approach in a Mediterranean Area. Landsc. Res. 40, 208–225. https://doi.org/10.1080/01426397.2013.774329

Ampoorter, E., Barbaro, L., Jactel, H., Baeten, L., Boberg, J., Carnol, M., Castagneyrol, B., Charbonnier, Y., Dawud, S.M., Deconchat, M., De Smedt, P., De Wandeler, H., Guyot, V., Hättenschwiler, S., Joly, F.-X., Koricheva, J., Milligan, H., Muys, B., Nguyen, D., Ratcliffe, S., Raulund-Rasmussen, K., Scherer-Lorenzen, M., van der Plas, F., Van Keer, J., Verheyen, K., Vesterdal, L., Allan, E., 2020. Tree diversity is key for promoting the diversity and abundance of forest-associated taxa in Europe. Oikos 129, 133–146. https://doi.org/10.1111/oik.06290

Andersen, E., Baldock, D., Bennett, H., Beaufoy, G., Bignal, E., Brouwer, F., Elbersen, B., Eiden, G., Godeschalk, F., Jones, G., Mccracken, D., Nieuwenhuizen, W., van Eupen, M., Hennekens, S., Zervas, G., 2004. Developing a High Nature Value Farming area indicator. Final Report. Institute for European Environmental Policy.

Anderson, R.P., Araújo, M.B., Guisan, A., Lobo, J.M., Martínez-Meyer, E., Peterson, A.T., Soberón, J.M., 2020. Optimizing biodiversity informatics to improve information flow, data quality, and utility for science and society. Front. Biogeogr. 12, 1–15. https://doi.org/10.21425/F5FBG47839

Antrop, M., 2004. Landscape change and the urbanization process in Europe. Landsc. Urban Plan. 67, 9–26. https://doi.org/10.1016/S0169-2046(03)00026-4

Araújo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33, 1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x

Arenas-Castro, S., Gonçalves, J., Alves, P., Alcaraz-Segura, D., Honrado, J.P., 2018. Assessing the multi-scale predictive ability of ecosystem functional attributes for species distribution modelling. PLoS One 13, e0199292. https://doi.org/10.1371/journal.pone.0199292

Arenas-Castro, S., Regos, A., Gonçalves, J.F., Alcaraz-Segura, D., Honrado, J., 2019. Remotely Sensed Variables of Ecosystem Functioning Support Robust Predictions of Abundance Patterns for Rare Species. Remote Sens. 11, 1–16. https://doi.org/10.3390/rs11182086

Arif, S., MacNeil, M.A., 2022. Predictive models aren’t for causal inference. Ecol. Lett. 25, 1741–1745. https://doi.org/10.1111/ele.14033

Aristeidou, M., Herodotou, C., Ballard, H.L., Higgins, L., Johnson, R.F., Miller, A.E., Young, A.N., Robinson, L.D., 2021. How Do Young Community and Citizen Science Volunteers Support Scientific Research on Biodiversity? The Case of iNaturalist. Diversity 13, 318.

Atauri, J.A., De Lucio, J. V., 2001. The role of landscape structure in species richness distribution of birds, amphibians, reptiles and lepidopterans in Mediterranean landscapes. Landsc. Ecol. 16, 147–159. https://doi.org/10.1023/A:1011115921050

Baddeley, A., Diggle, P.J., Hardegen, A., Lawrence, T., Milne, R.K., Nair, G., 2014. On tests of spatial pattern based on simulation envelopes. Ecol. Monogr. 84, 477–489.

Baddeley, A., Rubak, E., Turner, R., 2015. Spatial Point Patterns: Methodology and Applications with R. Chapman & Hall/CRC.

Baddeley, A., Turner, R., 2005. spatstat: An R Package for Analyzing Spatial Point Patterns. J. Stat. Softw. 12, 1–42. https://doi.org/10.1111/j.1540-4781.1939.tb01339.x

Baddeley, A., Turner, R., 2000. Practical maximum pseudolikelihood for spatial point patterns. Aust. New Zeal. J. Stat. 42, 283–322. https://doi.org/10.1111/1467-842X.00128

Bar-Massada, A., Wood, E.M., 2014. The richness-heterogeneity relationship differs between heterogeneity measures within and among habitats. Ecography (Cop.). 37, 528–535. https://doi.org/10.1111/j.1600-0587.2013.00590.x

Barbosa, A.M., 2015. fuzzySim: applying fuzzy logic to binary similarity indices in ecology. Methods Ecol. Evol. 6, 853–858. https://doi.org/10.1111/2041-210X.12372

Barclay-Estrup, P., 1971. The Description and Interpretation of Cyclical Processes in a Heath Community: III. Microclimate in Relation to the Calluna Cycle. J. Ecol. 59, 143–166.

Barton, K., 2019. MuMIn: Multi-Model Inference. R package version 1.43.15.

Bellis, L.M., Pidgeon, A.M., Radeloff, V.C., St-Louis, V., Navarro, J.L., Martella, M.B., 2008. Modeling Habitat Suitability for Greater Rheas Based on Satellite Image Texture. Ecol. Appl. 18, 1956–1966.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289–300.

Bergen, K.M., Goetz, S.J., Dubayah, R.O., Henebry, G.M., Hunsaker, C.T., Imhoff, M.L., Nelson, R.F., Parker, G.G., Radeloff, V.C., 2009. Remote sensing of vegetation 3-D structure for biodiversity and habitat: Review and implications for lidar and radar spaceborne missions. J. Geophys. Res. 114, G00E06. https://doi.org/10.1029/2008JG000883

Besnard, A.G., Davranche, A., Maugenest, S., Bouzillé, J.B., Vian, A., Secondi, J., 2015. Vegetation maps based on remote sensing are informative predictors of habitat selection of grassland birds across a wetness gradient. Ecol. Indic. 58, 47–54. https://doi.org/10.1016/j.ecolind.2015.05.033

Bink, F.A., 1992. Ecologische atlas van de dagvlinders van Noordwest-Europa. Schuyt & Co Uitgevers en Importeurs bv, Haarlem.

Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Stuart-Smith, R.D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J.F., Pecl, G.T., Barrett, N., Frusher, S., 2014. Statistical solutions for error and bias in global citizen science datasets. Biol. Conserv. 173, 144–154. https://doi.org/10.1016/j.biocon.2013.07.037

Boakes, E.H., Gliozzo, G., Seymour, V., Harvey, M., Smith, C., Roy, D.B., Haklay, M., 2016. Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. Sci. Rep. 6, 33051. https://doi.org/10.1038/srep33051

Bonari, G., Fajmon, K., Malenovský, I., Zelený, D., Holuša, J., Jongepierová, I., Kočárek, P., Konvička, O., Uřičář, J., Chytrý, M., 2017. Management of semi-natural grasslands benefiting both plant and insect diversity: The importance of heterogeneity and tradition. Agric. Ecosyst. Environ. 246, 243–252. https://doi.org/10.1016/j.agee.2017.06.010

Bonthoux, S., Lefèvre, S., Herrault, P.A., Sheeren, D., 2018. Spatial and temporal dependency of NDVI satellite imagery in predicting bird diversity over France. Remote Sens. 10, 1–22. https://doi.org/10.3390/rs10071136

Boria, R.A., Olson, L.E., Goodman, S.M., Anderson, R.P., 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecol. Modell. 275, 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012

Botella, C., Joly, A., Bonnet, P., Munoz, F., Monestiez, P., 2021. Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. Methods Ecol. Evol. 12, 933–945. https://doi.org/10.1111/2041-210X.13565

Box, G.E.P., 1966. Use and Abuse of Regression. Technometrics 8, 625–629.

Bradley, B.A., Mustard, J.F., 2006. Characterizing the landscape dynamics of an invasive plant and risk of invasion using remote sensing, Ecological Applications.

Bradley, B.A., Olsson, A.D., Wang, O., Dickson, B.G., Pelech, L., Sesnie, S.E., Zachmann, L.J., 2012. Species detection vs. habitat suitability: Are we biasing habitat suitability models with remotely sensed data? Ecol. Modell. 244, 57–64. https://doi.org/10.1016/j.ecolmodel.2012.06.019

Bradter, U., Mair, L., Jönsson, M., 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 Ecol. Evol. 9, 1667–1678. https://doi.org/10.1111/2041-210X.13012

Brereton, T., Botham, M., Middlebrook, I., Randle Z, D.N., Harris, S., Dennis, E., Robinson, A., Peck, K., Roy, D., 2019. United Kingdom Butterfly Monitoring Scheme report for 2018. Centre for Ecology & Hydrology, Butterfly Conservation, British Trust for Ornithology and Joint Nature Conservation Committee. Wareham, UK.

Broman, D.J.A., Litvaitis, J.A., Ellingwood, M., Tate, P., Reed, G.C., 2014. Modeling bobcat Lynx rufus habitat associations using telemetry locations and citizen-scientist observations: are the results comparable? Wildlife Biol. 20, 229–237. https://doi.org/10.2981/wlb.00022

Brose, U., 2003. Bottom-up control of carabid beetle communities in early successional wetlands: Mediated by vegetation structure or plant diversity? Oecologia 135, 407–413. https://doi.org/10.1007/s00442-003-1222-7

Brotons, L., Herrando, S., Pla, M., 2007. Updating bird species distribution at large spatial scales: applications of habitat modelling to data from long-term monitoring programs. Divers. Distrib. 13, 276–288. https://doi.org/10.1111/j.1472-4642.2007.00339.x

Burgess, H.K., Debey, L.B., Froehlich, H.E., Schmidt, N., Theobald, E.J., Ettinger, A.K., Hillerislambers, J., Tewksbury, J., Parrish, J.K., 2017. The science of citizen science: Exploring barriers to use as a primary research tool. Biol. Conserv. 208, 113–120. https://doi.org/10.1016/j.biocon.2016.05.014

Burnham, K.P., Anderson, D.R., Huyvaert, K.P., 2011. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav. Ecol. Sociobiol. 65, 23–35. https://doi.org/10.1007/s00265-010-1029-6

Burns, F., Eaton, M.A., Burfield, I.J., Klvaňová, A., Šilarová, E., Staneva, A., Gregory, R.D., 2021. Abundance decline in the avifauna of the European Union reveals cross-continental similarities in biodiversity change. Ecol. Evol. 11, 16647–16660. https://doi.org/10.1002/ece3.8282

Busby, J.R., 1991. BIOCLIM – a bioclimate analysis and prediction system. Plant Prot. Q. 6, 8–9.

Byriel, D.B., Ro-Poulsen, H., Kepfer-Rojas, S., Hansen, A.K., Hansen, R.R., Justesen, M.J., Kristensen, E., Møller, C.B., Schmidt, I.K., 2023. Contrasting responses of multiple insect taxa to common heathland management regimes and old-growth successional stages. Biodivers. Conserv. 32, 545–565. https://doi.org/10.1007/s10531-022-02511-9

Callaghan, C.T., Poore, A.G.B., Hofmann, M., Roberts, C.J., Pereira, H.M., 2021. Large-bodied birds are over-represented in unstructured citizen science data. Sci. Rep. 11, 19073. https://doi.org/10.1038/S41598-021-98584-7

Callaghan, C.T., Roberts, J.D., Poore, A.G.B., Alford, R.A., Cogger, H., Rowley, J.J.L., 2020. Citizen science data accurately predicts expert-derived species richness at a continental scale when sampling thresholds are met. Biodivers. Conserv. 29, 1323–1337. https://doi.org/10.1007/s10531-020-01937-3

Carrascal, L.M., Javier, S., Palomino, D., Alonso, C.L., Lobo, J.M., 2006. Species-specific features affect the ability of census-derived models to map winter avian distribution. Ecol. Res. 21, 681–691. https://doi.org/10.1007/s11284-006-0173-y

Carvalho, S.B., Gonçalves, J., Guisan, A., Honrado, J.P., 2016. Systematic site selection for multispecies monitoring networks. J. Appl. Ecol. 53, 1305–1316. https://doi.org/10.1111/1365-2664.12505

Catchpole, C.K., Slater, P.J.B., 2008. Bird Song, 2nd ed. ed. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511754791

Ceballos, G., Ehrlich, P.R., Raven, P.H., 2020. Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. Proc. Natl. Acad. Sci. U. S. A. 117, 13596–13602. https://doi.org/10.1073/pnas.1922686117

Chefaoui, R.M., Lobo, J.M., Hortal, J., 2011. Effects of species' traits and data characteristics on distribution models of threatened invertebrates. Anim. Biodivers. Conserv. 34, 229–247.

Chen, Q., Yin, Y., Zhao, R., Yang, Y., Teixeira da Silva, J.A., Yu, X., 2020. Incorporating Local Adaptation Into Species Distribution Modeling of Paeonia mairei, an Endemic Plant to China. Front. Plant Sci. 10, 1717. https://doi.org/10.3389/fpls.2019.01717

Cianfrani, C., Maiorano, L., Loy, A., Kranz, A., Lehmann, A., Maggini, R., Guisan, A., 2013. There and back again? Combining habitat suitability modelling and connectivity analyses to assess a potential return of the otter to Switzerland. Anim. Conserv. 16, 584–594. https://doi.org/10.1111/acv.12033

Clark, J.S., Gelfand, A.E., Woodall, C.W., Zhu, K., 2014. More than the sum of the parts: forest climate response from joint species distribution models. Ecol. Appl. 24, 990–999. https://doi.org/10.1890/13-1015.1

Colwell, R.K., Xuan Mao, C., Chang, J., 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology 85, 2717–2727.

Connor, T., Hull, V., Viña, A., Shortridge, A., Tang, Y., Zhang, J., Wang, F., Liu, J., 2017. Effects of grain size and niche breadth on species distribution modeling. Ecography (Cop.). 40, 001–012. https://doi.org/10.1111/ecog.03416

Coops, N.C., Wulder, M.A., 2019. Breaking the Habit(at). Trends Ecol. Evol. 34, 585–587. https://doi.org/10.1016/j.tree.2019.04.013

Cord, A., Rödder, D., 2011. Inclusion of habitat availability in species distribution models through multi-temporal remote-sensing data? Ecol. Appl. 21, 3285–3298.

Costa, H., Foody, G., Jiménez, S., Silva, L., 2015. Impacts of Species Misidentification on Species Distribution Modeling with Presence-Only Data. ISPRS Int. J. Geo-Information 4, 2496–2518. https://doi.org/10.3390/ijgi4042496

Couvreur, M., Menschaert, J., Sevenant, M., Ronse, A., Van Landuyt, W., De Blust, G., Antrop, M., Hermy, M., 2004. Ecodistricten en ecoregio's als instrument voor natuurstudie en milieubeleid. Natuur.Focus 3, 51–58.

Crall, A.W., Newman, G.J., Stohlgren, T.J., Holfelder, K.A., Graham, J., Waller, D.M., 2011. Assessing citizen science data quality: an invasive species case study. Conserv. Lett. 4, 433–442. https://doi.org/10.1111/J.1755-263X.2011.00196.X

Cribari-Neto, F., Zeileis, A., 2010. Beta Regression in R. J. Stat. Softw. 34, 1–24. https://doi.org/10.18637/jss.v034.i02

Cruickshank, S.S., Bühler, C., Schmidt, B.R., 2019. Quantifying data quality in a citizen science monitoring program: False negatives, false positives and occupancy trends. Conserv. Sci. Pract. 1, e54. https://doi.org/10.1111/CSP2.54

Danckaert, S., Carels, K., Van Gijseghem, D., Hens, M., 2009a. Indicatoren voor het opvolgen van de hoge natuurwaarden op landbouwgrond in het kader van de PDPO-monitoring. Een verkennende analyse., Beleidsdomein Landbouw en Visserij, afdeling Monitoring en Studie. Brussel.

Danckaert, S., Lenders, S., Oeyen, A., 2009b. De landbouwactiviteit in Vlaamse gemeenten, proeve van typologie. Departement Landbouw en Visserij afdeling Monitoring en Studie, Brussel.

Davies, A.B., Asner, G.P., 2014. Advances in animal ecology from 3D-LiDAR ecosystem mapping. Trends Ecol. Evol. 29, 681–691. https://doi.org/10.1016/j.tree.2014.10.005

De Blust, G., 2022. Heide en Heidebeheer, in: Van Uytvanck, J., Hermy, M., De Blust, G., Hoffmann, M. (Eds.), Natuurbeheer. Praktijk En Wetenschap Hand in Hand. Sterck & De Vreese, Gorredijk, Nederland, pp. 255–286.

De Bruyn, L., Belpaire, C., De Knijf, G., Gyselings, R., Lommelen, E., Maes, D., Packet, J., Speybroeck, J., Thomaes, A., Van Den Berge, K., Vanden Borre, J., Van Landuyt, W., Vermeersch, G., Vriens, L., 2019a. Advies over indicatorsoorten voor beheerovereenkomsten. INBO.A.3797. Instituuut voor Natuur en Bosonderzoek, Brussel.

De Bruyn, L., Maes, D., Leyssen, A., Thomaes, A., Wils, C., Belpaire, C., Vermeersch, G., Van Thuyne, G., Gouwy, J., Vanden Borre, J., Speybroek, J., Packet, J., Devos, K., Van Den Berge, K., Gyselings, R., 2019b. Advies over de afbakening van gebieden voor de inzet van beheerovereenkomsten. INBO.A.3847. Instituut voor Natuur- en Bosonderzoek, Brussel.

De Bruyn, L., Sierdsema, H., Van Dyck, H., van Swaay, C., Vermeersch, G., Anselin, A., Maes, D., 2009. Can we predict the distribution of heathland butterflies with heathland bird data? Anim. Biol. 59, 335–349. https://doi.org/10.1163/157075609X454962

De Frenne, P., Lenoir, J., Luoto, M., Scheffers, B.R., Zellweger, F., Aalto, J., Ashcroft, M.B., Christiansen, D.M., Decocq, G., De Pauw, K., Govaert, S., Greiser, C., Gril, E., Hampe, A., Jucker, T., Klinges, D.H., Koelemeijer, I.A., Lembrechts, J.J., Marrec, R., Meeussen, C., Ogée, J., Tyystjärvi, V., Vangansbeke, P., Hylander, K., 2021. Forest microclimates and climate change: Importance, drivers and future research agenda. Glob. Chang. Biol. 27, 2279–2297. https://doi.org/10.1111/gcb.15569

De Knijf, G., Paelinckx, D., 2013. Typische faunasoorten van de verschillende Natura 2000 habitattypes, in functie van de beoordeling van de staat van instandhouding op niveau Vlaanderen. INBO.A.2013.139. Instituut voor Natuur- en Bosonderzoek, Brussel.

De Knijf, G., Westra, T., Onkelinx, T., Quataert, P., Pollet, M., 2014. Monitoring Natura 2000-soorten en overige soorten prioritair voor het Vlaams beleid. Blauwdrukken soortenmonitoring in Vlaanderen. Rapporten van het Instituut voor Natuur- en Bosonderzoek 2014 (INBO.R.2014.2319355). Instituut voor Natuur- en Bosonderzoek, Brussel.

De Ro, A., Vanden broeck, A., Verschaeve, L., Van Dyck, H., Jacobs, I., T'Jollyn, F., Maes, D., 2021. Occasional long distance dispersal does not prevent inbreeding in a threatened butterfly. BMC Ecol. Evol. 21, 224.

De Saeger, S., Ameeuw, G., Berten, B., Bosch, H., Brichau, I., Knijf, D., Demolder, H., Erens, G., Guelinckx, R., Oosterlynck, P., Scheldeman, K., T'jollyn, F., Van Hove, M., Van Ormelingen, J., Vriens, L., Zwaenepoel, A., Van Dam, G., Verheirstraeten, M., Wils, C., Paelinckx, D., 2010. Biologische Waarderingskaart, versie 2.2. Rapporten van het Instituut voor Natuur- en Bosonderzoek 2010 (36). Instituut voor Natuur- en Bosonderzoek, Brussel.

De Saeger, S., Guelinckx, R., Oosterlynck, P., De Bruyn, A., Debusschere, K., Dhaluin, P., Erens, R., Hendrickx, P., Hennebel, D., Jacobs, I., Kumpen, M., Opdebeeck, J., Spanhove, T., Tamsyn, W., Van Oost, F., Van Dam, G., Van Hove, M., Wils, C., Paelinckx, D., 2020. Biologische Waarderingskaart en Natura 2000 Habitatkaart, uitgave 2020. Rapporten van het Instituut voor Natuur- en Bosonderzoek 2000 (35). Instituut voor Natuur- en Bosonderzoek, Brussel. https://doi.org/doi.org/10.21436/inbor.18840851

De Saeger, S., Oosterlynck, P., Paelinckx, D., 2017. The Biological Valuation Map (BVM): a field-driven survey of land cover and vegetation in the Flemish Region of Belgium. Doc. Phytosociologiques 6, 373–382.

De Solan, T., Renner, I., Cheylan, M., Geniez, P., Barnagaud, J.-Y., 2019. Opportunistic records reveal Mediterranean reptiles' scale-dependent responses to anthropogenic land use. Ecography (Cop.). 42, 608–620.

de Vries, J.P.R., Koma, Z., WallisDeVries, M.F., Kissling, W.D., 2021. Identifying fine-scale habitat preferences of threatened butterflies using airborne laser scanning. Divers. Distrib. 27, 1251–1264. https://doi.org/10.1111/DDI.13272

Decleer, K., 2007. Europees beschermde natuur in Vlaanderen en het Belgisch deel van de Noordzee: habitattypen, dier- en plantensoorten. Mededelingen van het Instituut voor Natuur- en Bosonderzoek (1). Instituut voor Natuur- en Bosonderzoek, Brussel.

Deliège, G., Neuteleers, S., 2015. Should Biodiversity be Useful? Scope and Limits of Ecosystem Services as an Argument for Biodiversity Conservation. Environ. Values 24, 165–182. https://doi.org/10.3197/096327114X13947900181275

Demolder, H., Schneiders, A., Spanhove, T., Maes, D., Van Landuyt, W., Adriaens, T., 2014. Hoofdstuk 4 - Toestand biodiversiteit. (INBO.R.2014.6194611). In Stevens, M. et al. (eds.), Natuurrapport - Toestand en trend van ecosystemen en ecosysteemdiensten in Vlaanderen. Technisch rapport. Mededelingen van het Instituut voor Natuur- en Bosonderzoek, Brussel.

Deschuytter, S., Somers, B., 2022. Using opportunistic citizen science data and species distribution modelling in support of biodiversity conservation policy in Flanders. Dissertation presented in fulfilment of the requirements for the degree of Master of Bioscience Engineering: Agro- and Ecosystems Engineering. KU Leuven, Leuven.

Devos, K., Anselin, A., Driessens, G., Herremans, M., Onkelinx, T., Spanoghe, G., Stienen, E., T'Jollyn, F., Vermeersch, G., Maes, D., 2016. De IUCN Rode-Lijst van de broedvogels in Vlaanderen (2016). Rapporten van het Instituut voor Natuur- en Bosonderzoek jaar (11485739). Instituut voor Natuur- en Bosonderzoek, Brussel. https://doi.org/dx.doi.org/10.21436/inbor.11485739

Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010. Citizen Science as an Ecological Research Tool: Challenges and Benefits. Annu. Rev. Ecol. Evol. Syst. 41, 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

Dicks, L. V., Hodge, I., Randall, N.P., Scharlemann, J.P.W., Siriwardena, G.M., Smith, H.G., Smith, R.K., Sutherland, W.J., 2014. A Transparent Process for "Evidence-Informed" Policy Making. Conserv. Lett. 7, 119–125. https://doi.org/10.1111/conl.12046

Diemont, W.H., Heijman, W.J.M., Siepel, H., Webb, N.R. (Eds.), 2015. Economy and Ecology of Heathlands. KNNV Publishing, Zeist, The Netherlands. https://doi.org/10.1163/9789004277946

Dobson, A.D.M., Milner-Gulland, E.J., Aebischer, N.J., Beale, C.M., Brozovic, R., Coals, P., Critchlow, R., Dancer, A., Greve, M., Hinsley, A., Ibbett, H., Johnston, A., Kuiper, T., Le Comber, S., Mahood, S.P., Moore, J.F., Nilsen, E.B., Pocock, M.J.O., Quinn, A., Travers, H., Wilfred, P., Wright, J., Keane, A., 2020. Making Messy Data Work for Conservation. One Earth 2, 455–465. https://doi.org/10.1016/J.ONEEAR.2020.04.012

Dochy, O., 2014. Verslag van de Frans-Belgische akkervogelinventarisatie 2013. Provincie West-Vlaanderen, Brugge. Brussels, Belgium. https://doi.org/10.13140/2.1.2065.4082

Dorazio, R.M., 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Glob. Ecol. Biogeogr. 23, 1472–1484. https://doi.org/10.1111/GEB.12216

Dorazio, R.M., Royle, J.A., Söderström, B., Glimskär, A., 2006. Estimating species richness and accumulation by modeling species occurrence and detectability. Ecology 87, 842–854. https://doi.org/10.1890/0012-9658(2006)87[842:ESRAAB]2.0.CO;2

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., Mcclean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography (Cop.). 36, 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Daniel Kissling, W., Kühn, I., Ohlemüller, R., R. Peres-Neto, P., Reineking, B., Schröder, B., Schurr, F.M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography (Cop.). 30, 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B., Singer, A., 2012. Correlation and process in species distribution models: Bridging a dichotomy. J. Biogeogr. 39, 2119–2131. https://doi.org/10.1111/j.1365-2699.2011.02659.x

Downey, H., Amano, T., Cadotte, M., Cook, C.N., Cooke, S.J., Haddaway, N.R., Jones, J.P.G., Littlewood, N., Walsh, J.C., Abrahams, M.I., Adum, G., Akasaka, M., Alves, J.A., Antwis, R.E., Arellano, E.C., Axmacher, J., Barclay, H., Batty, L., Benítez-López, A., Bennett, J.R., Berg, M.J., Bertolino, S., Biggs, D., Bolam, F.C., Bray, T., Brook, B.W., Bull, J.W., Burivalova, Z., Cabeza, M., Chauvenet, A.L.M., Christie, A.P., Cole, L., Cotton, A.J., Cotton, S., Cousins, S.A.O., Craven, D., Cresswell, W., Cusack, J.J., Dalrymple, S.E., Davies, Z.G., Diaz, A., Dodd, J.A., Felton, A., Fleishman, E., Gardner, C.J., Garside, R., Ghoddousi, A., Gilroy, J.J., Gill, D.A., Gill, J.A., Glew, L., Grainger, M.J., Grass, A.A., Greshon, S., Gundry, J., Hart, T., Hopkins, C.R., Howe, C., Johnson, A., Jones, K.W., Jordan, N.R., Kadoya, T., Kerhoas, D., Koricheva, J., Lee, T.M., Lengyel, S., Livingstone, S.W., Lyons, A., McCabe, G., Millett, J., Strevens, C.M., Moolna, A., Mossman, H.L., Mukherjee, N., Muñoz-Sáez, A., Negrões, N., Norfolk, O., Osawa, T., Papworth, S., Park, K.J., Pellet, J., Phillott, A.D., Plotnik, J.M., Priatna, D., Ramos, A.G., Randall, N., Richards, R.M., Ritchie, E.G., Roberts, D.L., Rocha, R., Rodríguez, J.P., Sanderson, R., Sasaki, T., Savilaakso, S., Sayer, C., Sekercioglu, C., Senzaki, M., Smith, G., Smith, R.J., Soga, M., Soulsbury, C.D., Steer, M.D., Stewart, G., Strange, E.F., Suggitt, A.J., Thompson, R.R.J., Thompson, S., Thornhill, I., Trevelyan, R.J., Usieta, H.O., Venter, O., Webber, A.D., White, R.L., Whittingham, M.J., Wilby, A., Yarnell, R.W., Zamora-Gutierrez, V., Sutherland, W.J., 2021. Training future generations to deliver evidence-based conservation and ecosystem management. Ecol. Solut. Evid. 2, e12032. https://doi.org/10.1002/2688-8319.12032

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J.P., Guisan, A., 2011. Predicting spatial patterns of plant species richness: A comparison of direct macroecological and species stacking modelling approaches. Divers. Distrib. 17, 1122–1131. https://doi.org/10.1111/j.1472-4642.2011.00792.x

Dujardin, S., Stas, M., Van Eupen, C., Aerts, R., Hendrickx, M., Delcloo, A.W., Duchêne, F., Hamdi, R., Nawrot, T.S., Van Nieuwenhuyse, A., Aerts, J.M., Van Orshoven, J., Somers, B., Linard, C., Dendoncker, N., 2022. Mapping abundance distributions of allergenic tree species in urbanized landscapes: A nation-wide study for Belgium using forest inventory and citizen science data. Landsc. Urban Plan. 218, 104286. https://doi.org/10.1016/j.landurbplan.2021.104286

Dumortier, M., Van Gossum, P., Van Calster, H., Adriaens, D., Adriaenssens, V., Alaerts, K., Brys, R., Cools, N., De Knijf, G., Denys, L., De Saeger, S., De Vos, B., Devos, K., Leyssen, A., Oosterlynck, P., Packet, J., Peymen, J., Pollet, M., Provoost, S., Scheppers, T., Spanhove, T., Thomaes, A., Vanden Borre, J., Van Den Broeck, A., Vanderhaeghe, F., Vandevoorde, B., Hoffmann, M., 2022. Voorstel voor een Meetnet Biodiversiteit Agrarisch Gebied. Adviezen van het Instituut voor Natuur- en Bosonderzoek, Nr. INBO.A.4387. Instituut voor Natuur- en Bosonderzoek, Brussel.

Dupont, Y.L., Overgaard Nielsen, B., 2006. Species composition, feeding specificity and larval trophic level of flower-visiting insects in fragmented versus continuous heathlands in Denmark. Biol. Conserv. 131, 475–485. https://doi.org/10.1016/j.biocon.2005.12.020

Early, R., Bradley, B.A., Dukes, J.S., Lawler, J.J., Olden, J.D., Blumenthal, D.M., Gonzalez, P., Grosholz, E.D., Ibañez, I., Miller, L.P., Sorte, C.J.B., Tatem, A.J., 2016. Global threats from invasive alien species in the twenty-first century and national response capacities. Nat. Commun. 7, 12485. https://doi.org/10.1038/ncomms12485

El-Gabbas, A., Dormann, C.F., 2017. Improved species-occurrence predictions in data-poor regions: using large-scale data and bias correction with down-weighted Poisson regression and Maxent. Ecography (Cop.). 40, 001–011. https://doi.org/10.1111/ecog.03149

Elith, J., Graham, C., Anderson, R., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. Overton, J., Townsend Peterson, A., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Shapire, R.E., Soberón, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography (Cop.). 29, 129–151. https://doi.org/10.1111/j.1432-1033.1987.tb13499.x

Elith, J., Leathwick, J., 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Divers. Distrib. 13, 265–275. https://doi.org/10.1111/j.1472-4642.2007.00340.x

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Elith, J., Leathwick, J.R., 2009. Species distribution models: Ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., En Chee, Y., Yates, C.J., 2010. A statistical explanation of MaxEnt for ecologists. Divers. Distrib. 1–15. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Escamilla Molgora, J.M., Sedda, L., Diggle, P.J., Atkinson, P.M., 2022. A taxonomic-based joint species distribution model for presence-only data. J. R. Soc. Interface 19, 20210681. https://doi.org/10.1098/rsif.2021.0681

European Commission, 2022. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on nature restoration. Vol. COM(2022). Brussels.

European Commission, 2020. EU Biodiversity Strategy for 2030, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Brussels.

European Commission, 2019. Guidance on a strategic framework for further supporting the deployment of EU-level green and blue infrastructure, Commission Staff Working Document. Brussels.

European Commission, 2011. Our life insurance, our natural capital: an EU biodiversity strategy to 2020. Communication from the commission to the European Parliament, the Council, the Economic and Social Committee and the Committee of the Regions, COM(2011) 244 final. Brussels. https://doi.org/10.5738/jale.20.37

Evens, R., Beenaerts, N., Neyens, T., Witters, N., Smeets, K., Artois, T., 2018. Proximity of breeding and foraging areas affects foraging effort of a crepuscular, insectivorous bird. Sci. Rep. 8, 11. https://doi.org/10.1038/s41598-018-21321-0

Evens, R., Jacot, A., Artois, T., Ulenaers, E., Neyens, T., Rappaz, L., Theux, C., Pradervand, J.-N., 2021. Improved ecological insights commission new conservation targets for a crepuscular bird species. Anim. Conserv. 24, 457–469. https://doi.org/10.1111/acv.12650

Ewers, R.M., Thorpe, S., Didham, R.K., 2007. Synergistic interactions between edge and area effects in a heavily fragmented landscape. Ecology 88, 96–106. https://doi.org/10.1890/0012-9658(2007)88[96:SIBEAA]2.0.CO;2

Eyre, A.C., Briscoe, N.J., Harley, D.K.P., Lumsden, L.F., McComb, L.B., Lentini, P.E., 2022. Using species distribution models and decision tools to direct surveys and identify potential translocation sites for a critically endangered species. Divers. Distrib. 28, 700–711. https://doi.org/10.1111/ddi.13469

Fagúndez, J., 2013. Heathlands confronting global change: drivers of biodiversity loss from past to future scenarios. Ann. Bot. 111, 151–172. https://doi.org/10.1093/aob/mcs257

Fahrig, L., 2003. Effects of Habitat Fragmentation on Biodiversity. Annu. Rev. Ecol. Evol. Syst. 34, 487–515. https://doi.org/10.1146/annurev.ecolsys.34.011802.132419

Fahrig, L., Baudry, J., Brotons, L., Burel, F.G., Crist, T.O., Fuller, R.J., Sirami, C., Siriwardena, G.M., Martin, J.-L., 2011. Functional landscape heterogeneity and animal biodiversity in agricultural landscapes. Ecol. Lett. 14, 101–112. https://doi.org/10.1111/j.1461-0248.2010.01559.x

Farmer, R.G., Leonard, M.L., Horn, A.G., 2012. Observer Effects and Avian-Call-Count Survey Quality: Rare-Species Biases and Overconfidence. Auk 129, 76–86. https://doi.org/10.1525/AUK.2012.11129

Farrell, S.L., Collier, B.A., Skow, K.L., Long, A.M., Campomizzi, A.J., Morrison, M.L., Hays, K.B., Wilkins, R.N., 2013. Using LiDAR-derived vegetation metrics for high-resolution, species distribution models for conservation planning. Ecosphere 4, 1–18. https://doi.org/10.1890/ES12-000352.1

Farwell, L.S., Elsen, P.R., Razenkova, E., Pidgeon, A.M., Radeloff, V.C., 2020. Habitat heterogeneity captured by 30-m resolution satellite image texture predicts bird richness across the United States. Ecol. Appl. 30, e02157. https://doi.org/10.1002/EAP.2157

Farwell, L.S., Gudex-Cross, D., Anise, I.E., Bosch, M.J., Olah, A.M., Radeloff, V.C., Razenkova, E., Rogova, N., Silveira, E.M.O., Smith, M.M., Pidgeon, A.M., 2021. Satellite image texture captures vegetation heterogeneity and explains patterns of bird richness. Remote Sens. Environ. 253, 112175. https://doi.org/10.1016/j.rse.2020.112175

Fernández, N., Román, J., Delibes, M., 2016. Variability in primary productivity determines metapopulation dynamics. Proc. R. Soc. B Biol. Sci. 283, 20152998. https://doi.org/10.1098/rspb.2015.2998

Ferrari, S., Cribari-Neto, F., 2004. Beta Regression for Modelling Rates and Proportions. J. Appl. Stat. 31, 799–815. https://doi.org/10.1080/0266476042000214501

Ferrier, S., Powell, G.V.N., Richardson, K.S., Manion, G., Overton, J.M., Allnutt, T.F., Cameron, S.E., Mantle, K., Burgess, N.D., Faith, D.R., Lamoreux, J.F., Kier, G., Hijmans, R.J., Funk, V.A., Cassis, G.A., Fisher, B.L., Flemons, P., Lees, D., Lovett, J.C., Van Rompaey, R.S.A.R., 2004. Mapping more of terrestrial biodiversity for global conservation assessment. Bioscience 54, 1101–1109. https://doi.org/10.1641/0006-3568(2004)054[1101:MMOTBF]2.0.CO;2

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315. https://doi.org/10.1002/joc.5086

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49. https://doi.org/10.1017/S0376892997000088

Fithian, W., Hastie, T., 2013. Finite-sample equivalence in statistical models for presence-only data. Ann. Appl. Stat. 7, 1917–1939. https://doi.org/10.1214/13-AOAS667

Fithian, W., Elith, J., Hastie, T., Keith, D.A., 2015. Bias correction in species distribution models: Pooling survey and collection data for multiple species. Methods Ecol. Evol. 6, 424–438. https://doi.org/10.1111/2041-210X.12242

Fitzpatrick, M.C., Preisser, E.L., Ellison, A.M., Elkinton, J.S., 2009. Observer Bias and the Detection of Low-Density Populations. Ecol. Appl. 19, 1673–1679.

Fletcher, R.J., Didham, R.K., Banks-Leite, C., Barlow, J., Ewers, R.M., Rosindell, J., Holt, R.D., Gonzalez, A., Pardini, R., Damschen, E.I., Melo, F.P.L., Ries, L., Prevedello, J.A., Tscharntke, T., Laurance, W.F., Lovejoy, T., Haddad, N.M., 2018. Is habitat fragmentation good for biodiversity? Biol. Conserv. 226, 9–15. https://doi.org/10.1016/j.biocon.2018.07.022

Fletcher, R.J., Hefley, T.J., Robertson, E.P., Zuckerberg, B., Mccleery, R.A., Dorazio, R.M., 2019. A practical guide for combining data to model species distributions. Ecology 100, e02710. https://doi.org/10.1002/ecy.2710

Frans, V.F., Augé, A.A., Edelhoff, H., Erasmi, S., Balkenhol, N., Engler, J.O., 2018. Quantifying apart what belongs together: A multi-state species distribution modelling framework for species using distinct habitats. Methods Ecol. Evol. 9, 98–108. https://doi.org/10.1111/2041-210X.12847

Gábor, L., Moudrý, V., Barták, V., Lecours, V., 2020. How do species and data characteristics affect species distribution models and when to use environmental filtering? Int. J. Geogr. Inf. Sci. 34, 1567–1584. https://doi.org/10.1080/13658816.2019.1615070

Gábor, L., Šímová, P., Keil, P., Zarzo-Arias, A., Marsh, C.J., Rocchini, D., Malavasi, M., Barták, V., Moudrý, V., 2022. Habitats as predictors in species distribution models: Shall we use continuous or binary data? Ecography (Cop.). 1–9. https://doi.org/10.1111/ecog.06022

Ghyselinck, N., 2021. Transnational Report PARTRIDGE. Interreg North Sea Region, European Development Fund, European Union.

Gibson, L.A., Wilson, B.A., Cahill, D.M., Hill, J., 2004. Spatial prediction of rufous bristlebird habitat in a coastal heathland: A GIS-based approach. J. Appl. Ecol. 41, 213–223. https://doi.org/10.1111/j.0021-8901.2004.00896.x

Giraud, C., Calenge, C., Coron, C., Julliard, R., 2016. Capitalizing on opportunistic data for monitoring relative abundances of species. Biometrics 72, 649–658. https://doi.org/10.1111/biom.12431

Goetz, S.J., Steinberg, D., Betts, M.G., Holmes, R.T., Doran, P.J., Dubayah, R., Hofton, M., 2010. Lidar remote sensing variables predict breeding habitat of a Neotropical migrant bird, Ecology.

Gottschalk, T.K., Aue, B., Hotes, S., Ekschmitt, K., 2011. Influence of grain size on species-habitat models. Ecol. Modell. 222, 3403–3412. https://doi.org/10.1016/j.ecolmodel.2011.07.008

Graf, R.F., Mathys, L., Bollmann, K., 2009. Habitat assessment for forest dwelling species using LiDAR remote sensing: Capercaillie in the Alps. For. Ecol. Manage. 257, 160–167. https://doi.org/10.1016/j.foreco.2008.08.021

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., Mccarthy, M.A., Tingley, R., Wintle, B.A., 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Glob. Ecol. Biogeogr. 24, 276–292. https://doi.org/10.1111/geb.12268

Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N.G., Lehmann, A., Zimmermann, N.E., 2006. Using niche-based models to improve the sampling of rare species. Conserv. Biol. 20, 501–511. https://doi.org/10.1111/j.1523-1739.2006.00354.x

Guisan, A., Edwards, T.C., Jr, Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Modell. 157, 89–100. https://doi.org/10.1111/j.1365-3040.1985.tb01209.x

Guisan, A., Thuiller, W., 2005. Predicting species distribution: Offering more than simple habitat models. Ecol. Lett. 8, 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., Brotons, L., Mcdonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H.P., Buckley, Y.M., 2013. Predicting species distributions for conservation decisions. Ecol. Lett. 16, 1424–1435. https://doi.org/10.1111/ele.12189

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Modell. 135, 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Gustafson, E.J., 1998. Quantifying landscape spatial pattern: what is the state of the art? Ecosystems 1, 143–156.

Haddad, N.M., Baum, K.A., 1999. An experimental test of corridor effects on butterfly densities. Ecol. Appl. 9, 623–633. https://doi.org/10.1890/1051-0761(1999)009[0623:AETOCE]2.0.CO;2

Hanberry, B.B., He, H.S., Dey, D.C., 2012. Sample sizes and model comparison metrics for species distribution models. Ecol. Modell. 227, 29–33. https://doi.org/10.1016/j.ecolmodel.2011.12.001

Hanski, I., 1998. Metapopulation dynamics. Nature 396, 41–49. https://doi.org/10.1038/23876

Hanspach, J., Pompe, S., Klotz, S., 2010. Predictive performance of plant species distribution models depends on species traits. Perspect. Plant Ecol. Evol. Syst. 12, 219–225. https://doi.org/10.1016/j.ppees.2010.04.002

Haralick, R.M., 1979. Statistical and structural approaches to texture. Proc. IEEE 67, 786–804.

Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. Textural Features for Image Classification. IEEE Trans. Syst. Man Cybern. SMC-3, 610–621. https://doi.org/10.1109/TSMC.1973.4309314

He, K.S., Bradley, B.A., Cord, A.F., Rocchini, D., Tuanmu, M.N., Schmidtlein, S., Turner, W., Wegmann, M., Pettorelli, N., 2015. Will remote sensing shape the next generation of species distribution models? Remote Sens. Ecol. Conserv. 1, 4–18. https://doi.org/10.1002/rse2.7

Henckel, L., Bradter, U., Jönsson, M., Isaac, N.J.B., Snäll, T., 2020. Assessing the usefulness of citizen science data for habitat suitability modelling: Opportunistic reporting versus sampling based on a systematic protocol. Divers. Distrib. 00, 1–15. https://doi.org/10.1111/ddi.13128

Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography (Cop.). 29, 773–785. https://doi.org/10.1111/j.0906-7590.2006.04700.x

Herremans, M., De Knijf, G., Hansen, K., Westra, T., Vanreusel, W., Martens, E., Van Gossum, H., Anselin, A., Vermeersch, G., Pollet, M., 2014. Monitoring van beleidsrelevante soorten in Vlaanderen met inzet van vrijwilligers. Rapporten van het Instituut voor Natuur- en Bosonderzoek 2014 (rapportnr. INBO.R.2014.1628917). Instituut voor Natuur- en Bosonderzoek, Brussel.

Herremans, M., Swinnen, K., Vanreusel, W., Vercayie, D., Veraghtert, W., Vanormelingen, P., 2018. www.waarnemingen.be. Een veelzijdig portaal voor natuurgegevens. Natuur.focus 17, 153–166.

Hesselbarth, M.H.K., Sciaini, M., With, K.A., Wiegand, K., Nowosad, J., 2019. landscapemetrics: an open-source R tool to calculate landscape metrics. Ecography (Cop.). 42, 1648–1657. https://doi.org/10.1111/ecog.04617

Hijmans, R.J., Philips, S., Leathwick, J., Elith, J., 2017. dismo: Species Distribution Modeling. R package version 1.1-4.

Hirzel, A.H., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. Ecol. Modell. 145, 111–121. https://doi.org/10.1016/S0304-3800(01)00396-9

Hogeweg, L., Schermer, M., Pieterse, S., Roeke, T., Gerritsen, W., 2019. Machine Learning Model for Identifying Dutch/ Belgian Biodiversity. Biodivers. Inf. Sci. Stand. 3. https://doi.org/10.3897/biss.3.39229

Huber, N., Kienast, F., Ginzler, C., Pasinelli, G., 2016. Using remote-sensing data to assess habitat selection of a declining passerine at two spatial scales. Landsc. Ecol. 31, 1919–1937. https://doi.org/10.1007/s10980-016-0370-1

Hubert-Moy, L., Rozo, C., Perrin, G., Bioret, F., Rapinel, S., 2022. Large-scale and fine-grained mapping of heathland habitats using open-source remote sensing data. Remote Sens. Ecol. Conserv. 8, 448–463. https://doi.org/10.1002/rse2.253

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83, 195–213. https://doi.org/10.1016/S0020-1693(00)85959-9

Husson, F., Josse, J., Pagès, J., 2010. Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data?, Applied Mathematics Department.

IPBES, 2019. Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. IPBES secretariat, Bonn, Germany. https://doi.org/10.1111/padr.12283

Isaac, N.J.B., Jarzyna, M.A., Keil, P., Dambly, L.I., Boersch-Supan, P.H., Browning, E., Freeman, S.N., Golding, N., Guillera-Arroita, G., Henrys, P.A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O.L., Schmucki, R., Simmonds, E.G., O'Hara, R.B., 2020. Data Integration for Large-Scale Models of Species Distributions. Trends Ecol. Evol. 35, 56–67. https://doi.org/10.1016/j.tree.2019.08.006

Isaac, N.J.B., Pocock, M.J.O., 2015. Bias and information in biological records. Biol. J. Linn. Soc. 115, 522–531. https://doi.org/10.1111/bij.12517/abstract

Isaac, N.J.B., van Strien, A.J., August, T.A., de Zeeuw, M.P., Roy, D.B., 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. Methods Ecol. Evol. 5, 1052–1060. https://doi.org/10.1111/2041-210X.12254

IUCN Standards and Petitions Committee, 2022. Guidelines for Using the IUCN Red List Categories and Criteria. Version 15.1. Gland, Switzerland.

Jacobs, C., Zipf, A., 2017. Completeness of citizen science biodiversity data from a volunteered geographic information perspective. Geo-spatial Inf. Sci. 20, 3–13. https://doi.org/10.1080/10095020.2017.1288424

Jiménez-Valverde, A., 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Glob. Ecol. Biogeogr. 21, 498–507. https://doi.org/10.1111/j.1466-8238.2011.00683.x

Jiménez-Valverde, A., Lobo, J., Hortal, J., 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. Community Ecol. 10, 196–205. https://doi.org/10.1556/ComEc.10.2009.2.9

Jiménez-Valverde, A., Lobo, J.M., 2007. Threshold criteria for conversion of probability of species presence to either-or presence-absence. Acta oecologica 31, 361–369. https://doi.org/10.1016/j.actao.2007.02.001

Jiménez, L., Soberón, J., 2020. Leaving the area under the receiving operating characteristic curve behind: An evaluation method for species distribution modelling applications based on presence-only data. Methods Ecol. Evol. 00, 1–16. https://doi.org/10.1111/2041-210X.13479

Johnston, A., Fink, D., Hochachka, W.M., Kelling, S., 2018. Estimates of observer expertise improve species distributions from citizen science data. Methods Ecol. Evol. 9, 88–97. https://doi.org/10.1111/2041-210X.12838

Johnston, A., Fink, D., Hochachka, W.M., Kelling, S., 2017. Estimates of observer expertise improve species distributions from citizen science data. Methods Ecol. Evol. 00, 1–10. https://doi.org/10.1111/2041-210X.12838

Johnston, A., Hochachka, W.M., Strimas-Mackey, M.E., Ruiz Gutierrez, V., Robinson, O.J., Miller, E.T., Auer, T., Kelling, S.T., Fink, D., 2021. Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. Divers. Distrib. 27, 1265–1277. https://doi.org/10.1111/DDI.13271

Johnston, A., Matechou, E., Dennis, E.B., 2023. Outstanding challenges and future directions for biodiversity monitoring using citizen science data. Methods Ecol. Evol. 14, 103–116. https://doi.org/10.1111/2041-210X.13834

Johnston, A., Newson, S.E., Risely, K., Musgrove, A.J., Massimino, D., Baillie, S.R., Pearce-Higgins, J.W., 2014. Species traits explain variation in detectability of UK birds. Bird Study 61, 340–350. https://doi.org/10.1080/00063657.2014.941787

Jones, L., Stevens, C., Rowe, E.C., Payne, R., Caporn, S.J.M., Evans, C.D., Field, C., Dale, S., 2017. Can on-site management mitigate nitrogen deposition impacts in non-wooded habitats? Biol. Conserv. 212, 464–475. https://doi.org/10.1016/j.biocon.2016.06.012

Kadykalo, A.N., Buxton, R.T., Morrison, P., Anderson, C.M., Bickerton, H., Francis, C.M., Smith, A.C., Fahrig, L., 2021. Bridging research and practice in conservation. Conserv. Biol. 35, 1725–1737. https://doi.org/10.1111/cobi.13732

Kaivanto, K., 2008. Maximization of the sum of sensitivity and specificity as a diagnostic cutpoint criterion. J. Clin. Epidemiol. https://doi.org/10.1016/j.jclinepi.2007.10.011

Kallimanis, A.S., Panitsa, M., Dimopoulos, P., 2017. Quality of non-expert citizen science data collected for habitat type conservation status assessment in Natura 2000 protected areas. Sci. Rep. 7, 1–10. https://doi.org/10.1038/s41598-017-09316-9

Kamp, J., Oppel, S., Heldbjerg, H., Nyegaard, T., Donald, P.F., 2016. Unstructured citizen science data fail to detect long-term population declines of common birds in Denmark. Divers. Distrib. 22, 1024–1035. https://doi.org/10.1111/ddi.12463

Kearney, M., Porter, W., 2009. Mechanistic niche modelling: Combining physiological and spatial data to predict species' ranges. Ecol. Lett. 12, 334–350. https://doi.org/10.1111/j.1461-0248.2008.01277.x

Kelling, S., Johnston, A., Bonn, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Fernandez, M., Hochachka, W.M., Julliard, R., Kraemer, R., Guralnick, R., 2019. Using Semistructured Surveys to Improve Citizen Science Data for Monitoring Biodiversity. Bioscience 69, 170–179. https://doi.org/10.1093/biosci/biz010

Kelling, S., Johnston, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Bonn, A., Fernandez, M., Hochachka, W.M., Julliard, R., Kraemer, R., Guralnick, R., 2018. Finding the signal in the Noise of Citizen Science Observations. bioRxiv 326314. https://doi.org/10.1101/326314

Kelling, S., Johnston, A., Hochachka, W.M., Iliff, M., Fink, D., Gerbracht, J., Lagoze, C., La Sorte, F.A., Moore, T., Wiggins, A., Wong, W.K., Wood, C., Yu, J., 2015. Can observation skills of citizen scientists be estimated using species accumulation curves? PLoS One 69, 170–179. https://doi.org/10.1371/journal.pone.0139600

Kelling, S., Lagoze, C., Wong, W., Yu, J., Damoulas, T., Gerbracht, J., Fink, D., Gomes, C., 2013. eBird: A Human / Computer Learning Network to Improve Biodiversity Conservation and Research. AI Mag. 10–20.

Kéry, M., Royle, J.A., Schmid, H., Schaub, M., Volet, B., Häfliger, G., Zbinden, N., 2009. Site-Occupancy Distribution Modeling to Correct Population-Trend Estimates Derived from Opportunistic Observations. Conserv. Biol. 24, 1388–1397. https://doi.org/10.1111/j.1523-1739.2010.01479.x

Kéry, M., Schmidt, B.R., 2008. Imperfect detection and its consequences for monitoring for conservation. Community Ecol. 9, 207–216. https://doi.org/10.1556/ComEc.9.2008.2.10

Klijn, F., de Haes, H.A.U., 1994. A hierarchical approach to ecosystems and its implications for ecological land classification. Landsc. Ecol. 9, 89–104. https://doi.org/10.1007/BF00124376

Kosmala, M., Wiggins, A., Swanson, A., Simmons, B., 2016. Assessing data quality in citizen science. Front. Ecol. Environ. 14, 551–560. https://doi.org/10.1002/fee.1436

Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A.K., Augeri, D.M., Cheyne, S.M., Hearn, A.J., Ross, J., Macdonald, D.W., Mathai, J., Eaton, J., Marshall, A.J., Semiadi, G., Rustam, R., Bernard, H., Alfred, R., Samejima, H., Duckworth, J.W., Breitenmoser-Wuersten, C., Belant, J.L., Hofer, H., Wilting, A., 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. Divers. Distrib. 19, 1366–1379. https://doi.org/10.1111/DDI.12096

Kühn, I., 2007. Incorporating spatial autocorrelation may invert observed patterns. Divers. Distrib. 13, 66–69. https://doi.org/10.1111/j.1472-4642.2006.00293.x

Laanisto, L., Tamme, R., Hiiesalu, I., Szava-Kovats, R., Gazol, A., Pärtel, M., 2013. Microfragmentation concept explains non-positive environmental heterogeneity-diversity relationships. Oecologia 171, 217–226. https://doi.org/10.1007/s00442-012-2398-5

Lahoz-monfort, J.J., Guillera-arroita, G., Wintle, B.A., 2014. Imperfect detection impacts the performance of species distribution models. Glob. Ecol. Biogeogr. 23, 504–515. https://doi.org/10.1111/geb.12138

Lakner, S., Holst, C., Dittrich, A., Hoyer, C., Pe'er, G., 2019. Impacts of the EU's Common Agricultural Policy on Biodiversity and Ecosystem Services, in: Schröther, M. (Ed.), Atlas of Ecosystem Services. Springer International Publishing AG, pp. 383–389. https://doi.org/10.1007/978-3-319-96229-0_58

Lawson, C.R., Hodgson, J.A., Wilson, R.J., Richards, S.A., 2014. Prevalence, thresholds and the performance of presence-absence models. Methods Ecol. Evol. 5, 54–64. https://doi.org/10.1111/2041-210X.12123

Le, S., Josse, J., Husson, F.F.F., Lê, S., Josse, J., Husson, F.F.F., 2008. FactoMineR: An R Package for Multivariate Analysis. J. Stat. Softw. 25, 1–18. https://doi.org/10.18637/jss.v025.i01

Leclère, D., Obersteiner, M., Barrett, M., Butchart, S.H.M., Chaudhary, A., De Palma, A., DeClerck, F.A.J., Di Marco, M., Doelman, J.C., Dürauer, M., Freeman, R., Harfoot, M., Hasegawa, T., Hellweg, S., Hilbers, J.P., Hill, S.L.L., Humpenöder, F., Jennings, N., Krisztin, T., Mace, G.M., Ohashi, H., Popp, A., Purvis, A., Schipper, A.M., Tabeau, A., Valin, H., van Meijl, H., van Zeist, W.J., Visconti, P., Alkemade, R., Almond, R., Bunting, G., Burgess, N.D., Cornell, S.E., Di Fulvio, F., Ferrier, S., Fritz, S., Fujimori, S., Grooten, M., Harwood, T., Havlík, P., Herrero, M., Hoskins, A.J., Jung, M., Kram, T., Lotze-Campen, H., Matsui, T., Meyer, C., Nel, D., Newbold, T., Schmidt-Traub, G., Stehfest, E., Strassburg, B.B.N., van Vuuren, D.P., Ware, C., Watson, J.E.M., Wu, W., Young, L., 2020. Bending the curve of terrestrial biodiversity needs an integrated strategy. Nature 585, 551–556. https://doi.org/10.1038/s41586-020-2705-y

Lee-Yaw, J.A., McCune, J.L., Pironon, S., Sheth, S.N., 2022. Species distribution models rarely predict the biology of real populations. Ecography (Cop.). e05877. https://doi.org/10.1111/ecog.05877

Leitão, P.J., Santos, M.J., 2019. Improving models of species ecological niches: A remote sensing overview. Front. Ecol. Evol. 7, 7. https://doi.org/10.3389/fevo.2019.00009

Lembrechts, J.J., Aalto, J., Ashcroft, M.B., De Frenne, P., Kopecký, M., Lenoir, J., Luoto, M., Maclean, I.M.D., Roupsard, O., Fuentes-Lillo, E., García, R.A., Pellissier, L., Pitteloud, C., Alatalo, J.M., Smith, S.W., Björk, R.G., Muffler, L., Ratier Backes, A., Cesarz, S., Gottschall, F., Okello, J., Urban, J., Plichta, R., Svátek, M., Phartyal, S.S., Wipf, S., Eisenhauer, N., Pușcaș, M., Turtureanu, P.D., Varlagin, A., Dimarco, R.D., Jump, A.S., Randall, K., Dorrepaal, E., Larson, K., Walz, J., Vitale, L., Svoboda, M., Finger Higgens, R., Halbritter, A.H., Curasi, S.R., Klupar, I., Koontz, A., Pearse, W.D., Simpson, E., Stemkovski, M., Jessen Graae, B., Vedel Sørensen, M., Høye, T.T., Fernández Calzado, M.R., Lorite, J., Carbognani, M., Tomaselli, M., Forte, T.G.W., Petraglia, A., Haesen, S., Somers, B., Van Meerbeek, K., Björkman, M.P., Hylander, K., Merinero, S., Gharun, M., Buchmann, N., Dolezal, J., Matula, R., Thomas, A.D., Bailey, J.J., Ghosn, D., Kazakis, G., de Pablo, M.A., Kemppinen, J., Niittynen, P., Rew, L., Seipel, T., Larson, C., Speed, J.D.M., Ardö, J., Cannone, N., Guglielmin, M., Malfasi, F., Bader, M.Y., Canessa, R., Stanisci, A., Kreyling, J., Schmeddes, J., Teuber, L., Aschero, V., Čiliak, M., Máliš, F., De Smedt, P., Govaert, S., Meeussen, C., Vangansbeke, P., Gigauri, K., Lamprecht, A., Pauli, H., Steinbauer, K., Winkler, M., Ueyama, M., Nuñez, M.A., Ursu, T.M., Haider, S., Wedegärtner, R.E.M., Smiljanic, M., Trouillier, M., Wilmking, M., Altman, J., Brůna, J., Hederová, L., Macek, M., Man, M., Wild, J., Vittoz, P., Pärtel, M., Barančok, P., Kanka, R., Kollár, J., Palaj, A., Barros, A., Mazzolari, A.C., Bauters, M., Boeckx, P., Benito Alonso, J.L., Zong, S., Di Cecco, V., Sitková, Z., Tielbörger, K., van den Brink, L., Weigel, R., Homeier, J., Dahlberg, C.J., Medinets, S., Medinets, V., De Boeck, H.J., Portillo-Estrada, M., Verryckt, L.T., Milbau, A., Daskalova, G.N., Thomas, H.J.D., Myers-Smith, I.H., Blonder, B., Stephan, J.G., Descombes, P., Zellweger, F., Frei, E.R., Heinesch, B., Andrews, C., Dick, J., Siebicke, L., Rocha, A., Senior, R.A., Rixen, C., Jimenez, J.J., Boike, J., Pauchard, A., Scholten, T., Scheffers, B., Klinges, D., Basham, E.W., Zhang, J., Zhang, Z., Géron, C., Fazlioglu, F., Candan, O., Sallo Bravo, J., Hrbacek, F., Laska, K., Cremonese, E., Haase, P., Moyano, F.E., Rossi, C., Nijs, I., 2020. SoilTemp: A global database of near-surface temperature. Glob. Chang. Biol. 26, 6616–6629. https://doi.org/10.1111/gcb.15123

Lembrechts, J.J., Nijs, I., Lenoir, J., 2019. Incorporating microclimate into species distribution models. Ecography (Cop.). 42, 1267–1279. https://doi.org/10.1111/ecog.03947

Lenoir, J., Hattab, T., Pierre, G., 2017. Climatic microrefugia under anthropogenic climate change: implications for species redistribution. Ecography (Cop.). 40, 253–266. https://doi.org/10.1111/ecog.02788

Lin, Y.-P., Lin, W.-C., Lien, W.-Y., Anthony, J., Petway, J.R., 2017. Identifying Reliable Opportunistic Data for Species Distribution Modeling: A Benchmark Data Optimization Approach. Environments 4, 81. https://doi.org/10.3390/environments4040081

Liu, C., Newell, G., White, M., 2019. The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. Ecography (Cop.). 42, 535–548. https://doi.org/10.1111/ecog.03188

Liu, C., White, M., Newell, G., 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. J. Biogeogr. 40, 778–789. https://doi.org/10.1111/jbi.12058

Liu, H.Q., Huete, A., 1995. Feedback based modification of the NDVI to minimize canopy background and atmospheric noise. IEEE Trans. Geosci. Remote Sens. 33, 457–465. https://doi.org/10.1109/36.377946

Lobo, J.M., Jiménez-Valverde, A., Hortal, J., 2010. The uncertain nature of absences and their importance in species distribution modelling. Ecography (Cop.). 33, 103–114. https://doi.org/10.1111/j.1600-0587.2009.06039.x

Lobo, J.M., Jiménez-valverde, A., Real, R., 2008. AUC: A misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. 17, 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

Louette, G., Adriaens, D., Paelinckx, D., Hoffmann, M., 2015. Implementing the habitats directive: How science can support decision making. J. Nat. Conserv. 23, 27–34. https://doi.org/10.1016/j.jnc.2014.12.002

Lustig, A., Stouffer, D.B., Doscher, C., Worner, S.P., 2017. Landscape metrics as a framework to measure the effect of landscape structure on the spread of invasive insect species. Landsc. Ecol. 32, 2311–2325. https://doi.org/10.1007/s10980-017-0570-3

MacArthur, R.H., Wilson, E.O., 1967. The Theory of Island Biogeography. Princeton University Press, Princeton.

Mace, G.M., Norris, K., Fitter, A.H., 2012. Biodiversity and ecosystem services: A multilayered relationship. Trends Ecol. Evol. 27, 19–25. https://doi.org/10.1016/j.tree.2011.08.006

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L., Hines, J.E., 2017. Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence. Elsevier.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L., Hines, J.E., 2006. Occupancy Estimation and Modeling, Analysis of Capture-Recapture Data. https://doi.org/10.1201/b17222-10

Maclean, I.M.D., Duffy, J.P., Haesen, S., Govaert, S., De Frenne, P., Vanneste, T., Lenoir, J., Lembrechts, J.J., Rhodes, M.W., Van Meerbeek, K., 2021. On the measurement of microclimate. Methods Ecol. Evol. 12, 1397–1410. https://doi.org/10.1111/2041-210X.13627

Maes, D., Adriaens, D., Van Der Meulen, M., Poelmans, L., Van Landuyt, W., Anselin, A., Casaer, J., De Knijf, G., Devos, K., Packet, J., Speybroeck, J., Stienen, E., Stuyck, J., Filiep, T., Van Daele, T., Van Den Berge, K., Van Elegem, B., Vermeersch, G., Wils, C., Pollet, M., 2015a. Afbakenen van potentiële leefgebiedenkaarten voor Europese en Vlaamse prioritaire soorten in het kader van de voortoets. Versie 2.0. Rapporten van het Instituut voor Natuur- en Bosonder- zoek 2015 (INBO.R.2015.10201559). Instituut voor Natuur- en Bosonderzoek, Brussel.

Maes, D., Adriaens, D., van der Meulen, M., Poelmans, L., Vandegehuchte, M., Everaert, J., Verhaeghe, F., Anselin, A., Casaer, J., Decleer, K., De Knijf, G., Devos, K., Engelen, G., Gouwy, J., Packet, J., Stienen, E., Stuyck, J., Thomaes, A., T'jollyn, F., Speybroeck, J., Van Den Berge, K., Van Elegem, B., Van Landuyt, W., Vermeersch, G., Wils, C., Pollet, M., 2017a. Potentiële leefgebieden voor bedreigde soorten. Mogelijke toepassingen in het Vlaamse natuurbeleid en -beheer. Natuur.Focus 16, 56–66.

Maes, D., Adriaens, T., Decleer, K., Foquet, B., Foquet, R., Lambrechts, J., Lock, K., Piesschaert, F., 2017b. IUCN Rode Lijst van de sprinkhanen en krekels in Vlaanderen. Rapporten van het Instituut voor Natuur- en Bosonderzoek 2017 (29). Instituut voor Natuur- en Bosonderzoek, Brussel.

Maes, D., Bauwens, D., De Bruyn, L., Anselin, A., Vermeersch, G., Van Landuyt, W., De Knijf, G., Gilbert, M., 2005. Species richness coincidence: Conservation strategies based on predictive modelling. Biodivers. Conserv. 14, 1345–1364. https://doi.org/10.1007/S10531-004-9662-X

Maes, D., Brosens, D., T'jollyn, F., Desmet, P., Piesschaert, F., Van Hoey, S., Adriaens, T., Dekoninck, W., Devos, K., Lock, K., Onkelinx, T., Packet, J., Speybroeck, J., Thomaes, A., Van Den Berge, K., Van Landuyt, W., Verreycken, H., 2019a. A database of threat statuses and life-history traits of Red List species in Flanders (northern Belgium). Biodivers. Data J. 7, 1–19. https://doi.org/10.3897/BDJ.7.e34089

Maes, D., Brosens, D., T'Jollyn, F., Desmet, P., Piesschaert, F., Van Hoey, S., Adriaens, T., Dekoninck, W., Devos, K., Lock, K., Onkelinx, T., Packet, J., Speybroeck, J., Thomaes, A., Van Den Berge, K., Van Landuyt, W., Verreycken, H., 2019b. Validated red lists of Flanders, Belgium. Research Institute for Nature and Forest (INBO), Brussels. https://doi.org/https://doi.org/10.15468/8tk3tk

Maes, D., Decleer, K., De Keersmaeker, L., Van Uytvanck, J., Louette, G., 2017c. Intensified habitat management to mitigate negative effects of nitrogen pollution can be detrimental for faunal diversity: A comment on Jones et al. (2017). Biol. Conserv. 212, 493–494. https://doi.org/10.1016/j.biocon.2017.03.001

Maes, D., Ellis, S., Goffart, P., Cruickshanks, K.L., van Swaay, C.A.M., Cors, R., Herremans, M., Swinnen, K.R.R., Wils, C., Verhulst, S., De Bruyn, L., Matthysen, E., O'Riordan, S., Hoare, D.J., Bourn, N.A.D., 2019c. The potential of species distribution modelling for reintroduction projects: the case study of the Chequered Skipper in England. J. Insect Conserv. 23, 419–431. https://doi.org/10.1007/s10841-019-00154-w

Maes, D., Everaert, J., Anselin, A., De Bruyn, L., Decleer, K., De Knijf, G., Gouwy, J., Pollet, M., Speybroeck, J., Thomaes, A., Van Den Berge, K., Verhaeghe, F., 2016. Afbakenen van actueel relevante potentiële leefgebieden voor een selectie van habitattypische Europese en Vlaamse prioritaire diersoorten. Rapporten van het Instituut voor Natuur- en Bosonderzoek 2016 (INBO.R.2016. 11534907). Instituut voor Natuur- en Bosonderzoek, Brussel.

Maes, D., Herremans, M., Vantieghem, P., Veraghtert, W., Jacobs, I., Fajgenblat, M., Dyck, H. Van, 2021. IUCN Rode Lijst van de dagvlinders in Vlaanderen 2021. Rapporten van het Instituut voor Natuur- en Bosonderzoek 2021 (10). Instituut voor Natuur- en Bosonderzoek, Brussel. https://doi.org/10.21436/inbor.34052968

Maes, D., Isaac, N.J.B., Harrower, C.A., Collen, B., van Strien, A.J., Roy, D.B., 2015b. The use of opportunistic data for IUCN Red List assessments. Biol. J. Linn. Soc. 115, 690–706. https://doi.org/10.1111/bij.12530

Maes, D., Titeux, N., Hortal, J., Anselin, A., Decleer, K., de Knijf, G., Fichefet, V., Luoto, M., 2010. Predicted insect diversity declines under climate change in an already impoverished region. J. Insect Conserv. 14, 485–498. https://doi.org/10.1007/s10841-010-9277-3

Maes, D., Van Calster, H., Herremans, M., Van Dyck, H., 2022. Challenges and bottlenecks for butterfly conservation in a highly anthropogenic region: Europe's worst case scenario revisited. Biol. Conserv. 274. https://doi.org/10.1016/j.biocon.2022.109732

Maes, D., van der Meulen, M., Verhaeghe, F., Bot, J., Defoort, T., Poelmans, L., Adriaens, D., De Knijf, G., Devos, K., Packet, J., Speybroeck, J., Stienen, E., T'jollyn, F., Van Den Berge, K., Van Landuyt, W., Vermeersch, G., Wils, C., 2019d. Inschatting van de effecten van habitatuitbreiding in het kader van de realisatie van de instandhoudingsdoelstellingen op potentiële leefgebieden van soorten. Rapporten van het Instituut voor Natuur- en Bosonderzoek 2019 (42). Instituut voor Natuur- en Bosonderzoek, Brussel.

Maes, D., Van Dyck, H., 2005. Habitat quality and biodiversity indicator performances of a threatened butterfly versus a multispecies group for wet heathlands in Belgium. Biol. Conserv. 123, 177–187. https://doi.org/10.1016/j.biocon.2004.11.005

Maes, D., Verhaeghe, F., Pollet, M., Piesschaert, F., Defoort, T., Hoffmann, M., 2018. Het gebruik van losse waarnemingen in het Vlaamse natuurbeleid. De cruciale rol van waarnemingen.be. Natuur.Focus 17, 178–184.

Mantilla-Contreras, J., Schirmel, J., Zerbe, S., 2012. Influence of soil and microclimate on species composition and grass encroachment in heath succession. J. Plant Ecol. 5, 249–259. https://doi.org/10.1093/jpe/rtr031

Maréchal, R., Tavernier, E., 1974. Atlas van België. Commentaar bij de bladen 11A en 11B uittreksels van de bodemkaart bodemassociaties. Pedologie. Commissie voor de Nationale Atlas, Gent.

McCarthy, M.A., Moore, J.L., Morris, W.K., Parris, K.M., Garrard, G.E., Vesk, P.A., Rumpff, L., Giljohann, K.M., Camac, J.S., Bau, S.S., Friend, T., Harrison, B., Yue, B., 2013. The influence of abundance on detectability. Oikos 122, 717–726. https://doi.org/10.1111/j.1600-0706.2012.20781.x

McPherson, J.M., Jetz, W., 2007. Effects of species' ecology on the accuracy of distribution models. Ecography (Cop.). 30, 135–151. https://doi.org/10.1111/j.0906-7590.2007.04823.x

McPherson, J.M., Jetz, W., Rogers, D.J., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? J. Appl. Ecol. 41, 811–823. https://doi.org/10.1111/J.0021-8901.2004.00943.X

Menard, S., 2001. Applied Logistic Regression Analysis. 2nd edition. SAGE Publications, Inc.

Merow, C., Smith, M.J., Silander, J.A., 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography (Cop.). 36, 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x

Merow, C., Wilson, A.M., Jetz, W., 2017. Integrating occurrence data and expert maps for improved species range predictions. Glob. Ecol. Biogeogr. 26, 243–258. https://doi.org/10.1111/geb.12539

Meynard, C.N., Leroy, B., Kaplan, D.M., 2019. Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing? Ecography (Cop.). 42, 1–16. https://doi.org/10.1111/ecog.04385

Milanesi, P., Della Rocca, F., Robinson, R.A., 2020. Integrating dynamic environmental predictors and species occurrences: Toward true dynamic species distribution models. Ecol. Evol. 10, 1087–1092. https://doi.org/10.1002/ece3.5938

Milanesi, P., Herrando, S., Pla, M., Villero, D., Keller, V., 2017. Towards continental bird distribution models: environmental variables for the second European breeding bird atlas and identification of priorities for further surveys. Vogelwelt 60, 53–60.

Miller, D.A., Nichols, J.D., McClintock, B.T., Campbell Grant, E.H., Bailey, L.L., Weir, L.A., 2011. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. Ecology 92, 1422–1428. https://doi.org/10.1890/10-1396.1

Miller, D.A.W., Pacifici, K., Sanderlin, J.S., Reich, B.J., 2019. The recent past and promising future for data integration methods to estimate species' distributions. Methods Ecol. Evol. 10, 22–37. https://doi.org/10.1111/2041-210X.13110

Miranda, E.B.P., Menezes, J.F.S., Farias, C.C.L., Munn, C., Peres, C.A., 2019. Species distribution modeling reveals strongholds and potential reintroduction areas for the world's largest eagle. PLoS One 14, 1–19. https://doi.org/10.1371/journal.pone.0216323

Mitchell, P.J., Monk, J., Laurenson, L., 2017. Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. Methods Ecol. Evol. 8, 12–21. https://doi.org/10.1111/2041-210X.12645

Molnár, Z., Babai, D., 2021. Inviting ecologists to delve deeper into traditional ecological knowledge. Trends Ecol. Evol. 36, 679–690. https://doi.org/10.1016/j.tree.2021.04.006

Moore, I.D., Gessler, P.E., Nielsen, G., Peterson, G.A., 1993. Soil Attribute Prediction Using Terrain Analysis. Soil Sci. Soc. Am. J. 57, 443–452. https://doi.org/10.2136/sssaj1993.572npb

Moquet, L., Laurent, E., Bacchetta, R., Jacquemart, A.L., 2018. Conservation of hoverflies (Diptera, Syrphidae) requires complementary resources at the landscape and local scales. Insect Conserv. Divers. 11, 72–87. https://doi.org/10.1111/icad.12245

Morelli, F., Pruscini, F., Santolini, R., Perna, P., Benedetti, Y., Sisti, D., 2013. Landscape heterogeneity metrics as indicators of bird diversity: Determining the optimal spatial scales in different landscapes. Ecol. Indic. 34, 372–379. https://doi.org/10.1016/j.ecolind.2013.05.021

Morton, E.S., 1975. Ecological Sources of Selection on Avian Sounds. Am. Nat. 109, 17–34.

Moudrý, V., Cord, A.F., Gábor, L., Laurin, G. V., Barták, V., Gdulová, K., Malavasi, M., Rocchini, D., Stereńczak, K., Prošek, J., Klápště, P., Wild, J., 2022. Vegetation structure derived from airborne laser scanning to assess species distribution and habitat suitability: The way forward. Divers. Distrib. 29, 39–50. https://doi.org/10.1111/ddi.13644

Mouillot, D., Graham, N.A.J., Villéger, S., Mason, N.W.H., Bellwood, D.R., 2013. A functional approach reveals community responses to disturbances. Trends Ecol. Evol. 28, 167–177. https://doi.org/10.1016/j.tree.2012.10.004

Nagendra, H., 2001. Using remote sensing to assess biodiversity. Int. J. Remote Sens. 22, 2377–2400. https://doi.org/10.1080/01431160117096

Nagendra, H., Lucas, R., Honrado, J.P., Jongman, R.H.G., Tarantino, C., Adamo, M., Mairota, P., 2013. Remote sensing for conservation monitoring: Assessing protected areas, habitat extent, habitat condition, species diversity, and threats. Ecol. Indic. 33, 45–59. https://doi.org/10.1016/j.ecolind.2012.09.014

Navid, D., 1984. International Cooperation for Wetland Conservation: The Ramsar Convention. Trans. North Am. Wildl. Nat. Resour. Conf. 49, 33–41.

Neilan, W.L., Barton, P.S., Mcalpine, C.A., Wood, J.T., Lindenmayer, D.B., 2019. Contrasting effects of mosaic structure on alpha and beta diversity of bird assemblages in a human-modified landscape. Ecography (Cop.). 42, 173–186. https://doi.org/10.1111/ecog.02981

Newbold, T., Hudson, L.N., Hill, S.L.L., Contu, S., Lysenko, I., Senior, R.A., Börger, L., Bennett, D.J., Choimes, A., Collen, B., Day, J., De Palma, A., Díaz, S., Echeverria-Londoño, S., Edgar, M.J., Feldman, A., Garon, M., Harrison, M.L.K., Alhusseini, T., Ingram, D.J., Itescu, Y., Kattge, J., Kemp, V., Kirkpatrick, L., Kleyer, M., Laginha, D., Correia, P., Martin, C.D., Meiri, S., Novosolov, M., Pan, Y., Phillips, H.R.P., Purves, D.W., Robinson, A., Simpson, J., Tuck, S.L., Weiher, E., White, H.J., Ewers, R.M., Mace, G.M., Scharlemann, J.P.W., Purvis, A., 2015. Global effects of land use on local terrestrial biodiversity. Nature 520, 45–50. https://doi.org/10.1038/nature14324

Nijssen, M.E., WallisDeVries, M.F., Siepel, H., 2017. Pathways for the effects of increased nitrogen deposition on fauna. Biol. Conserv. 212, 423–431. https://doi.org/10.1016/j.biocon.2017.02.022

Norberg, A., Abrego, N., Guillaume Blanchet, F., Adler, F.R., Anderson, B.J., Anttila, J., Araújo, M.B., Dallas, T., Dunson, D., Elith, J., Foster, S.D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., Hill, N.A., Holt, R.D., Hui, F.K.C., Husby, M., Kalas, J.A., Lehikoinen, A., Luoto, M., Mod, H.K., Newell, G., Renner, I., Roslin, T., Soininen, J., Thuiller, W., Vanhatalo, J., Warton, D., White, M., Zimmermann, N.E., Gravel, D., Ovaskainen, O., 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecol. Monogr. 89, e01370. https://doi.org/10.1002/ecm.1370

Oeser, J., Heurich, M., Senf, C., Pflugmacher, D., Belotti, E., Kuemmerle, T., 2020. Habitat metrics based on multi-temporal Landsat imagery for mapping large mammal habitat. Remote Sens. Ecol. Conserv. 6, 52–69. https://doi.org/10.1002/rse2.122

Oliveira, M.R., Tomas, W.M., Guedes, N.M.R., Peterson, A.T., Szabo, J.K., Júnior, A.S., Camilo, A.R., Padovani, C.R., Garcia, L.C., 2021. The relationship between scale and predictor variables in species distribution models applied to conservation. Biodivers. Conserv. 30, 1971–1990. https://doi.org/10.1007/s10531-021-02176-w

Olivier, T., Schmucki, R., Fontaine, B., Villemey, A., Archaux, F., 2016. Butterfly assemblages in residential gardens are driven by species' habitat preference and mobility. Landsc. Ecol. 31, 865–876. https://doi.org/10.1007/s10980-015-0299-9

Olmeda, C., Šefferová, V., Underwood, E., Millan, L., Gil, T., Naumann, S., 2020. Action plan to maintain and restore to favourable conservation status the habitat type 4030 European dry heaths. European Commission.

Ovaskainen, O., Hottola, J., Shtonen, J., 2010. Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. Ecology 91, 2514–2521. https://doi.org/10.1890/10-0173.1

Ovaskainen, O., Soininen, J., 2011. Making more out of sparse data: Hierarchical modeling of species communities. Ecology 92, 289–295. https://doi.org/10.1890/10-1251.1

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., Abrego, N., 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecol. Lett. 20, 561–576. https://doi.org/10.1111/ele.12757

Overloop, S.M., Van Gijseghem, D.E., Helming, J.F., 2001. Environmental scenarios for the future nitrogen policy in Flanders, Belgium. ScientificWorldJournal. 1 Suppl 2, 873–879. https://doi.org/10.1100/tsw.2001.289

Pacifici, K., Reich, B.J., Miller, D.A.W., Gardner, B., Stauffer, G., Singh, S., Mckerrow, A., Collazo, J.A., 2017. Integrating multiple data sources in species distribution modeling: a framework for data fusion*. Ecology 98, 840–850.

Packet, J., Scheers, K., Smeekens, V., Leyssen, A., Wils, C., Denys, L., 2018. Watervlakken versie 1.0: polygonenkaart van stilstaand water in Vlaanderen. Een nieuw instrument voor onderzoek, water-, milieu- en natuurbeleid. Rapporten van het Instituut voor Natuur- en Bosonderzoek 2018 (14). Instituut voor Natuur- en Bosonderzoek, Brussel. https://doi.org/10.21436/inbor.14178464

Paelinckx, D., Sannen, K., Goethals, V., Louette, G., Rutten, J., Hoffmann, M., 2009. Methode voor het opstellen van gewestelijke doelstellingen voor de habitats van de Europese Habitatrichtlijn. In: Paelinckx D., et al. (red.), Gewestelijke doelstellingen voor de habitats en soorten van de Europese Habitat- en Vogelrichtlijn voor Vlaanderen. Mededelingen van het Instituut voor Natuur- en Bosonder- zoek INBO.M.2009.6, Brussel, 16-45.

Palmer, M.W., Earls, P.G., Hoagland, B.W., White, P.S., Wohlgemuth, T., 2002. Quantitative tools for perfecting species lists. Environmetrics 13, 121–137. https://doi.org/10.1002/env.516

Paracchini, M.L., Petersen, J., Hoogeveen, Y., Bamps, C., Burfield, I., van Swaay, C., 2008. High Nature Value Farmland in Europe. An estimate of the distribution patterns on the basis of land cover and biodiversity data. JRC Scientific and Technical Reports. European Commission.

Parker, T.H., Forstmeier, W., Koricheva, J., Fidler, F., Hadfield, J.D., Chee, Y.E., Kelly, C.D., Gurevitch, J., Nakagawa, S., 2016. Transparency in Ecology and Evolution: Real Problems, Real Solutions. Trends Ecol. Evol. 31, 711–719. https://doi.org/10.1016/j.tree.2016.07.002

Parks, D., Al-Fulaij, N., Brook, C., Butchart, S.H.M., Collomb, J.G., Cope, D., Dowell, S., Falkingham, B., Frick, W.F., Gibbs, D., Gray, E.E., Heard, N., Leventis, A., Mastro, K., Meredith, H., Mickleburgh, S., Miller, F., Muir, M., Nuijten, R.J.M., Ockendon, N., Owen, N.R., Owens, J.R., Rodríguez, J.P., Tully, E., Vié, J.C., 2022. Funding evidence-based conservation. Conserv. Biol. 36, 12–14. https://doi.org/10.1111/cobi.13991

Parviainen, M., Zimmermann, N.E., Heikkinen, R.K., Luoto, M., 2013. Using unclassified continuous remote sensing data to improve distribution models of red-listed plant species. Biodivers. Conserv. 22, 1731–1754. https://doi.org/10.1007/s10531-013-0509-1

Pe'er, G., Dicks, L.V., Visconti, P., Arlettaz, R., Báldi, A., Benton, T.G., Collins, S., Dieterich, M., Gregory, R.D., Hartig, F., Henle, K., Hobson, P.R., Kleijn, D., Neumann, K., Robijns, T., Schmidt, J., Shwartz, A., Sutherland, W.J., Turbé, A., Wulf, F., Scott, A.V., 2014. EU agricultural reform fails on biodiversity. Science. 344, 1090–1092.

Pe'er, G., Zinngrebe, Y., Hauck, J., Schindler, S., Dittrich, A., Zingg, S., Tscharntke, T., Oppermann, R., Sutcliffe, L., Sirami, C., Schmidt, J., Hoyer, C., Schleyer, C., Lakner, S., 2017. Adding some green to the greening: improving the EU's Ecological Focus Areas for biodiversity and farmers. Conserv. Lett. 10, 517–530. https://doi.org/10.1111/conl.12333.This

Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Modell. 133, 225–245. https://doi.org/10.1016/S0304-3800(00)00322-7

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory, D.R., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J.P.W., Stuart, S.N., Turak, E., Walpole, M., Wegmann, M., 2013. Essential Biodiversity Variables. Science. 339, 277–278.

Pereira, H.M., Navarro, L.M., Martins, I.S., 2012. Global biodiversity change: The Bad, the good, and the unknown. Annu. Rev. Environ. Resour. 37, 25–50. https://doi.org/10.1146/annurev-environ-042911-093511

Peters, V.E., Campbell, K.U., Dienno, G., García, M., Leak, E., Loyke, C., Ogle, M., Steinly, B., Crist, T.O., 2016. Ants and plants as indicators of biodiversity, ecosystem services, and conservation value in constructed grasslands. Biodivers. Conserv. 25, 1481–1501. https://doi.org/10.1007/s10531-016-1120-z

Peterson, A.T., Soberón, J., Krishtalka, L., 2015. A global perspective on decadal challenges and priorities in biodiversity informatics. BMC Ecol. 15, 1–9. https://doi.org/10.1186/s12898-015-0046-8

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M., Araújo, M.B., 2011. Ecological Niches and Geographic Distributions, Monographs in Population Biology 49. Princeton University Press, Princeton. https://doi.org/10.23943/princeton/9780691136868.001.0001

Pettorelli, N., Vik, J.O., Mysterud, A., Gaillard, J.M., Tucker, C.J., Stenseth, N.C., 2005. Using the satellite-derived NDVI to assess ecological responses to environmental change. Trends Ecol. Evol. 20, 503–510. https://doi.org/10.1016/j.tree.2005.05.011

Pettorelli, N., Wegmann, M., Skidmore, A., Mücher, S., Dawson, T.P., Fernandez, M., Lucas, R., Schaepman, M.E., Wang, T., O'Connor, B., Jongman, R.H.G., Kempeneers, P., Sonnenschein, R., Leidner, A.K., Böhm, M., He, K.S., Nagendra, H., Dubois, G., Fatoyinbo, T., Hansen, M.C., Paganini, M., de Klerk, H.M., Asner, G.P., Kerr, J.T., Estes, A.B., Schmeller, D.S., Heiden, U., Rocchini, D., Pereira, H.M., Turak, E., Fernandez, N., Lausch, A., Cho, M.A., Alcaraz-Segura, D., McGeoch, M.A., Turner, W., Mueller, A., St-Louis, V., Penner, J., Vihervaara, P., Belward, A., Reyers, B., Geller, G.N., 2016. Framing the concept of satellite remote sensing essential biodiversity variables: challenges and future directions. Remote Sens. Ecol. Conserv. 2, 122–131. https://doi.org/10.1002/rse2.15

Pfeifer, M., Lefebvre, V., Peres, C.A., Banks-Leite, C., Wearn, O.R., Marsh, C.J., Butchart, S.H.M., Arroyo-Rodríguez, V., Barlow, J., Cerezo, A., Cisneros, L., D'Cruze, N., Faria, D., Hadley, A., Harris, S.M., Klingbeil, B.T., Kormann, U., Lens, L., Medina-Rangel, G.F., Morante-Filho, J.C., Olivier, P., Peters, S.L., Pidgeon, A., Ribeiro, D.B., Scherber, C., Schneider-Maunoury, L., Struebig, M., Urbina-Cardona, N., Watling, J.I., Willig, M.R., Wood, E.M., Ewers, R.M., 2017. Creation of forest edges has a global impact on forest vertebrates. Nature 551, 187–191. https://doi.org/10.1038/nature24457

Phillips, S.J., 2017. A Brief Tutorial on Maxent. Available from url: http://biodiversityinformatics.amnh.org/open_source/maxent/. Accessed on 2023-05-25.

Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E., Blair, M.E., 2017. Opening the black box: an open-source release of Maxent. Ecography (Cop.). 40, 887–893. https://doi.org/10.1111/ecog.03049

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Modell. 190, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. Ecol. Appl. 19, 181–197. https://doi.org/10.1890/07-2153.1

Piessens, K., Honnay, O., Devlaeminck, R., Hermy, M., 2006. Biotic and abiotic edge effects in highly fragmented heathlands adjacent to cropland and forest. Agric. Ecosyst. Environ. 114, 335–342. https://doi.org/10.1016/j.agee.2005.11.016

Piessens, K., Honnay, O., Hermy, M., 2005. The role of fragment area and isolation in the conservation of heathland species. Biol. Conserv. 122, 61–69. https://doi.org/10.1016/j.biocon.2004.05.023

Poelmans, L., Van Daele, T., 2014. Landgebruikskaart NARA-T 2014. Vlaams Instituut voor Technologisch Onderzoek (VITO), Mol.

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M., Vesk, P.A., Mccarthy, M.A., 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods Ecol. Evol. 5, 397–406. https://doi.org/10.1111/2041-210X.12180

Pöyry, J., Luoto, M., Heikkinen, R.K., Saarinen, K., 2008. Species traits are associated with the quality of bioclimatic models. Glob. Ecol. Biogeogr. 17, 403–414. https://doi.org/10.1111/j.1466-8238.2007.00373.x

Prendergast, J.R., Quinn, R.M., Lawton, J.H., Eversham, B.C., Gibbons, D.W., 1993. Rare species, the coincidence of diversity hotspots and conservation strategies. Nature 365, 335–337. https://doi.org/10.1038/365335a0

Price, P.W., Bouton, C.E., Gross, P., McPheron, B.A., Thompson, J.N., Weis, A.E., 1980. Interactions Among Three Throphic Levels: Influence of Plants on Interactions Between Insect Herbivores and Natural Enemies. Annu. Rev. Ecol. Syst. 11, 41–65. https://doi.org/https://doi.org/10.1146/annurev.es.11.110180.000353

R Core Team, 2021. R: A language and environment for statistical computing.

Randin, C.F., Ashcroft, M.B., Bolliger, J., Cavender-Bares, J., Coops, N.C., Dullinger, S., Dirnböck, T., Eckert, S., Ellis, E., Fernández, N., Giuliani, G., Guisan, A., Jetz, W., Joost, S., Karger, D., Lembrechts, J., Lenoir, J., Luoto, M., Morin, X., Price, B., Rocchini, D., Schaepman, M., Schmid, B., Verburg, P., Wilson, A., Woodcock, P., Yoccoz, N., Payne, D., 2020. Monitoring biodiversity in the Anthropocene using remote sensing in species distribution models. Remote Sens. Environ. 239, 111626. https://doi.org/10.1016/j.rse.2019.111626

Ratnieks, F.L.W., Schrell, F., Sheppard, R.C., Brown, E., Bristow, O.E., Garbuzov, M., 2016. Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers. Methods Ecol. Evol. 7, 1226–1235. https://doi.org/10.1111/2041-210X.12581

Regos, A., Gagne, L., Alcaraz-Segura, D., Honrado, J.P., Domínguez, J., 2019. Effects of species traits and environmental predictors on performance and transferability of ecological niche models. Sci. Rep. 9, 4211. https://doi.org/10.1038/s41598-019-40766-5

Regos, A., Gómez-Rodríguez, P., Arenas-Castro, S., Tapia, L., Vidal, M., Domínguez, J., 2020. Model-Assisted Bird Monitoring Based on Remotely Sensed Ecosystem Functioning and Atlas Data. Remote Sens. 12, 14. https://doi.org/10.3390/rs12162549

Reif, M.K., Theel, H.J., 2017. Remote sensing for restoration ecology: Application for restoring degraded, damaged, transformed, or destroyed ecosystems. Integr. Environ. Assess. Manag. 13, 614–630. https://doi.org/10.1002/ieam.1847

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G., Warton, D.I., 2015. Point process models for presence-only analysis. Methods Ecol. Evol. 6, 366–379. https://doi.org/10.1111/2041-210X.12352

Renner, I.W., Warton, D.I., 2013. Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. Biometrics 69, 274–281. https://doi.org/10.1111/j.1541-0420.2012.01824.x

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography (Cop.). 40, 913–929. https://doi.org/10.1111/ecog.02881

Robinson, O.J., Ruiz-Gutierrez, V., Reynolds, M.D., Golet, G.H., Strimas-Mackey, M., Fink, D., 2019. Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. Divers. Distrib. 26, 976–986. https://doi.org/10.1101/806547

Rocchini, D., Ricotta, C., Chiarucci, A., 2007. Using satellite imagery to assess plant species richness: The role of multispectral systems. Appl. Veg. Sci. 10, 325–331. https://doi.org/10.1111/j.1654-109X.2007.tb00431.x

Ruete, A., 2015. Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. Biodivers. Data J. 3. https://doi.org/10.3897/BDJ.3.e5361

Runge, T., Latacz-Lohmann, U., Schaller, L., Todorova, K., Daugbjerg, C., Termansen, M., Liira, J., Le Gloux, F., Dupraz, P., Leppanen, J., Fogarasi, J., Vigh, E.Z., Bradfield, T., Hennessy, T., Targetti, S., Viaggi, D., Berzina, I., Schulp, C., Majewski, E., Bouriaud, L., Baciu, G., Pecurul, M., Prokofieva, I., Velazquez, F.J.B., 2022. Implementation of Eco-schemes in Fifteen European Union Member States. EuroChoices 21, 19–27. https://doi.org/10.1111/1746-692X.12352

Rutten, A., Casaer, J., Swinnen, K.R.R., Herremans, M., Leirs, H., 2019. Future distribution of wild boar in a highly anthropogenic landscape: Models combining hunting bag and citizen science data. Ecol. Modell. 411, 108804. https://doi.org/10.1016/j.ecolmodel.2019.108804

Sagarin, R., Pauchard, A., 2010. Observational approaches in ecology open new ground in a changing world. Front. Ecol. Environ. 8, 379–386. https://doi.org/10.1890/090001

Sanczuk, P., De Lombaerde, E., Haesen, S., Van Meerbeek, K., Van der Veken, B., Hermy, M., Verheyen, K., Vangansbeke, P., De Frenne, P., 2022. Species distribution models and a 60-year-old transplant experiment reveal inhibited forest plant range shifts under climate change. J. Biogeogr. 49, 537–550. https://doi.org/10.1111/jbi.14325

Schellenberg, J., Bergmeier, E., 2020. Heathland plant species composition and vegetation structures reflect soil-related paths of development and site history. Appl. Veg. Sci. 23, 386–405. https://doi.org/10.1111/avsc.12489

Schiegg, K., 2000. Effects of dead wood volume and connectivity on saproxylic insect species diversity. Ecoscience 7, 290–298. https://doi.org/10.1080/11956860.2000.11682598

Schindler, S., von Wehrden, H., Poirazidis, K., Wrbka, T., Kati, V., 2013. Multiscale performance of landscape metrics as indicators of species richness of plants, insects and vertebrates. Ecol. Indic. 31, 41–48. https://doi.org/10.1016/j.ecolind.2012.04.012

Schirmel, J., Fartmann, T., 2014. Coastal heathland succession influences butterfly community composition and threatens endangered butterfly species. J. Insect Conserv. 18, 111–120. https://doi.org/10.1007/s10841-014-9619-7

Schirmel, J., Mantilla-Contreras, J., Blindow, I., Fartmann, T., 2011. Impacts of succession and grass encroachment on heathland Orthoptera. J. Insect Conserv. 15, 633–642. https://doi.org/10.1007/s10841-010-9362-7

Schmeller, D.S., Henry, P.-Y., Julliard, R., Gruber, B., Clobert, J., Dziock, F., Lengyel, S., Nowicki, P., Déri, E., Budrys, E., Kull, T., Tali, K., Bauch, B., Settele, J., Van Swaay, C., Kobler, A., Babij, V., Papastergiadou, E., Henle, K., 2008. Advantages of Volunteer-Based Biodiversity Monitoring in Europe. Conserv. Biol. 23, 307–316. https://doi.org/10.1111/j.1523-1739.2008.01125.x

Schmidt, J., Fassnacht, F.E., Förster, M., Schmidtlein, S., 2018. Synergetic use of Sentinel-1 and Sentinel-2 for assessments of heathland conservation status. Remote Sens. Ecol. Conserv. 4, 225–239. https://doi.org/10.1002/rse2.68

Schneiders, A., Alaerts, K., Michels, H., Stevens, M., Van Gossum, P., Van Reeth, W., Vught, I., 2020. Natuurrapport 2020: feiten en cijfers voor een nieuw biodiversiteitsbeleid. Mededelingen van het Instituut voor Natuur- en Bosonderzoek 2020. Instituut voor Natuur- en Bosonderzoek, Brussel.

Schneiders, A., Thoonen, M., Alaerts, K., 2016. Hoofdstuk 2 - 50 tinten groen. Naar een gemeenschappelijke beleidsstrategie voor groene infrastructuur (INBO.R.2016.12342848), in: Van Gossum et Al. (Eds.), Natuurrapport – Aan de Slag Met Ecosysteemdiensten. Technisch Rapport. Mededelingen van Het Instituut Voor Natuur- En Bosonderzoek, INBO.M.2016.12342456. Brussel. https://doi.org/doi.org/10.21436/inbor.12342848

Scholes, R.J., Walters, M., Turak, E., Saarenmaa, H., Heip, C.H.R., Tuama, É.Ó., Faith, D.P., Mooney, H.A., Ferrier, S., Jongman, R.H.G., Harrison, I.J., Yahara, T., Pereira, H.M., Larigauderie, A., Geller, G., 2012. Building a global observing system for biodiversity. Curr. Opin. Environ. Sustain. 4, 139–146. https://doi.org/10.1016/j.cosust.2011.12.005

Seavy, N.E., Viers, J.H., Wood, J.K., 2009. Riparian Bird Response to Vegetation Structure: A Multiscale Analysis Using LiDAR Measurements of Canopy Height. Ecol. Appl. 19, 1848–1857.

Segurado, P., Araújo, M.B., Kunin, W.E., 2006. Consequences of spatial autocorrelation for niche-based models. J. Appl. Ecol. 43, 433–444. https://doi.org/10.1111/j.1365-2664.2006.01162.x

Seoane, J., Carrascal, L.M., Alonso, L., Palomino, D., 2005. Species-specific traits associated to prediction errors in bird habitat suitability modelling. Ecol. Modell. 185, 299–308. https://doi.org/10.1016/j.ecolmodel.2004.12.012

Serra-Diaz, J.M., Enquist, B.J., Maitner, B., Merow, C., Svenning, J.C., 2017. Big data of tree species distributions: how big and how good? For. Ecosyst. 4. https://doi.org/10.1186/s40663-017-0120-0

Sheeren, D., Bonthoux, S., Balent, G., 2014. Modeling bird communities using unclassified remote sensing imagery: Effects of the spatial resolution and data period. Ecol. Indic. 43, 69–82. https://doi.org/10.1016/j.ecolind.2014.02.023

Sillero, N., 2011. What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. Ecol. Modell. 222, 1343–1346. https://doi.org/10.1016/j.ecolmodel.2011.01.018

Sillero, N., Gonçalves-Seco, L., 2014. Spatial structure analysis of a reptile community with airborne LiDAR data. Int. J. Geogr. Inf. Sci. 28, 1709–1722. https://doi.org/10.1080/13658816.2014.902062

Silvertown, J., 2009. A new dawn for citizen science. Chemosphere 24, 467–471.

Simmonds, E.G., Jarvis, S.G., Henrys, P.A., Isaac, N.J.B., O'Hara, R.B., 2020. Is more data always better? A simulation study of benefits and limitations of integrated distribution models. Ecography (Cop.). 43, 1413–1422. https://doi.org/10.1111/ECOG.05146

Sladonja, B., Damijanić, D., 2021. Remote Sensing in Invasive Species Detection and Monitoring. Int. J. Environ. Sci. Nat. Resour. 29, 5–7. https://doi.org/10.19080/ijesnr.2021.29.556255

Sólymos, P., Matsuoka, S.M., Stralberg, D., Barker, N.K.S., Bayne, E.M., 2018. Phylogeny and species traits predict bird detectability. Ecography (Cop.). 41, 1595–1603. https://doi.org/10.1111/ecog.03415

Soroye, P., Ahmed, N., Kerr, J.T., 2018. Opportunistic citizen science data transform understanding of species distributions, phenology, and diversity gradients for global change research. Glob. Chang. Biol. 24, 5281–5291. https://doi.org/10.1111/gcb.14358

Steen, V.A., Elphick, C.S., Tingley, M.W., 2019. An evaluation of stringent filtering to improve species distribution models from citizen science data. Biodivers. Res. 25, 1857–1869. https://doi.org/10.1111/ddi.12985

Stein, A., Gerstner, K., Kreft, H., 2014. Environmental heterogeneity as a universal driver of species richness across taxa, biomes and spatial scales. Ecol. Lett. 17, 866–880. https://doi.org/10.1111/ele.12277

Stephenson, P.J., 2020. Technological advances in biodiversity monitoring: applicability, opportunities and challenges. Curr. Opin. Environ. Sustain. 45, 36–41. https://doi.org/10.1016/j.cosust.2020.08.005

Stockwell, D.R.B., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecol. Modell. 148, 1–13.

Storchová, L., Hořák, D., 2018. Life-history characteristics of European birds. Glob. Ecol. Biogeogr. 27, 400–406. https://doi.org/10.1111/geb.12709

Stubbe, F., 2021. 'CAP Pillar II: management agreements'. Minutes of an online meeting on the application potential of species distribution models in agricultural policy in Flanders, Flemish Land Agency (VLM), 26 March 2021, Leuven.

Suhaimi, S.S.A., Blair, G.S., Jarvis, S.G., 2021. Integrated species distribution models: A comparison of approaches under different data quality scenarios. Divers. Distrib. 00, 1–10. https://doi.org/10.1111/ddi.13255

Sullivan, B.L., Aycrigg, J.L., Barry, J.H., Bonney, R.E., Bruns, N., Cooper, C.B., Damoulas, T., Dhondt, A.A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J.W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W.M., Iliff, M.J., Lagoze, C., La Sorte, F.A., Merrifield, M., Morris, W., Phillips, T.B., Reynolds, M., Rodewald, A.D., Rosenberg, K. V, Trautmann, N.M., Wiggins, A., Winkler, D.W., Wong, W., Wood, C.L., Yu, J., Kelling, S., 2014. The eBird enterprise: An integrated approach to development and application of citizen science. Biol. Conserv. 169, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S., 2009. eBird: A citizen-based bird observation network in the biological sciences. Biol. Conserv. 142, 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006

Sullivan, M.J.P., Pearce-Higgins, J.W., Newson, S.E., Scholefield, P., Brereton, T., Oliver, T.H., 2017. A national-scale model of linear features improves predictions of farmland biodiversity. J. Appl. Ecol. 54, 1776–1784. https://doi.org/10.1111/1365-2664.12912

Sundseth, K., 2008. Natura 2000: Protecting Europe’s biodiversity. European Commission, Directorate-General for the Environment, Brussels, Belgium.

Sutherland, W.J., Alvarez-Castañeda, S.T., Amano, T., Ambrosini, R., Atkinson, P., Baxter, J.M., Bond, A.L., Boon, P.J., Buchanan, K.L., Barlow, J., Bogliani, G., Bragg, O.M., Burgman, M., Cadotte, M.W., Calver, M., Cooke, S.J., Corlett, R.T., Devictor, V., Ewen, J.G., Fisher, M., Freeman, G., Game, E., Godley, B.J., Gortázar, C., Hartley, I.R., Hawksworth, D.L., Hobson, K.A., Lu, M.L., Martín-López, B., Ma, K., Machado, A., Maes, D., Mangiacotti, M., McCafferty, D.J., Melfi, V., Molur, S., Moore, A.J., Murphy, S.D., Norris, D., van Oudenhoven, A.P.E., Powers, J., Rees, E.C., Schwartz, M.W., Storch, I., Wordley, C., 2020. Ensuring tests of conservation interventions build on existing literature. Conserv. Biol. 34, 781–783. https://doi.org/10.1111/cobi.13555

Sutherland, W.J., Pullin, A.S., Dolman, P.M., Knight, T.M., 2004. The need for evidence-based conservation. Trends Ecol. Evol. 19, 305–308. https://doi.org/10.1016/j.tree.2004.03.018

Sutherland, W.J., Worldley, Cl.F., 2018. A fresh approach to evidence. Nature 558, 364–366.

Swinnen, K.R.R., Strubbe, D., Matthysen, E., Leirs, H., 2017. Reintroduced Eurasian beavers (Castor fiber): colonization and range expansion across human-dominated landscapes. Biodivers. Conserv. 26, 1863–1876. https://doi.org/10.1007/s10531-017-1333-9

Swinnen, K.R.R., Vercayie, D., Vanreusel, W., Barendse, R., Boers, K., Bogaert, J., Dekeukeleire, D., Driessens, G., Dupriez, P., Jooris, R., Steeman, R., van Asten, K., van den Neucker, T., van Dorsselaer, P., van Vooren, P., Wysmantel, N., Gielen, K., Desmet, P., Herremans, M., 2018. Waarnemingen.be-Non-native plant and animal occurrences in Flanders and the Brussels Capital Region, Belgium. BioInvasions Rec. 7, 335–342. https://doi.org/10.3391/bir.2018.7.3.17

Tamme, R., Hiiesalu, I., Laanisto, L., Szava-Kovats, R., Pärtel, M., 2010. Environmental heterogeneity, species diversity and co-existence at different spatial scales. J. Veg. Sci. 21, 796–801. https://doi.org/10.1111/j.1654-1103.2010.01185.x

Tarantino, C., Blonda, P., Adamo, M., 2015. Application of a semi-automatic unsupervised change detection to (SEMI-) natural grassland loss at very high resolution. Int. Geosci. Remote Sens. Symp. 2015-Novem, 1666–1669. https://doi.org/10.1109/IGARSS.2015.7326106

Tessarolo, G., Rangel, T.F., Araújo, M.B., Hortal, J., 2014. Uncertainty associated with survey design in Species Distribution Models. Divers. Distrib. 20, 1258–1269. https://doi.org/10.1111/ddi.12236

Tews, J., Brose, U., Grimm, V., Tielbörger, K., Wichmann, M.C., Schwager, M., Jeltsch, F., 2004. Animal species diversity driven by habitat heterogeneity/diversity: The importance of keystone structures. J. Biogeogr. 31, 79–92. https://doi.org/10.1046/j.0305-0270.2003.00994.x

Theobald, E.J., Ettinger, A.K., Burgess, H.K., DeBey, L.B., Schmidt, N.R., Froehlich, H.E., Wagner, C., HilleRisLambers, J., Tewksbury, J., Harsch, M.A., Parrish, J.K., 2015. Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. Biol. Conserv. 181, 236–244. https://doi.org/10.1016/j.biocon.2014.10.021

Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A.C., Guisan, A., 2014. Measuring the relative effect of factors affecting species distribution model predictions. Methods Ecol. Evol. 5, 947–955. https://doi.org/10.1111/2041-210X.12203

Thomaes, A., Kervyn, T., Maes, D., 2008. Applying species distribution modelling for the conservation of the threatened saproxylic Stag Beetle (Lucanus cervus). Biol. Conserv. 141, 1400–1410. https://doi.org/10.1016/j.biocon.2008.03.018

Thoonen, G., Spanhove, T., Vanden Borre, J., Scheunders, P., 2013. Classification of heathland vegetation in a hierarchical contextual framework. Int. J. Remote Sens. 34, 96–111. https://doi.org/10.1080/01431161.2012.708061

Titeux, N., Maes, D., Marmion, M., Luoto, M., Heikkinen, R.K., 2009. Inclusion of soil data improves the performance of bioclimatic envelope models for insect species distributions in temperate Europe. J. Biogeogr. 36, 1459–1473. https://doi.org/10.1111/j.1365-2699.2009.02088.x

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., Legendre, F., 2017. Taxonomic bias in biodiversity data and societal preferences. Sci. Rep. 7, 9132. https://doi.org/10.1038/s41598-017-09084-6

Truong, T.T.A., Hardy, G.E.S.J., Andrew, M.E., 2017. Contemporary remotely sensed data products refine invasive plants risk mapping in data poor regions. Front. Plant Sci. 8. https://doi.org/10.3389/fpls.2017.00770

Tscharntke, T., Klein, A.M., Kruess, A., Steffan-Dewenter, I., Thies, C., 2005. Landscape perspectives on agricultural intensification and biodiversity – ecosystem service management. Ecol. Lett. 8, 857–874. https://doi.org/10.1111/j.1461-0248.2005.00782.x

Tye, C.A., McCleery, R.A., Fletcher, R.J., Greene, D.U., Butryn, R.S., 2017. Evaluating citizen vs. professional data for modelling distributions of a rare squirrel. J. Appl. Ecol. 54, 628–637. https://doi.org/10.1111/1365-2664.12682

United Nations, 1992. Convention on Biological Diversity. https://doi.org/10.1016/B978-0-12-384719-5.00418-4

Urban, M.C., Bocedi, G., Hendry, A.P., Mihoub, J.B., Pe'er, G., Singer, A., Bridle, J.R., Crozier, L.G., De Meester, L., Godsoe, W., Gonzalez, A., Hellmann, J.J., Holt, R.D., Huth, A., Johst, K., Krug, C.B., Leadley, P.W., Palmer, S.C.F., Pantel, J.H., Schmitz, A., Zollner, P.A., Travis, J.M.J., 2016. Improving the forecast for biodiversity under climate change. Science. 353, aad8466. https://doi.org/10.1126/science.aad8466

Valavi, R., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G., 2019. blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol. Evol. 10, 225–232. https://doi.org/10.1111/2041-210X.13107

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., 2022. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecol. Monogr. 92, 1–27. https://doi.org/10.1002/ecm.1486

Valbuena, R., O'Connor, B., Zellweger, F., Simonson, W., Vihervaara, P., Maltamo, M., Silva, C.., Almeida, D.R.A., Danks, F., Morsdorf, F., Chirici, G., Lucas, R., Coomes, D., Coops, N.., 2020. Standardising ecosystem morphological traits from 3D information sources. Trends Ecol. Evol. 35, 656–667.

Van Daele, F., Honnay, O., De Kort, H., 2021. The role of dispersal limitation and reforestation in shaping the distributional shift of a forest herb under climate change. Divers. Distrib. 27, 1775–1791. https://doi.org/10.1111/ddi.13367

van den Berg, L.J.L., Bullock, J.M., Clarke, R.T., Langston, R.H.W., Rose, R.J., 2001. Territory selection by the Dartford warbler (Sylvia undata) in Dorset, England: The role of vegetation type, habitat fragmentation and population size. Biol. Conserv. 101, 217–228. https://doi.org/10.1016/S0006-3207(01)00069-6

Van Eupen, C., 2017. The Potential Impact of EFA Implementation on Biodiversity and Ecosystem Services. Masterproef ingediend tot het behalen van de graad van master of Science in de biowetenschappen: land- en tuinbouwkunde, afstudeerrichting natuur en milieu.

van Proosdij, A.S.J., Sosef, M.S.M., Wieringa, J.J., Raes, N., 2016. Minimum required number of specimen records to develop accurate species distribution models. Ecography (Cop.). 39, 542–552. https://doi.org/10.1111/ECOG.01509

Van Strien, A.J., Van Swaay, C.A.M., Termaat, T., 2013. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. J. Appl. Ecol. 50, 1450–1458. https://doi.org/10.1111/1365-2664.12158

van Swaay, C., Warren, M., Loïs, G., 2006. Biotope use and trends of European butterflies. J. Insect Conserv. 10, 189–209. https://doi.org/10.1007/s10841-006-6293-4

Vanden Borre, J., Paelinckx, D., Mücher, C.A., Kooistra, L., Haest, B., De Blust, G., Schmidt, A.M., 2011. Integrating remote sensing in Natura 2000 habitat monitoring: Prospects on the way forward. J. Nat. Conserv. 19, 116–125. https://doi.org/10.1016/j.jnc.2010.07.003

Vanden Broeck, A., Maes, D., Kelager, A., Wynhoff, I., WallisDeVries, M.F., Nash, D.R., Oostermeijer, J.G.B., Van Dyck, H., Mergeay, J., 2017. Gene flow and effective population sizes of the butterfly Maculinea alcon in a highly fragmented, anthropogenic landscape. Biol. Conserv. 209, 89–97. https://doi.org/10.1016/j.biocon.2017.02.001

Vanermen, N., Courtens, W., Daelemans, R., Lens, L., Müller, W., Van De Walle, M., Verstraete, H., Stienen, E.W.M., 2020. Attracted to the outside: A meso-scale response pattern of lesser black-backed gulls at an offshore wind farm revealed by GPS telemetry. ICES J. Mar. Sci. 77, 701–710. https://doi.org/10.1093/icesjms/fsz199

Vanreusel, W., Maes, D., Van Dyck, H., 2007. Transferability of species distribution models: A functional habitat approach for two regionally threatened butterflies. Conserv. Biol. 21, 201–212. https://doi.org/10.1111/j.1523-1739.2006.00577.x

Vanreusel, W., Swinnen, K., Herremans, M., 2018. Waarnemingen.be: Mesthoop of schatkist? Natuur.focus 4.

Vantieghem, P., Maes, D., Kaiser, A., Merckx, T., 2017. Quality of citizen science data and its consequences for the conservation of skipper butterflies (Hesperiidae) in Flanders (northern Belgium). J. Insect Conserv. 21, 451–463. https://doi.org/10.1007/s10841-016-9924-4

Varela, S., Anderson, R.P., García-Valdés, R., Fernández-González, F., 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography (Cop.). 37, 1084–1091. https://doi.org/10.1111/j.1600-0587.2013.00441.x

Veloz, S.D., 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. J. Biogeogr. 36, 2290–2299. https://doi.org/10.1111/j.1365-2699.2009.02174.x

Verdonckt, F., 2018. 'Management agreements'. Minutes of meeting on the application potential of species distribution models in agricultural policy in Flanders, Natuurpunt, 12 July 2018. Mechelen.

Vermeersch, G., Devos, K., Driessens, G., Evereaert, J., Feys, S., Herremans, M., Onkelinx, T., Stienen, E.W.M., T'Jollyn, F., Anselin, A., 2020. Broedvogels in Vlaanderen 2013-2018. Medelingen van het Instituut voor Natuur- en Bosonderzoek 2020 (1), Brussel. https://doi.org/10.21436/inbor.18794135

Vihervaara, P., Mononen, L., Auvinen, A.-P., Virkkala, R., Lü, Y., Pippuri, I., Packalen, P., Valbuena, R., Valkama, J., 2015. How to integrate remotely sensed data and biodiversity for ecosystem assessments at landscape scale. Landsc. Ecol. 30, 501–516. https://doi.org/10.1007/s10980-014-0137-5

Vila-Viçosa, C., Arenas-Castro, S., Marcos, B., Honrado, J., García, C., Vázquez, F.M., Almeida, R., Gonçalves, J., 2020. Combining satellite remote sensing and climate data in species distribution models to improve the conservation of iberian white oaks (Quercus l.). ISPRS Int. J. Geo-Information 9. https://doi.org/10.3390/ijgi9120735

Vogels, J.J., Verberk, W., Lamers, L., Siepel, H., 2017. Can changes in soil biochemistry and plant stoichiometry explain loss of animal diversity of heathlands? Biol. Conserv. 212, 432–447. https://doi.org/10.1016/j.biocon.2016.08.039

Vogels, J.J., Verberk, W.C.E.P., Kuper, J.T., Weijters, M.J., Bobbink, R., Siepel, H., 2021. How to Restore Invertebrate Diversity of Degraded Heathlands? A Case Study on the Reproductive Performance of the Field Cricket Gryllus campestris (L.). Front. Ecol. Evol. 9, 1–12. https://doi.org/10.3389/fevo.2021.659363

Vollering, J., Halvorsen, R., Auestad, I., Rydgren, K., 2019. Bunching up the background betters bias in species distribution models. Ecography (Cop.). 42, 1717–1727. https://doi.org/10.1111/ecog.04503

Wachendorf, M., Fricke, T., Möckel, T., 2018. Remote sensing as a tool to assess botanical composition, structure, quantity and quality of temperate grasslands. Grass Forage Sci. 73, 1–14. https://doi.org/10.1111/gfs.12312

Wang, Y., Stone, L., 2019. Understanding the connections between species distribution models for presence-background data. Theor. Ecol. 12, 73–88. https://doi.org/10.1007/s12080-018-0389-9

Warton, D.I., Renner, I.W., Ramp, D., 2013. Model-based control of observer bias for the analysis of presence-only data in ecology. PLoS One 8. https://doi.org/10.1371/journal.pone.0079168

Warton, D.I., Shepherd, L.C., 2010. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. Ann. Appl. Stat. 4, 1383–1402. https://doi.org/10.1214/10-AOAS331

Wätzold, F., Mewes, M., van Apeldoorn, R., Varjopuro, R., Chmielewski, T.J., Veeneklaas, F., Kosola, M.L., 2010. Cost-effectiveness of managing Natura 2000 sites: An exploratory study for Finland, Germany, the Netherlands and Poland. Biodivers. Conserv. 19, 2053–2069. https://doi.org/10.1007/s10531-010-9825-x

Webb, N.R., 1998. The traditional management of European heathlands. J. Appl. Ecol. 35, 987–990. https://doi.org/10.1111/j.1365-2664.1998.tb00020.x

Wehr, A., Lohr, U., 1999. Airborne laser scanning - An introduction and overview. ISPRS J. Photogramm. Remote Sens. 54, 68–82. https://doi.org/10.1016/S0924-2716(99)00011-8

Westra, T., De Knijf, G., Ledegen, H., De Bruyn, L., Maes, D., Onkelinx, T., Piesschaert, F., Vanreusel, W., Elegem, B. Van, Pollet, M., Quataert, P., 2016. Monitoring van prioritaire dier-en plantensoorten in Vlaanderen Opstart van nieuwe meetnetten. Natuur.Focus 15, 156–165.

Williams, K., Sader, S.A., Pryor, C., Reed, F., 2006. Application of geospatial technology to monitor forest legacy conservation easements. J. For. 104, 89–93.

Wilson, G.A., Hart, K., 2001. Farmer participation in agri-environmental schemes: Towards conservation-oriented thinking? Sociol. Ruralis 41, 254–274. https://doi.org/10.1111/1467-9523.00181

Wintle, B.A., Kujala, H., Whitehead, A., Cameron, A., Veloz, S., Kukkala, A., Moilanen, A., Gordon, A., Lentini, P.E., Cadenhead, N.C.R., Bekessy, S.A., 2019. Global synthesis of conservation studies reveals the importance of small habitat patches for biodiversity. Proc. Natl. Acad. Sci. U. S. A. 116, 909–914. https://doi.org/10.1073/pnas.1813051115

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., Group, N.P.S.D.W., 2008. Effects of sample size on the performance of species distribution models. Divers. Distrib. 14, 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

Wogan, G.O.U., 2016. Life history traits and niche instability impact accuracy and temporal transferability for historically calibrated distribution models of North American birds. PLoS One 11, e0151024. https://doi.org/10.1371/journal.pone.0151024

Wood, E.M., Pidgeon, A.M., Radeloff, V.C., Keuler, N.S., 2013. Image Texture Predicts Avian Density and Species Richness. PLoS One 8, e63211. https://doi.org/10.1371/journal.pone.0063211

Wood, E.M., Pidgeon, A.M., Radeloff, V.C., Keuler, N.S., 2012. Image texture as a remotely sensed measure of vegetation structure. Remote Sens. Environ. 121, 516–526. https://doi.org/10.1016/j.rse.2012.01.003

Wood, K.A., Stillman, R.A., Hilton, G.M., 2018. Conservation in a changing world needs predictive models. Anim. Conserv. 21, 87–88. https://doi.org/10.1111/acv.12371

Wood, S., 2017. Generalized Additive Models: An Introduction with R, 2nd ed. Chapman and Hall, London.

Worboys, G.L., Francis, W.L., Lockwood, M., 2010. Connectivity conservation management: A global guide. Earthscan, London, United Kingdom. https://doi.org/10.4324/9781849774727

Wu, B., Zhang, M., Zeng, H., Tian, F., Potgieter, A.B., Qin, X., Yan, N., Chang, S., Zhao, Y., Dong, Q., Boken, V., Plotnikov, D., Guo, H., Wu, F., Zhao, H., Deronde, B., Tits, L., Loupian, E., 2022. Challenges and opportunities in remote sensing-based crop monitoring: a review. Natl. Sci. Rev. 10, 1–17.

Yackulic, C.B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Campbell Grant, E.H., Veran, S., 2013. Presence-only modelling using MAXENT: When can we trust the inferences? Methods Ecol. Evol. 4, 236–243. https://doi.org/10.1111/2041-210x.12004

Yang, Y., Hillebrand, H., Lagisz, M., Cleasby, I., Nakagawa, S., 2022. Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology. Glob. Chang. Biol. 28, 969–989. https://doi.org/10.1111/gcb.15972

Yu, J., Wong, W.-K., Hutchinson, R.A., 2010. Modeling Experts and Novices in Citizen Science Data for Species Distribution Modeling, in: IEEE International Conference on Data Mining Modeling. pp. 1157–1162. https://doi.org/10.1109/ICDM.2010.103

Zellweger, F., De Frenne, P., Lenoir, J., Rocchini, D., Coomes, D., 2019. Advances in Microclimate Ecology Arising from Remote Sensing. Trends Ecol. Evol. 34, 327–341. https://doi.org/10.1016/j.tree.2018.12.012

Żmihorski, M., Dziarska-Pałac, J., Sparks, T.H., Tryjanowski, P., 2013. Ecological correlates of the popularity of birds and butterflies in Internet information resources. Oikos 122, 183–190. https://doi.org/10.1111/J.1600-0706.2012.20486.X

Zurell, D., Franklin, J., König, C., Bouchet, P.J., Dormann, C.F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J.J., Leitão, P.J., Park, D.S., Peterson, A.T., Rapacciuolo, G., Schmatz, D.R., Schröder, B., Serra-Diaz, J.M., Thuiller, W., Yates, K.L., Zimmermann, N.E., Merow, C., 2020. A standard protocol for reporting species distribution models. Ecography (Cop.). 43, 1261–1277. https://doi.org/10.1111/ECOG.04960

Zurell, D., Jeltsch, F., Dormann, C.F., Schröder, B., 2009. Static species distribution models in dynamically changing systems: How good can predictions really be? Ecography (Cop.). 32, 733–744. https://doi.org/10.1111/j.1600-0587.2009.05810.x

APPENDICES

CHAPTER II

Appendix A: Data selection and model evaluation procedure

The impact of applying data quality filters on opportunistic citizen science data was assessed for 255 species across four taxonomic groups, i.e. birds, butterflies, dragonflies and plants. Species occurrence records were extracted from the Belgian data platform *waarnemingen.be* (BOX 1). All submitted species records in *waarnemingen.be* were provided by Natuurpunt Studie as point coordinates with specified geographical precision, accompanied by all the details provided by the observer at data entry. We had access to the relevant metadata to define the degree of structure and received the validation status of each record at the moment of data extraction from the database. User ids were randomised for privacy law compliance. An initial selection retained only observations made from January 2014 until September 2019 in the Flemish region of Belgium and with a precision of 500 metres or less. Absences (zero-counts) and records with an 'incorrect' validation status were removed, and only birds that breed in Flanders were used (Vermeersch et al., 2020).

Figure A.1 presents a scheme of the formation of the different training and testing sets for species distribution modelling. In the process of selecting the model testing set, which is preferably an independent dataset of structured records (Araújo and Guisan, 2006), we were confronted with two major limitations. The first limitation was the unavailability of fully independent structured data for most species in Flanders. The second limitation was that random cross-validation, where species records are repeatedly split in a model training and model testing set and model performance is averaged across different folds (Fielding and Bell, 1997), has been discouraged for presence-only SDMs (van Proosdij et al., 2016) and when using auto-correlated data (Roberts et al., 2017), and it was therefore not suited to compare presence-only datasets of different quality. To counter these limitations, we separated the structured from the unstructured records and used the first for model testing and the latter for model training.

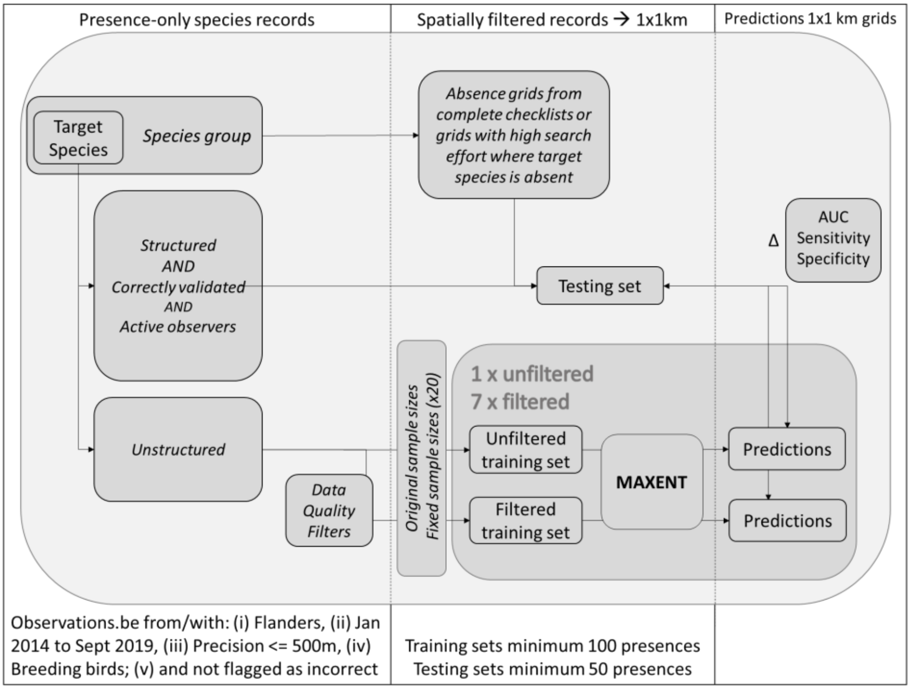

Presence-only species records
Spatially filtered records → 1x1km
Predictions 1x1 km grids
Target Species
Species group
Absence grids from complete checklists or grids with high search effort where target species is absent
AUC
Sensitivity
Specificity
Δ
Structured
AND
Correctly validated
AND
Active observers
Testing set
Original sample sizes
Fixed sample sizes (x20)
1 x unfiltered
7 x filtered
Unstructured
Unfiltered training set
MAXENT
Predictions
Data Quality Filters
Filtered training set
Predictions
Observations.be from/with: (i) Flanders, (ii) Jan 2014 to Sept 2019, (iii) Precision <= 500m, (iv) Breeding birds; (v) and not flagged as incorrect
Training sets minimum 100 presences
Testing sets minimum 50 presences

Figure A.1. *Scheme of the dataset generation and model evaluation procedure. First, presence-only records from a target species were restricted to match five (i-v) conditions and structured records were separated from unstructured records. The structured records were further reduced to match only records that were validated as correct and made by more active observers. These high-quality records formed the model testing set, together with absences from complete checklists and absences derived from 1x1km grids with high search effort for the associated taxonomic group and where the target species was not observed. All unstructured records were used for model training and were subjected to three data quality filters, as single filters or in combinations, forming seven filtered datasets. Unfiltered datasets were also kept to use as a baseline for the evaluation of a change in model performance by filtering. All eight datasets were aggregated to one presence per 1x1 km grid to use as input for Maxent. The resulting datasets of presence grids were modelled as such, with varying sample sizes, and they were also repeatedly (x20) and randomly reduced to six fixed levels of sample size (100, 250, 500, 1000, 2000 and 4000 presences), if possible. Only species with at least one filtered dataset of at least 100 presence grids and a testing set with at least 50 presence grids were retained. Model predictions were compared with the high-quality testing set, and model discrimination accuracy was measured by calculating the AUC (area under the receiver operating characteristic curve), Sensitivity and Specificity.*

The unstructured model training data was subjected to three filters and their combinations (Table A.1), forming one unfiltered and seven filtered datasets (Figure A.2). Species records were then aggregated by a 1x1 km grid in our study area. Using an unstructured dataset as the baseline for assessing the impact of filtering aided the transferability of our study, because large databases of volunteer-generated data often consist only of unstructured incidental observations.

Table A.1. *Definitions and motivations of the three data quality filters.*

Filter	Measurement	Definition of high quality	Motivation
ACTIVITY	Activity rate = the average number of active days of an observer per full year. *Threshold = the first quartile of the activity rates of observers that provided 80% of all the 20,676,308 observations in the study period 01-2014 to 09-2019 in waarnemingen.be.*	Activity rate >= 93 days *3% of the 28,855 observers met this threshold. First-year observers were classified as low quality.*	More active observers have more experience, leading to lower rates of both false-negative and false-positive errors *e.g. Farmer et al. (2012), Kallimanis et al. (2017), Kelling et al. (2015)*
DETAIL	More detailed information, beyond the default name, date and location, was given at data entry. *E.g. behaviour, sex, comments, …*	One or more extra 'information fields' were filled out.	An observer providing detailed information shows an increased effort *e.g. Steen et al. (2019)*
VALSTAT	The validation status of the record in the database. *Classification of validation codes in waarnemingen.be:* *A, J = Correct* *O, P, I, U = Uncertain* *Incorrect entries (N) were removed*	Correctly validated by auto-validation or expert verification	Correct data are meant to contain no misidentification errors *e.g. Vantieghem et al. (2017)*

The structured model testing data was further restricted to data that was also validated as correct and coming from more experienced observers. Here we also aggregated the species records by the 1x1 km grid. The high-quality presences were complemented with absences from complete checklists (Sullivan et al., 2014) and absences derived from grid cells with the highest search effort, based on the principle of species accumulation curves (Colwell et al., 2004), where an absence is noted when many species from a taxonomic group were recorded in a grid cell but not the target species. We first took derived absences from the 5% most frequently visited grid cells with checklist observations and supplemented these with derived absences from grid cells with a high search effort for the considered taxonomic group. The number of absences was chosen to match the number of high-quality presences, and we formed one presence-absence model testing set per species.

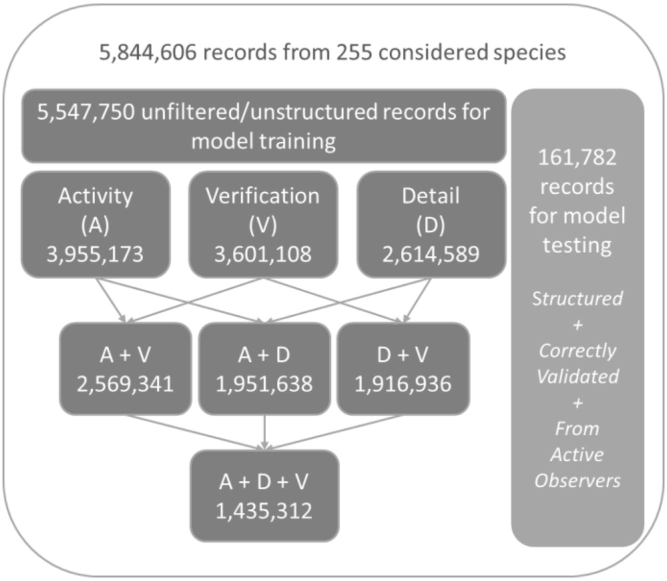

5,844,606 records from 255 considered species
5,547,750 unfiltered/unstructured records for model training
Activity (A) 3,955,173
Verification (V) 3,601,108
Detail (D) 2,614,589
A + V 2,569,341
A + D 1,951,638
D + V 1,916,936
A + D + V 1,435,312
161,782 records for model testing
Structured + Correctly Validated + From Active Observers

Figure A.2. *The total number of records in the filtered and unfiltered model training sets and in the model testing set. Unstructured data was used to generate model training sets, i.e. one unfiltered set and seven filtered sets per species. Structured data was never used for model training and was further reduced to records that were also validated as correct and collected by more active observers.*

Only species with at least one filtered training set of 100 presences and a testing set of at least 50 presences were selected. We ran Maxent models (Phillips et al., 2006) on the model training sets obtained after the data aggregation step. In addition, we repeatedly (x20) and randomly reduced the sample size to six fixed levels of 100, 250, 500, 1000, 2000 and 4000 presences, if the sample size of the species allowed it, and ran Maxent models for each of them. Figure A.3 shows an example of how many different training sets can be formed for one particular species.

We used the area under the receiver operating characteristic curve (AUC), Sensitivity and Specificity (Fielding and Bell, 1997) to assess Maxent's discrimination accuracy by comparing the predictions for a species' distribution based on a presence-only model training set to the species' presence-absence model testing set. An assessment of relative model performance was chosen because an accurate model assessment based on opportunistic presence-only data is impossible (Peterson et al., 2011) and because our main interest was to see whether model performance changes with the manipulation of the model training set. We evaluated the results of our analyses across all species and across species within one taxonomic group. We compared model performance (i) between models built with data of different quality (and constant sample size), (ii) between models built with data of different sample size (and constant data quality), and (iii) between models built with unfiltered data versus those built with filtered data (Figure A.3). Based on the differences in model performance of the latter comparison, we could assess the combined impact of data quality and sample size on model performance.

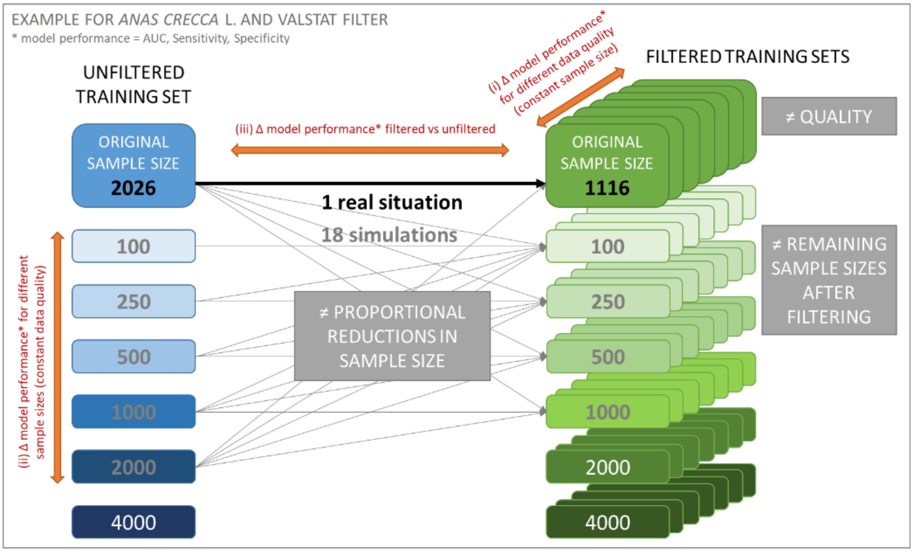

EXAMPLE FOR *ANAS CRECCA* L. AND VALSTAT FILTER
* model performance = AUC, Sensitivity, Specificity
UNFILTERED TRAINING SET
FILTERED TRAINING SETS
(i) Δ model performance* for different data quality (constant sample size)
(iii) Δ model performance* filtered vs unfiltered
ORIGINAL SAMPLE SIZE
2026
ORIGINAL SAMPLE SIZE
1116
≠ QUALITY
1 real situation
18 simulations
(ii) Δ model performance* for different sample sizes (constant data quality)
100
250
500
1000
2000
4000
≠ PROPORTIONAL REDUCTIONS IN SAMPLE SIZE
100
250
500
1000
2000
4000
≠ REMAINING SAMPLE SIZES AFTER FILTERING

Figure A.3. *An example of how many different training sets can be formed for one particular species (Anas crecca L.) and filter (VALSTAT) to assess the impact of quality filtering on model performance. 66,940 species records were aggregated to 2026 unfiltered presences (1x1 km grid cells) (see Table C.1). This allowed unfiltered training sets of 100, 250, 500, 1000 and 2000 presences to be used in the analysis of the impact of absolute sample size on model performance (ii). Filtering by VALSTAT retained 37,631 records, aggregated in 1116 filtered presences. This allowed filtered training sets of 100, 250, 500 and 1000 presences to be used in the analysis of the impact of absolute sample size on model performance (ii). The example can be repeated for the other filters, each resulting in a different number of presences. By comparing model performance metrics between training sets of different quality (unfiltered data and filtered data) but equal sample size, we could assess the impact of data quality (i). In this example, we could compare the quality of filtered data and unfiltered data at 100, 250, 500 and 1000 presences. By combining the different training sets of unfiltered and filtered data, we could assess the combined impact of data quality and sample size (iii). In this example, we could form 19 combinations (= 19 values for Δ AUC, Δ Sensitivity and Δ Specificity impacted by different proportional reductions in sample size and different remaining sample sizes after filtering).*

Appendix B: ODMAP protocol

Overview

Authorship

Contact: camille.vaneupen@kuleuven.be

Study link: https://doi.org/10.1016/j.ecolmodel.2021.109453

Model objective

Model objective: Mapping and interpolation

Target output: A relative occurrence rate per 1x1 km grid cell.

Focal Taxon – 255 species from 4 taxonomic groups: 54 bird species, 25 butterfly species, 14 dragonfly species and 162 vascular plant species

Location – Flanders

Scale of Analysis

Spatial extent: 2.51, 5.95, 50.67, 51.51 (xmin, xmax, ymin, ymax)

Spatial resolution: 1 km

Temporal extent: January 2014 to September 2019

Boundary: political, regional

Biodiversity data

Observation type: citizen science

Response data type: presence-only

Predictors – Predictor types: climatic, habitat, geographical

Hypotheses – We tested whether filtering by data quality improved model performance when model parameters and predictors were kept constant.

Assumptions – Model assumptions: We assumed that (i) the species are at (pseudo-) equilibrium with their environment, (ii) spatial thinning was sufficient to reduce the impact of sampling bias to a minimum, (iii) the included ecological drivers were sufficiently relevant for all considered species to ensure a constant predictor set throughout our study, (iv) the environmental response of a species was similar across the entire study area, i.e. with no adaptation to the local environment.

Algorithms

Modelling techniques: Maxent

Model complexity: Maxent models were built with linear, quadratic and product features only.

Workflow – Model workflow: Species records were subjected to three data quality filters, as single filters or in combinations, forming seven filtered datasets. Unfiltered datasets were also kept to use as a baseline for the evaluation of a change in model performance by filtering. All eight datasets were aggregated to one presence per 1x1 km grid to use as input for Maxent. The resulting datasets of presence grids were modelled as such, with varying sample sizes, and they were also repeatedly (x20) and randomly reduced to six fixed levels of sample size (100, 250, 500, 1000, 2000 and 4000 presences), if possible. Only species with at least one filtered dataset of at least 100 presence grids and a testing set with at least 50 presence grids were retained. Model predictions were compared with the high-quality testing set, and model discrimination accuracy was measured by calculating the AUC (area under the receiver operating characteristic curve), Sensitivity and Specificity.

Software

Software: Analyses were conducted in R version 4.0.1 (R Core Team, 2021), using Maxent version 3.4.1 (https://biodiversityinformatics.amnh.org/open_source/maxent/) implemented in the R package ´dismo` v1.1-4 (Hijmans et al., 2017).

Data availability: The full dataset (species presences for model training, model testing set, model predictors and quality tags) will be made available in Dryad Digital Repository at a 1x1 km resolution.

<u>Data</u>

Biodiversity data

Taxon names:

- 54 birds: Accipiter nisus, Actitis hypoleucos, Alcedo atthis, Alopochen aegyptiaca, Anas crecca, Anser anser, Ardea alba, Ardea cinerea, Athene noctua, Aythya farina, Aythya fuligula, Branta canadensis, Branta leucopsis, Buteo buteo, Carduelis carduelis, Chroicocephalus ridibundus, Ciconia ciconia, Circus aeruginosus, Circus cyaneus, Corvus frugilegus, Cuculus canorus, Cygnus olor, Delichon urbicum, Egretta garzetta, Falco tinnunculus, Fulica atra, Gallinago gallinago, Larus argentatus, Larus canus, Larus fuscus, Limosa limosa, Linaria cannabina, Luscinia svecica, Mareca strepera, Motacilla alba, Motacilla flava, Numenius arquata, Oenanthe oenanthe, Perdix perdix, Phalacrocorax carbo, Platalea leucorodia, Podiceps cristatus, Psittacula krameri, Rallus aquaticus, Recurvirostra avosetta, Riparia riparia, Spatula clypeata, Spinus spinus, Sterna hirundo, Tachybaptus ruficollis, Tadorna tadorna, Tringa tetanus, Turdus pilaris, Vanellus vanellus;

- 25 butterflies: Aglais io, Aglais urticae, Anthocharis cardamines, Aphantopus hyperantus, Araschnia Levana, Aricia agestis, Celastrina argiolus, Coenonympha pamphilus, Colias crocea, Favonius quercus, Gonepteryx rhamni, Issoria lathonia, Lycaena phlaeas, Maniola jurtina, Ochlodes sylvanus, Papilio machaon, Pararge aegeria, Pieris brassicae, Pieris napi, Pieris rapae, Polygonia c-album, Polyommatus icarus, Pyronia tithonus, Vanessa atalanta, Vanessa cardui;

- 14 dragonflies: Aeshna cyanea, Aeshna mixta, Anax imperator, Calopteryx splendens, Coenagrion puella, Enallagma cyathigerum, Ischnura elegans, Libellula depressa, Libellula quadrimaculata, Orthetrum cancellatum, Platycnemis pennipes, Pyrrhosoma nymphula, Sympetrum sanguineum, Sympetrum striolatum;

- 162 vascular plants: Acer pseudoplatanus, Achillea millefolium, Aegopodium podagraria, Ajuga reptans, Alliaria petiolata, Allium vineale, Alnus glutinosa, Anemone nemorosa, Angelica sylvestris, Anisantha sterilis, Anthriscus sylvestris, Arabidopsis thaliana, Artemisia vulgaris, Arum maculatum, Asplenium ruta-muraria, Bellis perennis, Betula pendula, Bromus hordeaceus, Calamagrostis epigejos, Calluna vulgaris, Capsella bursa-pastoris, Cardamine hirsute, Cardamine pratensis, Carex hirta, Centaurea jacea, Cerastium glomeratum, Chamerion angustifolium, Chelidonium majus, Chenopodium album, Cirsium arvense, Cirsium palustre, Cirsium vulgare, Convolvulus arvensis, Convolvulus sepium, Conyza Canadensis, Coronopus didymus, Corylus avellana, Crataegus monogyna, Crepis capillaris, Cytisus scoparius, Dactylis glomerata, Daucus carota, Draba verna, Dryopteris filix-mas, Echinochloa crus-galli, Epilobium hirsutum, Epipactis helleborine, Equisetum arvense, Erodium cicutarium, Eupatorium cannabinum, Euphorbia peplus, Fallopia japonica, Ficaria verna, Filipendula ulmaria, Fraxinus excelsior, Galium aparine, Galium palustre, Geranium dissectum, Geranium molle, Geranium robertianum, Geum urbanum, Glechoma hederacea, Gnaphalium luteoalbum, Gnaphalium uliginosum, Hedera helix, Heracleum sphondylium, Hieracium pilosella, Hieracium umbellatum, Holcus lanatus, Humulus lupulus, Hypericum perforatum, Hypochaeris radicata, Ilex aquifolium, Iris pseudacorus, Jacobaea vulgaris, Juncus effusus, Lactuca serriola, Lamium album, Lamium purpureum, Lapsana communis, Lathyrus pratensis, Leucanthemum vulgare, Linaria vulgaris, Lonicera periclymenum, Lotus corniculatus, Lotus pedunculatus, Luzula campestris, Lycopus europaeus, Lysimachia vulgaris, Lythrum salicaria, Malva sylvestris, Matricaria discoidea, Medicago lupulina, Melilotus albus, Mentha aquatica, Mercurialis annua, Ornithopus perpusillus, Papaver dubium, Papaver rhoeas, Persicaria amphibian, Persicaria maculosa, Phragmites australis, Plantago coronopus, Plantago lanceolata, Poa annua, Polygonatum multiflorum, Polygonum aviculare, Potentilla anserina, Potentilla reptans, Prunella vulgaris, Prunus avium, Prunus serotine, Prunus spinosa, Pulicaria dysenterica, Quercus robur, Ranunculus acris, Ranunculus repens, Ranunculus sceleratus, Rorippa palustris, Rumex acetosa, Rumex acetosella, Rumex obtusifolius, Sagina procumbens, Salix caprea, Sambucus nigra, Scrophularia nodosa, Sedum acre, Senecio inaequidens, Senecio vulgaris, Silene dioica, Silene flos-cuculi, Silene latifolia, Sinapis arvensis, Sisymbrium officinale, Solanum dulcamara, Sonchus asper, Sonchus oleraceus, Sorbus aucuparia, Stachys palustris, Stachys sylvatica, Stellaria holostea, Stellaria media, Symphytum officinale, Tanacetum vulgare, Tragopogon pratensis, Trifolium arvense, Trifolium dubium, Trifolium pratense, Trifolium repens, Tussilago farfara, Typha latifolia, Urtica dioica, Valeriana officinalis, Veronica arvensis, Veronica chamaedrys, Veronica hederifolia, Veronica persica, Veronica serpyllifolia, Vicia cracca, Vicia hirsuta, Vicia sativa, Viola arvensis.

Taxonomic reference system: We followed the taxonomy of the data repository ‘waarnemingen.be’

Ecological level: species

Data sources: waarnemingen.be

Sampling design: Model training data were unstructured opportunistic data (i.e. incidental observations that are not related to any survey project, and not supported by guidelines nor a protocol).

Sample size: We used the sample sizes of the original training sets (Table C.1) and six fixed sample sizes (100, 250, 500, 1000, 2000, 4000)

Clipping: Flanders

Scaling: We received the data as point coordinates and excluded records with a geographical precision > 500 m. Point records were aggregated in a 1x1 km grid, resulting in one presence per grid cell per species.

Cleaning: We retained only breeding birds and removed absences (zero-counts) and entries validated as incorrect. The aggregation of point records is also called ‘spatial thinning’ or ‘spatial filtering’, a common technique to reduce spatial bias.

Background data: The entire study area was used as background (1x1km resolution)

Errors and biases: Opportunistic data may suffer from biases and error (e.g. misidentification errors, imperfect detection, sampling bias), that can result in low-quality training data. The goal of the study was to evaluate whether stringent filtering can improve data quality, potentially removing some of the biases and errors.

Data partitioning

Training data: Model training presences: Unstructured species records

Test data:

- Model testing presences: structured species records, validated as correct in the database's internal validation system, made by more active observers;
- Model testing absences: absences from complete checklists and absences derived from 1x1 km grids with high search effort for the associated taxonomic group and where the target species was not observed.

Predictor variable

Predictor variables (Table C.3):

- 12 continuous predictors:
  - 10 land use classes (forest, semi-natural grassland, scrub, heathland, saltmarshes, wetlands, dunes, urban areas, water and other green areas)
  - 2 climate variables (mean annual temperature and mean annual precipitation)
- 2 factor variables: dominant soil texture and ecoregion

Data sources:

- Land use (Poelmans and Van Daele, 2014) and ecoregion (Couvreur et al., 2004): https://www.geopunt.be/catalogus
- Mean annual temperature and mean annual precipitation, BIO1 and BIO12 from WorldClim 2 respectively (Fick and Hijmans, 2017): https://www.worldclim.org/data/index.html
- Dominant soil texture class (Maréchal and Tavernier, 1974): https://www.dov.vlaanderen.be/

Spatial resolution:

- Rasters: Land use at 10m resolution and mean annual temperature and mean annual precipitation at approximately 1 km resolution

Coordinate reference system: all data was available in or transformed to Belge 1972 / Belgian Lambert 72 - Belgium - EPSG:31370

Data processing: Land use was aggregated in 11 classes: agriculture, forest, semi-natural grassland, scrub, heathland, saltmarshes, wetlands, dunes, urban areas, water and other green areas. The area of these classes in each 1x1 km cell was calculated. We used the mean temperature and precipitation and the modal value of the dominant soil texture class and ecoregion in each 1x1 km cell.

<u>Model</u>

Multicollinearity – Pearson correlations between predictors were calculated. We removed one class "agriculture" from the set because of the relatively high collinearity with other classes (maximum $|\rho| = 0.69$) and because of the problem with perfect multicollinearity in compositional data.

Model settings – maxent: featureSet (linear, quadratic, product), maximumbackground (13552)

Threshold selection – AUC is a threshold-independent measure of model performance. The threshold to calculate Sensitivity and Specificity was set to the value that maximizes the sum of Sensitivity and Specificity calculated on the species' testing set.

Assessment

Performance statistics – Performance on test data: AUC, Sensitivity, Specificity

Prediction

Prediction output – Prediction unit: relative occurrence rate

Appendix C: Summary species and predictor variables

Table C.1. Species occurrence records and presences (i.e. the number of 1x1km grid cells) in the model testing and model training sets for each of the 255 analysed species in the four considered taxonomic groups.

		Model Testing Data		Model Training Data											
				UNFILTERED		FILTERED									
						Records			*Presences (1x1 km grid cells)*						
TAXONOMIC GROUP species	**All observations**	*Testing records*	*Testing presences*	*Training records*	*Training presences*	ACTIVITY (A)	DETAIL (D)	VALSTAT (V)	A	D	V	A + D	A + V	D + V	A + D + V
BIRDS															
Anas crecca	66940	1075	189	64591	2026	46075	22002	37631	1724	1296	1116	1175	1020	855	807
Aythya ferina	48406	1081	108	46259	1067	31596	16680	28288	879	677	577	604	508	460	429
Ciconia ciconia	22415	205	63	22103	2940	12938	10266	17084	1734	1774	1754	1165	1215	1232	920
Circus aeruginosus	38929	152	81	38483	2197	25957	22132	15466	1740	1708	1310	1446	1135	1165	1046
Cygnus olor	53724	1285	183	51753	2084	35968	21176	40334	1781	1391	1412	1256	1244	1064	979
Fulica atra	122925	6169	635	114743	3463	81893	49185	102053	2966	2578	2884	2358	2548	2290	2128
Podiceps cristatus	80134	2094	268	77261	1922	53551	31455	68381	1650	1412	1536	1294	1377	1237	1159
Riparia riparia	12946	277	58	12471	977	8879	6281	5442	796	693	432	610	386	377	355
Spatula clypeata	55531	1257	158	53305	1416	36967	17781	33042	1213	900	793	821	727	637	600
Tadorna tadorna	72180	1345	265	69713	2547	50085	25486	45725	2127	1667	1469	1496	1337	1177	1106
Alcedo atthis	65360	837	139	63854	3734	40979	21966	43825	2761	2350	2074	1877	1698	1592	1380
Athene noctua	26215	191	89	23058	3705	15062	14101	8312	2489	2845	1077	2024	848	889	731
Limosa limosa	28061	238	66	27510	850	17589	10902	22137	686	634	574	553	514	475	445
Mareca strepera	88789	3587	346	82514	2387	58320	29276	59055	2097	1647	1530	1533	1404	1223	1165
Motacilla flava	35019	286	106	33173	3677	23394	14563	7525	2846	2395	1309	1997	1136	1051	946
Numenius arquata	36836	413	139	36102	2274	23724	15402	25897	1803	1676	1186	1434	1022	998	912
Phalacrocorax carbo	83573	2985	376	78682	3491	54935	26850	53965	2823	2040	2322	1828	2029	1677	1556
Tachybaptus ruficollis	73576	1623	205	70884	2315	49499	31470	45978	1960	1659	1417	1488	1279	1166	1084
Ardea cinerea	141335	5771	748	134133	6797	93723	46211	120460	5596	4343	5913	3820	4984	4008	3572
Circus cyaneus	20748	99	57	20505	2140	13330	10988	6656	1539	1561	944	1202	772	803	681
Anser anser	61632	2875	316	57965	2273	41169	23902	52791	1884	1494	1793	1361	1559	1308	1219
Cuculus canorus	51543	465	94	48910	4515	31855	36801	13066	3356	3992	920	3112	846	882	818
Delichon urbicum	36026	621	96	32982	3373	23041	18294	9715	2599	2359	1456	1927	1229	1208	1047
Psittacula krameri	20889	1488	74	17047	1135	12316	6994	5784	824	771	601	615	483	497	429
Aythya fuligula	94955	2011	231	90788	2279	65172	35121	52833	1975	1633	1303	1491	1220	1118	1067
Larus canus	28453	428	117	26630	3171	21565	9735	8798	2769	1717	1014	1611	939	735	700
Oenanthe oenanthe	30057	155	58	29626	2658	21249	8285	9455	2046	1274	1273	1070	1090	789	706
Branta canadensis	71715	2310	423	66905	3603	47572	29252	34469	2941	2503	2636	2174	2253	2094	1868
Ardea alba	102215	841	222	100758	3791	67725	29323	75265	2983	2336	2712	2019	2297	1910	1731
Recurvirostra avosetta	19421	198	55	19068	432	13009	7635	16016	361	297	292	270	257	228	216
Branta leucopsis	19420	203	66	18979	1261	12714	6980	14588	1038	740	635	657	568	478	445
Buteo buteo	226402	7694	1460	215145	10409	158255	74435	167866	9033	7595	8836	6899	7900	6946	6415
Sterna hirundo	23890	432	81	23251	847	16041	10056	17437	698	636	569	568	487	474	435
Larus argentatus	39999	1022	272	37065	3290	29000	15076	22325	2705	1926	1722	1747	1532	1286	1199
Actitis hypoleucos	40101	332	102	39445	1887	28533	11270	24666	1524	1110	1048	991	915	761	710
Perdix perdix	24772	108	51	24253	4047	17112	11682	8533	2976	2750	1668	2166	1339	1351	1131
Platalea leucorodia	29070	173	58	28756	661	18098	11121	25632	505	443	498	376	413	381	339
Corvus frugilegus	21340	307	139	20294	3248	15771	10718	7514	2664	1928	1462	1697	1288	1058	972
Tringa totanus	31689	221	58	31104	923	21612	10525	20393	787	643	413	589	380	343	327
Gallinago gallinago	41587	275	69	40787	2457	28343	14634	19646	1921	1573	737	1318	649	631	576

		Model Testing Data		Model Training Data											
				UNFILTERED		FILTERED									
						Records			*Presences (1x1 km grid cells)*						
TAXONOMIC GROUP species	**All observations**	*Testing records*	*Testing presences*	*Training records*	*Training presences*	ACTIVITY (A)	DETAIL (D)	VALSTAT (V)	A	D	V	A + D	A + V	D + V	A + D + V
Larus fuscus	37415	722	215	35484	3545	28109	14289	13957	3002	2052	1914	1853	1670	1368	1248
Egretta garzetta	34384	179	54	34062	1336	22324	10132	27712	981	728	681	594	584	486	446
Luscinia svecica	38297	814	102	35776	1650	25258	27833	17712	1403	1493	584	1320	550	562	537
Chroicocephalus ridibundus	90486	7034	1007	81536	5347	62309	30172	70663	4469	3287	3959	2991	3446	2714	2507
Accipiter nisus	56253	387	198	54627	6908	39659	27949	23002	5406	4913	3235	4036	2668	2697	2284
Carduelis carduelis	61030	2147	330	57315	4874	40949	25931	40278	3776	3261	3427	2706	2777	2609	2229
Motacilla alba	74384	2492	435	69438	6646	50595	26373	41294	5234	4030	4195	3425	3478	2998	2622
Rallus aquaticus	34313	402	60	33139	1233	23789	23703	16029	1025	1070	387	922	343	367	332
Alopochen aegyptiaca	69409	2779	557	63949	4840	45003	23925	43752	3863	3110	3729	2664	3077	2697	2354
Falco tinnunculus	133428	2001	724	129573	8168	96905	52431	84360	6970	5930	5563	5278	4963	4553	4182
Linaria cannabina	48356	1810	263	44924	3519	33086	19367	27945	2923	2229	1972	1968	1716	1477	1348
Spinus spinus	32863	668	163	30608	3595	22427	14390	12946	2800	2298	1659	1904	1362	1256	1073
Turdus pilaris	45436	140	51	43358	5061	31109	16749	5192	3893	2947	1077	2416	914	909	795
Vanellus vanellus	114506	1729	376	109499	6678	76958	51991	51679	5362	5153	3103	4382	2754	2768	2544
BUTTERFLIES															
Aglais io	87676	3459	720	83627	7194	61825	64229	83121	5848	6168	7186	5276	5843	6161	5272
Aglais urticae	31981	397	154	31512	4483	22995	21922	30641	3479	3433	4420	2829	3446	3391	2806
Araschnia levana	38277	1325	296	36438	4232	25154	27596	25782	3195	3531	3248	2809	2582	3130	2522
Coenonympha pamphilus	36060	1639	206	33297	2143	24032	25284	22923	1791	1748	1477	1523	1320	1438	1291
Colias crocea	9717	137	50	9518	2260	7082	6936	6084	1748	1836	1724	1486	1399	1673	1366
Papilio machaon	18248	339	128	17806	3552	10810	13316	13183	2181	2924	2885	1909	1894	2828	1881
Pieris brassicae	40766	2206	408	38026	4739	28888	29048	27866	3646	3926	3824	3194	3130	3794	3118
Pieris rapae	90958	5344	787	84437	7033	65904	71719	69611	5929	6257	6172	5470	5421	6131	5400
Vanessa atalanta	90433	2210	523	87614	7306	67118	67950	63624	5979	6333	6173	5421	5277	6088	5251
Vanessa cardui	40963	1470	342	39172	5625	28699	30093	28571	4293	4657	4609	3763	3721	4507	3676
Pieris napi	52476	1991	457	50042	4748	37461	41937	41974	3906	4126	4125	3503	3498	4068	3469
Aricia agestis	20238	1033	152	18845	2733	13908	15142	14009	2090	2328	2076	1872	1674	2006	1635
Anthocharis cardamines	34494	432	159	33352	4264	22160	23517	18068	3251	3409	2495	2774	2105	2458	2084
Issoria lathonia	7809	388	60	7349	1006	4704	5755	6350	652	842	824	578	554	752	522
Aphantopus hyperantus	21142	892	76	19322	1725	13051	14295	9392	1299	1369	806	1084	648	762	624
Maniola jurtina	92313	10375	699	79626	5588	59553	64757	64194	4513	4832	4793	4056	4039	4753	4018
Ochlodes sylvanus	33223	1514	208	31011	3800	22109	24325	21077	2903	3198	2667	2545	2113	2548	2053
Celastrina argiolus	35352	1071	229	33802	4554	24146	25923	22305	3495	3767	3190	3060	2607	3102	2566
Gonepteryx rhamni	88637	3610	587	84013	6604	58487	62352	83282	5342	5532	6578	4727	5331	5520	4719
Polyommatus icarus	55777	2122	305	52431	4461	38373	40668	38277	3584	3854	3544	3218	2980	3458	2934
Lycaena phlaeas	31449	833	223	30123	4025	21263	22980	22111	3075	3375	3097	2722	2506	2996	2452
Pararge aegeria	84092	2523	473	80296	6487	59858	63342	57532	5324	5622	4872	4814	4165	4796	4136
Favonius quercus	7105	270	55	6644	1290	4766	5475	4589	957	1108	921	868	720	896	708
Polygonia c-album	51957	1602	389	50040	5288	35995	38509	49759	3964	4421	5277	3499	3960	4413	3495
Pyronia tithonus	47206	3256	293	42763	4400	32439	34639	30664	3400	3719	3059	3022	2510	3002	2486
DRAGONFLIES															
Coenagrion puella	23061	567	112	21645	2477	13598	16216	15053	1835	2056	2009	1613	1541	1721	1377
Enallagma cyathigerum	13955	233	61	13378	1135	8927	10081	9319	890	938	826	777	665	708	592
Ischnura elegans	25035	440	132	23741	2633	15093	17526	16062	2040	2144	2035	1751	1611	1707	1410
Libellula depressa	12405	191	80	11988	2053	7482	8018	7390	1431	1545	1735	1159	1238	1333	1015
Orthetrum cancellatum	23908	482	156	22802	2632	15171	16965	15372	2112	2217	2103	1856	1713	1810	1539
Sympetrum striolatum	17363	293	60	16679	2433	11676	13175	8876	1845	2116	1600	1660	1224	1480	1154
Libellula quadrimaculata	13595	305	86	12534	1184	7427	8655	7471	918	968	1014	793	806	847	709
Anax imperator	20651	292	104	19722	2519	13054	14066	9599	1966	2029	1993	1679	1580	1634	1371

		Model Testing Data		**Model Training Data**											
				UNFILTERED		**FILTERED**									
						Records			*Presences (1x1 km grid cells)*						
TAXONOMIC GROUP species	**All observations**	*Testing records*	*Testing presences*	*Training records*	*Training presences*	ACTIVITY (A)	DETAIL (D)	VALSTAT (V)	A	D	V	A + D	A + V	D + V	A + D + V
Platycnemis pennipes	9184	225	71	8791	966	5406	6261	5939	733	773	870	625	665	702	572
Aeshna mixta	12716	187	79	12347	2124	9066	9518	7247	1672	1727	1705	1411	1357	1405	1156
Sympetrum sanguineum	22901	413	94	21803	2215	15191	17986	12361	1666	1908	1805	1489	1421	1606	1300
Aeshna cyanea	10002	113	50	9806	1901	7001	7181	5694	1411	1476	1502	1131	1136	1185	916
Pyrrhosoma nymphula	13673	170	58	13160	2036	7899	9232	8474	1417	1597	1566	1200	1131	1294	999
Calopteryx splendens	14466	481	77	13588	1598	8176	8970	9244	1169	1248	1275	968	937	1037	802
PLANTS															
Acer pseudoplatanus	7113	146	85	6704	1992	4716	1906	2967	1530	650	808	550	640	389	326
Achillea millefolium	13802	392	176	13069	3705	9581	4699	8090	2917	1722	2163	1509	1715	1248	1081
Alliaria petiolata	8051	209	126	7651	2431	4935	3052	5095	1859	1203	1579	1016	1216	927	784
Alnus glutinosa	7661	90	72	7284	2178	4526	1427	2498	1644	653	762	544	570	336	277
Arabidopsis thaliana	2900	197	52	2543	1219	1868	978	1233	950	431	521	353	382	251	200
Artemisia vulgaris	8437	209	105	7946	2676	5918	1856	4308	2106	926	1224	794	991	609	522
Bromus hordeaceus	4325	64	52	3969	1719	2729	1158	1535	1261	486	581	395	432	252	202
Calamagrostis epigejos	6810	93	50	6550	1203	5623	1160	2009	970	342	439	304	363	200	173
Capsella bursa-pastoris	5145	168	110	4755	2181	3345	1839	2374	1672	891	1065	748	814	567	473
Cardamine hirsuta	5508	268	103	4998	1973	3624	2149	2408	1476	844	1009	666	744	530	428
Carex hirta	6640	111	58	6318	1698	4895	1229	3637	1304	561	722	491	544	312	268
Centaurea jacea	14693	711	195	13733	3029	9770	4496	9174	2419	1457	2053	1257	1638	1178	1014
Cerastium glomeratum	4621	259	71	4132	1758	3093	1599	1982	1343	616	796	494	593	406	324
Chenopodium album	3801	123	76	3494	1675	2569	859	1482	1295	452	676	364	517	304	239
Cirsium arvense	25536	450	165	24241	3229	20778	3563	19588	2453	1326	1718	1124	1304	948	789
Cirsium vulgare	10888	318	171	10219	2837	7797	2655	7315	2193	1227	1745	1023	1359	951	789
Convolvulus arvensis	2810	110	73	2563	1100	1690	1165	1494	824	554	602	453	451	379	304
Convolvulus sepium	7068	221	112	6537	2101	4610	1934	3091	1525	873	1167	725	853	620	511
Conyza canadensis	5551	109	68	5194	1636	4151	1988	3381	1193	507	627	401	443	304	231
Dactylis glomerata	10014	114	65	9414	2354	6959	1516	2048	1785	693	756	581	588	338	285
Daucus carota	8923	255	109	8451	1867	6953	1715	6283	1472	756	933	632	734	506	421
Dryopteris filix-mas	7639	143	85	7255	2115	5004	2332	4417	1671	727	1192	595	930	501	396
Echinochloa crus-galli	1962	76	59	1756	1112	1339	471	790	891	314	432	246	335	211	155
Epilobium hirsutum	6958	147	93	6522	2082	4675	1962	3234	1549	927	1142	775	847	664	547
Epipactis helleborine	6910	115	76	6709	2182	4407	3273	5335	1571	1122	1684	836	1210	966	717
Equisetum arvense	6441	177	102	5915	2199	4223	1582	3070	1623	712	1124	577	822	462	365
Erodium cicutarium	3795	99	64	3585	1142	2935	1188	2177	917	520	667	439	542	385	328
Eupatorium cannabinum	10019	235	146	9380	2531	6304	2866	5915	1981	1155	1563	988	1238	882	755
Fraxinus excelsior	6109	90	53	5738	1890	3743	1312	1918	1425	602	603	518	476	280	239
Galium aparine	9425	225	132	8785	2811	5502	2469	4241	2089	1037	1353	875	1046	683	595
Geranium molle	5519	139	81	5180	1918	3743	2144	2926	1428	737	979	589	744	499	401
Hypericum perforatum	8394	276	72	7849	2517	5368	1998	3277	1844	900	758	725	587	354	305
Hypochaeris radicata	13699	403	137	13003	2972	9777	4753	6776	2298	1151	1496	991	1184	771	661
Jacobaea vulgaris	16578	422	163	15826	2302	14015	3329	11913	1761	1100	1379	903	1067	837	690
Lactuca serriola	3386	72	52	3134	1440	2174	1002	1570	1077	530	646	424	466	323	245
Lamium album	7498	355	178	6854	2472	4610	3350	4832	1857	1390	1682	1130	1263	1086	882
Lapsana communis	5432	145	91	5036	1998	3397	1846	2512	1494	748	953	595	700	508	391
Lotus corniculatus	5442	155	75	5116	1615	3307	1674	2060	1179	638	590	533	462	320	277
Medicago lupulina	5355	131	82	4994	1600	3755	1393	2039	1216	587	717	484	528	381	307
Melilotus albus	3515	124	67	3231	1173	2349	1156	2106	906	550	640	469	488	384	319
Mentha aquatica	10932	149	73	10593	1669	8171	1934	7015	1289	808	1037	689	823	619	538
Persicaria amphibia	3680	94	55	3402	1163	2515	1113	2199	885	442	639	368	495	333	274
Persicaria maculosa	3474	83	52	3157	1488	2157	814	1073	1038	481	445	371	304	221	167

		Model Testing Data		Model Training Data											
				UNFILTERED		FILTERED									
						Records			*Presences (1x1 km grid cells)*						
TAXONOMIC GROUP **species**	**All observations**	*Testing records*	*Testing presences*	*Training records*	*Training presences*	ACTIVITY (A)	DETAIL (D)	VALSTAT (V)	A	D	V	A + D	A + V	D + V	A + D + V
Phragmites australis	29787	150	80	29302	2443	26294	2075	4227	1917	919	1019	810	820	492	430
Plantago coronopus	3275	110	69	3083	1092	2236	1154	1828	888	347	575	300	460	252	214
Plantago lanceolata	19895	470	134	19074	3616	14504	5880	10729	2804	1510	1831	1307	1450	1010	876
Poa annua	13047	132	52	12581	2122	10925	7782	7947	1620	557	578	461	456	243	201
Polygonum aviculare	7322	113	66	6961	1832	5791	1936	4342	1389	526	655	430	516	298	247
Prunella vulgaris	7845	166	89	7473	2054	5162	2411	5410	1509	867	1320	727	991	693	578
Prunus spinosa	6951	151	88	6614	1780	4613	2044	3472	1417	869	946	733	780	567	483
Ranunculus acris	15931	262	121	15098	3019	11231	3826	6102	2267	1348	1454	1135	1093	830	687
Ranunculus repens	14588	543	132	13515	3048	9864	3327	4449	2279	1240	1389	1004	1047	759	618
Rorippa palustris	2126	89	66	1948	903	1426	486	949	715	317	482	251	376	242	192
Rumex obtusifolius	8316	151	81	7639	2736	4887	1751	2814	1984	834	1006	693	725	425	354
Sagina procumbens	3260	109	64	2982	988	2479	1485	1828	796	240	355	202	295	134	117
Sambucus nigra	9097	165	106	8571	2782	5838	2172	3318	2134	946	1139	802	892	551	465
Senecio vulgaris	5668	181	114	5293	2046	3904	1987	2758	1570	838	1017	689	768	560	455
Sinapis arvensis	1759	85	57	1554	876	1071	661	866	676	418	464	317	340	294	213
Sisymbrium officinale	3186	105	71	2910	1556	1976	945	1532	1152	562	818	450	580	397	311
Sonchus asper	5307	126	84	4941	1847	3654	1409	2977	1402	640	912	507	665	449	340
Sonchus oleraceus	4577	134	83	4202	1794	2915	1420	1863	1360	668	766	532	567	430	337
Stellaria holostea	10041	185	76	9753	1851	6646	3553	7604	1474	1043	1430	894	1139	884	762
Stellaria media	6031	236	131	5570	2311	4049	1986	2997	1788	868	1190	720	898	588	482
Tanacetum vulgare	9615	381	160	8877	2956	6155	3105	5313	2298	1357	1639	1158	1282	976	827
Trifolium dubium	5604	104	70	5172	1981	3615	1767	2407	1502	690	890	559	669	425	341
Trifolium repens	14178	205	111	13234	2742	10626	2550	8749	2024	968	1108	809	809	569	460
Tussilago farfara	6677	137	100	6427	1816	4518	3560	4936	1408	1120	1356	947	1053	915	778
Typha latifolia	5747	71	56	5478	1837	3720	1269	2816	1435	679	876	574	673	427	352
Urtica dioica	27564	558	174	26649	3938	20758	4802	7857	2948	1586	2017	1299	1522	1070	865
Veronica arvensis	4600	114	74	4317	1548	3434	2040	2687	1193	556	835	451	641	392	317
Vicia cracca	5426	154	89	4908	1837	3311	1485	2708	1354	701	913	561	662	474	377
Aegopodium podagraria	6310	142	90	5808	2134	4029	1741	2721	1643	801	1024	687	801	551	478
Allium vineale	2335	79	52	2199	892	1622	667	1413	730	333	556	284	442	253	216
Anisantha sterilis	3648	95	60	3329	1487	2321	1094	1301	1080	473	531	382	396	265	217
Anthriscus sylvestris	10372	246	154	9667	2877	6888	3544	6572	2204	1489	1820	1266	1409	1120	946
Bellis perennis	12269	261	175	11624	2981	8669	4501	6449	2210	1495	1941	1236	1392	1119	909
Chelidonium majus	6012	175	122	5661	2126	3842	2332	3778	1591	964	1319	798	980	721	592
Corylus avellana	8919	107	87	8567	2413	5441	2177	4450	1836	888	1220	743	972	640	545
Crataegus monogyna	7970	155	82	7495	2111	5152	1854	3143	1629	820	954	670	748	510	422
Crepis capillaris	6956	158	95	6498	2019	5199	2154	2747	1508	774	1006	613	746	542	420
Geranium dissectum	3487	68	54	3242	1373	2053	1204	1808	971	578	774	447	534	425	318
Geranium robertianum	9270	240	161	8747	2672	5794	3496	5778	2020	1236	1759	1006	1337	971	801
Geum urbanum	9388	234	112	8931	2185	6019	3462	5371	1692	1002	1244	873	983	707	617
Glechoma hederacea	15957	598	229	14968	3262	11160	4428	8903	2467	1564	2209	1299	1688	1228	1033
Hedera helix	10866	173	108	10457	2683	7504	3663	6111	2024	1062	1393	874	1025	727	574
Heracleum sphondylium	11541	311	151	10599	2866	7290	3063	5852	2208	1220	1595	1051	1249	856	726
Hieracium umbellatum	3991	122	71	3758	1337	2973	1225	2369	1133	539	761	480	642	389	339
Holcus lanatus	11285	140	63	10569	2497	7293	1822	2578	1827	731	774	602	583	352	296
Humulus lupulus	6092	116	79	5849	1939	4053	1662	3709	1557	752	1121	658	914	565	495
Ilex aquifolium	11132	106	80	10804	2610	7131	3325	5351	2085	1022	1288	899	1011	675	585
Iris pseudacorus	10649	215	83	10205	2372	6897	2986	5956	1863	1019	1309	891	1043	691	606
Lamium purpureum	7443	307	164	6911	2693	4875	3420	4867	2060	1431	1871	1177	1420	1142	937
Lathyrus pratensis	7196	181	69	6741	1622	4979	2363	4366	1288	731	996	651	774	523	466

		Model Testing Data		**Model Training Data**											
				UNFILTERED		**FILTERED**									
						Records			*Presences (1x1 km grid cells)*						
TAXONOMIC GROUP **species**	**All observations**	*Testing records*	*Testing presences*	*Training records*	*Training presences*	ACTIVITY (A)	DETAIL (D)	VALSTAT (V)	A	D	V	A + D	A + V	D + V	A + D + V
Leucanthemum vulgare	9202	253	95	8778	2454	6164	3285	4735	1961	1182	1217	1053	983	726	636
Lycopus europaeus	10583	156	91	10225	2256	6816	2367	6217	1765	896	1398	779	1068	675	573
Lythrum salicaria	8630	282	114	8050	2246	5415	2839	5507	1747	1147	1503	981	1166	918	778
Malva sylvestris	2757	113	76	2526	1184	1817	1155	1680	902	621	798	505	605	493	401
Matricaria discoidea	2750	104	73	2472	1199	1738	752	929	858	374	442	295	301	206	163
Potentilla anserina	13195	121	71	12846	1688	10967	1565	10813	1265	706	912	590	694	483	408
Potentilla reptans	3919	179	104	3587	1290	2382	1292	2192	984	556	742	476	569	412	354
Prunus avium	3986	88	54	3716	1411	2390	989	1523	1071	479	534	407	420	259	219
Pulicaria dysenterica	12965	270	119	12468	1840	10863	2389	10878	1459	978	1216	846	970	771	671
Rumex acetosa	11178	91	58	10550	2724	7000	2692	4252	2108	961	1078	834	823	526	462
Salix caprea	5160	140	62	4783	1740	3380	1107	1436	1291	548	631	431	472	288	215
Sedum acre	3803	95	68	3658	851	2746	1358	3027	673	392	578	333	453	314	273
Stachys sylvatica	6157	112	70	5903	1661	3977	2045	4052	1251	759	1079	632	822	592	488
Symphytum officinale	12343	384	140	11686	2319	8969	3261	6210	1766	1145	1573	969	1222	912	768
Trifolium pratense	13535	466	153	12661	2913	9505	3208	5313	2221	1285	1626	1098	1256	904	773
Veronica chamaedrys	5246	156	76	4913	1544	3237	1964	3327	1176	772	1075	655	823	602	515
Veronica hederifolia	3170	88	56	3006	1258	1957	1152	1940	946	568	797	482	587	426	367
Veronica serpyllifolia	3281	93	59	3081	1260	2250	1191	2073	978	500	783	420	604	391	331
Vicia hirsuta	4138	200	110	3775	1745	2471	1204	2037	1287	606	888	496	641	396	318
Vicia sativa	2996	110	65	2678	1342	1759	1161	1476	931	600	729	454	509	437	329
Chamerion angustifolium	3281	114	64	3030	1330	1890	1131	1741	997	536	696	441	516	363	301
Papaver dubium	1579	71	53	1400	748	971	465	628	584	252	320	216	257	149	126
Lotus pedunculatus	8911	272	71	8418	2010	5525	1999	3750	1602	751	860	650	698	465	409
Betula pendula	7539	89	57	7150	1805	5034	1173	1784	1338	465	550	368	422	188	149
Sorbus aucuparia	7986	114	85	7684	2235	5334	1836	3026	1775	766	925	647	740	434	358
Trifolium arvense	3126	89	67	2990	1088	2141	1038	2271	874	468	767	393	611	372	317
Cardamine pratensis	15497	482	144	14818	3285	10301	6598	11138	2574	2032	2450	1746	1912	1644	1420
Fallopia japonica	28640	102	63	28401	2384	23355	6621	20500	1679	1242	1190	840	847	713	513
Solanum dulcamara	7574	106	86	7336	1810	4663	2072	4564	1413	758	1134	648	897	603	522
Viola arvensis	2762	58	52	2619	1234	1837	1087	1602	945	544	774	443	586	406	326
Juncus effusus	11535	84	66	10985	2664	6171	1775	4348	2013	800	1019	690	773	444	384
Quercus robur	14304	156	92	13658	3248	9100	2956	5503	2524	1063	1386	881	1119	601	504
Coronopus didymus	1915	108	65	1659	976	1223	488	915	766	334	506	276	375	225	177
Mercurialis annua	2505	106	62	2292	1055	1654	819	1223	840	424	576	371	427	281	237
Ranunculus sceleratus	3338	105	73	3138	1277	2243	1103	2276	1000	579	843	493	659	445	378
Asplenium ruta-muraria	4979	100	63	4815	987	3462	1293	3443	831	368	621	330	525	293	270
Cirsium palustre	12684	286	78	12182	2102	8088	3213	7478	1688	911	1219	821	989	680	614
Euphorbia peplus	1516	80	50	1339	646	985	482	704	508	235	308	200	225	144	120
Filipendula ulmaria	15026	216	80	14544	2220	9840	4375	9022	1732	1054	1330	911	1040	776	667
Papaver rhoeas	3306	127	91	3056	1495	2128	1303	1808	1123	695	886	562	656	505	403
Draba verna	3321	106	50	3114	1276	2392	1330	1894	996	605	704	522	529	377	322
Stachys palustris	3234	81	53	3077	1137	1920	1191	2207	858	542	793	440	579	439	354
Linaria vulgaris	5745	205	129	5408	2120	3614	2245	3531	1614	985	1356	818	1017	762	627
Silene dioica	7153	163	75	6865	1882	4195	3051	5155	1382	1070	1398	882	1050	897	743
Veronica persica	3482	145	98	3233	1495	2064	1561	2277	1104	823	1073	648	769	669	522
Angelica sylvestris	9227	249	58	8673	1898	6056	2292	4277	1494	716	924	619	732	471	403
Ficaria verna	10374	357	107	9858	2683	6346	5212	7596	2020	1761	2076	1486	1561	1446	1222
Valeriana officinalis	6881	114	63	6660	1654	4206	2045	4311	1245	771	1088	657	815	580	493
Gnaphalium uliginosum	2649	65	51	2476	1208	1795	577	1177	937	352	566	290	425	236	187
Senecio inaequidens	6026	182	99	5697	1693	4315	1750	4146	1327	727	983	625	777	525	443

		Model Testing Data		**Model Training Data**											
				UNFILTERED		**FILTERED**									
						Records			*Presences (1x1 km grid cells)*						
TAXONOMIC GROUP **species**	**All observations**	*Testing records*	*Testing presences*	*Training records*	*Training presences*	ACTIVITY (A)	DETAIL (D)	VALSTAT (V)	A	D	V	A + D	A + V	D + V	A + D + V
Polygonatum multiflorum	9648	124	54	9433	1647	6472	3327	6353	1331	814	1006	720	841	601	539
Cytisus scoparius	10091	230	125	9485	2350	6896	2917	5465	1966	1074	1418	962	1188	795	710
Prunus serotina	9113	90	64	8723	2228	4868	2527	3558	1732	790	870	662	689	417	349
Scrophularia nodosa	3483	84	57	3254	1334	2121	1023	1898	1024	506	760	414	586	405	326
Anemone nemorosa	22356	225	56	22050	1849	16014	7732	18040	1475	1252	1406	1076	1128	1030	895
Hieracium pilosella	7306	89	50	7137	1478	4672	1899	4479	1211	576	848	521	711	428	393
Silene flos-cuculi	9481	597	56	8777	1764	5714	3686	6781	1326	1013	1203	857	930	800	681
Gnaphalium luteoalbum	2652	72	57	2516	991	1873	756	1836	774	411	647	354	498	330	286
Tragopogon pratensis	2369	67	50	2248	857	1502	978	1460	646	416	519	346	369	307	250
Lonicera periclymenum	14500	104	80	14252	2310	10008	4020	8906	1906	1023	1312	926	1103	739	672
Calluna vulgaris	18793	231	76	18418	1725	11558	4447	12878	1528	799	1148	752	1025	632	600
Galium palustre	7480	325	69	7055	1634	4699	1692	4310	1310	638	947	567	777	482	428
Rumex acetosella	8336	100	65	8070	2207	5406	2137	3970	1803	781	1009	703	830	450	402
Ajuga reptans	7861	138	53	7535	1624	5077	2410	5334	1271	760	1139	660	902	644	561
Arum maculatum	10290	194	51	10019	1226	7267	4203	7170	1015	733	840	644	715	570	505
Luzula campestris	6754	116	75	6467	2097	4643	1750	3402	1703	750	1131	656	926	529	471
Lysimachia vulgaris	13534	179	90	13190	2249	8417	3026	8862	1795	990	1432	868	1141	755	657
Ornithopus perpusillus	4317	70	53	4204	1296	2903	1167	2795	1017	431	837	375	654	337	294
Silene latifolia	2373	70	50	2247	888	1434	1081	1631	608	471	611	349	393	387	277

Table C.2. Summary of the predictor set.

<table>
<tr><th><b><i>Predictor</i></b></th><th><b><i>minimum</i></b></th><th><b><i>maximum</i></b></th><th><b><i>mean</i></b></th><th><b><i>sd</i></b></th><th><b><i>q25</i></b></th><th><b><i>median</i></b></th><th><b><i>q75</i></b></th></tr>
<tr><td><b>Dunes (%)</b></td><td>0.0</td><td>93.0</td><td>0.1</td><td>2.0</td><td>0.0</td><td>0.0</td><td>0.0</td></tr>
<tr><td><b>Forest (%)</b></td><td>0.0</td><td>100.0</td><td>10.2</td><td>16.4</td><td>0.3</td><td>3.0</td><td>12.4</td></tr>
<tr><td><b>Grassland (%)</b></td><td>0.0</td><td>32.9</td><td>1.1</td><td>2.3</td><td>0.0</td><td>0.1</td><td>1.2</td></tr>
<tr><td><b>Green (%)</b></td><td>0.0</td><td>37.2</td><td>1.9</td><td>2.1</td><td>0.6</td><td>1.3</td><td>2.6</td></tr>
<tr><td><b>Heathland (%)</b></td><td>0.0</td><td>100.0</td><td>0.6</td><td>5.2</td><td>0.0</td><td>0.0</td><td>0.0</td></tr>
<tr><td><b>Saltmarshes (%)</b></td><td>0.0</td><td>58.0</td><td>0.1</td><td>1.5</td><td>0.0</td><td>0.0</td><td>0.0</td></tr>
<tr><td><b>Scrubs & Hassle (%)</b></td><td>0.0</td><td>72.3</td><td>1.0</td><td>2.6</td><td>0.0</td><td>0.1</td><td>0.9</td></tr>
<tr><td><b>Urban (%)</b></td><td>0.0</td><td>100.0</td><td>30.3</td><td>23.3</td><td>12.2</td><td>23.7</td><td>43.0</td></tr>
<tr><td><b>Water (%)</b></td><td>0.0</td><td>98.6</td><td>2.3</td><td>6.4</td><td>0.1</td><td>0.4</td><td>1.5</td></tr>
<tr><td><b>Wetlands (%)</b></td><td>0.0</td><td>31.1</td><td>0.1</td><td>0.7</td><td>0.0</td><td>0.0</td><td>0.0</td></tr>
<tr><td><b>Mean annual precipitation (mm)</b></td><td>65.9</td><td>90.8</td><td>76.7</td><td>4.4</td><td>73.9</td><td>77.9</td><td>79.8</td></tr>
<tr><td><b>Mean annual temperature (°C)</b></td><td>9.4</td><td>10.3</td><td>10.0</td><td>0.2</td><td>9.8</td><td>10.1</td><td>10.2</td></tr>
<tr><td><b>Ecoregions</b></td><td colspan="2">1 = <i>Kustduinen</i>
2 = <i>Polders en getijdenschelde</i>
3 = <i>Pleistocene riviervalleien</i>
4 = <i>Cuesta's</i></td><td colspan="3">5 = <i>Kempen</i>
6 = <i>Westelijke interfluvia</i>
7 = <i>Midden-Vlaamse overgangsgebieden</i>
8 = <i>Zuidwestelijke heuvelzone</i></td><td colspan="2">9 = <i>Zuidoostelijke heuvelzone</i>
10 = <i>Krijt-leemgebieden</i>
11 = <i>Krijtgebieden</i>
12 = <i>Grindrivieren</i></td></tr>
<tr><td><b>Soil texture</b></td><td colspan="2">1 = Dunes
2 = Sand
3 = loamy sand
4 = sandy loam</td><td colspan="3">5 = light sandy loam
6 = loam
7 = clay
8 = heavy clay</td><td colspan="2">9 = peat
10 = chalk
11 = paved soils</td></tr>
</table>

Appendix D: The impact of data quality on Sensitivity and Specificity

The appendix shows the impact of data quality on Sensitivity (Figure D.1) and Specificity (Figure D.2) for all species and per taxonomic group, when absolute sample size is constant at six levels: 100, 250, 500, 1000, 2000 and 4000 presences. Per level, species were limited to those that could be modelled with all filters at the considered level, including the 3-filter combination ACTIVITY-DETAIL-VALSTAT. Species were subsequently classified at the highest level possible, meaning that model performance cannot be compared between sample size levels, because species are different. The number of species in each comparison is presented in the top left corner of the graphic areas. Not all levels could be assessed for all taxonomic groups, because for example for butterflies there were no species with less than 500 presences in our dataset, so all species were classified at level 500 or higher. Boxplots represent medians, upper and lower quartiles with whiskers extending to the minimum and maximum values. Asterisks show significant differences in model performance compared to the unfiltered data, tested by a multiple comparison test with Benjamini and Hochberg (1995) correction (*** $p<0.001$, ** $p<0.01$, * $p<0.05$). Colours indicate a positive (green) or negative (red) change in model performance.

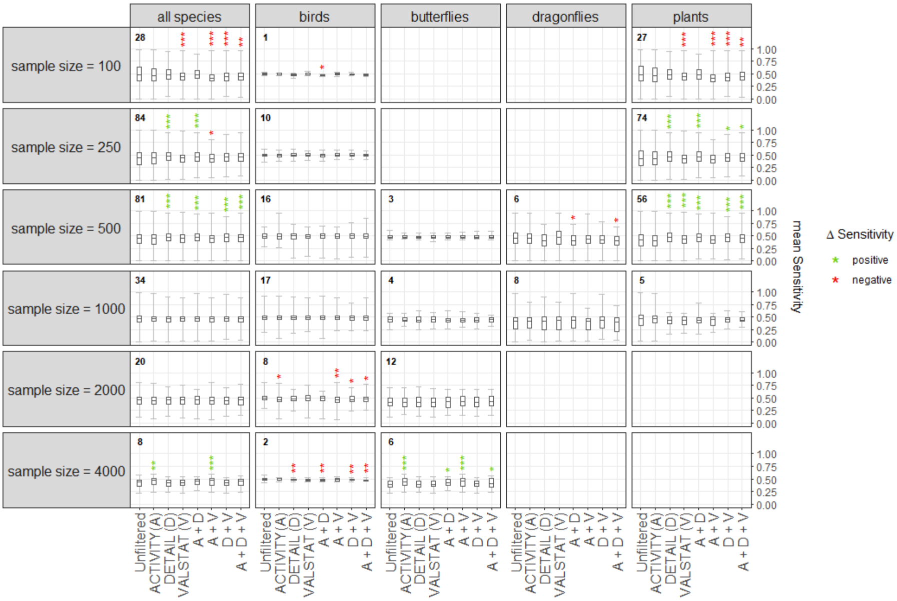


Figure D.1. *The impact of data quality on Sensitivity*

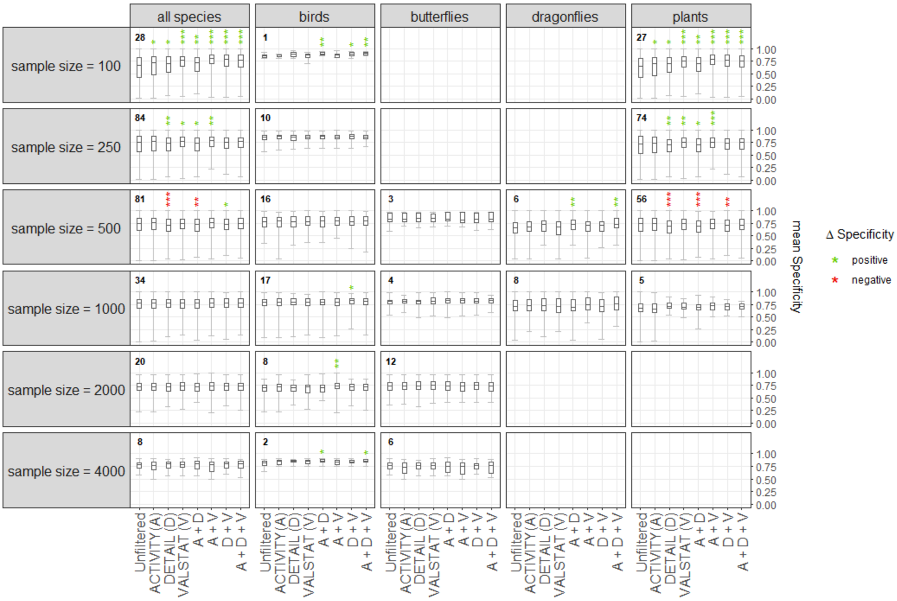


Figure D.2. *The impact of data quality on Specificity*

Appendix E: The impact of absolute sample size on Sensitivity and Specificity and the impact on AUC, Sensitivity and Specificity per taxonomic group

The appendix shows the impact of absolute sample size on Sensitivity (Figure E.1) and Specificity (Figure E.2) when data quality is constant for all species combined. It also shows the results for birds, butterflies, dragonflies and plants separately for AUC (Figures E.3-E.7), Sensitivity (Figures E.7-E.10) and Specificity (Figures E.11-E14). Per filter, species were grouped in one of the six specified intervals of sample size (left) that indicate the available sample sizes of the original training sets. Model performance was compared between models resulting from a repeated and random selection of different fixed sample sizes. Because species differ, results can only be compared within the graphic areas, i.e. between fixed sample sizes, but not between filters (horizontal) or intervals (vertical). The number of species in each comparison is presented in the top left corner of the graphic areas. Boxplots represent medians, upper and lower quartiles with whiskers extending to the minimum and maximum values. Asterisks show significant differences in model performance compared to the highest sample size, tested by a multiple comparison test with Benjamini & Hochberg (1995) correction (*** p<0.001, ** p<0.01, * p<0.05). Colours indicate a positive (green) or negative (red) change in model performance.

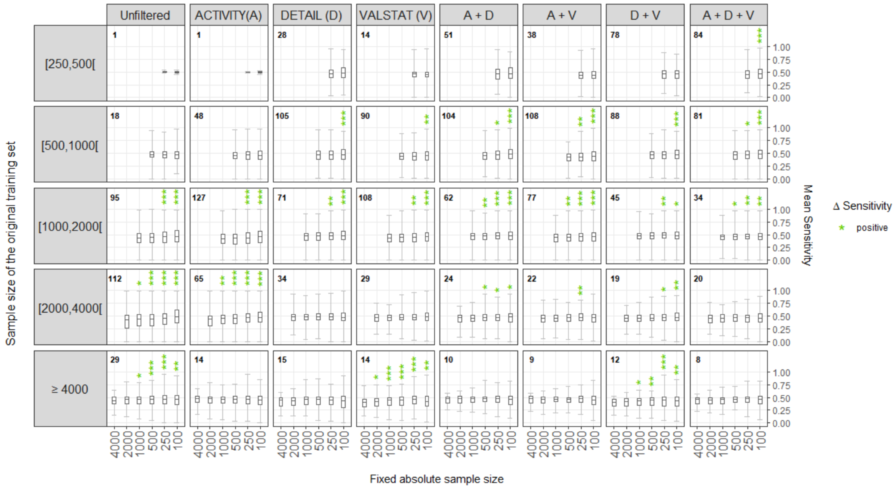


Figure E.1. *The impact of absolute sample size on Sensitivity*

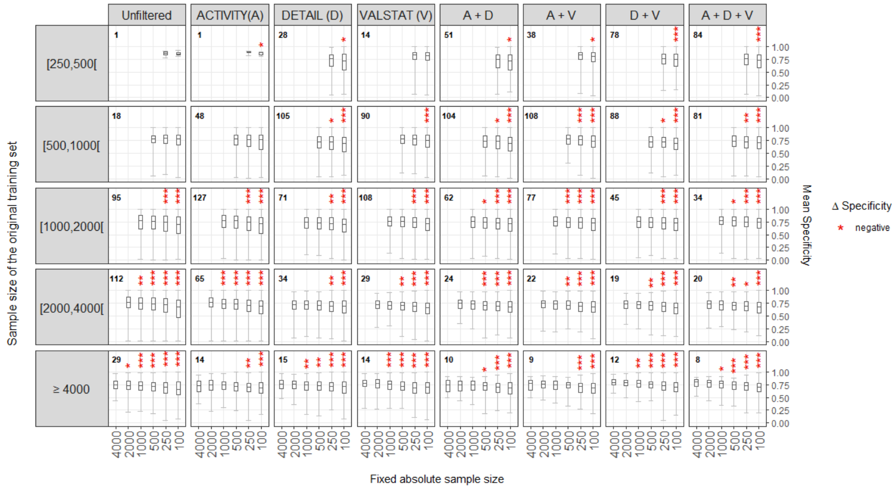


Figure E.2. *The impact of absolute sample size on Specificity*

Results per taxonomic group for AUC

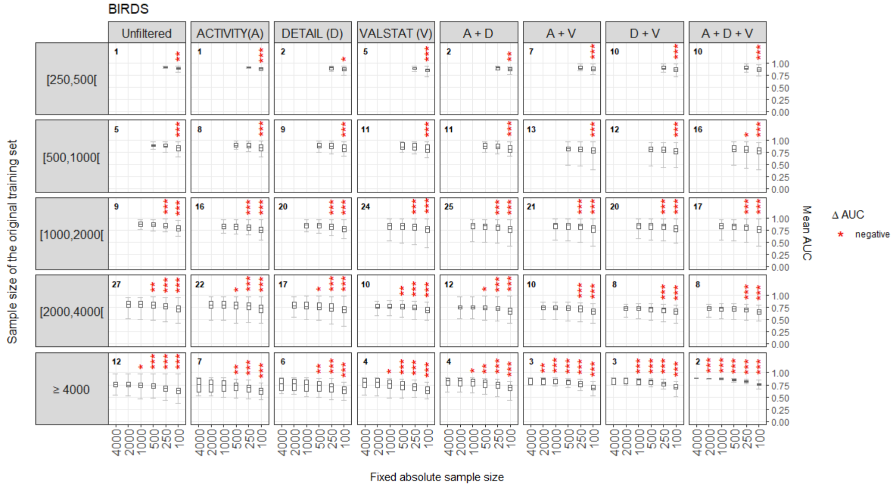


Figure E.3. *The impact of sample size on AUC when data quality is constant for birds.*

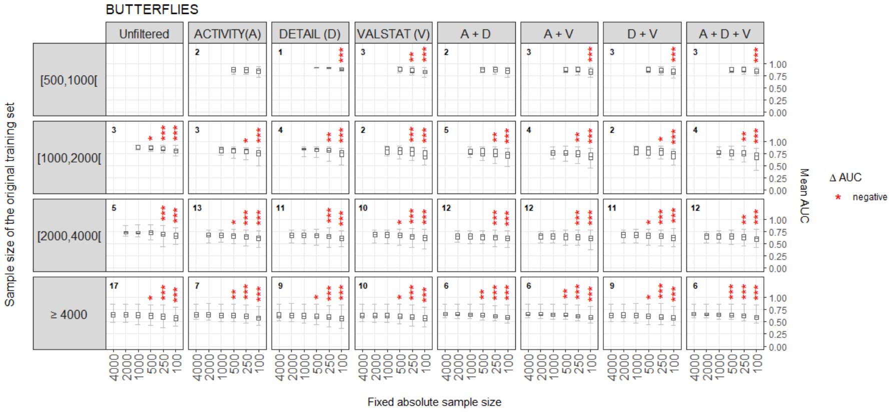


Figure E.4. The impact of sample size on AUC when data quality is constant for butterflies.

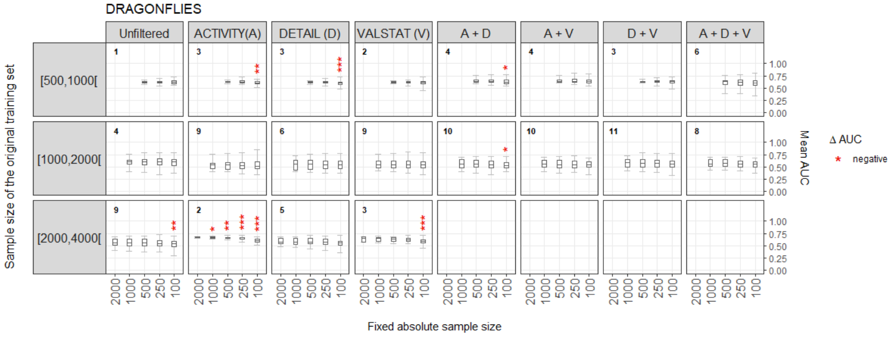


Figure E.5. The impact of sample size on AUC when data quality is constant for dragonflies.

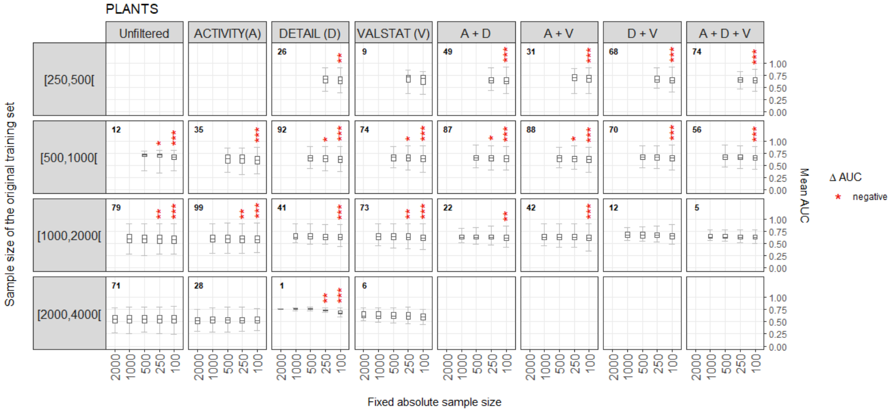


Figure E.6. The impact of sample size on AUC when data quality is constant for plants.

Results per taxonomic group for Sensitivity

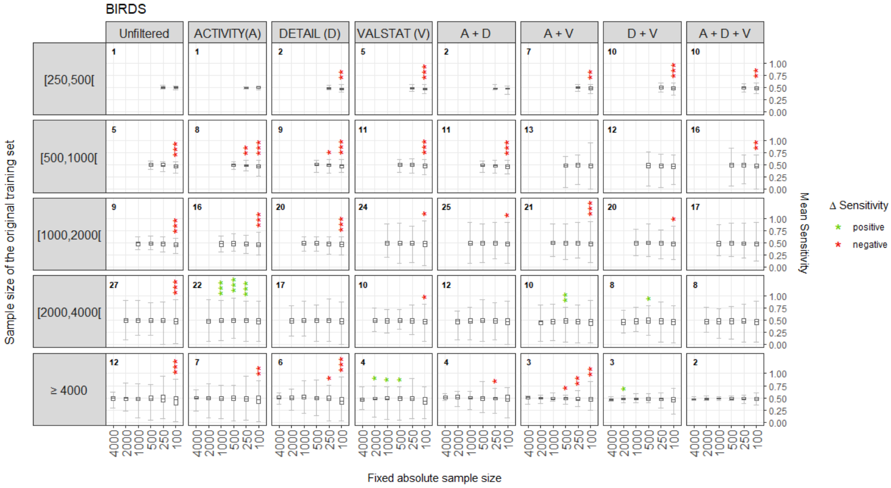


Figure E.7. *The impact of sample size on Sensitivity when data quality is constant for birds.*

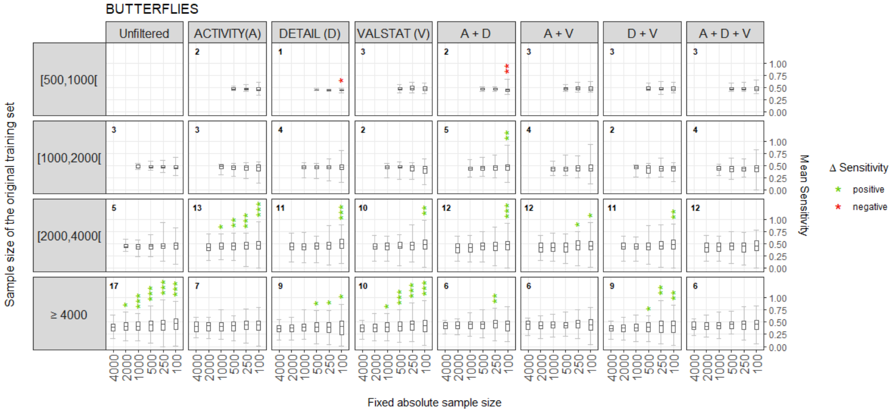


Figure E.8. *The impact of sample size on Sensitivity when data quality is constant for butterflies.*

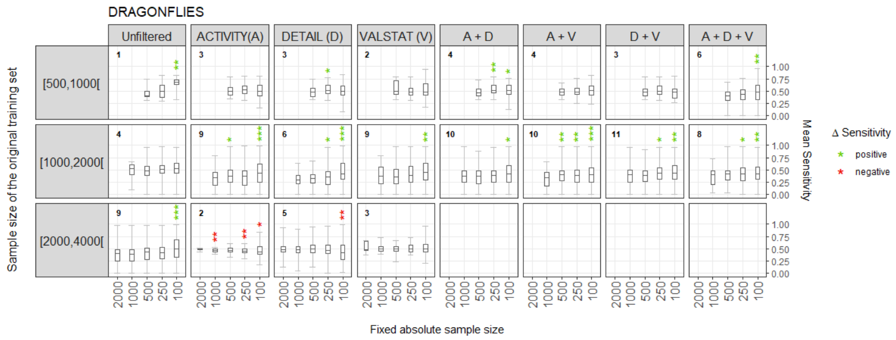


Figure E.9. *The impact of sample size on Sensitivity when data quality is constant for dragonflies.*

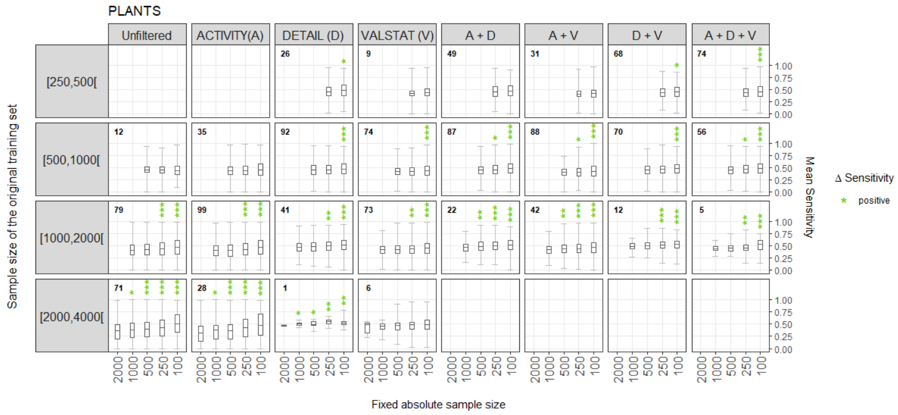


Figure E.10. *The impact of sample size on Sensitivity when data quality is constant for plants.*

Results per taxonomic group for Specificity

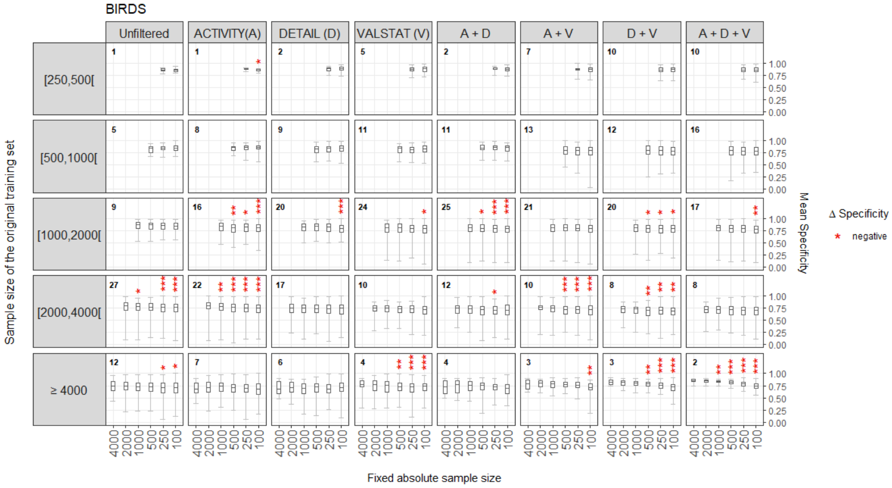


Figure E.11. *The impact of sample size on Specificity when data quality is constant for birds.*

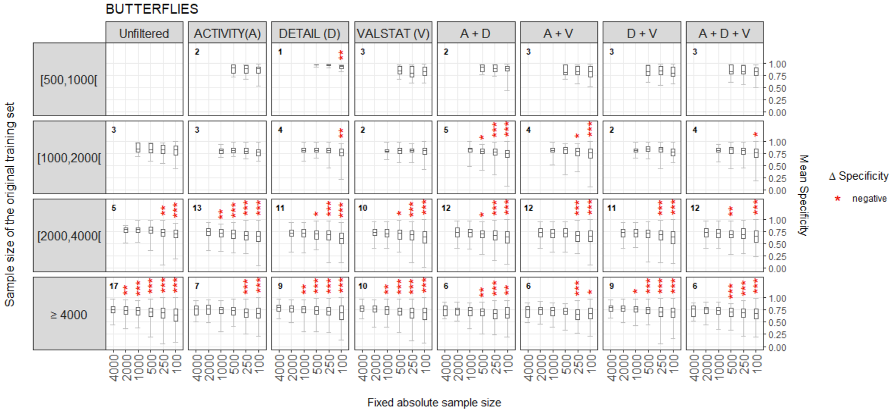


Figure E.12. *The impact of sample size on Specificity when data quality is constant for butterflies.*

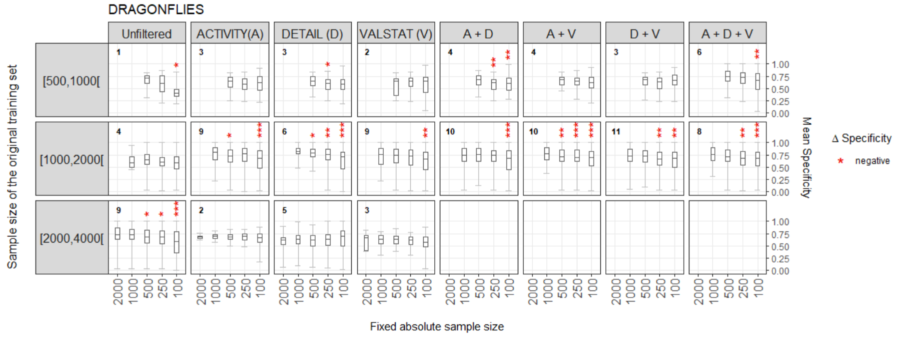


Figure E.13. The impact of sample size on Specificity when data quality is constant for dragonflies.

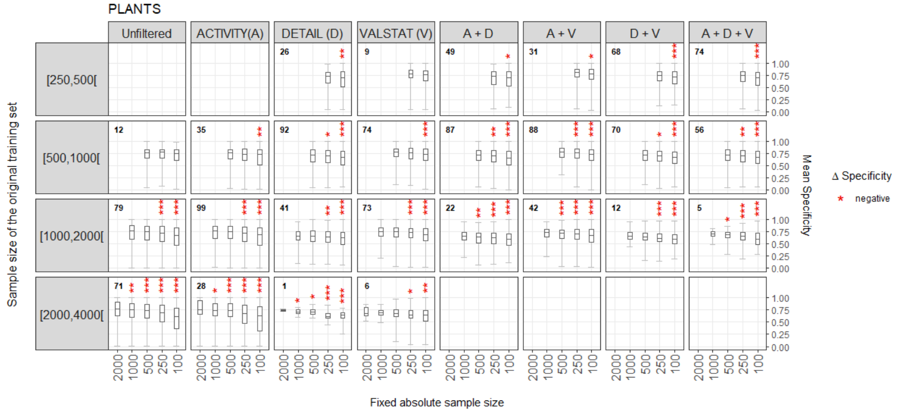


Figure E.14. The impact of sample size on Specificity when data quality is constant for plants.

Appendix F: GAMM model selection

Table F.1. Model selection results of the GAMMs (Generalized Additive Mixed Models) that model the impact of data quality and sample size on Δ AUC, Δ Sensitivity and Δ Specificity, per taxonomic group. Columns show the parameters in the model, the number of parameters (df), the log-likelihood (LogLik), the Akaike Information Criterion (AIC) and the difference in AIC (ΔAIC) compared to the top model, and the model weight (i.e. the relative likelihood of the model). We selected the model with the least parameters (bold) and a small difference in AIC (ΔAIC < 1) (highlighted in grey) compared to the top-ranked model.

Model parameters	*df*	*logLik*	*AIC*	*ΔAIC*	*weight*
BIRDS (ΔAUC ~)					
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	**91**	**23906.2**	**-47628.9**	**0**	**0.51**
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	92	23906.9	-47628.8	0.1	0.49
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + te(samplesize,reduction) + s(species)	85	23821.5	-47472.4	156.5	0
filter + s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	78	23800.9	-47445.9	183	0
s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	72	23350.7	-46556.5	1072.3	0
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + s(species)	82	23144.9	-46125	1503.8	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + s(species)	82	23145.1	-46124.7	1504.1	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(species)	75	23089.9	-46028.3	1600.6	0
filter + s(reduction) + s(samplesize) + s(species)	68	23071.7	-46007.2	1621.7	0
filter + s(reduction) + s(reduction, by = filter) + s(species)	69	22830.5	-45521.2	2107.6	0
filter + s(reduction) + s(species)	64	22818.2	-45508.3	2120.6	0
s(reduction) + s(samplesize) + s(species)	61	22681.7	-45239.4	2389.5	0
s(reduction) + s(species)	57	22444.1	-44772.3	2856.6	0
filter + s(samplesize) + s(samplesize, by = filter) + s(species)	75	20684.6	-41217.7	6411.2	0
filter + s(samplesize) + s(species)	63	20641.8	-41156.1	6472.8	0
s(samplesize) + s(species)	57	20447.9	-40780.5	6848.4	0
s(species)	53	19576.1	-39046.1	8582.7	0
BUTTERFLIES (ΔAUC ~)					
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	**52**	**14919.2**	**-29734**	**0**	**0.495**
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	52	14919.2	-29733.9	0.1	0.473
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + te(samplesize,reduction) + s(species)	49	14912.9	-29727.5	6.5	0.019
filter + s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	47	14911.3	-29726.7	7.3	0.013
s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	42	14669.4	-29254.6	479.4	0
filter + s(reduction) + s(samplesize) + s(species)	39	14308.4	-28537.3	1196.7	0

Model parameters	*df*	*logLik*	*AIC*	*ΔAIC*	*weight*
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(species)	40	14308.5	-28536.8	1197.2	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + s(species)	41	14309.9	-28536.4	1197.6	0
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + s(species)	41	14310	-28536.3	1197.7	0
s(reduction) + s(samplesize) + s(species)	33	14109.4	-28151.4	1582.6	0
filter + s(reduction) + s(species)	35	14074.8	-28078.2	1655.8	0
filter + s(reduction) + s(reduction, by = filter) + s(species)	36	14075.1	-28077.3	1656.7	0
s(reduction) + s(species)	29	13897.5	-27735.7	1998.3	0
filter + s(samplesize) + s(species)	35	13238.7	-26406.1	3327.9	0
filter + s(samplesize) + s(samplesize, by = filter) + s(species)	38	13240.9	-26405	3329	0
s(samplesize) + s(species)	29	13121.9	-26184.5	3549.5	0
s(species)	25	12429.5	-24808.2	4925.8	0
DRAGONFLIES (ΔAUC ~)					
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	29	5865.3	-11672.6	0	0.264
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	29	5865.3	-11672.6	0	0.264
filter + s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	**27**	**5863.8**	**-11672.3**	**0.2**	**0.236**
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + te(samplesize,reduction) + s(species)	27	5863.8	-11672.3	0.2	0.235
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + s(species)	28	5855	-11653.5	19	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(species)	27	5853.3	-11652.1	20.4	0
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + s(species)	26	5850.3	-11646.6	26	0
filter + s(reduction) + s(samplesize) + s(species)	26	5848.7	-11645.3	27.2	0
filter + s(reduction) + s(species)	23	5839.4	-11630.9	41.7	0
filter + s(reduction) + s(reduction, by = filter) + s(species)	23	5839.4	-11630.8	41.7	0
s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	21	5836.5	-11629.6	42.9	0
s(reduction) + s(samplesize) + s(species)	20	5822.2	-11604.4	68.2	0
s(reduction) + s(species)	17	5814.4	-11593.2	79.4	0
filter + s(samplesize) + s(samplesize, by = filter) + s(species)	25	5807	-11562.6	109.9	0
filter + s(samplesize) + s(species)	24	5805.7	-11561.8	110.8	0
s(samplesize) + s(species)	18	5779.5	-11521.6	151	0
s(species)	14	5716.1	-11402.4	270.1	0
PLANTS (ΔAUC ~)					
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	**196**	**49308.3**	**-98222.9**	**0**	**1**
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	192	49273.2	-98161.5	61.4	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + s(species)	189	49260.8	-98141.8	81.1	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + te(samplesize,reduction) + s(species)	188	49238	-98098.6	124.4	0
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + s(species)	186	49231.5	-98091	131.9	0

Model parameters	*df*	*logLik*	*AIC*	*ΔAIC*	*weight*
filter + s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	183	49228.1	-98089.1	133.8	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(species)	181	49195.2	-98028.4	194.6	0
filter + s(reduction) + s(samplesize) + s(species)	176	49186.1	-98020.2	202.8	0
filter + s(samplesize) + s(samplesize, by = filter) + s(species)	182	49057.9	-97750	472.9	0
filter + s(reduction) + s(reduction, by = filter) + s(species)	178	49049.3	-97741.9	481.1	0
filter + s(reduction) + s(species)	172	49036.1	-97727	495.9	0
filter + s(samplesize) + s(species)	172	49015.1	-97685	537.9	0
s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	181	44300.1	-88236.3	9986.6	0
s(reduction) + s(samplesize) + s(species)	170	44207.8	-88074.9	10148	0
s(samplesize) + s(species)	166	44121.2	-87910.1	10312.9	0
s(reduction) + s(species)	166	43923.2	-87514.3	10708.6	0
s(species)	162	43788.3	-87252.1	10970.8	0
BIRDS (ΔSensitivity ~)					
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	**80**	**16309.1**	**-32457.9**	**0**	**0.513**
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	80	16309	-32457.8	0.1	0.487
s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	68	16286.4	-32435.1	22.8	0
filter + s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	74	16289.5	-32429.2	28.7	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + te(samplesize,reduction) + s(species)	75	16289.6	-32428.8	29.1	0
filter + s(samplesize) + s(samplesize, by = filter) + s(species)	71	16224.1	-32305.4	152.5	0
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + s(species)	72	16224.7	-32305.2	152.7	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + s(species)	72	16224.7	-32305.2	152.7	0
s(samplesize) + s(species)	57	16201.4	-32287	170.9	0
s(reduction) + s(samplesize) + s(species)	59	16202.2	-32285.3	172.6	0
filter + s(samplesize) + s(species)	63	16204.1	-32280.5	177.4	0
filter + s(reduction) + s(samplesize) + s(species)	65	16204.9	-32278.7	179.2	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(species)	66	16205.3	-32278.2	179.7	0
s(reduction) + s(species)	56	16167	-32221.8	236.1	0
filter + s(reduction) + s(species)	62	16169.7	-32215.3	242.6	0
filter + s(reduction) + s(reduction, by = filter) + s(species)	63	16170.1	-32213.8	244.1	0
s(species)	53	16157.1	-32206.3	251.6	0
BUTTERFLIES (ΔSensitivity ~)					
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + te(samplesize,reduction) + s(species)	42	9903.1	-19722	0	0.396
filter + s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	**42**	**9903.1**	**-19721.4**	**0.6**	**0.3**
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	45	9906.1	-19721.3	0.6	0.289
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	42	9900	-19714.9	7	0.012

Model parameters	*df*	*logLik*	*AIC*	*ΔAIC*	*weight*
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + s(species)	39	9895	-19710.6	11.4	0.001
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + s(species)	39	9894.9	-19710.6	11.4	0.001
filter + s(reduction) + s(samplesize) + s(species)	37	9891.1	-19707.1	14.9	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(species)	37	9891.1	-19706.9	15	0
filter + s(reduction) + s(species)	35	9888.9	-19706.8	15.1	0
filter + s(reduction) + s(reduction, by = filter) + s(species)	35	9888.9	-19705.9	16.1	0
s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	36	9889	-19704.8	17.2	0
s(reduction) + s(samplesize) + s(species)	31	9876.3	-19689.3	32.7	0
s(reduction) + s(species)	29	9873.9	-19688.9	33.1	0
filter + s(samplesize) + s(samplesize, by = filter) + s(species)	39	9753.4	-19427.6	294.3	0
filter + s(samplesize) + s(species)	34	9743.9	-19418.9	303.1	0
s(samplesize) + s(species)	28	9729	-19401.1	320.8	0
s(species)	25	9709.1	-19367	355	0
DRAGONFLIES (ΔSensitivity ~)					
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	31	2790.8	-5519.1	0	0.394
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	**31**	**2790.8**	**-5519.1**	**0**	**0.394**
filter + s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	30	2788.3	-5516.4	2.6	0.105
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + te(samplesize,reduction) + s(species)	30	2788.3	-5516.4	2.6	0.105
s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	23	2777.6	-5507.2	11.9	0.001
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + s(species)	26	2753.7	-5454.5	64.5	0
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + s(species)	26	2753.7	-5454.5	64.5	0
filter + s(reduction) + s(samplesize) + s(species)	25	2751.5	-5452.9	66.2	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(species)	25	2751.5	-5452.9	66.2	0
s(reduction) + s(samplesize) + s(species)	19	2740.1	-5441.8	77.2	0
filter + s(samplesize) + s(samplesize, by = filter) + s(species)	24	2742.9	-5436.2	82.8	0
filter + s(samplesize) + s(species)	23	2740.9	-5434.6	84.5	0
filter + s(reduction) + s(species)	23	2737.7	-5428.6	90.4	0
filter + s(reduction) + s(reduction, by = filter) + s(species)	23	2737.7	-5428.6	90.4	0
s(samplesize) + s(species)	17	2729.8	-5424.2	94.9	0
s(reduction) + s(species)	17	2725.6	-5416.5	102.6	0
s(species)	14	2689.8	-5349.8	169.3	0
PLANTS (ΔSensitivity ~)					
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	**185**	**26346.9**	**-52321.8**	**0**	**0.636**
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	187	26348	-52320.7	1.1	0.364
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + te(samplesize,reduction) + s(species)	189	26305.4	-52232.1	89.7	0
filter + s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	184	26295.9	-52222.5	99.3	0

Model parameters	*df*	*logLik*	*AIC*	*ΔAIC*	*weight*
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + s(species)	183	26290.2	-52214.3	107.6	0
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + s(species)	181	26288	-52212.5	109.3	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(species)	181	26248	-52133.2	188.6	0
filter + s(reduction) + s(samplesize) + s(species)	175	26236.4	-52121.9	200	0
filter + s(reduction) + s(reduction, by = filter) + s(species)	178	26199.7	-52041.8	280	0
filter + s(reduction) + s(species)	171	26172.3	-52001.8	320	0
s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	178	26048.7	-51740.7	581.1	0
s(reduction) + s(samplesize) + s(species)	169	25994.6	-51650.5	671.3	0
s(reduction) + s(species)	165	25918.7	-51506.9	814.9	0
filter + s(samplesize) + s(samplesize, by = filter) + s(species)	175	25908.6	-51466.7	855.1	0
filter + s(samplesize) + s(species)	171	25855	-51366.8	955.1	0
s(samplesize) + s(species)	165	25619.6	-50908.2	1413.6	0
s(species)	161	25508.8	-50694.5	1627.3	0
BIRDS (ΔSpecificity ~)					
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	**80**	**15103.7**	**-30046.6**	**0**	**0.505**
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	80	15103.7	-30046.5	0	0.495
filter + s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	73	15087.6	-30028.5	18.1	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + te(samplesize,reduction) + s(species)	75	15087.8	-30025.1	21.5	0
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + s(species)	73	15073.9	-30001.1	45.5	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + s(species)	73	15073.9	-30001	45.5	0
filter + s(reduction) + s(samplesize) + s(species)	67	15059.8	-29985.4	61.2	0
s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	67	15060.1	-29984.9	61.7	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(species)	68	15060.1	-29982.9	63.6	0
filter + s(reduction) + s(species)	63	15044.6	-29961.7	84.9	0
filter + s(reduction) + s(reduction, by = filter) + s(species)	65	15044.9	-29959.5	87.1	0
s(reduction) + s(samplesize) + s(species)	61	15033.4	-29944.8	101.8	0
s(reduction) + s(species)	57	15017.6	-29919.7	126.9	0
filter + s(samplesize) + s(samplesize, by = filter) + s(species)	69	14924.2	-29710.1	336.5	0
filter + s(samplesize) + s(species)	61	14908.4	-29692.9	353.7	0
s(samplesize) + s(species)	55	14883.4	-29655	391.6	0
s(species)	54	14875.6	-29643	403.6	0
BUTTERFLIES (ΔSpecificity ~)					
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	**47**	**9315**	**-18534.9**	**0**	**0.404**
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	47	9315	-18534.9	0	0.401
filter + s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	44	9309.9	-18531.7	3.2	0.081
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + te(samplesize,reduction) + s(species)	44	9310.1	-18531.3	3.5	0.069

Model parameters	*df*	*logLik*	*AIC*	*ΔAIC*	*weight*
s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	38	9303.3	-18530.5	4.4	0.044
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + s(species)	39	9297.1	-18515	19.9	0
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + s(species)	39	9297.1	-18515	19.9	0
filter + s(reduction) + s(species)	35	9292.1	-18513.2	21.7	0
filter + s(reduction) + s(samplesize) + s(species)	35	9292.1	-18512.9	22	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(species)	36	9292.3	-18512	22.8	0
filter + s(reduction) + s(reduction, by = filter) + s(species)	36	9292.3	-18512	22.9	0
s(reduction) + s(species)	29	9285.5	-18511.9	22.9	0
s(reduction) + s(samplesize) + s(species)	29	9285.5	-18511.7	23.2	0
filter + s(samplesize) + s(samplesize, by = filter) + s(species)	40	8887	-17693.1	841.8	0
s(samplesize) + s(species)	29	8875.1	-17691	843.9	0
filter + s(samplesize) + s(species)	35	8880.3	-17689.4	845.5	0
s(species)	25	8713.1	-17375.3	1159.6	0
DRAGONFLIES (ΔSpecificity ~)					
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	31	2660.7	-5258.4	0	0.394
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	**31**	**2660.7**	**-5258.4**	**0**	**0.392**
filter + s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	30	2658	-5255.8	2.6	0.107
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + te(samplesize,reduction) + s(species)	30	2658	-5255.8	2.6	0.107
s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	24	2646.3	-5244.5	13.9	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + s(species)	26	2620.4	-5187.6	70.8	0
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + s(species)	26	2620.4	-5187.6	70.8	0
filter + s(reduction) + s(samplesize) + s(species)	25	2618.2	-5185.4	73	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(species)	25	2618.2	-5185.4	73	0
s(reduction) + s(samplesize) + s(species)	19	2605.7	-5172.2	86.2	0
filter + s(reduction) + s(species)	23	2601.4	-5156	102.4	0
filter + s(reduction) + s(reduction, by = filter) + s(species)	23	2601.4	-5156	102.4	0
filter + s(samplesize) + s(samplesize, by = filter) + s(species)	24	2602.2	-5154.7	103.7	0
filter + s(samplesize) + s(species)	23	2600.1	-5152.9	105.5	0
s(reduction) + s(species)	17	2588.2	-5141.5	116.9	0
s(samplesize) + s(species)	17	2588	-5140.5	117.9	0
s(species)	14	2532.6	-5035.5	222.9	0
PLANTS (ΔSpecificity ~)					
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	186	25170.8	-49968.1	0	0.534
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + te(samplesize,reduction) + s(species)	**186**	**25170.7**	**-49967.8**	**0.3**	**0.466**
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + te(samplesize,reduction) + s(species)	187	25143.7	-49912.8	55.3	0
filter + s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	184	25137.3	-49905.1	63	0

Model parameters	*df*	*logLik*	*AIC*	Δ*AIC*	*weight*
filter + s(reduction) + s(samplesize) + s(samplesize, by = filter) + s(species)	180	25091.9	-49822	146.1	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(samplesize, by = filter) + s(species)	181	25092.4	-49821.8	146.3	0
filter + s(reduction) + s(samplesize) + s(reduction, by = filter) + s(species)	180	25064	-49767.7	200.3	0
filter + s(reduction) + s(samplesize) + s(species)	175	25055.7	-49760.9	207.2	0
filter + s(reduction) + s(reduction, by = filter) + s(species)	177	25042.8	-49731.3	236.8	0
filter + s(reduction) + s(species)	171	25022.6	-49702.7	265.3	0
s(reduction) + s(samplesize) + te(samplesize,reduction) + s(species)	177	24779.3	-49203.8	764.3	0
s(reduction) + s(samplesize) + s(species)	169	24697.4	-49056.4	911.7	0
s(reduction) + s(species)	165	24675	-49019.6	948.5	0
filter + s(samplesize) + s(samplesize, by = filter) + s(species)	175	24600.8	-48850.7	1117.4	0
filter + s(samplesize) + s(species)	171	24563.1	-48783.9	1184.2	0
s(samplesize) + s(species)	165	24229.2	-48128.3	1839.7	0
s(species)	161	24110.7	-47898.9	2069.1	0

Appendix G: The relative variable importance in the best GAMM model per taxonomic group for Sensitivity and Specificity

The appendix shows the relative variable importance, based on the proportion of the percentage of deviance explained (%DE) by the different explanatory variables in the best GAMM (Generalized Additive Mixed Model) per taxonomic group (orange dots), and the relative variable importance across species, in the GAMs (Generalized Additive Models) where the random species effect was excluded (boxplots) for Sensitivity (Figure G.1) and Specificity (Figure G.2). The proportional %DE is the decrease in %DE between the full model and the model where the variable was excluded (but with identical smoothing parameters), relative to the %DE of the full model to summarize effects across n species. Species for which the full model could not be estimated due to convergence issues were excluded from the summary.

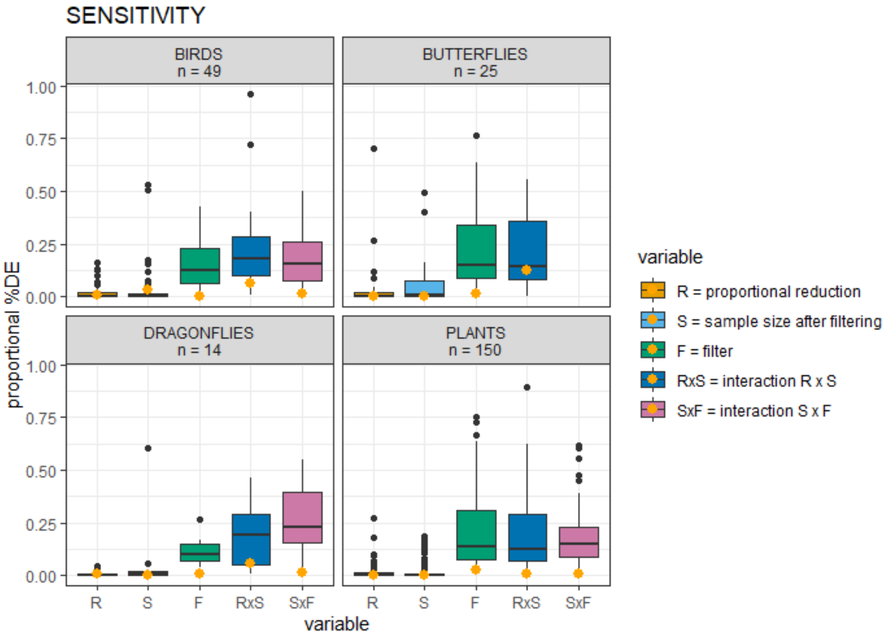


Figure G.1. *The relative variable importance in the best GAMM model per taxonomic group for Sensitivity.*

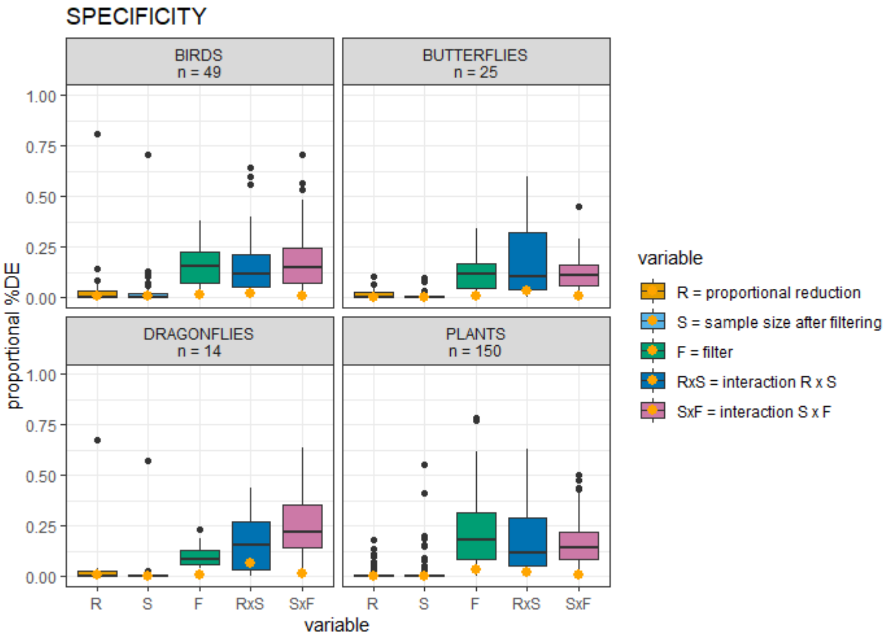


Figure G.2. *The relative variable importance in the best GAMM model per taxonomic group for Specificity*

Appendix H: The combined impact of data quality and sample size on model performance.

The appendix shows the combined impact of data quality and sample size on Δ Sensitivity (Figure H.1) and Δ Specificity (Figure H.2) per taxonomic group. The full lines are the predictions for the difference in model performance (Δ = filtered data - unfiltered data) from the 'best' GAMM (Generalized Additive Mixed Model) along a continuous scale of proportional reduction in sample size and for three sample sizes after filtering that we chose based on data availability: 100, 500 and 1000 presences. Colours represent the different filters (data quality). The red dotted line equals a zero difference, i.e. filtering did not impact model performance. We used the REML method (restricted maximum likelihood) in the 'gam' function of the ´mgcv` R package v 1.8-31 (Wood, 2017) to model our data. Filter type was modelled as factor variable and species as random effect. Smoothing functions were used to fit both sample size variables (proportional reduction and sample size after filtering), with cubic spline method and k = 5. Δ Sensitivity and Specificity were rescaled to fall between 0 and 1, so that we could use the 'betareg' family with logit-link, because of the double-bounded character of the response variable.

Figures H.3 to H.6 present the Maxent model predictions (i.e. *cloglog* transformed raw output) based on the unfiltered data and three situations of reduced sample size when using the best filter (i.e. the filter that caused the largest positive difference in AUC). The actual reduction in sample size refers to the situation where the filter was applied to the data as extracted from the *waarnemingen.be* database. The resulting sample size of 100 presences was the maximum reduction analyzed. The third situation is a reduction to a sample size in between the other two situations. We selected one species per taxonomic group, i.e. the species where the highest positive change in AUC was observed).

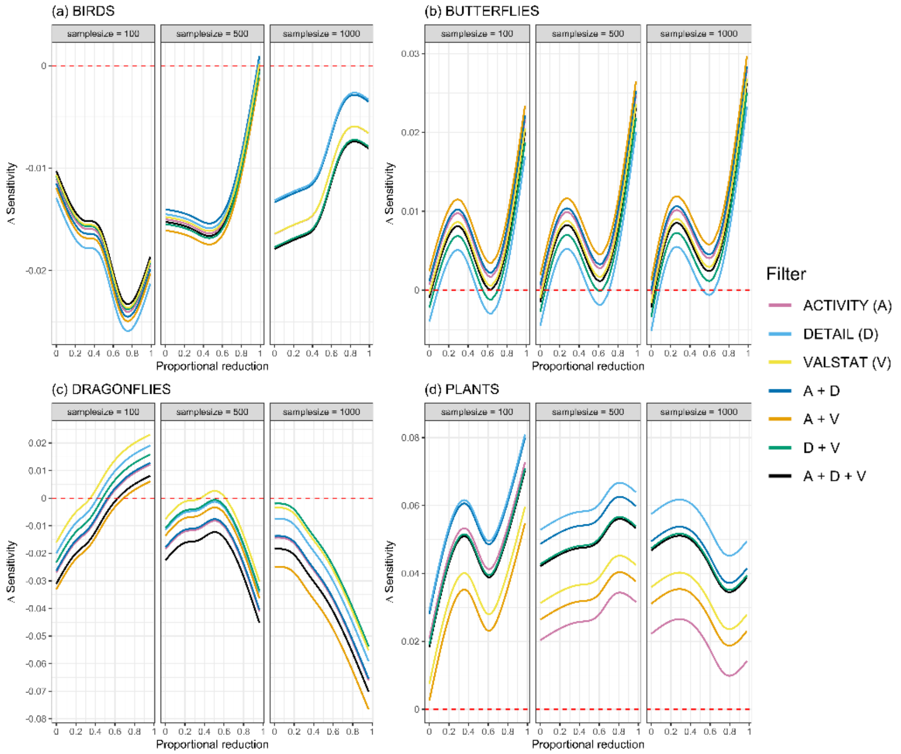


Figure H.1. *The combined impact of data quality and sample size on Δ Sensitivity per taxonomic group.*

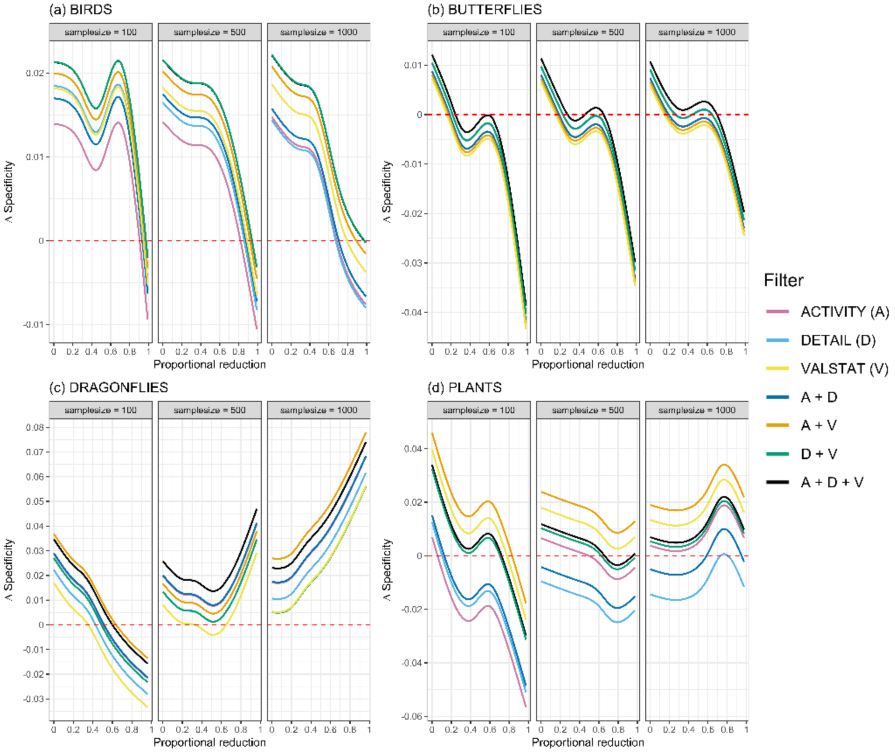


Figure H.2. *The combined impact of data quality and sample size on Δ Specificity per taxonomic group.*

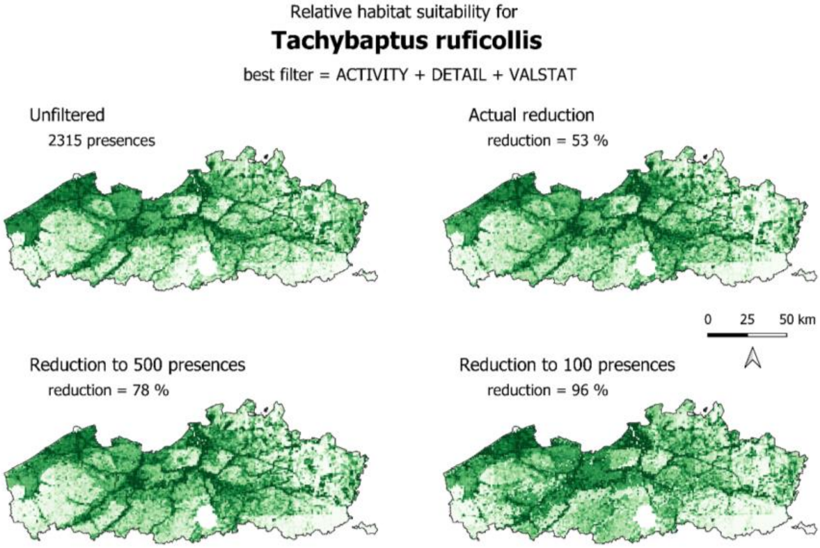

Relative habitat suitability for
Tachybaptus ruficollis
best filter = ACTIVITY + DETAIL + VALSTAT
Unfiltered
2315 presences
Actual reduction
reduction = 53 %
0
25
50 km
Reduction to 500 presences
reduction = 78 %
Reduction to 100 presences
reduction = 96 %

Figure H.3: *Example of the best filter and impact of sample size for birds: Relative habitat suitability (darker colour = more suitable) for Tachybaptus ruficollis.*

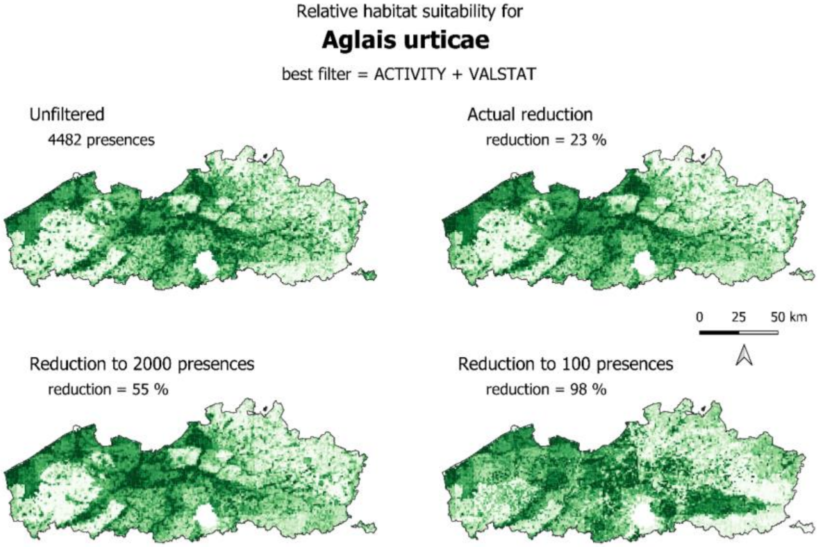

Relative habitat suitability for
Aglais urticae
best filter = ACTIVITY + VALSTAT
Unfiltered
4482 presences
Actual reduction
reduction = 23 %
0
25
50 km
Reduction to 2000 presences
reduction = 55 %
Reduction to 100 presences
reduction = 98 %

Figure H.4*: Example of the best filter and impact of sample size for butterflies: Relative habitat suitability (darker colour = more suitable) for Aglais urticae.*

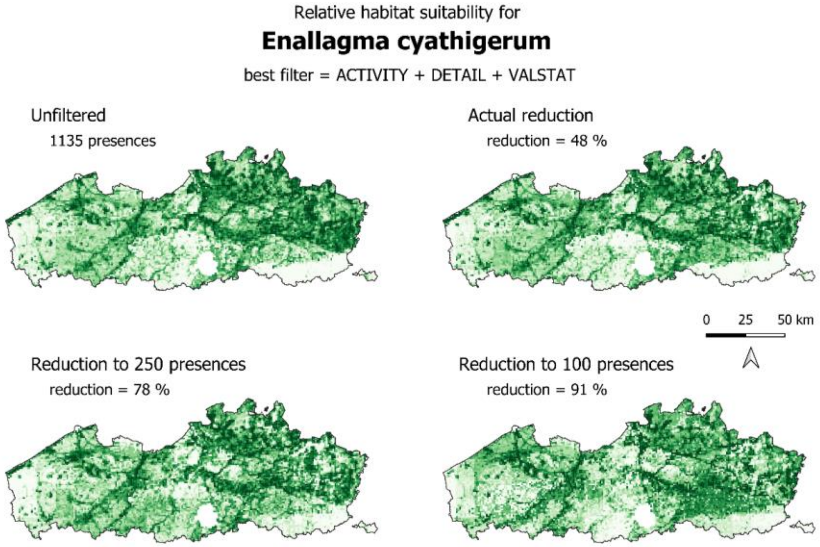

Relative habitat suitability for
Enallagma cyathigerum
best filter = ACTIVITY + DETAIL + VALSTAT
Unfiltered
1135 presences
Actual reduction
reduction = 48 %
0 25 50 km
Reduction to 250 presences
reduction = 78 %
Reduction to 100 presences
reduction = 91 %

Figure H.5*: Example of the best filter and impact of sample size for dragonflies: Relative habitat suitability (darker colour = more suitable) for Enallagma cyathigerum.*

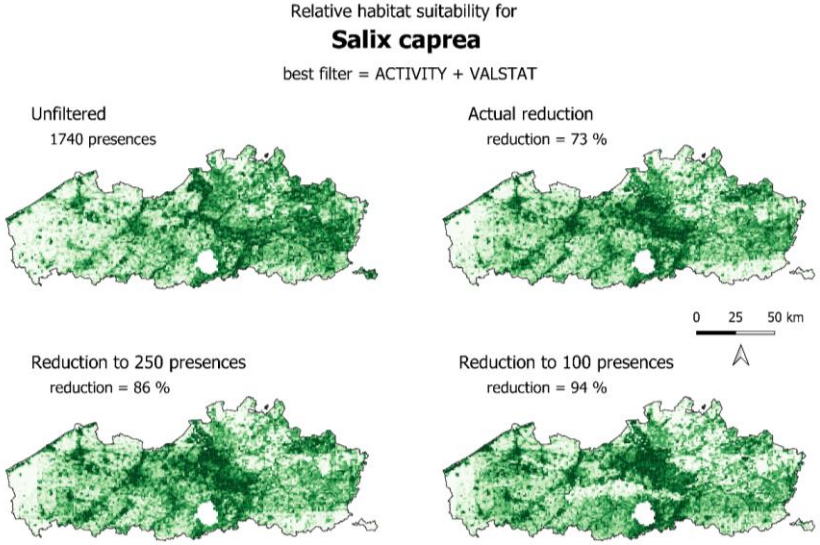
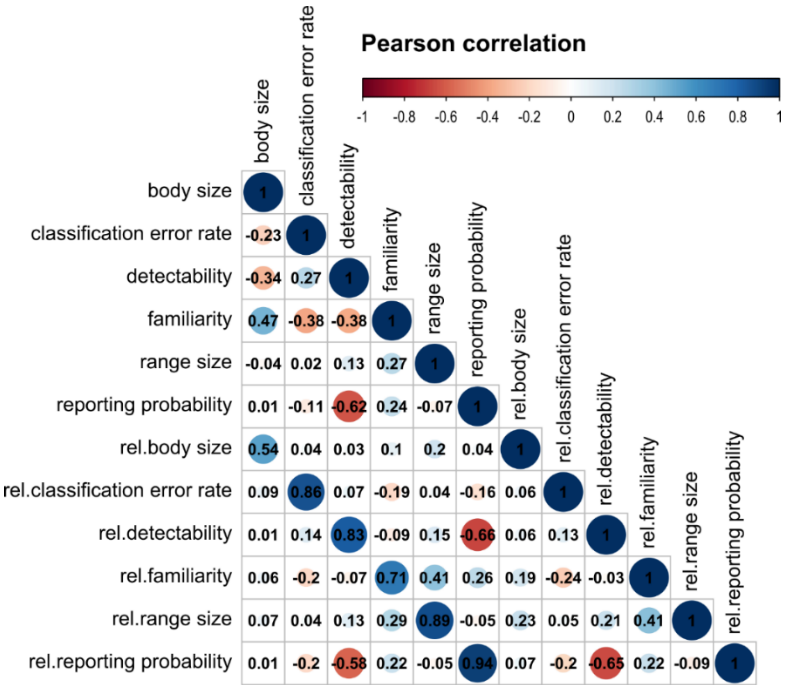

Relative habitat suitability for
Salix caprea
best filter = ACTIVITY + VALSTAT
Unfiltered
1740 presences
Actual reduction
reduction = 73 %
0 25 50 km
Reduction to 250 presences
reduction = 86 %
Reduction to 100 presences
reduction = 94 %

Figure H.6*: Example of the best filter and impact of sample size for plants: Relative habitat suitability (darker colour = more suitable) for Salix caprea.*

CHAPTER III

Appendix I: Data summary

Figure I.1. Pearson correlation matrix of absolute and relative (rel.) species traits.



	body size	classification error rate	detectability	familiarity	range size	reporting probability	rel.body size	rel.classification error rate	rel.detectability	rel.familiarity	rel.range size	rel.reporting probability
body size	1											
classification error rate	-0.23	1										
detectability	-0.34	0.27	1									
familiarity	0.47	-0.38	-0.38	1								
range size	-0.04	0.02	0.13	0.27	1							
reporting probability	0.01	-0.11	-0.62	0.24	-0.07	1						
rel.body size	0.54	0.04	0.03	0.1	0.2	0.04	1					
rel.classification error rate	0.09	0.86	0.07	-0.19	0.04	-0.16	0.06	1				
rel.detectability	0.01	0.14	0.83	-0.09	0.15	-0.66	0.06	0.13	1			
rel.familiarity	0.06	-0.2	-0.07	0.71	0.41	0.26	0.19	-0.24	-0.03	1		
rel.range size	0.07	0.04	0.13	0.29	0.89	-0.05	0.23	0.05	0.21	0.41	1	
rel.reporting probability	0.01	-0.2	-0.58	0.22	-0.05	0.94	0.07	-0.2	-0.65	0.22	-0.09	1

Table I.6. Species traits per species, the species profile they are associated with and the percentage of opportunistic observations accompanied by photographs (photo rate) or sound fragments (sound rate) in the model training sets for unfiltered and filtered data using the VALSTAT filter (i.e. approved data only). As trait values show taxonomic differences, continuous values (absolute traits) were rescaled per taxonomic group (relative traits) to detect patterns across taxonomic groups that could go unnoticed otherwise. For a full description of species traits and profiles see Table 1 and section 13.2, and Table 2 and section 14.2 respectively in the main text.

Species name (scientific)	**Species name (Dutch)**	**Body size**	**Classification error rate**	**Detectability**	**Familiarity**	**Reporting probability**	**Range size**	**Profile**	**Photo rate**		**Sound rate**	
		Wing length (birds & butterflies), head-to-tail length (dragonflies)	*Relative number of photographic misidentifications*	*Detection probability*	*Number of Google search results*	*Reporting probability divided by the detection probability*	*number of grid cells in which a species has been recorded*		*Unfiltered data (%)*	*Approved data (%)*	*Unfiltered data (%)*	*Approved data (%)*
BIRDS	VOGELS											
Accipiter nisus	Sperwer	222	0.065	0.518	5970	0.547	5829	**1**	10.7	25.4	3.6E-02	3.4E-02
Alcedo atthis	IJsvogel	76	0.000	0.767	11500	0.712	3349	**5**	22.1	32.1	2.0E-02	1.6E-02
Alopochen aegyptiaca	Nijlgans	384	0.007	0.737	16800	0.306	4969	**5**	9.7	14.2	4.3E-03	6.4E-03
Anas crecca	Wintertaling	184	0.021	0.816	4100	0.176	1492	**2**	7.5	12.8	6.0E-03	7.7E-03
Anser anser	Grauwe Gans	454	0.035	0.735	3460	0.222	2189	**3**	8.7	9.5	9.8E-03	3.6E-03
Ardea alba	Grote Zilverreiger	440	0.008	0.858	4100	0.526	3495	**3**	13.1	17.5	9.8E-04	n/a

Species name (scientific)	Species name (Dutch)	Body size	Classification error rate	Detectability	Familiarity	Reporting probability	Range size	Profile	Photo rate		Sound rate	
Ardea cinerea	Blauwe Reiger	450	0.006	0.688	7200	0.373	7064	**4**	12.1	13.4	6.4E-03	4.7E-03
Athene noctua	Steenuil	162	0.004	0.298	5890	1.748	2454	**5**	12.7	35.1	1.9E-01	3.6E-01
Aythya ferina	Tafeleend	212	0.026	0.859	3980	0.186	817	**2**	8.8	14.4	2.1E-03	3.4E-03
Aythya fuligula	Kuifeend	202	0.018	0.813	5330	0.194	1876	**2**	8.3	14.2	3.2E-03	5.4E-03
Branta canadensis	Grote Canadese Gans	443	0.019	0.791	1010	0.209	3549	**3**	9.4	17.9	4.2E-03	5.4E-03
Branta leucopsis	Brandgans	401	0.018	0.559	2220	0.314	790	**3**	14.9	19.3	1.5E-02	6.7E-03
Buteo buteo	Buizerd	392	0.035	0.795	10900	0.520	10581	**4**	9.8	12.5	1.1E-02	7.4E-03
Carduelis carduelis	Putter	77	0.005	0.770	12100	0.332	4931	**5**	10.4	14.8	2.3E-02	1.6E-02
Chroicocephalus ridibundus	Kokmeeuw	300	0.030	0.858	3460	0.246	5465	**3**	8.2	9.4	4.5E-03	5.1E-03
Ciconia ciconia	Ooievaar	570	0.004	0.458	16800	1.390	2228	**5**	30.8	39.6	4.5E-03	5.8E-03
Circus aeruginosus	Bruine Kiekendief	403	0.031	0.761	5130	0.530	1880	**3**	20.3	50.1	2.1E-02	3.2E-02
Circus cyaneus	Blauwe Kiekendief	357	0.029	0.521	2770	0.879	1678	**3**	23.0	70.0	n/a	n/a
Corvus frugilegus	Roek	315	0.047	0.438	3130	1.315	2846	**3**	7.9	21.3	1.9E-02	5.1E-02
Cuculus canorus	Koekoek	222	0.011	0.507	10600	1.240	1032	**5**	5.0	18.9	9.9E-02	2.7E-01
Cygnus olor	Knobbelzwaan	606	0.010	0.755	3500	0.296	1902	**3**	12.7	16.2	7.5E-03	7.2E-03
Delichon urbicum	Huiszwaluw	111	0.039	0.701	3350	0.493	2622	**2**	5.2	17.4	1.4E-02	2.9E-02
Egretta garzetta	Kleine Zilverreiger	288	0.012	0.959	2980	0.334	931	**3**	18.6	22.8	n/a	n/a
Falco tinnunculus	Torenvalk	251	0.021	0.637	6910	0.700	7964	**4**	9.9	15.1	3.8E-03	2.3E-03
Fulica atra	Meerkoet	212	0.015	0.891	5250	0.189	3536	**2**	5.6	6.2	1.1E-02	8.2E-03
Gallinago gallinago	Watersnip	134	0.011	0.799	4910	0.290	1348	**2**	10.5	21.6	2.6E-02	2.5E-02
Larus argentatus	Zilvermeeuw	432	0.088	0.795	4960	0.246	2341	**1**	14.4	24.3	1.0E-02	4.3E-03
Larus canus	Stormmeeuw	363	0.100	0.788	2730	0.256	1617	**1**	9.1	27.8	n/a	n/a

Species name (scientific)	Species name (Dutch)	Body size	Classification error rate	Detectability	Familiarity	Reporting probability	Range size	Profile	Photo rate		Sound rate	
Larus fuscus	Kleine Mantelmeeuw	420	0.084	0.669	2890	0.329	3140	**1**	10.6	26.4	5.3E-03	1.4E-02
Limosa limosa	Grutto	214	0.020	0.745	5100	0.411	680	**2**	16.0	19.8	1.4E-02	8.8E-03
Linaria cannabina	Kneu	78	0.024	0.708	3930	0.322	3403	**2**	9.0	14.4	3.5E-02	4.0E-02
Luscinia svecica	Blauwborst	105	0.005	0.522	4870	0.421	673	**2**	13.0	26.4	9.1E-02	1.4E-01
Mareca strepera	Krakeend	260	0.032	0.843	3820	0.166	2118	**3**	6.5	9.1	2.3E-03	1.6E-03
Motacilla alba	Witte Kwikstaart	90	0.014	0.748	2180	0.318	6433	**2**	8.2	13.7	6.8E-03	9.1E-03
Motacilla flava	Gele Kwikstaart	80	0.050	0.843	4580	0.395	2338	**2**	10.8	46.2	4.9E-02	1.4E-01
Numenius arquata	Wulp	302	0.020	0.370	5870	0.539	1680	**3**	9.8	13.6	2.4E-02	2.3E-02
Oenanthe oenanthe	Tapuit	96	0.016	0.564	3180	0.704	2251	**2**	19.0	58.6	3.3E-03	1.0E-02
Perdix perdix	Patrijs	159	0.016	0.306	8570	1.353	3089	**5**	9.4	26.4	4.1E-02	5.8E-02
Phalacrocorax carbo	Aalscholver	348	0.010	0.779	6470	0.289	3303	**3**	11.5	16.6	n/a	n/a
Platalea leucorodia	Lepelaar	382	0.001	0.565	4580	0.756	608	**3**	27.1	30.3	6.9E-03	3.9E-03
Podiceps cristatus	Fuut	190	0.003	0.864	6280	0.295	1926	**2**	12.7	14.3	2.5E-03	2.8E-03
Rallus aquaticus	Waterral	120	0.008	0.617	3100	0.583	780	**2**	7.2	15.0	1.4E-01	1.6E-01
Recurvirostra avosetta	Kluut	226	0.002	0.512	2960	0.401	390	**2**	16.5	19.7	1.0E-02	6.1E-03
Riparia riparia	Oeverzwaluw	107	0.029	0.746	2480	0.438	593	**2**	10.2	22.9	7.7E-03	1.7E-02
Spatula clypeata	Slobeend	237	0.008	0.969	4430	0.141	1177	**3**	8.3	13.3	n/a	n/a
Spinus spinus	Sijs	73	0.012	0.719	4820	0.323	3147	**2**	13.7	32.7	3.1E-02	4.4E-02
Sterna hirundo	Visdief	271	0.011	0.912	3050	0.348	825	**3**	21.7	28.8	2.1E-02	1.7E-02
Tachybaptus ruficollis	Dodaars	100	0.008	0.720	4040	0.329	2004	**2**	8.8	13.5	1.2E-02	1.5E-02
Tadorna tadorna	Bergeend	318	0.006	0.838	4220	0.257	2214	**3**	7.0	10.6	n/a	n/a
Tringa totanus	Tureluur	162	0.039	0.744	3390	0.247	606	**2**	11.4	17.4	1.6E-02	2.4E-02

Species name (scientific)	Species name (Dutch)	Body size	Classification error rate	Detectability	Familiarity	Reporting probability	Range size	Profile	Photo rate		Sound rate	
Turdus pilaris	Kramsvogel	145	0.026	0.629	3620	0.374	1390	**2**	10.1	85.1	2.4E-02	1.5E-01
Vanellus vanellus	Kievit	226	0.003	0.841	7620	0.445	5289	**4**	6.4	13.6	1.7E-03	1.8E-03
BUTTERFLIES	**VLINDERS**											
Aglais io	Dagpauwoog	28	0.014	0.891	4220	0.424	7629	**4**	12.4	12.4	n/a	n/a
Aglais urticae	Kleine Vos	24	0.024	0.724	4220	0.470	4645	**4**	15.2	15.5	n/a	n/a
Anthocharis cardamines	Oranjetipje	20	0.009	0.872	2930	0.616	2892	**2**	17.2	30.8	n/a	n/a
Aphantopus hyperantus	Koevinkje	21	0.030	0.875	1280	0.318	916	**2**	15.4	29.2	n/a	n/a
Araschnia levana	Landkaartje	18	0.019	0.819	2440	0.377	3484	**2**	24.5	33.8	n/a	n/a
Aricia agestis	Bruin Blauwtje	13	0.097	0.737	1950	0.373	2432	**1**	41.8	52.6	n/a	n/a
Celastrina argiolus	Boomblauwtje	14	0.026	0.760	2190	0.363	3748	**2**	16.3	23.5	n/a	n/a
Coenonympha pamphilus	Hooibeestje	16	0.056	0.793	1770	0.602	1706	**2**	12.3	16.8	n/a	n/a
Colias crocea	Oranje Luzernevlinder	24	0.049	0.747	1070	0.860	2495	**2**	24.8	35.7	n/a	n/a
Favonius quercus	Eikenpage	16	0.013	0.808	937	0.441	1027	**2**	49.3	69.1	n/a	n/a
Gonepteryx rhamni	Citroenvlinder	28	0.007	0.870	2920	0.554	6793	**4**	7.6	7.5	n/a	n/a
Issoria lathonia	Kleine Parelmoervlinder	21	0.021	0.760	1270	0.462	1082	**2**	48.6	54.5	n/a	n/a
Lycaena phlaeas	Kleine Vuurvlinder	14	0.010	0.818	1840	0.423	3492	**2**	34.9	46.3	n/a	n/a
Maniola jurtina	Bruin Zandoogje	24	0.062	0.941	1540	0.325	5156	**1**	9.8	11.5	n/a	n/a
Ochlodes sylvanus	Groot Dikkopje	14	0.045	0.821	1220	0.364	2996	**2**	27.7	38.5	n/a	n/a
Papilio machaon	Koninginnenpage	36	0.005	0.781	2240	0.682	3110	**4**	34.2	45.0	n/a	n/a
Pararge aegeria	Bont Zandoogje	20	0.013	0.839	1860	0.436	5523	**4**	12.2	16.5	n/a	n/a
Pieris brassicae	Groot Koolwitje	30	0.170	0.874	2130	0.238	4182	**1**	9.9	12.2	n/a	n/a
Pieris napi	Klein Geaderd Witje	22	0.087	0.887	1530	0.277	4476	**1**	14.7	16.8	n/a	n/a

Species name (scientific)	Species name (Dutch)	Body size	Classification error rate	Detectability	Familiarity	Reporting probability	Range size	Profile	Photo rate		Sound rate	
Pieris rapae	Klein Koolwitje	24	0.132	0.916	2750	0.385	6769	**1**	8.7	9.2	n/a	n/a
Polygonia c-album	Gehakkelde Aurelia	23	0.009	0.815	3000	0.335	5582	**4**	18.8	18.7	n/a	n/a
Polyommatus icarus	Icarusblauwtje	14	0.070	0.897	2020	0.373	4014	**1**	30.0	39.2	n/a	n/a
Pyronia tithonus	Oranje Zandoogje	18	0.050	0.939	2420	0.393	3541	**2**	13.8	17.9	n/a	n/a
Vanessa atalanta	Atalanta	29	0.016	0.797	3950	0.451	6836	**4**	10.7	14.1	n/a	n/a
Vanessa cardui	Distelvlinder	28	0.013	0.856	2760	0.396	5113	**4**	19.8	25.3	n/a	n/a
DRAGONFLIES	**LIBELLEN**											
Aeshna cyanea	Blauwe Glazenmaker	62	0.077	0.775	777	0.617	1664	**1**	31.0	50.8	n/a	n/a
Aeshna mixta	Paardenbijter	60	0.060	0.864	972	0.801	1906	**1**	32.0	51.7	n/a	n/a
Anax imperator	Grote Keizerlibel	74	0.038	0.906	941	0.430	2299	**1**	22.7	45.8	n/a	n/a
Calopteryx splendens	Weidebeekjuffer	47	0.045	0.956	2470	0.634	1396	**2**	27.6	38.7	n/a	n/a
Coenagrion puella	Azuurwaterjuffer	33	0.078	0.898	773	0.420	2239	**1**	35.4	47.1	n/a	n/a
Enallagma cyathigerum	Watersnuffel	33	0.108	0.865	656	0.310	948	**1**	28.7	38.9	n/a	n/a
Ischnura elegans	Lantaarntje	32	0.058	0.892	1180	0.430	2263	**1**	27.8	38.1	n/a	n/a
Libellula depressa	Platbuik	44	0.046	0.890	1210	0.522	1913	**1**	36.7	58.4	n/a	n/a
Libellula quadrimaculata	Viervlek	44	0.028	0.911	837	0.300	1101	**2**	29.0	46.9	n/a	n/a
Orthetrum cancellatum	Gewone Oeverlibel	47	0.051	0.932	1120	0.470	2301	**1**	33.8	48.0	n/a	n/a
Platycnemis pennipes	Blauwe Breedscheenjuffer	36	0.045	0.643	964	1.322	912	**2**	39.0	56.3	n/a	n/a
Pyrrhosoma nymphula	Vuurjuffer	34	0.017	0.743	793	0.509	1694	**2**	34.3	53.3	n/a	n/a
Sympetrum sanguineum	Bloedrode Heidelibel	37	0.089	0.881	1280	0.534	2130	**1**	46.0	69.4	n/a	n/a
Sympetrum striolatum	Bruinrode Heidelibel	40	0.092	0.925	1010	0.680	2030	**1**	55.3	93.0	n/a	n/a

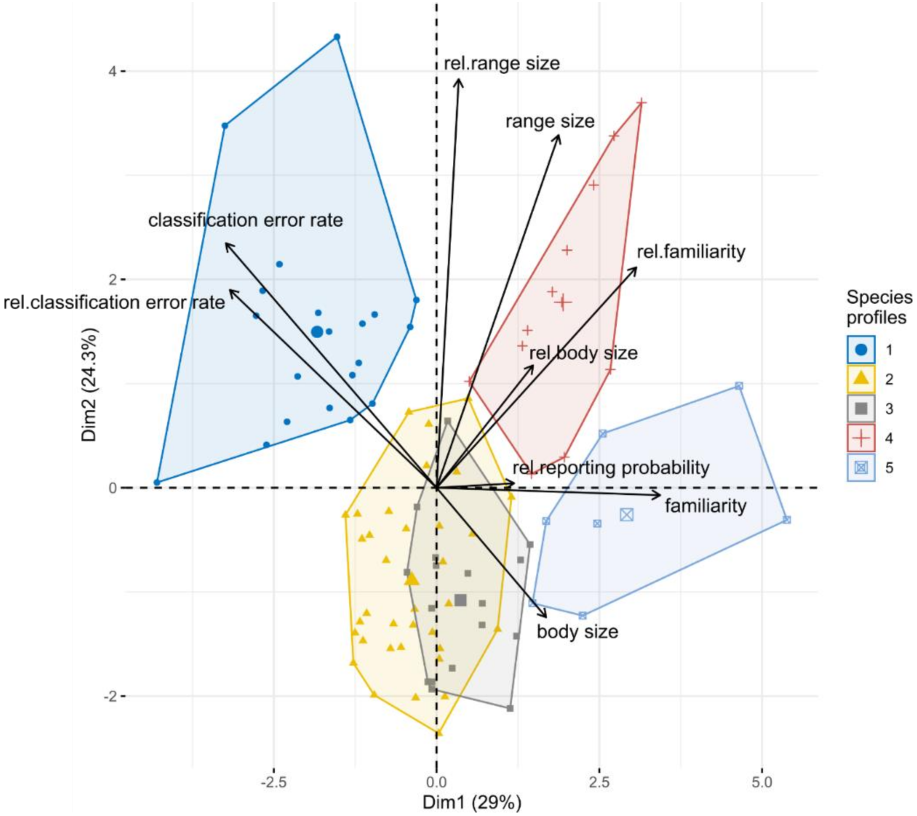
Appendix J: Principal component analysis (PCA) and clustering results

Table J.1. Description of the different dimensions (Dim.) of the PCA: (i) in terms of eigenvalues and percentage of the total variance explained; and in terms of coordinate values of the different variables (coord), their contributions to the dimensions (ctr) and the square cosine (cos2), which is a measure of the quality of the representation in the dimension (you can e.g. sum the cos2 values from the first two dimensions to look at the quality of representation in the first plane of the PCA). The results of the PCA are used in an agglomerative hierarchical clustering of individuals (i.e. species), which are described by the principal components. Note that supplementary variables do not contribute to the different dimensions. For supplementary categorical variables, there are also v-test results, with extreme values indicating coordinate values that are significantly different to 0.

	Dim.1			**Dim.2**			**Dim.3**			**Dim.4**			**Dim.5**			**Dim.6**			**Dim.7**			**Dim.8**			**Dim.9**		
Eigenvalue	2.614			2.189			1.555			0.992			0.863			0.429			0.213			0.111			0.033		
% of total variance explained	29.050			24.317			17.278			11.025			9.593			4.770			2.368			1.228			0.371		
Cumulative %	*29.050*			*53.367*			*70.645*			*81.670*			*91.263*			*96.033*			*98.401*			*99.629*			*100*		
Active variables	**coord**	**ctr**	**cos2**	**coord**	**ctr**	**cos2**	**coord**	**ctr**	**cos2**	**coord**	**ctr**	**cos2**	**coord**	**ctr**	**cos2**	**coord**	**ctr**	**cos2**	**coord**	**ctr**	**cos2**	**coord**	**ctr**	**cos2**	**coord**	**ctr**	**cos2**
body size	0.378	5.461	0.143	-0.279	3.564	0.078	0.834	44.730	0.696	-0.026	0.068	0.001	0.003	0.001	0.000	0.209	10.182	0.044	0.085	3.358	0.007	0.174	27.489	0.030	0.041	5.147	0.002
relative body size	0.333	4.232	0.111	0.264	3.183	0.070	0.656	27.693	0.431	0.255	6.543	0.065	-0.497	28.614	0.247	-0.242	13.657	0.059	-0.062	1.814	0.004	-0.119	12.730	0.014	-0.023	1.535	0.001
range size	0.421	6.781	0.177	0.760	26.416	0.578	-0.046	0.136	0.002	-0.232	5.439	0.054	-0.076	0.671	0.006	0.329	25.195	0.108	-0.273	34.854	0.074	0.011	0.110	0.000	-0.012	0.397	0.000
relative range size	0.076	0.223	0.006	0.881	35.478	0.776	-0.156	1.571	0.024	-0.151	2.285	0.023	-0.232	6.240	0.054	0.070	1.151	0.005	0.334	52.343	0.112	-0.002	0.005	0.000	0.015	0.703	0.000
relative reporting probability	0.268	2.737	0.072	0.010	0.004	0.000	-0.318	6.519	0.101	0.878	77.757	0.772	-0.057	0.374	0.003	0.226	11.926	0.051	0.023	0.251	0.001	0.021	0.403	0.000	0.003	0.029	0.000
familiarity	0.771	22.750	0.595	-0.016	0.012	0.000	0.215	2.964	0.046	-0.016	0.024	0.000	0.554	35.606	0.307	0.081	1.540	0.007	0.086	3.497	0.007	-0.193	33.606	0.037	-0.001	0.003	0.000
relative familiarity	0.689	18.162	0.475	0.475	10.322	0.226	-0.151	1.460	0.023	0.107	1.156	0.011	0.296	10.123	0.087	-0.388	35.118	0.151	-0.040	0.763	0.002	0.159	22.886	0.025	-0.002	0.009	0.000
classification error rate	-0.727	20.203	0.528	0.527	12.706	0.278	0.262	4.405	0.069	0.196	3.884	0.039	0.247	7.058	0.061	-0.052	0.638	0.003	-0.076	2.677	0.006	-0.040	1.484	0.002	0.125	46.946	0.016
relative classification error rate	-0.713	19.450	0.509	0.427	8.315	0.182	0.405	10.522	0.164	0.168	2.843	0.028	0.313	11.312	0.098	0.050	0.594	0.003	0.031	0.443	0.001	0.038	1.288	0.001	-0.123	45.232	0.015

	Dim.1			Dim.2			Dim.3			Dim.4			Dim.5			Dim.6			Dim.7			Dim.8			Dim.9		
Suppl. continuous variables	**coord**	**cos2**		**coord**	**cos2**		**coord**	**cos2**		**coord**	**cos2**		**coord**	**cos2**		**coord**	**cos2**		**coord**	**cos2**		**coord**	**cos2**		**coord**	**cos2**	
detectability	-0.296	0.088		0.336	0.113		-0.074	0.005		-0.402	0.162		-0.197	0.039		-0.357	0.127		-0.021	0.000		-0.036	0.001		0.026	0.001	
relative detectability	-0.169	0.029		0.172	0.030		0.231	0.053		-0.545	0.297		-0.007	0.000		-0.193	0.037		0.163	0.027		-0.025	0.001		-0.031	0.001	
reporting probability	0.265	0.070		-0.037	0.001		-0.193	0.037		0.807	0.651		0.087	0.008		0.283	0.080		0.216	0.047		0.023	0.001		0.036	0.001	
Δ AUC ACTIVITY	-0.062	0.004		0.398	0.159		0.010	0.000		0.145	0.021		-0.182	0.033		-0.351	0.123		-0.036	0.001		0.129	0.017		0.032	0.001	
Δ AUC DETAIL	-0.007	0.000		0.043	0.002		-0.020	0.000		0.100	0.010		-0.117	0.014		-0.036	0.001		0.154	0.024		-0.094	0.009		-0.114	0.013	
Δ AUC VALSTAT	-0.061	0.004		-0.059	0.003		0.072	0.005		0.133	0.018		0.039	0.002		-0.148	0.022		0.073	0.005		0.053	0.003		0.067	0.005	
Δ Sensitivity ACTIVITY	0.100	0.010		-0.065	0.004		-0.044	0.002		-0.009	0.000		-0.116	0.013		0.042	0.002		-0.179	0.032		0.051	0.003		0.175	0.031	
Δ Sensitivity DETAIL	0.051	0.003		-0.035	0.001		-0.007	0.000		0.016	0.000		-0.039	0.002		0.183	0.033		-0.042	0.002		0.007	0.000		-0.062	0.004	
Δ Sensitivity VALSTAT	0.083	0.007		-0.097	0.009		-0.069	0.005		0.066	0.004		0.070	0.005		0.054	0.003		-0.082	0.007		0.202	0.041		0.052	0.003	
Δ Specificity ACTIVITY	-0.140	0.020		0.095	0.009		0.005	0.000		0.014	0.000		0.068	0.005		-0.079	0.006		0.194	0.038		-0.031	0.001		-0.153	0.023	
Δ Specificity DETAIL	-0.069	0.005		0.048	0.002		-0.019	0.000		0.015	0.000		0.002	0.000		-0.204	0.042		0.071	0.005		-0.030	0.001		0.034	0.001	
Δ Specificity VALSTAT	-0.119	0.014		0.065	0.004		0.047	0.002		-0.053	0.003		-0.052	0.003		-0.119	0.014		0.090	0.008		-0.157	0.025		-0.034	0.001	
Suppl. categories	**coord**	**cos2**	**v.test**	**coord**	**cos2**	**v.test**	**coord**	**cos2**	**v.test**	**coord**	**cos2**	**v.test**	**coord**	**cos2**	**v.test**	**coord**	**cos2**	**v.test**	**coord**	**cos2**	**v.test**	**coord**	**cos2**	**v.test**	**coord**	**cos2**	**v.test**
birds	0.476	0.223	3.224	-0.635	0.397	-4.699	0.504	0.250	4.424	-0.192	0.036	-2.109	0.236	0.055	2.779	0.194	0.037	3.240	0.013	0.000	0.302	0.011	0.000	0.366	-0.023	0.001	-1.393
butterflies	-0.102	0.006	-0.368	0.876	0.410	3.456	-0.826	0.365	-3.870	0.228	0.028	1.338	-0.360	0.069	-2.264	-0.250	0.033	-2.229	-0.402	0.086	-5.084	-0.030	0.000	-0.529	0.058	0.002	1.861
dragonflies	-1.586	0.633	-3.967	0.793	0.158	2.169	-0.395	0.039	-1.281	0.305	0.023	1.237	-0.232	0.014	-1.011	-0.273	0.019	-1.686	0.671	0.113	5.876	0.013	0.000	0.152	-0.018	0.000	-0.392

Figure J.1. Biplot of the first two dimensions of the PCA with active variables, individuals (species) and species profiles.



CHAPTER IV

Appendix K: Exploratory data analysis

Table K.1. List of considered species with their Red List Status in Flanders (LC = Least Concern, NT = Near Threatened, EN = Endangered, CR = Critically Endangered) (Devos et al., 2016; Jooris et al., 2012; Maes et al., 2017, 2021), Conservation Interest (Habitats (HD) or Birds (BD) Directive, Flemish Priority Species (FPS), Habitat Specific Species (HSS) with Natura 2000 habitat types), species occurrence in the different landscape contexts (number of presences, spatially thinned per observation date at 50 metres and Intensity of the point process), r and sat values used in the Geyer interaction process of the Gibbs point process model, and the p values of the Diggle-Cressie-Loosmore-Ford (DCLF) goodness-of-fit test. For the DCLF-test, simulation envelopes were run based on the fitted models per landscape context. (NA = species were excluded, zero presences or models did not converge).

Species English name Dutch name	**Red List status in Flanders**	**Conservation Interest**	**Landscape context**	**n° pres**	**Intensity**	**r**	**sat**	**p.DCLF**
Birds								
Anthus trivialis Tree Pipit Boompieper	NT	HSS ^{2310}	open	907	4.70E-05	50	2	0.026
			closed	2683	6.54E-05	50	2	0.026
			anthropogenic	137	1.32E-05	50	1	0.103
Caprimulgus europaeus European Nightjar Nachtzwaluw	NT	BD Annex I HSS ^{4030}	open	151	1.25E-05	50	2	0.026
			closed	462	1.59E-05	100	2	0.026
			anthropogenic	26	4.71E-06	100	2	0.026
Lullula arborea Woodlark Boomleeuwerik	NT	BD Annex I HSS $^{2310, 4030}$	open	492	2.95E-05	50	2	0.026
			closed	1213	3.24E-05	50	2	0.026
			anthropogenic	80	8.89E-06	250	1	0.564
Oenanthe oenanthe Northern wheatear Tapuit	CR	HSS ^{2310}	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
Saxicola rubicola European Stonechat Roodborsttapuit	LC	HSS $^{2310, 4030}$	open	935	5.04E-05	50	2	0.026
			closed	1542	3.87E-05	50	2	0.026
			anthropogenic	130	1.41E-05	150	1	0.026
Butterflies								
Callophrys rubi Green Hairstreak Groentje	EN	HSS $^{2310, 4030}$	open	265	1.99E-05	100	2	0.026
			closed	321	1.13E-05	50	2	0.026
			anthropogenic	33	1.13E-05	450	0	0.026
Hesperia comma Silver-spotted Skipper Kommavlinder	EN	FPS HSS $^{2310, 2330, 4030}$	open	6	3.79E-06	250	1	0.051
			closed	85	5.44E-06	100	2	0.026
			anthropogenic	35	5.22E-06	450	1	0.103
Hipparchia semele Grayling Heivlinder	EN	FPS HSS $^{2310, 2330, 4030}$	open	302	1.98E-05	50	2	0.026
			closed	330	1.00E-05	100	2	0.026
			anthropogenic	485	6.15E-05	100	2	0.026
Plebejus argus	EN	HSS ^{4030}	open	621	4.34E-05	50	2	0.026

Species English name Dutch name	**Red List status in Flanders**	**Conservation Interest**	**Landscape context**	**n° pres**	**Intensity**	**r**	**sat**	**p.DCLF**
Silver-studded Blue Heideblauwtje			closed	483	1.70E-05	50	2	0.026
			anthropogenic	23	3.48E-06	50	2	0.026
Pyrgus malvae Grizzled skipper Aardbeivlinder	EN	FPS	open	NA	NA			
			closed	44	1.04E-06	*NA*	*NA*	*NA*
			anthropogenic	NA	NA			
Grasshoppers								
Chorthippus mollis Lesser Field Grasshopper Snortikker	LC	HSS $^{2310, 2330, 4030}$	open	10	1.98E-05	350	2	
			closed	9	1.23E-06	400	2	*NA*
			anth	3	4.11E-06	300	2	
Ephippiger diurnus Bush cricket Zadelsprinkhaan	EN	FPS HSS $^{2310, 4030}$	open	52	1.87E-05	300	2	
			closed	51	6.65E-06	350	1	*NA*
			anth	30	7.89E-05	150	1	
Gryllus campestris Field Cricket Veldkrekel	EN	HSS $^{2310, 2330}$	open	118	7.94E-06	100	2	0.026
			closed	324	8.84E-06	100	2	0.026
			anth	135	1.66E-05	450	1	0.026
Metrioptera brachyptera Bog bush cricket Heidesabelsprinkhaan	LC	FPS	open	30	2.83E-06	100	2	0.026
			closed	27	1.19E-06	400	2	0.487
			anth	13	6.65E-06	150	1	0.026
Myrmeleotettix maculatus Mottled grasshopper Knopsprietje	LC	HSS $^{2310, 2330}$	open	68	4.39E-06	50	2	0.026
			closed	243	7.14E-06	100	2	0.026
			anth	82	8.88E-06	100	1	0.026
Oedipoda caerulescens Blue Winged Grasshopper Blauwvleugelsprinkhaan	LC	HSS $^{2310, 4030}$	open	112	1.49E-05	200	2	0.026
			closed	296	9.97E-06	50	2	0.026
			anth	189	1.99E-05	50	2	0.026
Omocestus rufipes Woodland Grasshopper Zwart Wekkertje	NT	FPS HSS ^{2310}	open	59	8.34E-06	200	2	0.026
			closed	35	1.70E-06	100	2	0.333
			anth	18	7.24E-06	250	1	0.026
Reptiles								
Coronella austriaca Smooth Snake Gladde Slang	EN	HD Annex IV HSS $^{2310, 4030}$	open	32	6.46E-06	150	2	0.026
			closed	51	2.33E-06	50	2	0.051
			anth	0	0	*NA*	*NA*	*NA*
Zootoca vivipara Common Lizard Levendbarende Hagedis	LC	HSS ^{4030}	open	18	1.61E-06	50	1	0.103
			closed	37	1.41E-06	50	1	0.205
			anth	17	2.23E-06	350	2	0.026

Figure K.1. Values of the model predictors in the three landscape contexts (open, closed and anthropogenic). Correlations between landscape context and each of the predictors were tested in an ANOVA test (results under predictor) and correlations between the values in each landscape context with a Wilcoxon rank test (results in graph area). (**** = p < 0.0001, *** = p < 0.001, ** = p < 0.01, * = p < 0.05).

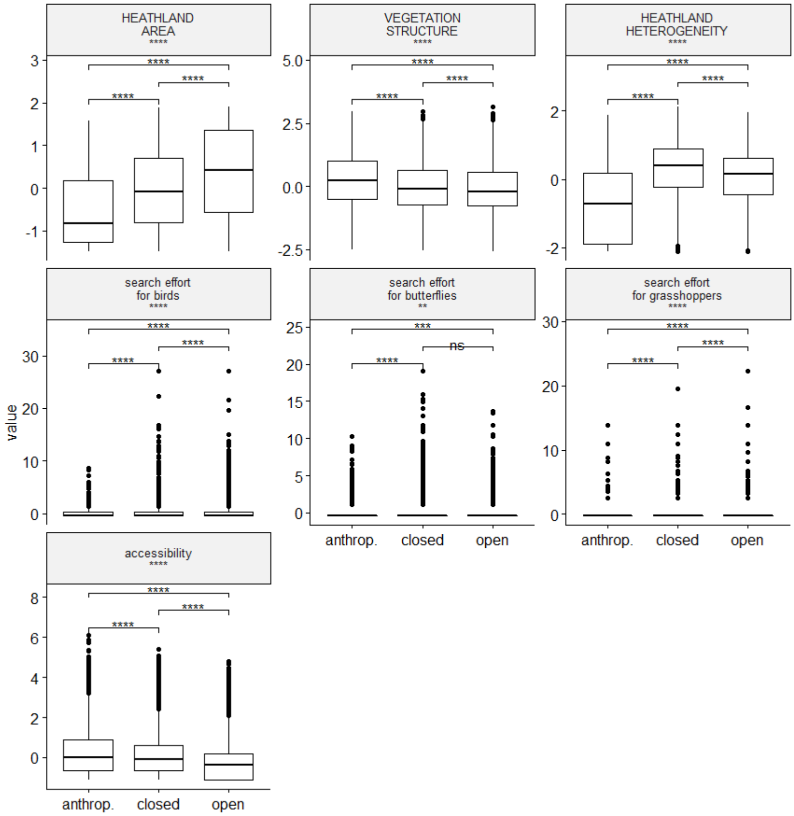


Table K.2. Pearson correlations between the predictor variables per species (only the selected species in the final analysis) and landscape context.

species	open	closed	anth
heathland size - vegetation structure			
Anthus trivialis	-0.41	-0.22	-0.20
Callophrys rubi	-0.12	-0.20	
Caprimulgus europaeus	-0.17	-0.22	
Gryllus campestris	-0.46	-0.41	
Hipparchia semele	-0.35	-0.21	-0.34
Lullula arborea	-0.37	-0.21	-0.37
Myrmeleotettix maculatus	-0.45	-0.19	-0.48
Oedipoda caerulescens	-0.06	-0.03	-0.46
Plebejus argus	-0.14	-0.30	
Saxicola rubicola	-0.45	-0.27	-0.21
heathland size - heathland heterogeneity			
Anthus trivialis	0.01	0.34	0.19
Callophrys rubi	0.32	0.36	
Caprimulgus europaeus	0.08	0.31	
Gryllus campestris	0.14	0.04	
Hipparchia semele	0.43	0.34	0.39
Lullula arborea	0.16	0.34	0.22
Myrmeleotettix maculatus	0.28	0.36	0.35
Oedipoda caerulescens	0.05	0.10	0.16
Plebejus argus	-0.01	0.06	
Saxicola rubicola	-0.02	0.27	0.15
vegetation structure - heathland heterogeneity			
Anthus trivialis	0.05	0.03	-0.16
Callophrys rubi	0.03	0.12	
Caprimulgus europaeus	0.15	-0.04	
Gryllus campestris	-0.17	0.07	
Hipparchia semele	-0.26	-0.03	-0.25
Lullula arborea	0.01	-0.01	-0.01
Myrmeleotettix maculatus	-0.07	0.02	-0.16
Oedipoda caerulescens	-0.05	0.14	-0.10
Plebejus argus	0.27	0.22	
Saxicola rubicola	0.11	-0.01	-0.09
heathland size - accessibility			
Anthus trivialis	-0.08	-0.06	-0.03
Callophrys rubi	0.08	-0.16	
Caprimulgus europaeus	-0.21	-0.28	
Gryllus campestris	-0.16	-0.32	
Hipparchia semele	0.20	-0.34	-0.69
Lullula arborea	0.15	-0.13	-0.27
Myrmeleotettix maculatus	0.02	-0.22	-0.59
Oedipoda caerulescens	-0.02	-0.03	-0.65
Plebejus argus	0.09	-0.10	
Saxicola rubicola	0.10	-0.12	-0.19
heathland size - search effort			
Anthus trivialis	0.00	-0.08	-0.15
Caprimulgus europaeus	0.32	0.10	
Lullula arborea	-0.12	-0.06	-0.21
Saxicola rubicola	-0.21	-0.20	-0.36
Callophrys rubi	-0.20	-0.08	
Hipparchia semele	-0.09	0.04	-0.23
Plebejus argus	-0.44	-0.26	
Gryllus campestris	-0.08	-0.31	
Myrmeleotettix maculatus	0.02	-0.03	0.01
Oedipoda caerulescens	-0.14	-0.10	-0.11

species	open	closed	anth
vegetation structure - accessibility			
Anthus trivialis	0.18	0.17	0.20
Callophrys rubi	0.05	0.21	
Caprimulgus europaeus	0.24	0.26	
Gryllus campestris	0.35	0.31	
Hipparchia semele	-0.05	0.29	0.41
Lullula arborea	-0.10	0.25	0.29
Myrmeleotettix maculatus	0.24	0.32	0.39
Oedipoda caerulescens	0.22	0.20	0.48
Plebejus argus	-0.02	0.19	
Saxicola rubicola	0.16	0.15	0.30
vegetation structure - search effort			
Anthus trivialis	0.05	0.16	-0.07
Caprimulgus europaeus	0.07	0.20	
Lullula arborea	0.09	0.20	0.07
Saxicola rubicola	0.12	0.29	0.08
Callophrys rubi	0.13	0.05	
Hipparchia semele	-0.19	-0.16	-0.06
Plebejus argus	0.07	0.30	
Gryllus campestris	0.21	0.24	
Myrmeleotettix maculatus	0.10	0.17	-0.19
Oedipoda caerulescens	0.14	0.21	-0.12
heathland heterogeneity - accessibility			
Anthus trivialis	-0.17	0.04	-0.16
Callophrys rubi	0.02	0.07	
Caprimulgus europaeus	0.05	-0.17	
Gryllus campestris	0.02	0.03	
Hipparchia semele	0.02	-0.04	-0.30
Lullula arborea	-0.27	-0.01	0.18
Myrmeleotettix maculatus	-0.08	0.03	-0.25
Oedipoda caerulescens	0.01	-0.07	-0.10
Plebejus argus	-0.10	0.03	
Saxicola rubicola	-0.04	-0.03	-0.03
heathland heterogeneity - search effort			
Anthus trivialis	0.01	-0.15	-0.32
Caprimulgus europaeus	-0.01	-0.18	
Lullula arborea	-0.16	-0.14	0.13
Saxicola rubicola	-0.07	-0.18	0.06
Callophrys rubi	0.02	-0.11	
Hipparchia semele	0.05	-0.04	-0.12
Plebejus argus	-0.04	0.10	
Gryllus campestris	0.05	0.21	
Myrmeleotettix maculatus	0.00	0.01	-0.02
Oedipoda caerulescens	-0.16	0.12	-0.15
accessibility - search effort			
Anthus trivialis	-0.01	0.14	0.01
Caprimulgus europaeus	0.17	0.11	
Lullula arborea	0.03	0.13	0.18
Saxicola rubicola	0.15	0.26	0.08
Callophrys rubi	0.51	0.02	
Hipparchia semele	0.14	-0.13	0.07
Plebejus argus	0.46	0.16	
Gryllus campestris	0.27	-0.01	
Myrmeleotettix maculatus	0.14	0.04	0.20
Oedipoda caerulescens	-0.06	-0.04	0.00

Figures K.2 to K.11. Inhomogeneous pair correlation functions $g_{inhom}(r)$ in each landscape context for the point processes of the selected model training sets per species. Upward deviations from the horizontal intercept y = 1 indicate clustering and downward deviations indicate inhibition. Values near zero cannot be interpreted (see page 229 in Baddeley et al., 2015).

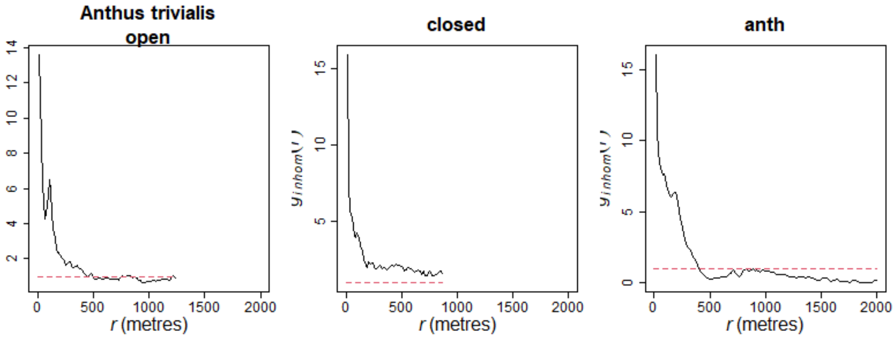


Figure K.2

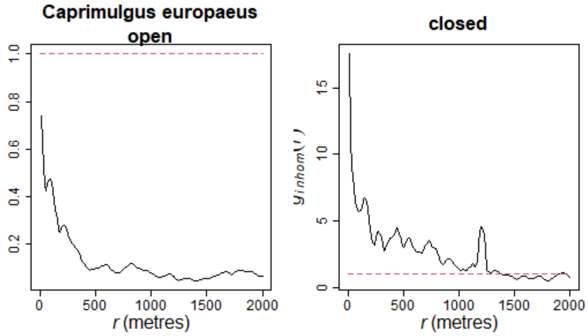


Figure K.3

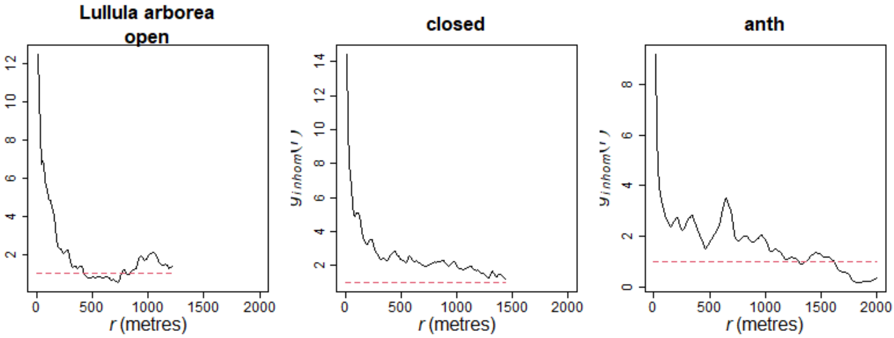


Figure K.4

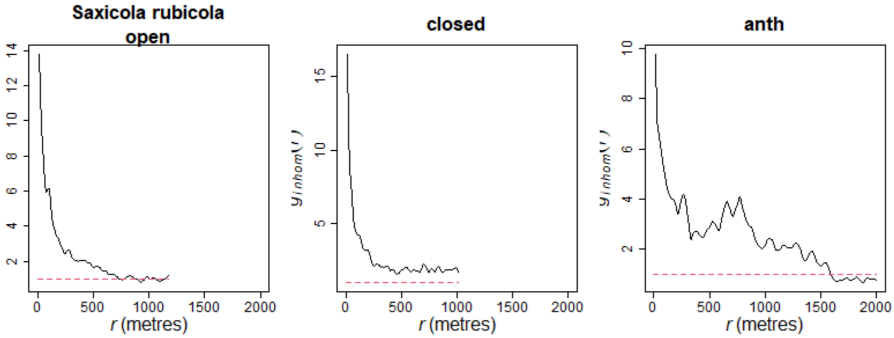


Figure K.5

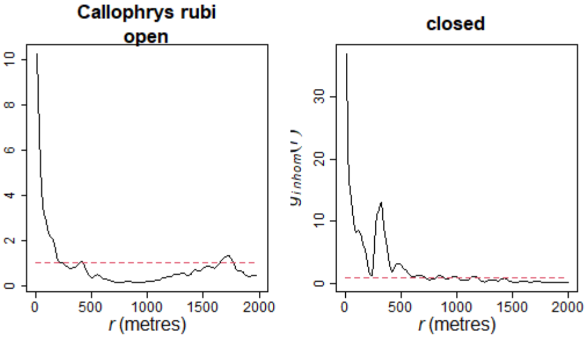


Figure K.6

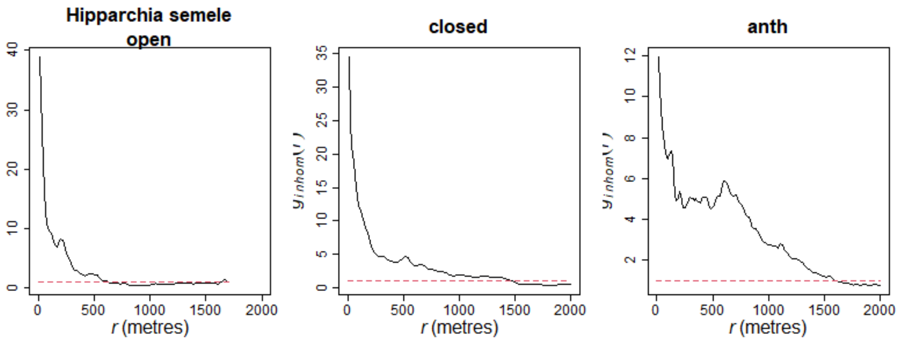


Figure K.7

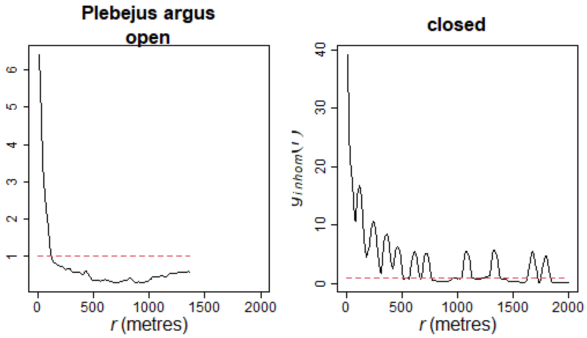


Figure K.8

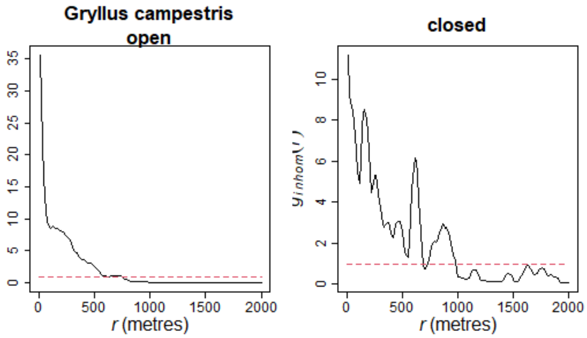


Figure K.9

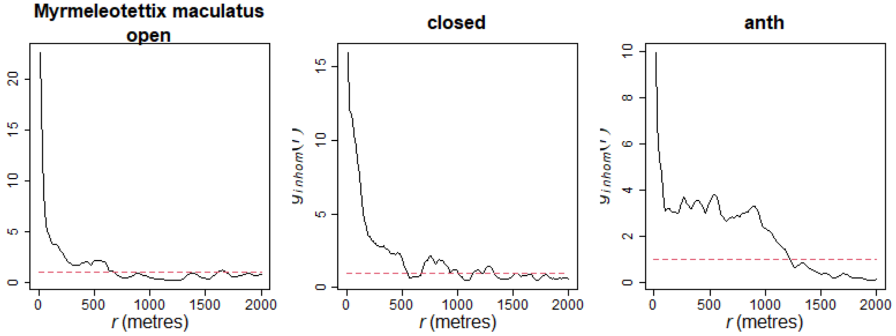


Figure K.10

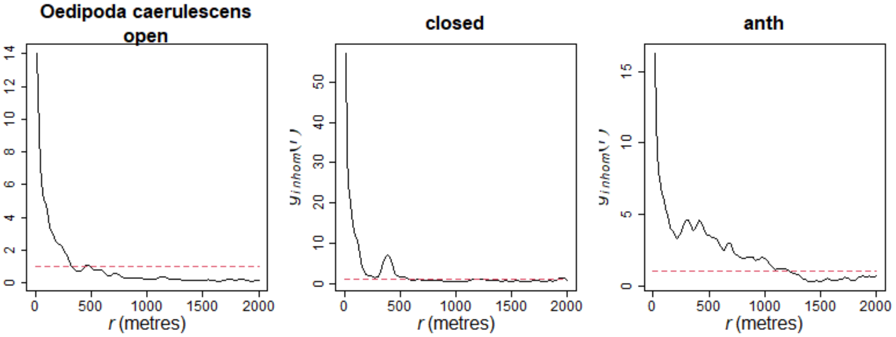


Figure K.11

Appendix L: Model performance per species

Table L.7. Model evaluation results from spatial-block cross-validation (AUC = Area Under the receiver operating Curve, CBI = Continuous Boyce Index, SENS = sensitivity, NA = models based on subsets of the training data could not be fitted).

	AUC	**SENS**	**CBI**	**AUC**	**SENS**	**CBI**	**AUC**	**SENS**	**CBI**
SPECIES	**open**			**closed**			**anthropogenic**		
ALL	**0.76 ± 0.12**	**0.71 ± 0.18**	**0.61 ± 0.44**	**0.71 ± 0.1**	**0.68 ± 0.18**	**0.75 ± 0.30**	**0.81 ± 0.09**	**0.77 ± 0.14**	**0.74 ± 0.32**
BIRDS	**0.82 ± 0.07**	**0.75 ± 0.13**	**0.75 ± 0.12**	**0.78 ± 0.06**	**0.73 ± 0.09**	**0.79 ± 0.17**	**0.83 ± 0.1**	**0.76 ± 0.17**	**0.81 ± 0.19**
ANTHUS TRIVIALIS	0.83 ± 0.07	0.74 ± 0.09	0.80 ± 0.06	0.77 ± 0.03	0.7 ± 0.06	0.88 ± 0.06	0.84 ± 0.10	0.76 ± 0.20	0.87 ± 0.10
CAPRIMULGUS EUROPAEUS	0.79 ± 0.09	0.77 ± 0.15	0.73 ± 0.11	0.79 ± 0.06	0.74 ± 0.09	0.72 ± 0.30	*NA*	*NA*	*NA*
LULLULA ARBOREA	0.83 ± 0.07	0.81 ± 0.12	0.77 ± 0.12	0.77 ± 0.06	0.73 ± 0.11	0.77 ± 0.09	0.82 ± 0.10	0.76 ± 0.14	0.75 ± 0.25
SAXICOLA RUBICOLA	0.82 ± 0.03	0.67 ± 0.13	0.72 ± 0.18	0.80 ± 0.07	0.74 ± 0.09	0.78 ± 0.09	*NA*	*NA*	*NA*
BUTTERFLIES	**0.69 ± 0.11**	**0.69 ± 0.21**	**0.42 ± 0.57**	**0.64 ± 0.09**	**0.61 ± 0.25**	**0.67 ± 0.42**	**0.76 ± 0.06**	**0.78 ± 0.09**	**0.60 ± 0.46**
CALLOPHRYS RUBI	0.74 ± 0.07	0.66 ± 0.20	0.69 ± 0.24	0.72 ± 0.08	0.72 ± 0.20	0.70 ± 0.28	*NA*	*NA*	*NA*
HIPPARCHIA SEMELE	0.71 ± 0.11	0.84 ± 0.14	0.05 ± 0.68	0.69 ± 0.05	0.59 ± 0.21	0.65 ± 0.41	0.76 ± 0.06	0.78 ± 0.09	0.60 ± 0.46
PLEBEJUS ARGUS	0.62 ± 0.11	0.59 ± 0.20	0.52 ± 0.54	0.56 ± 0.07	0.58 ± 0.30	0.68 ± 0.51	*NA*	*NA*	*NA*
GRASSHOPPERS	**0.72 ± 0.16**	**0.67 ± 0.22**	**0.60 ± 0.55**	**0.66 ± 0.1**	**0.65 ± 0.21**	**0.75 ± 0.33**	***NA***	***NA***	***NA***
GRYLLUS CAMPESTRIS	0.60 ± 0.29	0.60 ± 0.36	0.03 ± 0.93	0.59 ± 0.03	0.54 ± 0.20	0.94 ± 0.11	*NA*	*NA*	*NA*
MYRMELEOTETTIX MACULATUS	0.80 ± 0.08	0.80 ± 0.17	0.75 ± 0.15	0.67 ± 0.13	0.70 ± 0.23	0.70 ± 0.25	*NA*	*NA*	*NA*
OEDIPODA CAERULESCENS	0.75 ± 0.06	0.64 ± 0.16	0.81 ± 0.16	0.69 ± 0.07	0.64 ± 0.20	0.68 ± 0.53	*NA*	*NA*	*NA*

Appendix M: The impact of vegetation structure and heathland heterogeneity on the relative habitat suitability per species

For every species, we show the impact of vegetation structure and heathland heterogeneity on the relative habitat suitability of species in different landscape contexts and for different (classes of) heathland sizes.

Figures M.1 to M.3 show the log-transformed predictions of the Gibbs point process models with Geyer saturation process, including two-way interactions between heathland size and vegetation structure/heathland heterogeneity. Heathland size is divided into three classes: (1) small patches (≤ 10 hectares), i.e. mostly small and isolated patches with an occasional heathland patch edge largely surrounded by different land cover, (2) intermediate patches/patch edges (10-30 hectares), i.e. mostly edges of large heathland patches with an occasional medium-sized patch, and (3) large patches (> 30 hectares), i.e. core areas of large heathland patches.

Figures M.4 to M.6 show the model coefficients, with dots and bars representing mean estimates of the trend coefficients and Johnson-Neyman intervals for 2-way interactions (calculated in the R package 'interactions' version 1.1.5). Heathland size ranges from 5 to 50 hectares in steps of 5 hectares, with colours indicating the three classes: purple = small patches (≤ 10 hectares); yellow = intermediate patches/patch edges (10-30 hectares); green = large patches (> 30 hectares).

Figure M.1. Model predictions for birds (Relative habitat suitability: blue = low, orange = high)

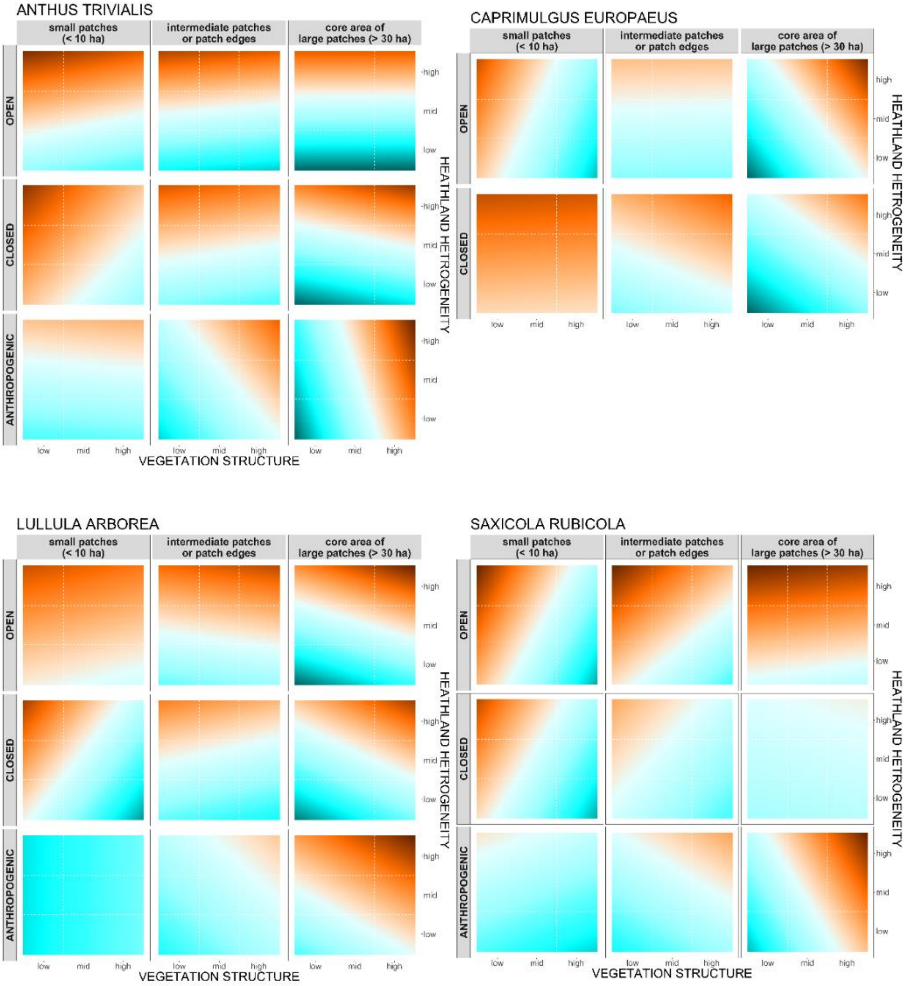


Figure M.2. Model predictions for butterflies (Relative habitat suitability: blue = low, orange = high)

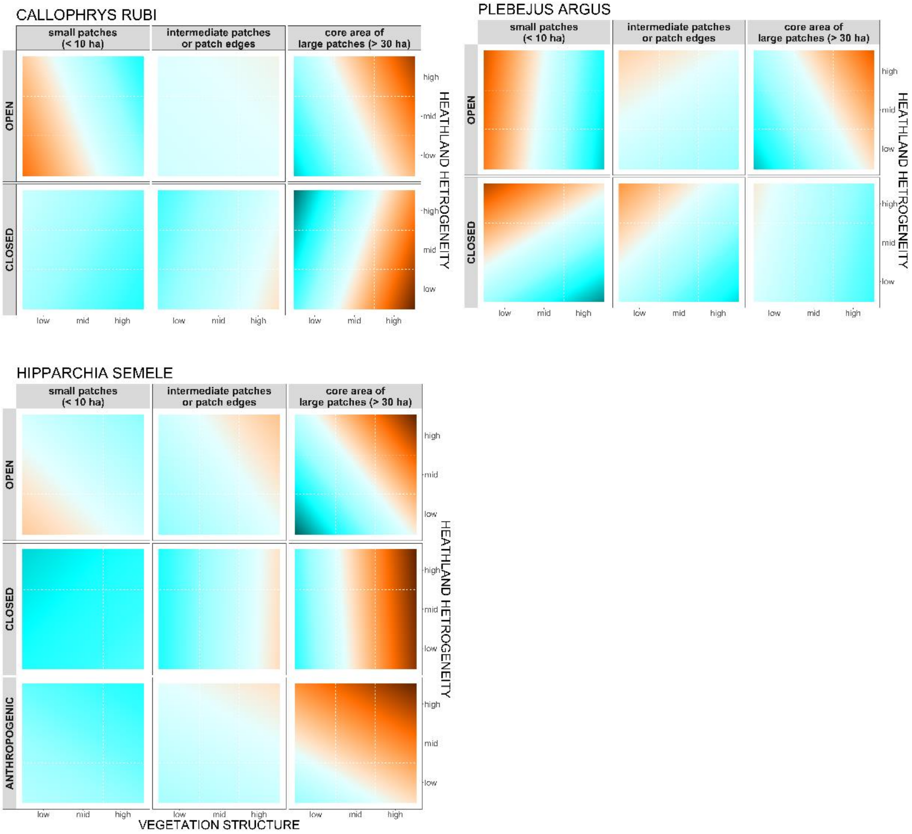


Figure M.3. Model predictions for grasshoppers (Relative habitat suitability: blue = low, orange = high)

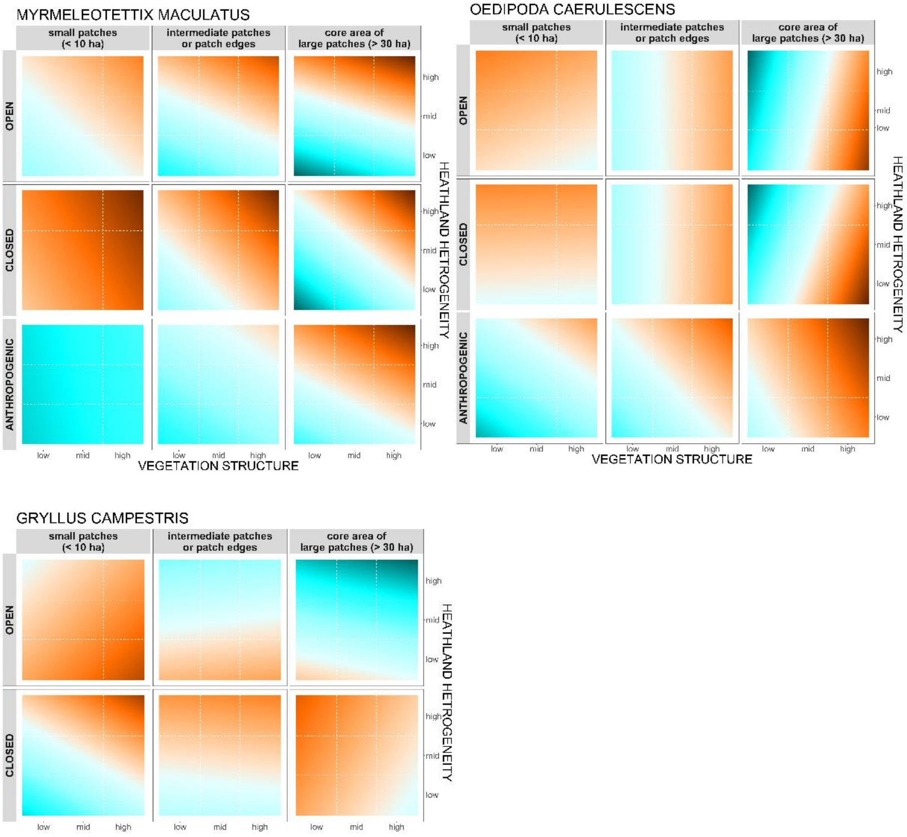


Figure M.4. Model coefficients for birds

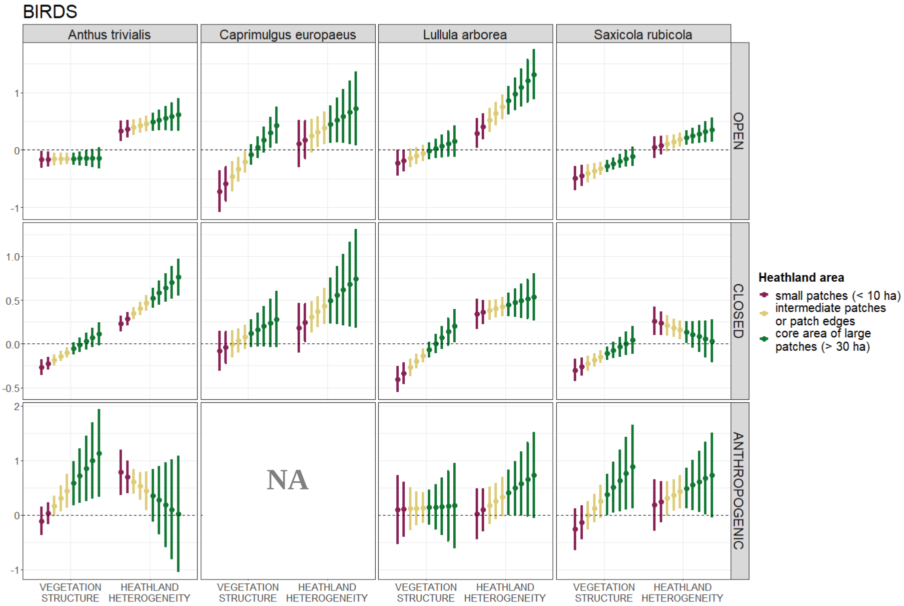


Figure M.5. Model coefficients for butterflies

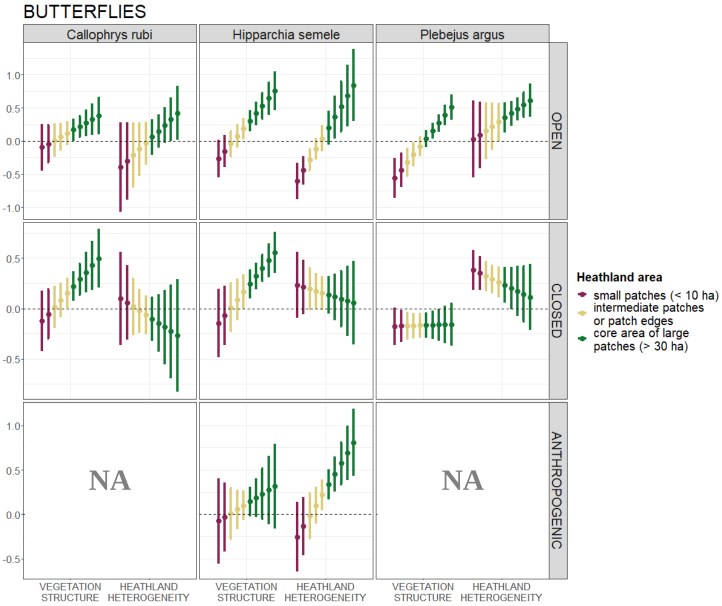
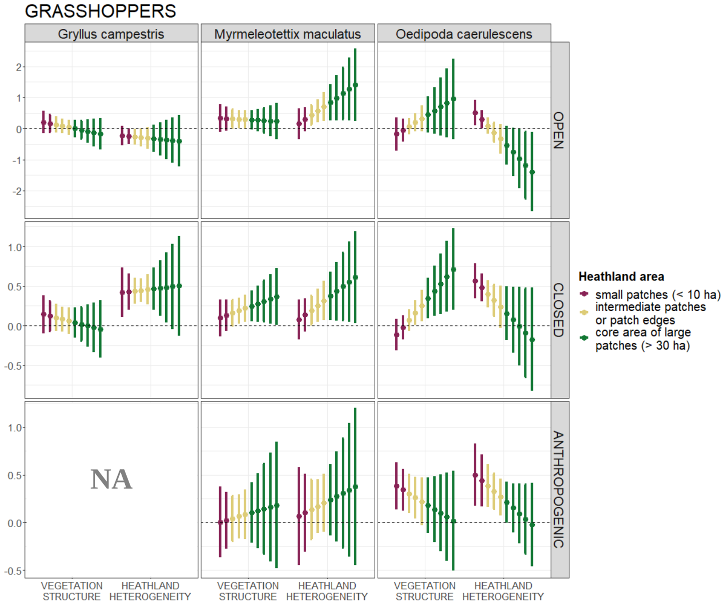


Figure M.6. Model coefficients for grasshoppers



Appendix N: Recommendations for vegetation management per species

Table N.1. Vegetation management recommendations per species to increase habitat suitability in heathlands of different sizes in different landscape contexts (HH = heathland heterogeneity, VS = vegetation structure)

LANDSCAPE CONTEXT:	OPEN	CLOSED	ANTHROPOGENIC
BIRDS			
Anthus trivialis Tree Pipit Boompieper	High HH	High HH Low VS in small patches	High HH in small patches High VS in large patches
Caprimulgus europaeus Nightjar Nachtzwaluw	Low VS in small patches High HH in intermediate patches/patch edges High HH and high VS in large patches	High HH in intermediate patches/patch edges and large patches	NA
Lullula arborea Woodlark Boomleeuwerik	High HH	High HH Low VS in small patches High VS in large patches	High VS and high HH in large patches
Saxicola rubicola European Stonechat Roodborsttapuit	Low VS and high HH in small and intermediate patches/patch edges High HH in large patches	Low VS and high HH in small patches	High VS and high HH in large patches
BUTTERFLIES			
Callophrys rubi Green Hairstreak Groentje	Low VS in small patches High VS in large patches	High VS in large patches	NA
Hipparchia semele Grayling Heivlinder	High VS and HH in large patches	High VS in small patches	High HH in large patches
Plebejus argus Silver-studded Blue Heideblauwtje	Low VS in small patches High VS and HH in large patches	Low VS and high HH in small patches	NA
GRASSHOPPERS			
Gryllus campestris Field cricket Veldkrekel	High VS in small patches Low HH	High HH	NA
Myrmeleotettix maculatus Mottled grasshopper Knopsprietje	High HH in large patches	High HH and high VS in intermediate patches/patch edges and large patches	High HH and high VS in large patches
Oedipoda caerulescens Blue Winged Grasshopper Blauwvleugelsprinkhaan	High VS in large patches	High VS in large patches	High VS and HH, especially in large patches

CHAPTER VI

Appendix O: Application potential in biodiversity conservation policy in Flanders.

Table O.1. Species used in the case studies for illustrating the application potential of the research. The list includes 14 randomly selected Flemish priority species and 19 farmland species, arranged by their IUCN Red list status. The weights were used to generate biodiversity scores, i.e. by multiplying the predicted relative occurrence rate by the weighting factor and summing the result per 500 x 500 m grid in the study area (Flanders). The number of presences is the number of cleansed (removing bad coordinates, wrong observations and observations with a precision of more than 250 metres), spatially thinned and quality filtered (chapters II and III) opportunistic presence-only records retrieved from the *waarnemingen.be* database for the period 2018-2022.

group	*Scientific name*	*Dutch name*	*IUCN Red List Status*	*weight*	Number of presences
14 FLEMISH PRIORITY SPECIES					
Amphibians	Hyla arborea	Boomkikker	CR	80	220
Butterflies	Melitaea cinxia	Veldparelmoervlinder	CR	80	143
Dragonflies	Leucorrhinia pectoralis	Gevlekte witsnuitlibel	CR	80	122
Reptiles	Coronella austriaca	Gladde Slang	EN	50	164
Amphibians	Epidalea calamita	Rugstreeppad	VU	30	375
Amphibians	Rana arvalis	Heikikker	VU	30	241
Amphibians	Salamandra salamandra	Vuursalamander	VU	30	129
Amphibians	Triturus cristatus	Kamsalamander	VU	30	173
Butterflies	Cyaniris semiargus	Klaverblauwtje	VU	30	130
Mammals	Castor fiber	Europese Bever	VU	30	1541
Mammals	Meles meles	Das	VU	30	312
Amphibians	Pelophylax lessonae	Poelkikker	NT	20	471
Breeding Birds	Limosa limosa	Grutto	LC	1	886
Butterflies	Lasiommata megera	Argusvlinder	LC	1	350
19 FARMLAND SPECIES					
Breeding Birds	Circus pygargus	Grauwe Kiekendief	CR	80	191
Breeding Birds	Emberiza calandra	Grauwe gors	CR	80	211
Breeding Birds	Circus aeruginosus	Bruine kiekendief	EN	50	2984
Breeding Birds	Emberiza citrinella	Geelgors	EN	50	2406
Breeding Birds	Passer montanus	Ringmus	EN	50	620
Breeding Birds	Vanellus vanellus	Kievit	EN	50	7287
Breeding Birds	Alauda arvensis	Veldleeuwerik	VU	30	4324
Breeding Birds	Hirundo rustica	Boerenzwaluw	VU	30	7423
Breeding Birds	Perdix perdix	Patrijs	VU	30	3536
Mammals	Meles meles	Das	VU	30	312
Breeding Birds	Motacilla flava	Gele kwikstaart	NT	20	2897
Mammals	Lepus europaeus	Haas	NT	20	3665
Mammals	Microtus subterraneus	Ondergrondse woelmuis	NT	20	109
Breeding Birds	Athene noctua	Steenuil	LC	1	2119
Breeding Birds	Columba palumbus	Houtduif	LC	1	6932
Breeding Birds	Coturnix coturnix	Kwartel	LC	1	1553
Breeding Birds	Falco tinnunculus	Torenvalk	LC	1	10707
Breeding Birds	Haematopus ostralegus	Scholekster	LC	1	4165
Breeding Birds	Tyto alba	Kerkuil	LC	1	482