Preprint submitted to Ecological Modelling. The final publication appeared in vol. 467 (2022) and is available at https://doi.org/10.1016/j.ecolmodel.2022.109910

Species Profiles Support Recommendations for Quality Filtering of Opportunistic Citizen Science Data

Camille Van Eupen*[,a,b], Dirk Maes[c,d], Marc Herremans[e], Kristijn R.R. Swinnen[e], Ben Somers**[,b] and Stijn Luca**[,a]

* corresponding author

** joint last authors

[a] Ghent University, Department of Data Analysis and Mathematical Modelling, Coupure Links 653, B-9000 Ghent, Belgium; camille.vaneupen@kuleuven.be (ORCID 0000-0002-0924-8892); stijn.luca@ugent.be (ORCID 0000-0002-6781-7870)

[b] KU Leuven, Department of Earth and Environmental Sciences, Division Forest Nature and Landscape, Celestijnenlaan 200E, B-3001 Heverlee, Belgium; ben.somers@kuleuven.be (ORCID 0000-0002-7875-107X)

[c] Research Institute for Nature and Forest (INBO), Herman Teirlinckgebouw, Havenlaan 88 box 73, B-1000 Brussels, Belgium; dirk.maes@inbo.be (ORCID 0000-0002-7947-3788)

[d] Radboud Institute for Biological and Environmental Sciences (RIBES), Radboud University, PO Box 9010, NL-6500 GL Nijmegen, The Netherlands; dirk.maes@ru.nl

[e] Natuurpunt Studie, Coxiestraat 11, 2800 Mechelen, Belgium; ; marc.herremans@natuurpunt.be (ORCID: 0000-0002-9719-8732); kristijn.swinnen@natuurpunt.be (ORICD 0000-0002-1910-9247)

Preprint submitted to Ecological Modelling. The final publication appeared in vol. 467 (2022) and is available at https://doi.org/10.1016/j.ecolmodel.2022.109910

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 were 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.

KEYWORDS

data quality filtering, filtering recommendations, opportunistic data, presence-only data, species distribution models, species traits

1. Introduction

Biodiversity conservation needs adequate monitoring of species (Lindenmayer et al., 2020), especially in times of rapid changes to the environment that threaten species and reduce abundances at alarming rates (Newbold et al., 2015; Urban et al., 2016). Structured surveys, where species are recorded in a standardized manner, are commonly put forward as the most desirable strategy for biodiversity monitoring because of their high information content (Dobson et al., 2020). It is challenging, however, to organize structured surveys for a large variety of species over broad spatial and temporal scales, leading to spatial and temporal data gaps (Urban et al., 2016). When information on the potential distributions of species is needed, e.g. for assigning protected areas (Thomaes et al., 2008) or areas of high potential nature value (Maes et al., 2005), spatial gaps can be filled by using species distribution models (SDMs). These models link environmental parameters, such as landscape and climate variables, to species occurrence records (Guisan and Zimmermann, 2000).

Species occurrence records with high information content (e.g. collected in structured surveys) are preferably used as input for SDMs, but as aforementioned, such data are sparse. Therefore, SDMs are increasingly built with data collected by citizen scientists in a semi-structured manner (termed semi-structured data) or in an opportunistic unstructured manner (termed opportunistic data). Large citizen science initiatives that have online data platforms either focus on one data type (e.g. eBird contains only semi-structured data (Sullivan et al., 2009), iNaturalist contains only opportunistic data (https://www.inaturalist.org/)) or they combine both data types (e.g. iRecord (https://www.brc.ac.uk/irecord/), waarnemingen.be (https://www.waarnemingen.be) and Observation.org (https://observation.org)). The majority of opportunistic data consist of species presence records with some basic information such as the date and geographical precision of the observation (termed opportunistic presence-only data). The main advantage of opportunistic data is the availability in large amounts, over large geographical areas and potentially long periods (Kosmala et al., 2016). A major disadvantage, however, is the prevalence of different types of bias and error (Bird et al., 2014; Isaac and Pocock, 2015) caused by a lack of standardised design and accompanying metadata. Opportunistic citizen science data is therefore associated with uncertainty and scepticism towards its use for SDMs (Burgess et al., 2017), and has led to many publications on how and when to use this type of data for biodiversity research (e.g. Henckel et al., 2020; Isaac & Pocock, 2015; Maes et al., 2015; Van Strien et al., 2013). Ideally, the benefits of both unstructured opportunistic data (quantity) and (semi-)structured survey data (information content) are exploited simultaneously. This can be in the form of model-based data integration (for a recent review see Isaac et al., 2020) or when structured data is used as external validation data for SDMs (e.g. Matutini et al., 2021; Van Eupen et al., 2021).

In preparation for an SDM study, data are expected to be cleansed (Zurell et al., 2020). At a minimum, this 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 validation (Steen et al., 2019). Even though drawing direct ecological inferences from opportunistic observations is not recommended (Dobson et al., 2020), clear recommendations on when and how to filter opportunistic data remain sparse (but see e.g. Kamp et al., 2016; Steen et al., 2019; Van Eupen et al., 2021; Vantieghem et al., 2017). The study of Van Eupen et al. (2021) 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 paper, 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 most variation can usually (but not exhaustively) be compiled to 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 on 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 (Van Eupen et al., 2021). 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).

2. Material and methods

2.1. Species data and impact of quality filtering

We used a dataset from a previous study on data quality filtering (Van Eupen et al., 2021, Van Eupen et al., 2021b), where three dichotomous filters were applied to opportunistic species observations belonging to four well-studied taxonomic groups in Flanders, i.e. birds, butterflies, dragonflies and plants. Plant observations were not used in the present analysis because their traits are not directly comparable to animal species traits. Data were collected from the '*waarnemingen.be*' database, an online citizen science platform that contains both semi-structured and opportunistic records (Swinnen et al., n.d.). The selected dataset included year-round records from January 2014 to September 2019 with a geographical precision of 500 metres or smaller (records can be submitted to the platform as point locations with specified precision or as observations made within a larger area). Potential seasonal changes in habitat occupancy or range size were not considered. The selection excluded records verified as incorrect by species experts or the data platform's auto-validation system (Swinnen et al., 2018) and records from non-native or non-breeding birds in Flanders (Vermeersch et al., 2020). Opportunistic presence-only data were used for model training, also excluding absences (zero-counts), and semi-structured data for model testing (Van Eupen et al., 2021). 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.

Table 1: Overview and definitions of the used variables in this study.

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	number of search results retrieved from the Google search engine
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
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		
real	the actual reduction in the number of presences after data quality filtering	
r50	a relative reduction in the number of presences after data quality filtering of more than 50%	
ss100	a reduction to 100 presences after data quality filtering	

After filtering, records were aggregated to a 1x1 km resolution to reduce spatial bias (Kramer-Schadt et al., 2013). Species were selected based on opportunistic data availability for model training as well as semi-structured data availability for model validation (Van Eupen et al., 2021). The impact of filtering on the performance of Maxent (Phillips et al., 2006) was assessed by evaluating the difference in three commonly used evaluation metrics of model discrimination before and after filtering: the area under the receiver operating characteristic (AUC), sensitivity and specificity (Fielding and Bell, 1997). Their use was justified because models were run for the same geographical extent and model predictions were evaluated on a testing set that contained the same presences and absences per species (Jiménez-Valverde, 2012; Lobo et al., 2008). Model performance could therefore be interpreted in a relative manner, where an increase in AUC after filtering implied that the filtered data produced models that could better distinguish between testing presences and absences, and increases in sensitivity and specificity implied a higher predicted positive and negative fraction respectively. 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. 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 the supplementary information 1 (Table C.1) and supplementary information 2 respectively in the study of Van Eupen et al. (2021).

2.2. 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 of time and space. All trait values can be found in the supplementary information (Table S1).

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.

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, supporting 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.

2.3. 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).

2.3.1. 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 et al., 2004). As trait values showed taxonomic differences (Figure 1), 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).

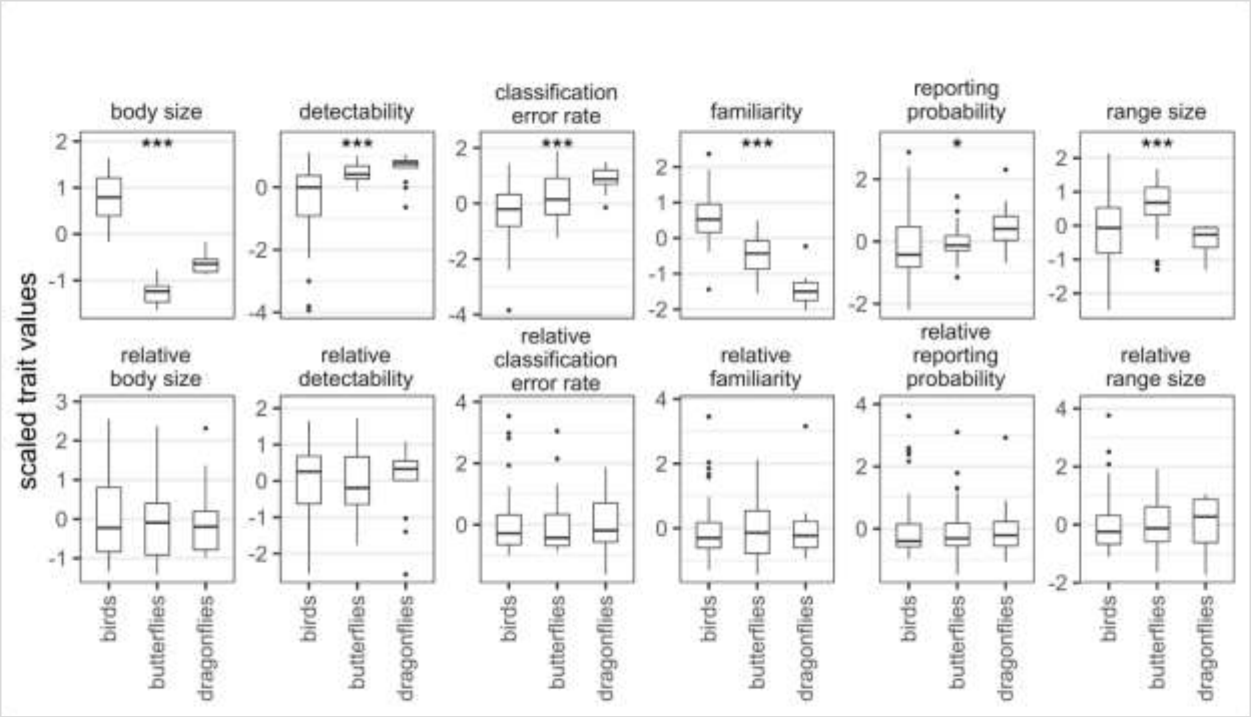

*Figure 1: Summary of the species traits per taxonomic group after value transformation and standardisation. Absolute traits (top row) were rescaled to relative traits (bottom row) per taxonomic group to assess patterns across taxonomic groups. Stars indicate differences in the medians of the trait values between taxonomic groups (*** = p < 0.001, * = p < 0.05).*

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 S1). We also quantified variable importance by leaving out each trait one by one and calculating the decrease in pseudo-R² 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).

2.3.2. Species profiles

To test whether groups of species with similar traits can improve recommendations for data quality filtering, we delineated species profiles. Species were clustered into groups with similar traits using the FactoMiner package v1.34 (Le et al., 2008). This package performs a principal component analysis (PCA) on a set of variables (i.e. species traits) followed by an agglomerative hierarchical clustering of individuals (i.e. species) each described by principal components of those variables (Husson et al., 2010). The active variables, included in the PCA and clustering, were those variables that contributed most to the change in model performance across filters, resulting from the multi-trait regression model selection. 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 2.3.1) 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 experimentally changed the number of clusters before selecting the final profiles, by choosing different heights in the hierarchical tree.

2.3.3. 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 (Van Eupen et al., 2021), 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 **real** situation, 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 **r50** situation, i.e. the filter effects when sample size was reduced by 50% or more and (3) the **ss100** situation, 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).

3. Results

3.1. Multi-trait analysis

Figure 2 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 2a), 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 2b), 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 1 and Figure S1). 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 S1 and section 2.2).

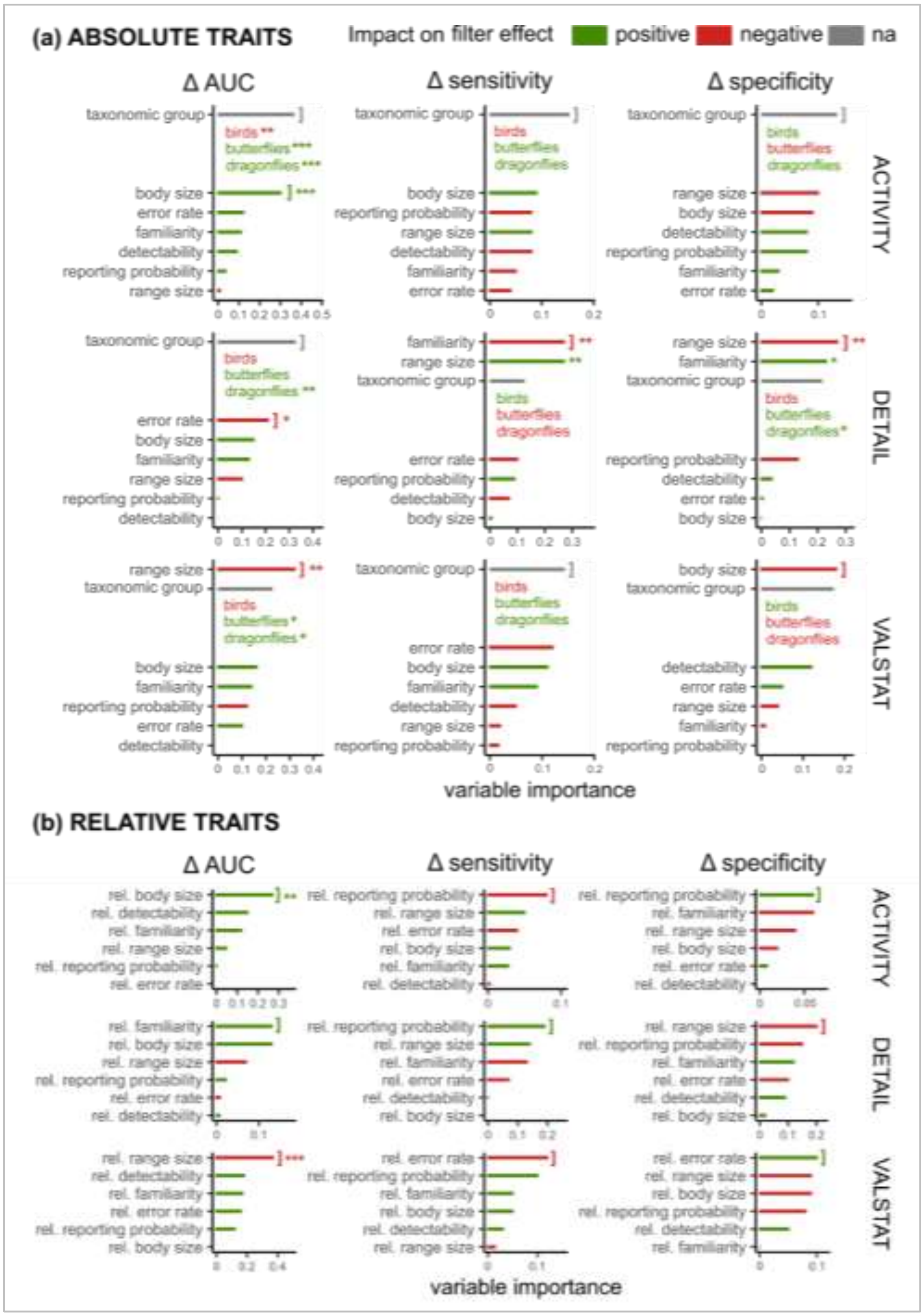

Figure 2: Variable importance in the multi-trait regressions for absolute (a) and relative (b) species traits per filter (ACTIVITY, DETAIL and VALSTAT) and change in model evaluation metric (Δ AUC, Δ sensitivity, Δ specificity). Variable importance is expressed as the square root of the change in pseudo-R² 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 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).*

3.2. 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 best captured the variation in species traits, but the impact on model performance still showed large variation within profiles. Ecologically, these profiles also separated species into quite general groups and we cross-checked the clustering of the species into four or more different profiles with species experts. Five clusters appeared the best outcome while keeping cluster size at a reasonable level (minimum cluster size equalled 7 species) (Table 2). The full results of the PCA and clustering analysis are presented in the supplementary information (Table S2 and Figure S2).

Positive or negative recommendations were only noted when the goal was to increase AUC. When the goal was to increase sensitivity or specificity, recommendations were either cautious or alarming 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).

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 (*** = 0.001, ** = 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*

Absolute traits appeared highly associated with the most represented taxonomic group for four out of five profiles (Table 2 and Figure 1). 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 2). 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 (profile 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).

3.3. 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 3). 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. 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).

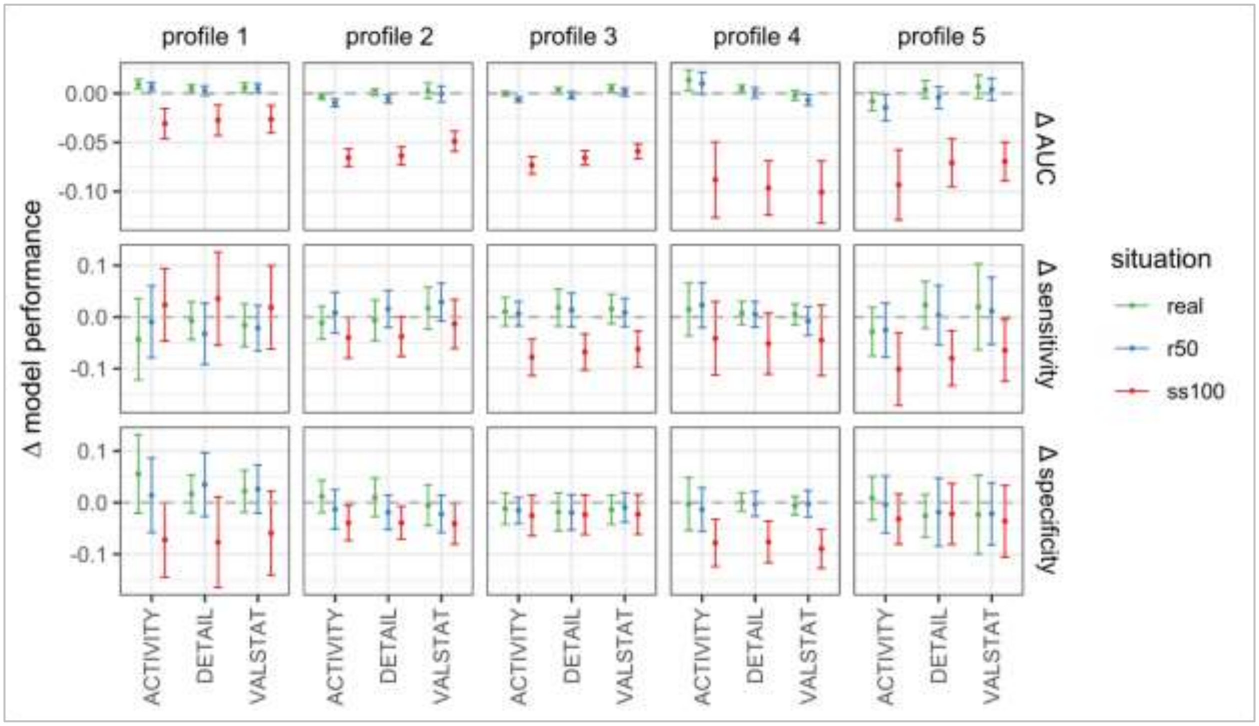

Figure 3: Recommendations for data quality filtering for the five species profiles in the three sample size situations (real = actual reduction in sample size when filtering, r50 = sample size reduced by at least 50%, ss100 = sample size reduced to 100 presences). Dots and error bars are the means and 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 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).

3.4. 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 2). 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 2), 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 real data situation, even at high reductions (Figure 3). 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 2).

Preprint submitted to Ecological Modelling. The final publication appeared in vol. 467 (2022) and is available at https://doi.org/10.1016/j.ecolmodel.2022.109910

4. 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 3.4). The analysis of the species profiles mostly agreed with the results of a regression analysis 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 previous findings based on the same dataset (Van Eupen et al., 2021) 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 study, 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 S1; evaluated by species experts). Moreover, they lead to less consistent results (sections 2.3.1 and 2.3.2) or less explicit filtering recommendations (i.e. larger confidence intervals in Figure 3). This confirms the expectation of Van Eupen et al. (2021) who concluded that filtering recommendations can differ between taxonomic groups, but that there might also be common traits among these groups that can refine them. The selected approach indeed revealed that it is possible to formulate recommendations based on taxonomic group and relative traits only (Table 2 and section 3.4). 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 for identifying an invasive plant species (Crall et al., 2011). Here, observers' self-identified that 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). A decent photograph or sound record can thus easily lead to multiple approved records and, by 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 S1), 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., 2019), 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 (Kramer-Schadt et al., 2013; Van Eupen et al., 2021). 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 (Van Eupen et al., 2021), 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 (Van Eupen et al., 2021).

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). 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 S1). 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 (Van Eupen et al., 2021) 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. heathland) 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 Van Eupen et al. (2021) for those species to be able to formulate generic recommendations. Based on the available data, 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.

Conclusions

Many have attempted to disentangle the relationships between species ecology and model performance, and this study adds to that knowledge with some basic recommendations for data quality filtering for three commonly studied taxonomic groups. Clustering species in 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 largely supported on: (1) the general species knowledge among citizen scientists, with high importance of data quality for widespread and familiar species in general and, more specifically, high importance of observer experience for less known taxonomic groups; and (2) the mechanism of record verification in an opportunistic data platform, with the 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 model discrimination capacity, sensitivity and/or specificity) and to keep an eye on the change in sample size caused by stringent filtering.

ACKNOWLEDGEMENTS

We foremost thank the thousands of volunteers for collecting the millions of records that supported this study; *Natuurpunt Studie* for making the data available for this research, in particular Joeri Belis and Karin Gielen for composing the dataset; and Tim Adriaens for his help with the interpretation of the data. We would also like to thank the anonymous reviewers for kindly improving the first version of this manuscript. This work was supported by the Flemish Research Foundation FWO–SB [grant number 1S92118N].

5. References

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

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

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

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

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

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

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.

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

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

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., Benedikt, |, Schmidt, 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

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.

Demolder, H., Schneiders, A., Spanhove, T., Maes, D., Van Landuyt, W., Adriaens, T., 2014. Hoofdstuk 4 Toestand biodiversiteit (INBO.R.2014.6194611), Natuurrapport - Toestand en trends van de ecosystemen en ecosysteemdiensten in Vlaanderen. Instituut voor Natuur- en Bosonderzoek, Brussels.

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

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

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

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

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

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

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

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.

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.

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.

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.12532

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

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., Gutierrez, V.R., 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., 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

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

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

Le, S., Josse, J., Husson, F., 2008. FactoMineR: An R Package for Multivariate Analysis. J. Stat. Softw. 25, 1–18.

Lindenmayer, D., Woinarski, J., Legge, S., Southwell, D., Lavery, T., Robinson, N., Scheele, B., Wintle, B., 2020. A checklist of attributes for effective monitoring of threatened species and threatened ecosystems. J. Environ. Manage. 262, 110312. https://doi.org/10.1016/j.jenvman.2020.110312

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

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

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.

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., Desmet, P., Piesschaert, F., Van Hoey, S., Adriaens, T., Dekoninck, W., Devos, K., Lock, K., Onkelinx, T., Packet, J., 2019. A database of threat statuses and life-history traits of Red List species in Flanders (northern Belgium). Biodivers. Data J. 7, e34089. https://doi.org/10.3897/BDJ.7.e34089

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

Matutini, F., Baudry, J., Pain, G., Sineau, M., Pithon, J., 2021. How citizen science could improve species distribution models and their independent assessment. Ecol. Evol. 11, 3028–3039. https://doi.org/10.1002/ece3.7210

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., 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.

Miller, D.A., Nichols, J.D., McClintock, B.T., Grant, E.H.C., 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

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

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

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

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

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

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

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

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

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

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

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

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

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

Swinnen, K.R.R., Jacobs, A., Claus, K., Ruyts, S., Vercayie, D., Lambrechts, J., Herremans, M., n.d. 'Animals under wheels': wildlife roadkill data collection by citizen scientists as a part of their nature recording activities. Nat. Conserv.

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

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

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

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

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 (80-.). 353, aad8466. https://doi.org/10.1126/science.aad8466

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

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

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

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

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

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

[dataset] Van Eupen, C., Maes, D., Herremans, M., Swinnen, K.R.R., Somers, B., Luca, S., 2021b. The impact of data quality filtering of opportunistic citizen science data on species distribution model performance: dataset used for Maxent modelling., Dryad, Dataset, https://doi.org/10.5061/dryad.jwstqjq83