Recognizability bias in citizen science photographs

Wouter Koch[1,2*] Laurens Hogeweg[3,4]
Erlend B. Nilsen[5] Robert B. O'Hara[6]
Anders G. Finstad[1]

[1]Department of Natural History, Norwegian University of Science and Technology, Erling Skakkes gate 47b, Trondheim, Norway
[2]Norwegian Biodiversity Information Centre, Havnegata 9, 7010 Trondheim, Norway
[3]Intel Benelux, High Tech Campus 83, 5656 AE Eindhoven, The Netherlands
[4]Naturalis Biodiversity Center, PO Box 9517, 2300 RA, Leiden, The Netherlands
[5]Norwegian Institute for Nature Research, Postboks 5685 Torgarden, 7485 Trondheim, Norway
[6] Department of Mathematical Sciences, Norwegian University of Science and Technology, Alfred Getz' vei 1, 7034 Trondheim, Norway

*Corresponding author. E-mail: wouter.koch@artsdatabanken.no
Contributing authors: laurens.hogeweg@naturalis.nl erlend.nilsen@nina.no bob.ohara@ntnu.no anders.finstad@ntnu.no

Abstract

Citizen science initiatives and automated collection methods increasingly depend on image recognition in order to provide the amounts of observational data research and management needs. Training recognition models, meanwhile, also requires large amounts of data from these sources, creating a feedback loop between the methods and the tools. Species that are harder to recognize, both for humans and machine learning algorithms, are likely to be underreported, and thus be less prevalent in the training data. As a result, the feedback loop may hamper training mostly for species that already pose the greatest challenge. In this study, we trained recognition models for various taxa, and found evidence for a *"recognizability bias"*, where species that models struggle with are also generally underreported. This has implications for the kind of performance one can expect from future models that are trained with more data, including such challenging species. We consider identification methods that rely on more than photographs alone to be important in improving future identification tools.

Introduction

There is an ever growing need for large amounts of biodiversity observation data. With an increasing awareness of the multiple crises biodiversity faces [1–3], substantial amounts of such data are essential if humanity is to monitor trends and address these issues [4–6]. Occurrence data are typically subject to spatial, temporal and taxonomic bias [7, 8], and traditional manual methods of data collection are insufficient to gather the data volume needed, or address these biases. Alternative data collection methods, ranging from citizen science (non-professional volunteers reporting observations [9]) to camera-traps automating insect monitoring [10, 11] are being deployed to gather large amounts of data. With the increased output from such initiatives, manual management and quality control become infeasible. Automated image recognition tools for species identification are increasingly used to facilitate this [12–15]. Training image recognition models, however, also requires large amounts of pictures [16]. This creates a mutual reliance between large scale image data collection and image recognition models [17].

Visual identification of species is a complex task, and taxa vary in their recognizability; while some species are unmistakable, many others are very challenging or even outright impossible to identify, regardless of picture quality [18]. As models are trained using training data reported and identified by humans, species with low recognizability among humans will be underreported and be underrepresented in the training data. This affects recognition models, as these are then being trained with data biased towards higher recognizability, consisting mostly of pictures of species that are easier to recognize. If this is the case, training models will be hampered not only by the lower recognizability of particularly challenging species, but also by their higher absence from the training data.

To evaluate the existence of this possible bias and its consequences, we evaluated how data availability, picture quality, biological traits and data collection differs across species within 3 orders of birds, and how these differences relate to recognition model performance. All data came from a large Norwegian citizen science project, where recognition tools are not a part of the reporting or validation process. Birds are the most well-represented orders per species, allowing for the most detailed analysis. We also trained models for 9 other orders of plants, animals and fungi, to test for a general correlation between data availability and model performance, and to evaluate what this means for future recognition models.

We find evidence for a *"recognizability bias"*, where species that are more readily identified by humans and recognition models alike are more prevalent in the available image data. This pattern is present across multiple taxa, and does not appear to relate to a difference in picture quality, biological traits, or data collection metrics other than recognizability.

Methods

We trained image recognition models using convolutional neural networks on pictures retrieved from the Norwegian citizen science platform Species Observation Service [19] for 12 orders: Agaricales, Anseriformes, Asparagales, Asterales, Charadriiformes, Coleoptera, Diptera, Lecanorales, Lepidoptera, Odonata, Passeriformes, and Polyporales [20]. A separate model was trained for each order, using 200 documented observations per species for training and validation, and a minimum of 20 for the test set. See Koch *et al.* [21] for details. From these models and various external datasets, several relevant metrics were collected (table 1).

Metric	**Definition**
Data availability	The total number of citizen science observations from the Norwegian citizen science platform Species Observation Service [19] for a species, containing one or more pictures. This is a more meaningful measure than simply the total number of pictures, as multiple pictures within an observation are not independent from one another and therefore do not add as much information as unique observations.
F_1-score	The performance obtained for a species in a recognition model, defined as the harmonic mean of the precision and recall [16]

Species in Norway	The number of species within an order that are present in Norway, according to the Norwegian Species Nomenclature Database [22].

Table 1: Metrics collected for species within all orders

More detailed analyses were done on the included bird orders; waterfowl (Anseriformes), shorebirds (Charadriiformes), and passerines (Passeriformes), as bird orders have the highest proportion of species in Norway represented in the dataset, and ample standardized available data on a range of biological traits allowing for a deeper analysis. For these analyses, a number of additional metrics were collected for the included bird species (table 2).

Metric	**Definition**
Picture quality	Using Label Studio v1.4 [23], ≥50 pictures per species were annotated by drawing rectangles approximately equal in surface area to the visible part of each individual bird. From this, we took the percentage of the picture occupied by the largest depiction of an individual of the target species, minus the percentage of the picture occupied by all individuals of other bird species. Per species, the median log value was used as a proxy for picture quality.
Urbanness	The proportion of 100 documented observations from the Species Observation Service with a location within a cell tagged as "urban" in the ESA CCI landcover dataset [24].
Hand-wing index	Wing length minus wing width, a measure positively correlated with flight efficiency and dispersal ability of a species. Retrieved from the Global-HWI dataset [25].
Body mass	The average log-transformed body mass of a species, retrieved from the Global-HWI dataset [25].
Habitat openness	A three-step scale of the openness of the habitat of a species, retrieved from the Global-HWI dataset [25].
Documentation rate	The proportion, per species, of observations in the Species Observation Service that have one or more pictures.
Picture density	The average number of pictures per observation from the Species Observation Service, from those with at least one picture.
Observation rate	The number of observations in the Species Observation Service dataset per observation in the TOV-e bird monitoring scheme [26]

Table 2: Metrics collected for species within the bird orders

LASSO multiple regression models were trained using Scikit-learn [27] to evaluate the effect of the biological traits, picture quality measurement, and data collection process from table 2 on the F_1-scores for birds. All LASSO models have the order as a factor. The full model for biological traits is given by

$$F_1 = \beta_0 + \beta_1 HWI + \beta_2 BM + \beta_3 H + \beta_4 U + \beta_5 DA + \epsilon + (1|Order)$$

where HWI is the hand-wing index, BM is the body mass, H is the habitat openness, U is the urbanness, and DA is the log data availability. The full model for picture quality is given by

$$F_1 = \beta_0 + \beta_1 Q + \beta_2 DA + \epsilon + (1|Order)$$

where Q is the picture quality, and DA is the log data availability. The full model for data collection parameters is given by

$$F_1 = \beta_0 + \beta_1 OR + \beta_2 DR + \beta_3 PD + \beta_4 DA + \epsilon + (1|Order)$$

where OR is the observation rate, DR is the documentation rate, PD is the picture density, and DA is the log data availability.

Results

There is a strong positive linear correlation between log data availability and the F_1-score for bird species (figure 1). Note that data availability does not affect training, as all models were trained and evaluated using 220 documented observations per species, regardless of the total availability. A positive linear correlation was also evident in 7 of the 9 other orders (figure 2), in particular Asterales and Odonata. The beetles (Coleoptera) and lichens (Lecanorales) exhibited no apparent correlation, with an R^2 of 0.06 and 0.12, and P-values of 0.27 and 0.18, respectively.

In each bird order, there is a linear relationship between species' picture density and documentation rate ($R^2 \geq 0.52$, $p \leq 1.51\times10^{-7}$, see table S2). We also find a negative linear correlation between picture density and F_1-scores ($R^2 \geq 0.23$, $p \leq 2.1\times10^{-4}$, see table S2), and some negative linear correlation between documentation rate and F_1-scores ($R^2 \geq 0.11$, $p \leq 4.64\times10^{-3}$, see table S2). For passerines, there is a negative linear relationship between habitat openness and picture quality ($R^2 = 0.26$, $p = 3.53\times10^{-8}$, see table S2). Waterfowl and shorebirds could not be evaluated as they only occur in open habitats.

LASSO models trained on biological traits, collection process parameters, and picture quality, all having and log data availability as an additional parameter and order as a factor, had R^2 values of 0.60, 0.57 and 0.63, respectively.

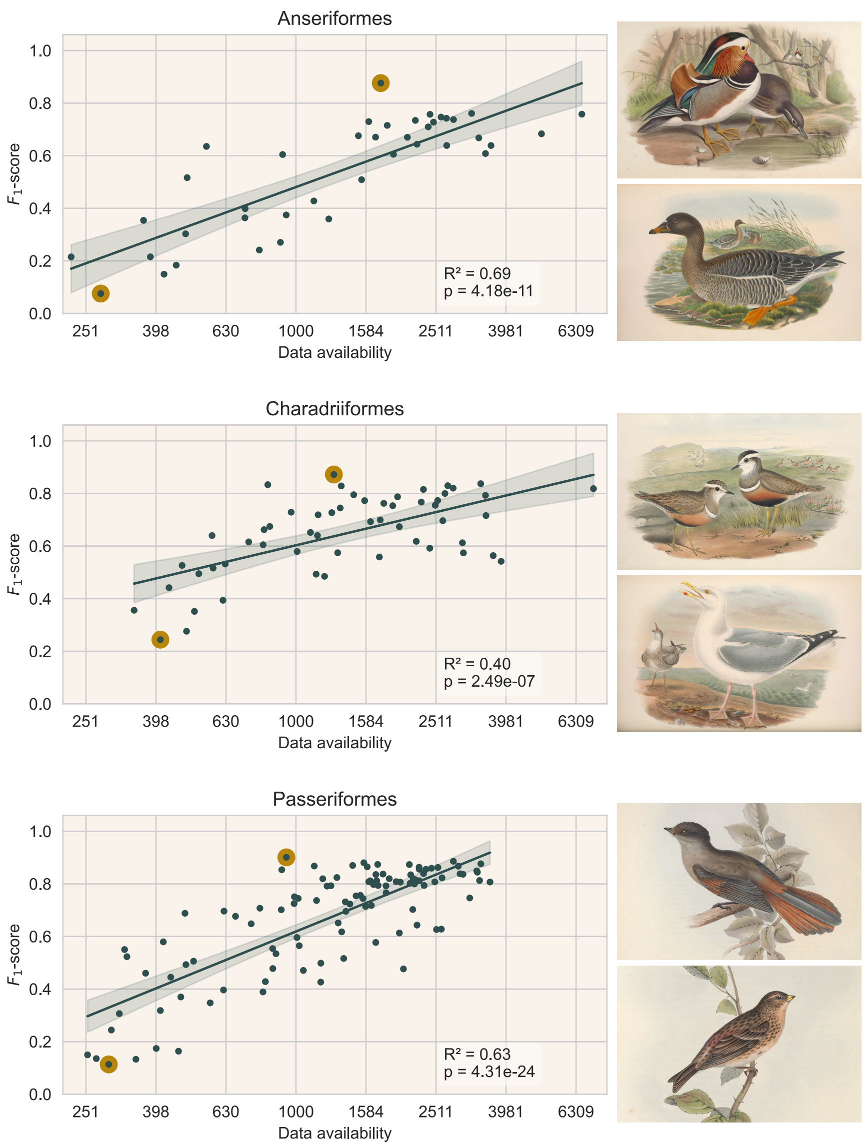

Figure 1: Effect of the total data availability per species on their F_1-scores, in models trained with 200 documented observations, for three bird orders. The top- and bottom-performing species per order (highlighted dots) are depicted, see table S1. Regressions are Ordinary Least Squares with 95% confidence intervals.

With that, none of the full model performances were substantial improvements from a LASSO model with log data availability as its only parameter (R^2 = 0.57).

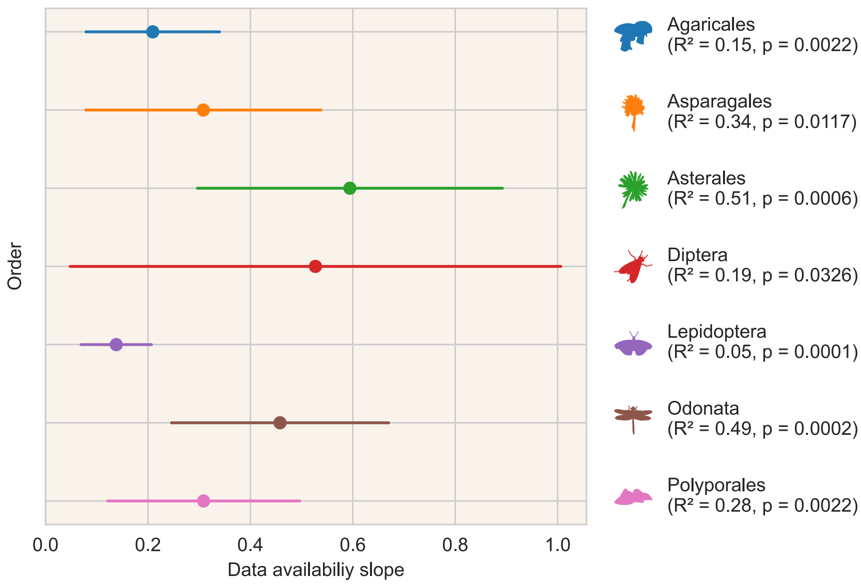

Figure 2: The slopes of the correlations between total data availability per species and their F_1-scores, in models trained with 200 documented observations, for non-bird orders with a correlation $p<0.05$. Regressions are Ordinary Least Squares, lines indicate the 95% confidence intervals.

Discussion

We find a conspicuous pattern where recognition models attain higher performances for species that are reported with pictures more frequently. It is probable that the recognizability of the species influences both their likelihood of being reported with pictures, as well as recognition model performances. The citizen science project used as a data source here does not include any recognition tools in its reporting or validation process, allowing a distinction between human and algorithm recognition biases. Unmistakable species can be recognized and reported by more citizen scientists, resulting in greater data availability for such species. A recognition model, dealing with the same information as human observers, is also proportionally more likely to reliably recognize these species.

This is supported by a qualitative comparison between species with the highest and lowest recognition model performances, where easy to recognize, characteristic species are reported more often than hard to recognize species (e.g. nondescript species or species similar to other related species) (see figure 1 and table S1). Further support comes from the fact that most of the correlation is explained by the data availability for a species, rather than the documentation rate or the picture density. Thus, there is more data available mainly when a species is recognized and reported more, rather than it being disproportionately more likely to be reported with pictures, or with many pictures when reported with pictures.

An alternative explanation to recognizability for increased model performance might be a difference in the kind of pictures, but we find no evidence for this. Species traits, habitat use, and image quality could affect recognition model performance if pictures of more photographed birds are taken more up close, with higher zoom, or were cropped more. We found no evidence, however, for a link between model performance and either picture quality or biological traits in birds. For the passerines, where habitat openness varies among species, we do find that picture quality decreases for species associated with more open habitats. It makes intuitive sense that birds in open habitats are photographed from a greater distance than their forest dwelling counterparts, which will be hidden from view unless in close proximity. While this intercorrelation supports the validity of the picture quality metric, neither habitat nor picture quality affect recognition model performance. We conclude that differences in model performance are caused by the recognizability of the species, rather than by how, or how large species are generally depicted.

Since multiple pictures connected to a single observation are not truly independent, training data are generated based on the number of documented observations, rather than the total number of pictures. One might expect that species with a higher picture density will perform better, as observations with more pictures can provide some additional information in the training process. We find a reverse effect however, where performance for such species is substantially lower. A likely explanation is that species with high picture densities are rarities in Norway (e.g. the top 3 species being Caspian gull, Blyth's reed warbler, and Pine bunting). Species with the lowest picture density, meanwhile, are typical common, well-known species such as corvids and titmice. Rarities are reported not because they are easy to find or identify by casual observers, but due to their popularity among avid birdwatchers, who are likely to document their observations. A strong correlation between picture density and documentation rate supports this; rarities are more often reported with pictures, and in such cases relatively often with several pictures.

While we investigated the bird orders in detail, the link between data availability and model performance is present in other orders too (figure 2). Some orders are notoriously difficult to identify to species level, e.g. flies (Diptera) and beetles (Coleoptera), but our models for these perform surprisingly well. The list of species with sufficient observations with pictures for inclusion in the experiment reveals that only relatively easy to recognize species, often with distinct colorations (e.g. ladybugs for beetles) are represented in this subset.

More generally, the requirement that species must have at least 220 citizen science observations with pictures generates a non-random subset of species, and it differs greatly per order how selective this criterion is. Bird species are most frequently reported; 48% of the species present in Norway [22] within the bird orders examined here meet the selection criterion. One of the other orders for which the pattern was found, the dragonflies and damselflies (Odonata), have only 52 species in Norway, of which 44% met the criteria for inclusion. This is in stark contrast to the beetles (1% inclusion), and lichens (2% inclusion), where no clear correlation is found. It is reasonable to assume that for these taxa, the experiment only considers the most recognizable species. If observations were thousandfold, more challenging species could be included, giving a broader range in performances and possibly a similar positive correlation between model performance and data availability.

The consequence of the recognizability bias found here is that as more data is collected, ultimately providing the numbers of pictures needed to train models also on less reported, harder to recognize species, current performance of recognition models cannot be extrapolated to these expanded models. In other words, data that are lacking now are in part lacking because such species are harder to recognize. When such data is added in the future, the performance increase will not be as great as in the past. Besides citizen science, even methods that have no inherent reporting bias, such as automated insect camera traps and trail cameras, can still be subject to recognizability bias. There too, species that are less readily identified will result in more unidentifiable pictures, providing relatively less training data.

Image recognition tools play an important role in maintaining the quality of the large amounts of biodiversity data science and management require. There are limits to what can be identified from a picture however, and identification tools are needed that rely on more than just pixel information. Models that take into account season, location, sound, etc. can be especially beneficial for difficult species. Still, there is no substitute for the taxonomic knowledge of experts. Preserving this knowledge, and making it available in the form of identification keys is vital. These can be powerful tools to more reliably identify challenging species, in tandem with automatic identification.

Acknowledgements

We are grateful to Rune Sørås, Ingeborg H. Bringslid, and Rienk W. Fokkema for their help in annotating pictures.

Data accessibility

All code is available through Zenodo at XXXX. Bird illustrations in figure 1 are works in the Public Domain made by John Gould (1804-1881), obtained through the Biodiversity Heritage Library [28–30]

Authors' contributions

WK: conception, experimental design, code, analysis, writing. LH: code, text revision. EBN: conception, text revision. RBOH: analysis, text revision. AGF: conception, analysis, text revision.

References

[1] IPBES. 2022 *Thematic assessment of the sustainable use of wild species of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. Bonn, Germany: IPBES Secretariat. (doi:10.5281/zenodo.6448567).

[2] IPCC. 2022 *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. In Press.

[3] Secretariat of the Convention on Biological Diversity. 2020. Global biodiversity outlook 5.

[4] Xu H, Cao Y, Yu D, Cao M, He Y, Gill M, Pereira HM. 2021 Ensuring effective implementation of the post-2020 global biodiversity targets. *Nature Ecology & Evolution* **5**, 4, 411–418. (doi:10.1038/s41559-020-01375-y).

[5] Wetzel FT, Bingham HC, Groom Q, Haase P, Kõljalg U, Kuhlmann M, Martin CS, Penev L, Robertson T, Saarenmaa H, *et al.* 2018 Unlocking biodiversity data: Prioritization and filling the gaps in biodiversity observation data in europe. *Biological Conservation* **221**, 78–85. (doi:10.1016/j.biocon.2017.12.024).

[6] Scholes RJ, Mace GM, Turner W, Geller GN, Jürgens N, Larigauderie A, Muchoney D, Walther BA, Mooney HA. 2008 Toward a global biodiversity observing system. *Science* **321**, 5892, 1044–1045. (doi:10.1126/science.1162055).

[7] Boakes EH, McGowan PJK, Fuller RA, Chang-qing D, Clark NE, O'Connor K, Mace GM. 2010 Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biology* **8**, 6, e1000385. (doi:10.1371/journal.pbio.1000385).

[8] Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F. 2017 Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* **7**, 1. (doi:10.1038/s41598-017-09084-6).

[9] Silvertown J. 2009 A new dawn for citizen science. *Trends in Ecology & Evolution* **24**, 9, 467–471. (doi:10.1016/j.tree.2009.03.017).

[10] Hansen OLP, Svenning JC, Olsen K, Dupont S, Garner BH, Iosifidis A, Price BW, Høye TT. 2019 Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecology and Evolution* **10**, 2, 737–747. (doi:10.1002/ece3.5921).

[11] Kirkeby C, Rydhmer K, Cook SM, Strand A, Torrance MT, Swain JL, Prangsma J, Johnen A, Jensen M, Brydegaard M, *et al.* 2021 Advances in

automatic identification of flying insects using optical sensors and machine learning. *Scientific Reports* **11**, 1. (doi:10.1038/s41598-021-81005-0).

[12] Christin S, Hervet É, Lecomte N. 2019 Applications for deep learning in ecology. *Methods in Ecology and Evolution* **10**, 10, 1632–1644. (doi:10.1111/2041-210x.13256).

[13] Weinstein BG. 2017 A computer vision for animal ecology. *Journal of Animal Ecology* **87**, 3, 533–545. (doi:10.1111/1365-2656.12780).

[14] Wäldchen J, Rzanny M, Seeland M, Mäder P. 2018 Automated plant species identification—trends and future directions. *PLOS Computational Biology* **14**, 4, e1005993. (doi:10.1371/journal.pcbi.1005993).

[15] Ceccaroni L, Bibby J, Roger E, Flemons P, Michael K, Fagan L, Oliver JL. 2019 Opportunities and risks for citizen science in the age of artificial intelligence. *Citizen Science: Theory and Practice* **4**, 1. (doi:10.5334/cstp.241).

[16] Goodfellow I, Bengio Y, Courville A. 2016 *Deep Learning*. MIT Press.

[17] Lotfian M, Ingensand J, Brovelli MA. 2021 The partnership of citizen science and machine learning: Benefits, risks, and future challenges for engagement, data collection, and data quality. *Sustainability* **13**, 14, 8087. (doi:10.3390/su13148087).

[18] Lukhtanov VA. 2019 Species delimitation and analysis of cryptic species diversity in the XXI century. *Entomological Review* **99**, 4, 463–472. (doi:10.1134/s0013873819040055).

[19] The Norwegian Biodiversity Information Centre. 2022. Norwegian species observation service. (doi:10.15468/zjbzel).

[20] GBIForg. 2021. Gbif occurrence download. (doi:10.15468/DL.TC4W55).

[21] Koch W, Hogeweg L, Nilsen EB, Finstad AG. 2022 Maximizing citizen scientists' contribution to automated species recognition. *Scientific Reports* **12**, 1. (doi:10.1038/s41598-022-11257-x).

[22] Norwegian Biodiversity Information Centre. 2021. Species nomenclature database.

[23] Tkachenko M, Malyuk M, Holmanyuk A, Liubimov N. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

[24] ESA. 2017 Land cover cci product user guide version 2. *Tech. Rep.* .

[25] Sheard C, Neate-Clegg MHC, Alioravainen N, Jones SEI, Vincent C, MacGregor HEA, Bregman TP, Claramunt S, Tobias JA. 2020 Ecological drivers of global gradients in avian dispersal inferred from wing morphology. *Nature Communications* **11**, 1. (doi:10.1038/s41467-020-16313-6).

[26] Kålås JA, Øien IJ, Stokke B, Vang R. 2022. Tov-e bird monitoring sampling data. version 1.6. (doi:10.15468/6jmw2e).

[27] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, *et al.* 2011 Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830.

[28] Gould E, Gould J, Lear E. 1837 *The Birds of Europe*. London, Printed by R. and J.E. Taylor, published by the author. (doi:10.5962/bhl.title.65989).

[29] Gould J, Sharpe RB, Richter HC, Hart WM, Wolf J, Francis T, Walton H. 1850 *The Birds of Asia*. London, Printed by Taylor and Francis, published by the author. (doi:10.5962/bhl.title.54727).

[30] Gould J, Wolf J, Richter HC, Hart WM, Walter of England, Archbishop of Palermo. 1873 *The birds of Great Britain*. London, Printed by Taylor and Francis, published by the author. (doi:10.5962/bhl.title.127814).

Supplementary information

Order	**Species**	**F_1-score**
Passeriformes	*Perisoreus infaustus*	0.901
Passeriformes	*Cinclus cinclus*	0.886
Passeriformes	*Periparus ater*	0.88
Passeriformes	*Bombycilla garrulus*	0.876
Anseriformes	*Aix galericulata*	0.876
Passeriformes	*Certhia familiaris*	0.874
Passeriformes	*Aegithalos caudatus*	0.873
Charadriiformes	*Charadrius morinellus*	0.872
Passeriformes	*Regulus regulus*	0.87
Passeriformes	*Lophophanes cristatus*	0.868
Passeriformes	*Emberiza citrinella*	0.867
Passeriformes	*Garrulus glandarius*	0.865
Passeriformes	*Pyrrhula pyrrhula*	0.863
Passeriformes	*Pinicola enucleator*	0.863
Passeriformes	*Cyanistes caeruleus*	0.86
Passeriformes	*Sitta europaea*	0.856
Passeriformes	*Turdus merula*	0.855
Passeriformes	*Phylloscopus sibilatrix*	0.854
Passeriformes	*Coccothraustes coccothraustes*	0.85
Passeriformes	*Carduelis carduelis*	0.845
Passeriformes	*Motacilla cinerea*	0.839
Passeriformes	*Erithacus rubecula*	0.838
Passeriformes	*Parus major*	0.837
Charadriiformes	*Haematopus ostralegus*	0.837
Passeriformes	*Motacilla alba*	0.837
Passeriformes	*Prunella modularis*	0.836
Passeriformes	*Lanius collurio*	0.835
Charadriiformes	*Phalaropus lobatus*	0.834
Charadriiformes	*Calidris maritima*	0.83
Charadriiformes	*Arenaria interpres*	0.829
Passeriformes	*Phylloscopus inornatus*	0.824
Passeriformes	*Saxicola rubicola*	0.823
Charadriiformes	*Charadrius hiaticula*	0.82
Passeriformes	*Sylvia atricapilla*	0.82
Passeriformes	*Turdus philomelos*	0.819
Charadriiformes	*Gallinago gallinago*	0.819
Passeriformes	*Luscinia svecica*	0.818
Passeriformes	*Plectrophenax nivalis*	0.816
Charadriiformes	*Tringa totanus*	0.815
Passeriformes	*Lanius excubitor*	0.813

Passeriformes	*Turdus viscivorus*	0.812
Passeriformes	*Fringilla coelebs*	0.812
Passeriformes	*Nucifraga caryocatactes*	0.812
Passeriformes	*Passer montanus*	0.808
Passeriformes	*Turdus iliacus*	0.808
Passeriformes	*Emberiza schoeniclus*	0.808
Passeriformes	*Oenanthe oenanthe*	0.807
Passeriformes	*Saxicola rubetra*	0.807
Passeriformes	*Chloris chloris*	0.802
Charadriiformes	*Calidris alpina*	0.8
Passeriformes	*Anthus petrosus*	0.8
Passeriformes	*Motacilla flava*	0.798
Charadriiformes	*Cepphus grylle*	0.795
Passeriformes	*Fringilla montifringilla*	0.794
Passeriformes	*Troglodytes troglodytes*	0.794
Charadriiformes	*Vanellus vanellus*	0.793
Passeriformes	*Carpodacus erythrinus*	0.793
Passeriformes	*Turdus torquatus*	0.793
Passeriformes	*Eremophila alpestris*	0.792
Charadriiformes	*Actitis hypoleucos*	0.788
Charadriiformes	*Pluvialis apricaria*	0.773
Charadriiformes	*Tringa glareola*	0.773
Charadriiformes	*Limosa lapponica*	0.767
Passeriformes	*Ficedula hypoleuca*	0.766
Charadriiformes	*Charadrius dubius*	0.762
Anseriformes	*Cygnus olor*	0.761
Anseriformes	*Mergellus albellus*	0.758
Anseriformes	*Clangula hyemalis*	0.757
Passeriformes	*Muscicapa striata*	0.756
Charadriiformes	*Calidris pugnax*	0.755
Passeriformes	*Phoenicurus ochruros*	0.754
Charadriiformes	*Tringa nebularia*	0.754
Passeriformes	*Acrocephalus schoenobaenus*	0.751
Anseriformes	*Branta leucopsis*	0.747
Passeriformes	*Sturnus vulgaris*	0.746
Charadriiformes	*Limosa limosa*	0.745
Passeriformes	*Calcarius lapponicus*	0.745
Passeriformes	*Phoenicurus phoenicurus*	0.744
Anseriformes	*Mergus merganser*	0.742
Anseriformes	*Bucephala clangula*	0.738
Passeriformes	*Curruca communis*	0.737
Anseriformes	*Tadorna tadorna*	0.734
Passeriformes	*Poecile montanus*	0.732

Anseriformes	*Melanitta fusca*	0.73
Charadriiformes	*Tringa erythropus*	0.729
Anseriformes	*Anas acuta*	0.728
Charadriiformes	*Calidris minuta*	0.727
Passeriformes	*Poecile palustris*	0.725
Passeriformes	*Passer domesticus*	0.723
Charadriiformes	*Calidris temminckii*	0.719
Passeriformes	*Hirundo rustica*	0.718
Charadriiformes	*Numenius arquata*	0.716
Anseriformes	*Mareca penelope*	0.715
Passeriformes	*Corvus frugilegus*	0.714
Anseriformes	*Mergus serrator*	0.71
Passeriformes	*Sylvia curruca*	0.708
Passeriformes	*Turdus pilaris*	0.703
Passeriformes	*Pica pica*	0.702
Charadriiformes	*Uria aalge*	0.699
Charadriiformes	*Chroicocephalus ridibundus*	0.697
Passeriformes	*Hippolais icterina*	0.696
Passeriformes	*Loxia leucoptera*	0.696
Charadriiformes	*Calidris canutus*	0.693
Passeriformes	*Emberiza pusilla*	0.688
Anseriformes	*Cygnus cygnus*	0.683
Passeriformes	*Panurus biarmicus*	0.677
Anseriformes	*Somateria spectabilis*	0.676
Charadriiformes	*Alca torda*	0.675
Charadriiformes	*Rissa tridactyla*	0.674
Anseriformes	*Branta canadensis*	0.671
Anseriformes	*Aythya fuligula*	0.67
Anseriformes	*Anas platyrhynchos*	0.668
Charadriiformes	*Alle alle*	0.663
Charadriiformes	*Calidris alba*	0.652
Passeriformes	*Alauda arvensis*	0.651
Passeriformes	*Corvus monedula*	0.648
Anseriformes	*Anas crecca*	0.644
Passeriformes	*Phylloscopus collybita*	0.643
Charadriiformes	*Pluvialis squatarola*	0.64
Charadriiformes	*Fratercula arctica*	0.64
Anseriformes	*Somateria mollissima*	0.639
Anseriformes	*Anser anser*	0.639
Anseriformes	*Anser indicus*	0.635
Passeriformes	*Acanthis flammea*	0.628
Passeriformes	*Anthus pratensis*	0.627
Charadriiformes	*Sterna hirundo*	0.618

Passeriformes	*Corvus cornix*	0.618
Charadriiformes	*Stercorarius parasiticus*	0.616
Passeriformes	*Phylloscopus trochilus*	0.613
Charadriiformes	*Larus hyperboreus*	0.613
Anseriformes	*Anser brachyrhynchus*	0.608
Anseriformes	*Aythya marila*	0.605
Charadriiformes	*Calidris ferruginea*	0.605
Anseriformes	*Aythya ferina*	0.605
Passeriformes	*Corvus corax*	0.596
Charadriiformes	*Larus fuscus*	0.592
Charadriiformes	*Tringa ochropus*	0.58
Passeriformes	*Locustella naevia*	0.579
Passeriformes	*Carduelis spinus*	0.577
Charadriiformes	*Numenius phaeopus*	0.575
Charadriiformes	*Larus canus*	0.574
Passeriformes	*Carduelis flavirostris*	0.565
Charadriiformes	*Larus glaucoides*	0.564
Charadriiformes	*Larus marinus*	0.559
Passeriformes	*Anthus trivialis*	0.554
Passeriformes	*Regulus ignicapilla*	0.55
Charadriiformes	*Larus argentatus*	0.542
Passeriformes	*Riparia riparia*	0.534
Charadriiformes	*Scolopax rusticola*	0.532
Charadriiformes	*Phalaropus fulicarius*	0.526
Passeriformes	*Sylvia nisoria*	0.523
Anseriformes	*Polysticta stelleri*	0.517
Passeriformes	*Acanthis hornemanni*	0.516
Charadriiformes	*Stercorarius skua*	0.516
Anseriformes	*Melanitta nigra*	0.509
Passeriformes	*Sylvia borin*	0.506
Passeriformes	*Loxia pytyopsittacus*	0.498
Charadriiformes	*Calidris falcinellus*	0.495
Charadriiformes	*Sterna paradisaea*	0.493
Passeriformes	*Lullula arborea*	0.493
Charadriiformes	*Hydrocoloeus minutus*	0.485
Passeriformes	*Pastor roseus*	0.478
Passeriformes	*Loxia curvirostra*	0.477
Passeriformes	*Acanthis cabaret*	0.471
Passeriformes	*Turdus atrogularis*	0.46
Passeriformes	*Ficedula parva*	0.445
Charadriiformes	*Stercorarius longicaudus*	0.442
Passeriformes	*Corvus corone*	0.428
Anseriformes	*Anas clypeata*	0.428

Passeriformes	*Carduelis cannabina*	0.426
Anseriformes	*Anser albifrons*	0.399
Passeriformes	*Acrocephalus dumetorum*	0.397
Charadriiformes	*Calidris melanotos*	0.394
Passeriformes	*Acrocephalus palustris*	0.389
Anseriformes	*Anas strepera*	0.375
Passeriformes	*Delichon urbicum*	0.37
Anseriformes	*Mareca strepera*	0.364
Anseriformes	*Anser fabalis*	0.359
Charadriiformes	*Thalasseus sandvicensis*	0.356
Anseriformes	*Branta bernicla*	0.354
Charadriiformes	*Larus melanocephalus*	0.352
Passeriformes	*Acrocephalus scirpaceus*	0.347
Passeriformes	*Motacilla citreola*	0.318
Passeriformes	*Luscinia luscinia*	0.307
Anseriformes	*Aythya collaris*	0.303
Charadriiformes	*Lymnocryptes minimus*	0.276
Anseriformes	*Spatula clypeata*	0.271
Passeriformes	*Emberiza leucocephalos*	0.244
Charadriiformes	*Larus cachinnans*	0.244
Anseriformes	*Anas querquedula*	0.241
Anseriformes	*Anas carolinensis*	0.215
Anseriformes	*Tadorna ferruginea*	0.215
Anseriformes	*Cygnus columbianus*	0.184
Passeriformes	*Anthus richardi*	0.174
Passeriformes	*Spinus spinus*	0.164
Passeriformes	*Anthus hodgsoni*	0.15
Anseriformes	*Spatula querquedula*	0.149
Passeriformes	*Anthus cervinus*	0.136
Passeriformes	*Linaria cannabina*	0.133
Passeriformes	*Linaria flavirostris*	0.113
Anseriformes	*Anser serrirostris*	0.076

Table S1: Metrics collected for species within the bird orders

Dependent variable	**Parameters**	**Slope**	**Intercept**	$\mathbf{R}^2$	**P-value**
Agaricales F_1-score	Data availability (log)	0.21	0.27	0.15	2.15×10^{-3}
Anseriformes documentation rate	Picture density	0.38	-0.47	0.52	1.51×10^{-7}

Anseriformes F_1-score	Data availability (log)	0.48	-0.97	0.69	4.18×10^{-11}
Anseriformes F_1-score	Documentation rate	-1.52	0.63	0.19	4.64×10^{-3}
Anseriformes F_1-score	Picture density	-1.02	1.95	0.31	2.10×10^{-4}
Asparagales F_1-score	Data availability (log)	0.31	0.01	0.34	0.0117
Asterales F_1-score	Data availability (log)	0.59	-0.71	0.51	5.92×10^{-4}
Charadriiformes documentation rate	Picture density	0.34	-0.43	0.76	5.51×10^{-18}
Charadriiformes F_1-score	Data availability (log)	0.32	-0.34	0.4	2.49×10^{-7}
Charadriiformes F_1-score	Documentation rate	-0.9	0.7	0.28	3.45×10^{-5}
Charadriiformes F_1-score	Picture density	-0.42	1.25	0.39	3.52×10^{-7}
Coleoptera F_1-score	Data availability (log)	0.13	0.57	0.06	0.273
Diptera F_1-score	Data availability (log)	0.53	-0.63	0.19	0.0326
Lecanorales F_1-score	Data availability (log)	0.2	0.31	0.12	0.118
Lepidoptera F_1-score	Data availability (log)	0.14	0.49	0.05	9.50×10^{-5}
Odonata F_1-score	Data availability (log)	0.46	-0.53	0.49	2.02×10^{-4}
Passeriformes documentation rate	Picture density	0.3	-0.36	0.55	1.59×10^{-19}
Passeriformes F_1-score	Data availability (log)	0.54	-1	0.63	4.31×10^{-24}
Passeriformes F_1-score	Documentation rate	-0.85	0.71	0.11	5.19×10^{-4}
Passeriformes F_1-score	Picture density	-0.5	1.37	0.23	1.68×10^{-7}
Passeriformes picture quality	Habitat openness	-0.12	5.93	0.26	5.53×10^{-8}
Polyporales F_1-score	Data availability (log)	0.31	-0.11	0.28	2.18×10^{-3}

Table S2: Metrics collected for species within the bird orders