No Bird Database is Perfect: Citizen Science and Professional Datasets Contain Different and Complementary Biodiversity Information

Authors: Galván, Sofía, Barrientos, Rafael, and Varela, Sara

Source: Ardeola, 69(1) : 97-114

Published By: Spanish Society of Ornithology

URL: https://doi.org/10.13157/arla.69.1.2022.ra6
NO BIRD DATABASE IS PERFECT:
CITIZEN SCIENCE AND PROFESSIONAL DATASETS
CONTAIN DIFFERENT AND COMPLEMENTARY
BIODIVERSITY INFORMATION

LAS BASES DE DATOS DE CIENCIA CIUDADANA
Y PROFESIONALES POSEEN INFORMACIÓN DIFERENTE
Y COMPLEMENTARIA SOBRE LA AVIFAUNA

Sofía Galván1 *, Rafael Barrientos2 and Sara Varela1, 3

SUMMARY. — Citizen science has become a powerful tool for collecting big data on biodiversity. However, concerns have been raised about potential biases in these new datasets. We aimed to test whether citizen science bird databases have more biases than professional scientific databases. Our hypotheses were 1) citizen science databases will have more data on “easy to spot” species, that are widely distributed and have large body sizes; whereas 2) professional databases will have more endangered species and species of special interest for research. We analysed six Spanish bird databases: three professional, two citizen science and one mixed database. Our results show that, in general, occurrences in citizen science databases are better explained by the studied variables than professional databases, but no clear differences were found when analysed individually. Both citizen science and professional databases contain invaluable information on biodiversity but every database comes with a particular history and its stored data is the result of years of field sampling with heterogeneous goals, sampling methods and sampling effort. Consequently, raw observations should not be used directly as an ideal survey of the distribution or abundance of birds. We need to uncover these biases and develop new methods to properly incorporate the extensive and heterogeneous biodiversity data that is readily available to research. — Galván, S., Barrientos, R. & Varela, S. (2022). No bird database is perfect: citizen science and professional datasets contain different and complementary biodiversity information. Ardeola, 69: 97-114.

Key words: big data, biodiversity monitoring, birdwatching, citizen science, macroecology, ornithology, sampling biases.

1 Centro de Investigación Mariña, University of Vigo, Animal Ecology Group (GEA), MAPAS lab, Vigo, Spain.
2 Road Ecology Lab, Department of Biodiversity, Ecology and Evolution, Faculty of Biology, Complutense University of Madrid, Madrid, Spain.
3 Museum für Naturkunde, Leibniz-Institut für Evolutions und Biodiversitätsforschung, Berlin, Germany.

* Corresponding author: sofia.galvan@uvigo.es

Twitter: @sofiaGA96 / @_Sara_Varela / @rafa_barri
INTRODUCTION

The study of patterns in nature, such as the distribution or abundance of organisms, often requires large-scale approaches (Bonney et al., 2009). Consequently, data must be collected over long periods of time and at large spatial scales (Kelling et al., 2009). Unfortunately, this often involves an effort that is logistically or economically unfeasible for individual research teams, which usually have limited financial and human resources. To address these limitations, different initiatives have been carried out. For example, the Global Biodiversity Information Facility (GBIF) aggregates multiple scientific databases from museums and collections around the world (https://www.gbif.org). Other projects, like the New and Old Worlds (NOW) database or PaleobioDB, collect records of fossil occurrences from scientific publications and make them available for further research (https://nowdatabase.org/; https://paleobiodb.org). Another way of collecting large amounts of information on biodiversity is through amateur naturalists. These citizen science projects allow the participation of the lay public in the observation, classification and collection of data, which can be used for research by scientists (Kullenberg & Kasperowski, 2016).

Birds are mostly diurnal, abundant and behaviourally and morphologically conspicuous (Sullivan et al., 2009). In addition, applying a phylogenetic species concept, there are about 18,000 species (Barrowclough et al., 2016) occupying all ecosystems on Earth, making them an ideal group for citizen science projects. This collaborative data gathering is already an established global tool used to record changes in species’ ranges, migration patterns, population trends and impacts of processes such as climate change (Dickinson et al., 2010). In fact, the number of records of species occurrences in citizen science databases are now larger than those in museum or scientific collections. For instance, eBird, an international birdwatching data-
base (https://ebird.org), already has over 100 million records; whereas the Natural History Museum (UK) has ~750,000 records and the Museo Nacional de Ciencias Naturales (Spain) has ~3,000 (via GBIF, searched in April 2021). This difference will continue to grow, as natural history collections are not expanding at the same rate as volunteer bird-watching projects.

Despite their size and utility, the quality of the data stored in citizen science databases has been questioned. It has been noted that their inequalities in sampling intensity over time, spatial coverage and sampling effort, and their uneven detection of rare species, could bias the data they contain (Isaac et al., 2014). Some studies show that volunteers easily identify common or iconic species but have more difficulty identifying rare species (Crall et al., 2011; Kelling et al., 2015; Swanson et al., 2016). Also, amateur observers are more successful at identifying those birds with easily recognisable songs but are less successful with inconspicuous species or those only present in certain, relatively inaccessible, habitats (Kelling et al., 2015). In addition, records are mostly aggregated in readily accessible/urban zones and in protected areas, showing a marked preference for observations of threatened species (Ferrer et al., 2006). For that reason, this information needs to be carefully analysed and validated in order to generate high-quality datasets for scientific purposes (Crall et al., 2011; Isaac et al., 2014; Johnston et al., 2019; Swanson et al., 2016).

We aimed to compare citizen science databases with professional databases. In particular, we studied six available databases of Spanish avifauna: three professional, two citizen science and one mixed. We expected that databases collected by citizen science projects (citizen databases) would contain a large proportion of records of species: a) with extensive geographical ranges, as they are widespread (Kelling et al., 2015; Swanson et al., 2016); b) of larger size, as they are more easily detected (Kamp et al., 2016), and 3) from habitats closer to human settlements (e.g. farmlands) or from frequently visited biodiversity hotspots, such as wetlands (Ferrer et al., 2006; Kelling et al., 2015). We also expected that databases compiled by scientists, specialists or professionals (professional databases) would have a greater representation of threatened species, as scientists often focus on endangered taxa (Ferrer et al., 2006).

METHODS

Databases

We used three professional databases (Estación Biológica de Doñana, Museo Nacional de Ciencias Naturales and Inventario Español de Especies Terrestres), two citizen science databases (eBird and Proyecto AVIS) and one mixed; the ringing database of the Spanish Ornithological Society (SEO/BirdLife), that includes contributions from both professional and amateur ringers. All data were downloaded in January 2020. The Proyecto AVIS database was obtained from its official website (https://proyectoavis.com/) and the remainder from the GBIF portal (https://gbif.org).

The Estación Biológica de Doñana (hereafter “EBD”) collection contains bird specimens collected between 1930 and 2007 (Cezón, 2018a), with 4,515 records belonging to 116 species. The database of the Museo Nacional de Ciencias Naturales (MNCN) is a collection of 2,980 records corresponding to 206 species, with records from 1841 to 2003 (Cezón, 2018b). The Inventario Español de Especies Terrestres (IEET) contains a total of 414,108 records of 331 bird species. Data were recorded between 2004 and 2012, based on a systematic record of occurrences in 10km² UTM grids throughout Spain (Villares, 2018).
The eBird database is a programme launched by the Cornell Lab of Ornithology and the National Audubon Society in 2002, which allows any user to enter data on bird observations using a standardised protocol (Sullivan et al., 2009; https://ebird.org/). The subset for Spain contained, when downloaded, 5,095,285 observations, comprising 589 species records collected from 1901 to 2020 (Levatich & Padilla, 2019). Proyecto AVIS is also a collaborative project on Spanish avifauna with a standardised protocol (Varela et al., 2014a). It had 106,694 observations of 426 species in January 2020, and its records range from 1973 to the present (https://proyectoavis.com).

The ringing database of the SEO/BirdLife NGO (SEO) contains 9,435,714 records belonging to 533 species, from 1905 to the present (SEO/BirdLife, 2020).

Data pre-processing

Datasets with several taxa were filtered to extract only bird records. For those databases containing records from more than one country (EBD and eBird), we selected those from Spain. We filtered the databases to select terrestrial breeding species following D’Amico et al. (2019), thus excluding wintering, non-breeding, invasive and marine species. The information comprising the resulting databases is shown in Table A1 (Supplementary Material).

Morphological and ecological variables

We used the following explanatory variables from D’Amico et al. (2019): 1) “geographical distribution” i.e., percentage of 10km² UTM grids where the species is present; 2) “weight” (in grams) and 3) “wingspan” (in centimetres), used as proxies of body size; 4) total number of “nesting habitats” (farmland, agroforest, forest, scrub-land, wetland and cliffs) where the species can breed (species can select more than one habitat, so percentage totals may exceed 100%) and 5) “conservation status”, with the CR, EN and VU categories (according to the IUCN Red List criteria; IUCN Standards and Petitions Committee, 2019) grouped together as “threatened”.

Descriptive statistics

We first described the spatio-temporal and taxonomic coverage of the databases. We used a Venn diagram to show the taxonomic uniqueness of each database and, for the spatio-temporal coverage, we utilised the year of collection and the geographical coordinates of each observation (EBD and MNCN databases do not include coordinates, so we excluded them for these tests). In addition, we carried out a case study aimed to explore the record density in each database along Spain for two species: the evenly-distributed and highly-abundant House Sparrow *Passer domesticus* and the Spanish Sparrow *Passer hispaniolensis* (which ranges throughout the peninsular Southwest and the Canary Islands, with higher abundance in Extremadura region) (Carrascal & Weykam, 2006; Martí & del Moral, 2003). To do so, we performed a 2D kernel density estimation (Silverman, 1986), using the `stat_density_2d` function of the `ggplot2` version 3.3.5 package for building the maps (this function uses the `kde2d` function from the MASS library (Venables & Ripley, 2002; Wickham, 2016).

We then used univariate descriptive statistics (boxplots and histograms) to show the distribution of the target continuous variables (geographical distribution, weight and wingspan) and calculated the percentage distribution for the categorical ones (nesting habitat and conservation status). We log-transformed weight and wingspan for visualisation purposes.
We built generalized linear models (GLMs) to link the number of observations per species with species’ traits and to complement the analyses aimed to understand the differences between databases. We included geographical distribution, weight and conservation status (in binary terms: threatened/non-threatened), and we excluded wingspan as it is correlated with weight ($r = 0.82$). From the nesting habitats, we selected only wetland and ‘accessible habitats’ (containing agroforest and farmland variables) as binary variables (yes/no), as they were the most selected habitats and also adequate for testing our hypothesis regarding the accessibility/proximity and attractiveness of these habitats for birdwatchers. As our explanatory variable “geographical distribution” was directly based on the second Atlas of Breeding Birds of Spain (Martí & del Moral, 2003), we excluded the IEET database from this analysis.

Finally, we used decision trees (“Regression trees”) to identify the explanatory variables that best predict the number of observations per species in each database, including interactions between them. This recursive partition algorithm (Steinberg, 2009) allowed us to classify information on the different databases based on our explanatory variables; in our case, species traits. For this analysis, we transformed nesting habitat and conservation status into binary variables (yes/no; threatened/non-threatened), and we excluded the IEET database for the same reason as above. We built decision trees using the `rpart` version 4.1-15 and `rpart.plot` version 3.0.8 packages (Milborrow, 2019; Therneau et al., 2019).

All analyses were performed using R version 3.6.2 (R-Core-Team, 2019), and the largest databases (SEO and eBird) were loaded using the `data.table` version 1.12.8 package (Dowle et al., 2019). R scripts for data filtering and analyses are available at https://github.com/SofiaGalv/SpanishBird Databases.git.

**RESULTS**

**Taxonomic, spatial and temporal heterogeneity in the databases**

The distribution of the species collected in each database is shown in Figure 1A. Only 32 species are included in all databases, whereas the greater number ($n = 137$) corresponds to species gathered in all databases except the EBD. Similarly, there are 30 species shared between all databases except for EBD and MNCN, indicating the taxonomic uniqueness of these two databases.

Temporally, the MNCN database includes the oldest records (median = 1981, interquartile range = 1934-1995, $N = 2,829$) (Figure 1B; Supplementary Material, Table A2 and Figure A4), followed by EBD (median = 1980, interquartile range = 1978-1986, $N = 3,888$) and with their most recent records in 2003 and 2007, respectively. On the other hand, IEET information was mainly recorded around 2004 ($N = 404,210$). Lastly, observations in the remaining databases are mainly post-2000 up to the present (2019), with the eBird database standing out as collecting the most modern information (SEO: median = 2004, interquartile range = 1997-2010, $N = 8,583,541$; eBird: median = 2017, interquartile range = 2016-2018, $N = 4,434,772$; AVIS: median = 2008, interquartile range = 2006-2012, $N = 98,468$).

In the spatial context, our case study pointed out that the distribution of records is different in each database (Figure 2). In the case of the House Sparrow (which is evenly-distributed and shows high abundance throughout the Iberian Peninsula), in general, the highest density of observations is concentrated in the centre (Madrid) and the south (Andalucía) in all databases. However, the SEO database shows high density on the Mediterranean coast, whereas the eBird database also shows concentrations in the north (mainly País Vasco and Navarra).
Fig. 1.—Taxonomic and temporal heterogeneity between databases. (A) Taxonomic coverage shared between databases. The central figure in each circle represents the number of species shared between the databases indicated in parenthesis. (B) Median ± first and third quartiles, with whiskers at 1.5 times interquartile range (IQR) for collection years in each database. Professional databases in black (MNCN, EBD and IEET), citizen databases in light grey (eBird and AVIS) and mixed database in dark grey (SEO).

[Heterogeneidad taxonómica y temporal entre bases de datos. (A) Cobertura taxonómica compartida entre las bases de datos. En cada círculo se representa el número de especies compartidas entre las bases de datos indicadas entre paréntesis. (B) Mediana ± primer y tercer cuartiles, con bigotes a 1,5 veces el rango intercuartílico (RIQ) para los años de recolección de las observaciones. Las bases de datos profesionales se muestran en negro (MNCN, EBD y IEET), las ciudadanas en gris claro (eBird y AVIS) y la base de datos mixta en gris oscuro (SEO).]
Fig. 2.—Maps of observations high-density zones for four databases (MNCN, SEO, eBird and AVIS) and two species (*Passer domesticus* and *Passer hispaniolensis*). Legends represent the product of the Kernel density estimate and the number of observations in each group.

[Mapas de las zonas con alta densidad de observaciones para cuatro bases de datos (MNCN, SEO, eBird y AVIS) y dos especies (*Passer domesticus* y *Passer hispaniolensis*). Las leyendas representan el producto de la estimación de densidad Kernel y el número de observaciones en cada grupo.]

*Ardeola* 69(1), 2022, 97-114
In the case of the less-widespread species (the Spanish Sparrow), observations are equally aggregated in the centre and the south of Spain, although eBird and AVIS also showed a marked density of observations in Extremadura and in Córdoba province (Andalucía), respectively. The IEET and eBird databases also show concentrations of Spanish Sparrow observations in the Canary Islands. In summary, all databases tend to accumulate observations in certain zones of the Spanish territory. Finally, we detected a peak in the abundance of occurrences in the limits between UTM zones in the IEET map. UTM is the original coordinate system of this database, and the overlap between UTM zones in Iberian Peninsula (29, 30 and 31) after conversion to latitude-longitude might not have been taken into consideration when uploading this database to GBIF (Figure 2).

Description of species traits in the databases

Preliminary exploration of the data showed that no database fulfils normality and homoscedasticity assumptions and that each has its own particular frequencies of species traits (Supplementary Material, Figure A1-A4), which prevented any statistic testing to compare them directly.

Most records in the databases are of birds that are widely distributed in Spain, covering around 60% of the territory (Figure 3A; Supplementary Material, Table A2; MNCN: median = 58.3%, N = 2,829; IEET: median = 64.4%, N = 404,210; SEO: median = 65.1%, N = 8,583,541; eBird: median = 61.2%, N = 4,434,772; AVIS: median = 53.3%, N = 98,468). In particular, the SEO and IEET databases have a larger number of records of widely distributed species (SEO: interquartile range = 30.7%-91.5%, IEET: interquartile range = 40.8%-83.9%). However, the EBD database includes birds with narrower geographical distributions (median = 3.7%, N = 3,888).

In relation to weight (Figure 3B; Supplementary Material, Table A2), professional databases (MNCN and IEET) collect birds with low body weights (MNCN: median = 152.0g, interquartile range = 22.9g-315.5g, N = 2,829; IEET: median = 40.5g, interquartile range = 18.6g-204.0g, N = 404,210). The mixed-database (SEO) contains records for the specimens with the lowest body weights (median = 18.6g, N = 8,583,541), while citizen database medians are 77.6g for eBird (interquartile range = 18.9g-502.7g, N = 4,434,772) and 180.5g for AVIS (interquartile range = 22.6g-828.5g, N = 98,468). Finally, the EBD database stands out with respect to the other datasets as it has heavier than average birds (median = 946.2g, interquartile range = 615.7g-2,750.0g, N = 3,888).

Regarding wingspan (Figure 3C; Supplementary material, Table A2), SEO has the smallest individuals (median = 23.3cm, interquartile range = 21.0cm-27.8cm, N = 8,583,541), followed by IEET (median = 33.3cm, interquartile range = 23.8cm-58.0cm, N = 404,210). On the other hand, eBird has a median of 40.0cm (interquartile range = 24.3cm-77.5cm, N = 4,434,772). MNCN and AVIS databases show identical medians, although the species collected by AVIS are larger (MNCN: median = 52.5cm, interquartile range = 26.0cm-85.5cm, N = 2,829; AVIS: median = 52.5cm, interquartile range = 26.5cm-110.5cm, N = 98,468). The EBD database has the largest birds (median = 120.5cm, interquartile range = 77.0cm-152.5cm, N = 3,888).

Lastly, our analyses indicate that most of the records in the non-citizen databases are of species that breed in agroforest habitats (Table 1, MNCN: 74.5%, IEET: 74.2%, SEO: 70.3%). In contrast, most records in the EBD database correspond to wetland species (71.2%). In citizen databases, obser-
Observations mainly comprise agroforest-breeding species (eBird: 63.0%, AVIS: 64.7%). Lastly, the percentage occurrence of threatened species is lowest in IEET, SEO and eBird (5.5%, 6.0% and 6.7%, respectively), has intermediate values in MNCN and AVIS (9.2% and 8.0%, respectively) and is highest in the EBD database (34.3%).

**Explained variance**

Our GLMs show that models performed with citizen datasets are able to explain the highest percentage of total variance, with 55.2% for eBird and 54.8% for AVIS (Table 2). However, the same variables explained less than 30% of variance in professional databases, with the EBD model being the least explanatory (only 14.0%). Positive coefficients were obtained for weight, geographical distribution, wetlands and accessible systems in all databases except for SEO (where species’ weight is negatively correlated with the number of records). In the case of conservation status, all coefficients were negative or not significant except for EBD (where records of threatened species are positively correlated with the total number of records).

**Types of records in the databases**

Regression trees show that, regarding professional databases, most records in the MNCN database are related to weight and geographical distribution (involving mainly large species with extensive distribution areas; Supplementary Material, Figure A5A), while the EBD records are mostly explained by conservation status and, for non-threatened species, by wingspan (Supplementary Material, Figure A5B). The mixed SEO tree also includes geographical distribution and weight, as does the MCMN database, but the second variable focuses on low weight species (<20.2 grams) instead of large species (Supplementary Material, Figure A6). Regarding citizen databases, eBird has the most complex tree (Supplementary Material, Figure A7A), showing divisions based on geographical distribution, weight, wetlands as breeding habitat and conservation status.

### Table 1

<table>
<thead>
<tr>
<th>Wetland</th>
<th>Farmland</th>
<th>Agroforest</th>
<th>Forest</th>
<th>Scrubland</th>
<th>Cliffs</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNCN</td>
<td>15.8</td>
<td>51.8</td>
<td>74.5</td>
<td>52.4</td>
<td>31.9</td>
</tr>
<tr>
<td>EBD</td>
<td>71.2</td>
<td>15.7</td>
<td>27.6</td>
<td>28.2</td>
<td>2.7</td>
</tr>
<tr>
<td>IEET</td>
<td>13.9</td>
<td>53.1</td>
<td>74.2</td>
<td>52.9</td>
<td>40.8</td>
</tr>
<tr>
<td>SEO</td>
<td>22.6</td>
<td>40.7</td>
<td>70.3</td>
<td>58.7</td>
<td>42.8</td>
</tr>
<tr>
<td>eBird</td>
<td>31.7</td>
<td>48.1</td>
<td>63.0</td>
<td>44.8</td>
<td>31.1</td>
</tr>
<tr>
<td>AVIS</td>
<td>32.7</td>
<td>48.5</td>
<td>64.7</td>
<td>46.1</td>
<td>26.9</td>
</tr>
</tbody>
</table>

Proportion of nesting habitats selected by the species occurring in each database. The highest value is highlighted in bold.

[Proporción de hábitats de nidificación seleccionados por las especies recogidas en cada base de datos. Se resalta el valor más alto.]
GLM for the number of observations according to: weight, geographical distribution, conservation status, wetlands and accessible habitats. The assigned coefficient, the F-value and its significance are shown for each variable in each database. '****' = p < 0.001, '***' = p < 0.01, '**' = p < 0.05, '*' = p < 0.1, NS = not significant. For each model, the percentage of explained variance is indicated.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficients</th>
<th>F-value</th>
<th>p-value</th>
<th>Explained variance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MNCN</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>1.4e-4</td>
<td>9.4</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Geographical distribution</td>
<td>0.017</td>
<td>20.5</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Conservation status</td>
<td>-0.021</td>
<td>-0.3</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>Wetland</td>
<td>0.061</td>
<td>1.1</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>Accessible habitats</td>
<td>0.34</td>
<td>6.2</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Null deviance: 3,064.5</td>
<td>Residual deviance: 2,185.1</td>
<td>Explained variance: 879.4</td>
<td>28.7%</td>
<td></td>
</tr>
<tr>
<td><strong>EBD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>9.3e-5</td>
<td>9.5</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Geographical distribution</td>
<td>0.015</td>
<td>15.5</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Conservation status</td>
<td>0.14</td>
<td>3.5</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Wetland</td>
<td>0.87</td>
<td>19.5</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>Accessible habitats</td>
<td>0.26</td>
<td>5.9</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Null deviance: 3,782.2</td>
<td>Residual deviance: 3,253.3</td>
<td>Explained variance: 528.9</td>
<td>14.0%</td>
<td></td>
</tr>
<tr>
<td><strong>SEO</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>-0.0016</td>
<td>-834.6</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Geographical distribution</td>
<td>0.025</td>
<td>1726.5</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Conservation status</td>
<td>-0.18</td>
<td>-117.4</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Wetland</td>
<td>0.56</td>
<td>587.1</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Accessible habitats</td>
<td>0.023</td>
<td>22.8</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Null deviance: 24,603,284</td>
<td>Residual deviance: 17,219,211</td>
<td>Explained variance: 7,384,073</td>
<td>30.0%</td>
<td></td>
</tr>
<tr>
<td><strong>eBird</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>2.1e-4</td>
<td>473.4</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Geographical distribution</td>
<td>0.026</td>
<td>1247.3</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Conservation status</td>
<td>-0.61</td>
<td>-292.3</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Wetland</td>
<td>0.85</td>
<td>683.6</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Accessible habitats</td>
<td>0.040</td>
<td>29.5</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Null deviance: 4,958,574</td>
<td>Residual deviance: 2,220,601</td>
<td>Explained variance: 2,737,973</td>
<td>55.2%</td>
<td></td>
</tr>
<tr>
<td><strong>AVIS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>3.3e-4</td>
<td>144.0</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Geographical distribution</td>
<td>0.025</td>
<td>176.3</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Conservation status</td>
<td>-0.70</td>
<td>-51.2</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Wetland</td>
<td>1.0</td>
<td>124.0</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Accessible habitats</td>
<td>0.29</td>
<td>32.2</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Null deviance: 125,230</td>
<td>Residual deviance: 56,660</td>
<td>Explained variance: 68,570</td>
<td>54.8%</td>
<td></td>
</tr>
</tbody>
</table>
This database is best explained in terms of species with wide geographical distributions that select wetlands for nesting. Finally, the AVIS tree splits first based on geographical distribution, and then on weight and wing-span (Supplementary Material, Figure A7B). In this case, the species with more observations are those with extensive geographical distributions and large body weights.

In summary, our results do not fully support the hypotheses presented. Occurrences in citizen science databases are better explained by the studied variables but there are no clear differences in species traits between the groups when analysed individually. Thus, the particularity of each database does not seem to be related to the group in charge of collecting the data, whether professionals or volunteers.

DISCUSSION

Biodiversity data has been stored in museum collections for well over a hundred years, and facilities like GBIF now enable researchers to easily access and download these datasets. Furthermore, so-called citizen science projects, a new way of collecting information, have gained importance in recent decades. Today, there are numerous initiatives, such as Proyecto AVIS at national scales (Varela et al., 2014a) and Vertnet or eBird at global scales (http://vertnet.org/; https://ebird.org/). Citizen science platforms have already overtaken the amount of data collected by professionals and museums (Figure 1B; Supplementary Material, Figure A4; Spear et al., 2017), which means that key information on recent biodiversity changes may not have been collected by professionals.

Studies of biases in citizen science data have found that volunteers make similar mistakes to professionals (Kosmala et al., 2016). However, biases in professional databases have rarely been discussed. Here, we explored six databases of Spanish birds (three professional, two citizen-science based and one mixed) to understand their potential problems and strengths. We started by showing the differences between databases regarding their taxonomic, temporal and spatial coverage. We then tested whether bird records differ between databases on the bases of species’ ranges, body sizes, habitat types or conservation status.

Firstly, it is important to emphasise that, although we categorised the databases into three groups, the divisions between them are diffuse. In this regard, IEET is the result of a structured survey, while citizen science projects, usually based on more informal data collection (Kelling et al., 2019), can also incorporate certain protocols to refine the uploaded data. For example, projects such as eBird, and AVIS to a less extent, could be considered semi-structured because they incorporate information for controlling some of the possible biases related to volunteer-based programs (date, time, location or observation period) (Sullivan et al., 2009; Varela et al., 2014a). eBird also presents other options to improve data collection, by gathering information on species absence or distance travelled by the observer, passing data through an automated filter and incorporating validation to detect unusual (e.g. out of range) species records (Sullivan et al., 2009).

Temporally, MNCN and EBD databases are based on museum collections dating back to the early 1900s (Figure 1B). This means that, for historical records of species presence, it is necessary to explore museum collections. However, nowadays, museums and research centres are not updating their collections, and citizen science projects are the main and sometimes the only source of information about species occurrence and abundance. In addition, although museum collections are available on online platforms like GBIF, important information such as geographical coordinates may be not included, which pre-
FIG. 3.—Median ± first and third quartiles, with whiskers at 1.5 times interquartile range (IQR) and notches (+/-1.58 IQR/sqrt(N)), for (A) geographical distribution (%), (B) weight (grams) (log-transformed) and (C) wingspan (centimetres) (log-transformed). Professional databases in black (MNCN, EBD and IEET), citizen databases in light grey (eBird and AVIS) and mixed database in dark grey (SEO).

[Median ± primer y tercer cuartiles, con bigotes a 1,5 veces el rango intercuartílico (RIQ) y muestras (+/-1,58 RIQ/raíz(N)) para (A) la distribución geográfica (%), (B) el peso (gramos) (log-transformado) y (C) envergadura (centímetros) (log-transformado). Las bases de datos profesionales se muestran en negro (MNCN, EBD y IEET), las ciudadanas en gris claro (eBird y AVIS) y la base de datos mixta en gris oscuro (SEO).]
vents their use in numerous scientific studies. The taxonomic uniqueness of museum collections was also detected (Figure 1A), which has been related to a narrower geographical coverage of their records (Boakes et al., 2010). This could not be tested due to the lack of coordinates, although the low geographical coverage of the species recorded in the EBD database seems to support this (Figure 3A).

The spatial analysis of the data from citizen and mixed databases also shows marked record concentration in some Spanish regions (Figure 2). Whereas this could be due to changes in species density, this may also reflect greater birdwatching activity or a higher impact of citizen science projects in these areas: the Madrid urban area, ornithological destinations such as Andalucía (Doñana National Park) or main cities on the east coast. These results agree with previous studies showing a prevalence of birdwatching in urban, high human population density or protected areas (Ferrer et al., 2006; Martin et al., 2012; Reddy & Dávalos, 2003; Sumner et al., 2019), as well as with results showing the habitat preferences of the species recorded (Table 1). Thus, all the above points and the potential spatial biases of the open access biodiversity databases need to be taken into account when using these data for research. For example, one specific use of big data on biodiversity is forecasting species distributions across time. In this case, biases in species occurrences can modify the output of niche models (van Eupen et al., 2021; Varela et al., 2014b).

Analysing databases separately, our models also show that the selected species traits (weight, geographical distribution, conservation status, wetland and accessible habitats) can better explain the number of records per species in citizen science databases than in professional databases (Table 2). This could indicate that there are other constraints than general detectability/interest and commonness explaining the number of records per species stored in the historical bird collections at research institutions. This result can also indicate that bird occurrences in citizen science databases may be more homogeneous and predictable than those stored in historical collections.

Regression trees allow us to see interactions between variables, and here suggest that both weight and wingspan are key predictors of the number of observations in all databases and not only in citizen science ones (Supplementary Material, Figure A5-A7); although the trend is more pronounced for these. Thus, in general, records of large species are more likely than those of smaller birds, with the exception of the SEO database (based on ringing records). Mist nets are the most widespread sampling method for ringing birds, allowing bird-ringers to record small species. However, these species are not so easy to collect/observe with the observational methodologies used in other databases, explaining this pattern. In general, there are no clear differences between citizen and professional databases for species’ size (Figure 3B-3C). However, other studies detected some discrepancies, as they found that citizen science data tended to be less precise in capturing trends in small-sized birds (Kamp et al., 2016). This pattern also occurs with other taxa; for example, in a study on road-killed mammals, citizens recorded heavier species than trained groups (Périquet et al., 2018). The exception to this pattern was the EBD database (Figure 3B-3C; Supplementary Material, Figure A5B). This database is managed by a research centre located in a Spanish national park, which includes wetlands as one of its more representative biotopes. Thus, waterfowl and herons (i.e. heavy species of conservation concern) are strongly represented in their database.

The ubiquity of common species can increase the likelihood of being recorded by volunteers (Horns et al., 2018). However,
even if our analyses show the importance of geographical coverage in predicting the number of observations (Table 2; Supplementary Material, Figure A5-A7), this pattern is not restricted to citizen databases, but applies to all database types (except EBD) (Figure 3A). EBD has a high number of large, narrow-range, endangered species; setting this database apart (Figure 3, Table 1). Thus, in general, species are more frequently recorded simply because they are common or abundant. Conversely, species with restricted distributions may be less detected and may even need to be proactively tracked in some cases, which may explain the scarcity of their records and may sometimes lead to underestimation of their abundance (Bird et al., 2014).

Volunteers and professional ornithologists tend to record similar numbers of endangered species (after excluding EBD) (see Results). It might be expected that a group of volunteers with varying skills would register fewer threatened species, but amateur birdwatchers seem to detect endangered species in a similar ratio to professional ornithologists (Cordell & Herbert, 2002; Galloway et al., 2006; Snäll et al., 2011). This was also noted by Ferrer et al. (2006), who found that preferred areas for birdwatchers were those where endangered species were most likely to be spotted.

Finally, the general trend to include birds from more accessible and nearby habitats, as well as those more attractive to tourists, was confirmed in all databases (Table 1, Table 2). Birds from agroforest systems, defined as a set of wooded areas embedded in an agricultural matrix, had more records in all databases, with the exception of EBD. Heterogeneous habitats can potentially support a greater number of species (Kelling et al., 2015) and higher abundances (Pickett & Siriwardena, 2011). Nevertheless, these differences may also reveal biases, suggesting that these areas may have been better explored due to their greater accessibility. In a study on the delimitation of priority conservation areas, researchers detected a marked trend towards sampling near cities (Reddy & Dávalos, 2003). Similarly, other authors have found a preference by people to remain close to inhabited areas, natural parks, research facilities and tourist destinations (Ferrer et al., 2006; McKinley et al., 2017).

In summary, our results indicate that none of the databases is perfect. In this way, databases such as EBD, which contains species with unique characteristics, are extremely useful for local studies on wetland species in southern Spain. On the other hand, other databases such as the IEET can be particularly useful for developing maps and distribution models, because it uses a standardised protocol to cover the entire Spanish territory. However, all databases may underestimate the presence of certain species in the ecosystems. In conclusion, amateur and professional ornithologist databases may have different biases in relation to the geographical coverage of their records (e.g. towards common or rare species or towards species with limited ranges) and to species traits (e.g. biases towards large or small species). The sampling method (e.g. mist nets), or the opportunities to observe certain species (e.g. accessibility to hotspots of wetlands diversity in the EBD dataset), result in some databases storing a disproportionately large amount of “rare” occurrences. Thus, none of these databases can be used as a realistic sample of the composition or abundance of species, and analysing their raw data without clear knowledge on sampling methods or potential taxonomic/trait biases could lead to a misestimation of the actual patterns of biodiversity. Solutions to mitigate these biases include, for instance, designing stratified sub-sampling (e.g. Rosenberg et al., 2019) or standardised protocols, training volunteers (Frigerio et al., 2018), filtering the data (Wiggins et al., 2011) or using other statistical tools such as modelling methods (Bird et al., 2014; Isaac et al., 2014).
We are now entering a new paradigm for studying life on Earth, where we can use big data techniques to explore biodiversity patterns and gain knowledge on the processes that regulate life. To do this, we will increasingly need to combine data collected by citizen scientists and professional ornithologists and to overcome associated biases to build adequate training sets for answering our questions reliably.

ACKNOWLEDGMENTS.—This study comprised S. Galván Master’s thesis. It is also part of the MAPAS project and has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant agreement No. 947921) to S. Varela. R. Barrientos was financed by the Comunidad de Madrid through 2018T1/AMB10374. E. Casabella provided us the raw dataset from Proyecto AVIS.

AUTHOR CONTRIBUTIONS.—Study conception S.G., R.B., S.V.; Methodology S.G., R.B., S.V.; Computation S.G., S.V.; Formal analysis S.G.; Data collection S.G.; Data curation S.G., R.B.; Writing initial draft S.G.; Writing revision R.B., S.V.; Writing visualization S.G.; Supervision R.B., S.V.; Project Administration S.G., R.B., S.V. and Funding Acquisition S.V., R.B.

REFERENCES


**SUPPLEMENTARY ELECTRONIC MATERIAL**

Additional supporting information may be found in the online version of this article. See the volumen 69(1) on www.ardeola.org

[Información adicional sobre este artículo en su versión en línea en www.ardeola.org, volumen 69(1).]

<table>
<thead>
<tr>
<th><strong>Table A1:</strong></th>
<th>Number of observations and species collected in each database.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table A2:</strong></td>
<td>Database summaries.</td>
</tr>
<tr>
<td><strong>Figure A1:</strong></td>
<td>Species’ geographical distribution in each database.</td>
</tr>
<tr>
<td><strong>Figure A2:</strong></td>
<td>Distribution of species’ weight in each database.</td>
</tr>
<tr>
<td><strong>Figure A3:</strong></td>
<td>Distribution of species’ wingspan in each database.</td>
</tr>
<tr>
<td><strong>Figure A4:</strong></td>
<td>Distribution of collection years in each database.</td>
</tr>
<tr>
<td><strong>Figure A5:</strong></td>
<td>Regression trees for the number of observations in professional databases.</td>
</tr>
<tr>
<td><strong>Figure A6:</strong></td>
<td>Regression trees for the number of observations in SEO database.</td>
</tr>
<tr>
<td><strong>Figure A7:</strong></td>
<td>Regression trees for the number of observations in citizen databases.</td>
</tr>
</tbody>
</table>

Submitted: May 03, 2021

Major revision: July 07, 2021

Second version arrived: August 20, 2021

Minor revision: November 11, 2021

Accepted: November 10, 2021

Editor: J. Ramos