



OPEN

Assessing citizen science data quality for bird monitoring in the Iberian Peninsula

J. Prenda^{1✉}, J. L. Domínguez-Olmedo², E. López-Lozano¹, R. Fernández de Villarán³ & J. J. Negro⁴

Citizen science data on biodiversity has experienced significant growth, largely driven by advancements in technology that facilitate data sharing. In recent years, mobile applications have provided a substantial boost to both the collection and sharing of this data. While this accessible information is undoubtedly valuable, we must consider the numerous biases present in this data when used for high-quality research. In this study, we analyse citizen science data for the birds of the Iberian Peninsula, comprising over 23 million unique records after filtering for duplicates (i.e., instances where the same observation was uploaded more than once). Using a 10 × 10 km square grid, we present information on well-surveyed cells (completeness) as well as temporal, taxonomic, geographical, and land use coverages. We found that the completeness of information is generally high, with better coverage around densely populated areas such as major cities and coastal regions, as well as popular birding destinations, which are frequently visited. The Mediterranean ecoregion and wetlands exhibit the highest levels of surveying. Furthermore, we observed an increase in temporal coverage since the 1980s and broad taxonomic coverage across all bird orders in the Iberian region. Our results underscore the utility of citizen science for many locations, as expressed in 10 × 10 km cells. However, they also highlight the inadequate data completeness across part of the territory, particularly in forested or sparsely inhabited areas. These findings not only identify cells suitable for bird diversity or conservation studies but also indicate areas where citizen-driven bird recording efforts should be encouraged.

Keywords Avifauna, Digital repositories, Data completeness, Conservation, Biodiversity management, Iberian Peninsula

New technologies have significantly advanced various scientific disciplines, and biodiversity is no exception¹. Citizen science platforms have emerged as prolific sources of valuable data contributed by amateur observers, facilitating increased collaboration and data sharing among scientists². Notably, Observation.org and eBird stand out as the most well-established platforms among users in Western Europe. Furthermore, GBIF compiles data from these platforms and various other sources, including additional platforms, natural science museums, government data, etc.

This information, or citizen science (CS) data, is experiencing rapid growth, amassing an extensive and booming dataset. For the Iberian Peninsula alone, the combined bird records from these three platforms exceed 40 million, excluding duplicates, as of June 2024. Each record typically represents the observation of a confidently identified species at specific coordinates on a precise date. Although this information currently serves as the primary source of biodiversity data³, its full utilization is yet to be realized due to data quality, data integration, management, and new analytical techniques, as well as societal factors such as collaborations between citizen scientists and professional researchers. In addition, Callaghan et al.² identify three areas to leverage CS data: increasing sampling efforts in undersampled areas and taxa, estimating species abundances over space and time, and utilizing secondary data in CS observations.

The significance of having up-to-date distribution data over large spatiotemporal scales cannot be overstated. However, no single institution, governmental, academic or NGO, possesses the means and funds to monitor all biodiversity taxa across all territories. This knowledge proves especially valuable for biodiversity hotspots or megadiverse countries, typically in areas where data remains incomplete^{4,5}. Species distribution models, crucial

¹Departamento de Ciencias Integradas, Universidad de Huelva, Huelva, Spain. ²Departamento de Tecnologías de la Información, Universidad de Huelva, Huelva, Spain. ³Departamento de Ciencias Agroforestales, Universidad de Huelva, Huelva, Spain. ⁴Estación Biológica de Doñana (CSIC), Sevilla, Spain. ✉email: jprenda@uhu.es

for understanding extinction processes, heavily rely on sample size. CS data, by providing information from diverse locations, contributes to the creation of accurate models.

The utility of CS data in defining and characterizing biodiversity patterns hinges on the completeness of species inventories at specific locations and the even distribution of locations across temporal, spatial, and environmental dimensions^{6,7}. Both the completeness (defined as the ratio between observed and expected richness^{4,5}) and suitability for various macro-scale biodiversity studies are pivotal before comparing species assemblages, analysing distribution patterns, or conducting other basic ecological assessments. Unfortunately, these aspects are often insufficiently analysed in current studies.

While well-covered CS data is apt for representing distributional patterns and the ecological dynamics of species, encompassing patterns of extinction and colonization, incomplete data (taxonomic, spatial, and/or temporal dimensions) may yield misleading results, thus fostering false conclusions at different scales^{8,9}. In general, biases in CS can arise from several factors that do not mutually exclude one another. For instance, users' place of origin (e.g., urban areas) and their preferences (spatial: e.g., wetlands, taxonomic: e.g., raptors, temporal: e.g., summer and/or good weather conditions)¹⁰. Additionally, an inadequate or incomplete review of uploaded records by platform experts, coupled with users' preference for searching rare species, could further amplify taxonomic biases¹¹. Incomplete, biased, and/or outdated taxonomic coverage may impact the accurate inference of species richness and the definition of multispecies assemblages within a specific range¹². On the other hand, uneven geographical coverage may influence the performance of species distribution models based on species occurrence records, potentially underestimating the actual or potential distribution patterns of focal species^{13,14}. Inadequate coverage of biogeographical regions or ecoregions may adversely affect inferences of species' regional ranges, complicating the accurate determination or revision of the division of zoogeographical regions or ecoregions¹⁵. Simultaneously, incomplete coverage by land uses may introduce biases in taxonomic coverage. Finally, insufficient temporal coverage of data may impede the accurate assessment of temporal dynamics, especially those associated with extinction processes¹⁶.

Completeness analyses have been conducted for select taxonomic groups in a few countries worldwide, such as Brazil (plants)⁶, Mexico (birds)⁴, the United States (vertebrates and some invertebrates)⁷, India (birds)⁵, and West Africa (birds)¹⁷. However, the Iberian Peninsula has seen coverage assessments only for Scarabaeidae and bees¹⁸, as well as terrestrial mammals¹⁹, leaving birds without comprehensive evaluation so far.

The Iberian Peninsula stands as a territory of remarkable biodiversity, situated within the Mediterranean basin hotspot²⁰. Its geographical placement has given rise to three distinct zoological components: (1) Iberian endemic elements, fostered by glacial isolation during the Pleistocene²¹; (2) European biotas; and (3) African elements, owing to the proximity of North Africa and historical connections with this continent during the Messinian Crisis six million years ago²², or attributed to the volcanic arc in the Alborán Sea, as described by Booth-Rea et al.²³ during the same epoch. These factors collectively render this territory highly intriguing for biodiversity studies, thereby elevating the relevance of CS data. This region hosts 456 bird species^{24–26}, some of which, especially those prevalent in agricultural landscapes, are undergoing a rapid and pronounced decline due to habitat transformation, climate change, or pollution²⁷. Consequently, the scrutiny of these spatiotemporal processes through extensive CS data offers an opportunity to unravel the intricacies surrounding these issues^{28,29}.

The primary goal of this study is to evaluate the extent of coverage and completeness within the CS datasets pertaining to birds of the Iberian Peninsula, sourced from eBird, Observation.org, and GBIF. Initially, we assessed completeness in 10 × 10 km cells, aiming to gauge the efficacy of data compilation efforts in analysing spatial variation in species richness. Subsequently, we delineated the coverage level of bird censuses derived from these databases from five distinct perspectives: (1) geographic (assessing bias in the total distribution of records), (2) temporal (examining well-sampled cells over time), (3) ecoregional (evaluating coverage across terrestrial ecoregions in the Iberian Peninsula as defined by the World Wildlife Fund), (4) environmental (considering annual mean temperature, annual precipitation, and altitude), and (5) land use (exploring forested areas, agricultural lands, wetlands, urban spaces, water bodies, and mixed-use regions). The scrutiny of coverage along environmental gradients represents a crucial yet often overlooked aspect of biodiversity data planning. Ultimately, by ascertaining completeness and coverage within extensive bird CS databases and across ecologically relevant gradients, our objective is to illuminate the factors contributing to data gaps.

Materials and methods

Study area

The Iberian Peninsula (583,254 km²) (Figure 1), including Andorra, Portugal, Spain, and Gibraltar, is located in south-western Europe and is separated from Africa only by the 14 km wide Gibraltar Strait. Serving as the western link between the two vast continents of Eurasia and Africa, it occupies the space between the Atlantic Ocean and the Mediterranean Sea. Furthermore, endowed with geological, physiographic, climatic, and ecological diversity, the Iberian Peninsula functions akin to a small continent. This distinctive character, coupled with its strategic location, historical human activities, and its role as a glacial refuge during the Pleistocene, bestows upon it extraordinary biodiversity values. These combined factors have fostered the coexistence of a rich blend of European, African, and endemic biotas^{30–32}. In the realm of avifauna, the Iberian Peninsula serves as a crossroads for the Occidental European migratory route, giving rise to a significant number of species in transit³³ (Fig. 1).

Biogeographically, the Iberian Peninsula predominantly falls within the Mediterranean ecoregion, with a narrow extension into the Euro-Siberian domain in the north. The northern region is characterized by Cantabrian mixed forests, encompassing the Cantabrian Mountain Range, Pyrenees, and Galician Massif. The Mediterranean ecoregion itself is further subdivided into seven major ecological units (Iberian conifer forests, Iberian sclerophyllous and semi-deciduous forests, Northeastern Spain and Southern France Mediterranean forests, Northwest Iberian montane forests, Pyrenees conifer and mixed forests, Southeastern Iberian shrubs and

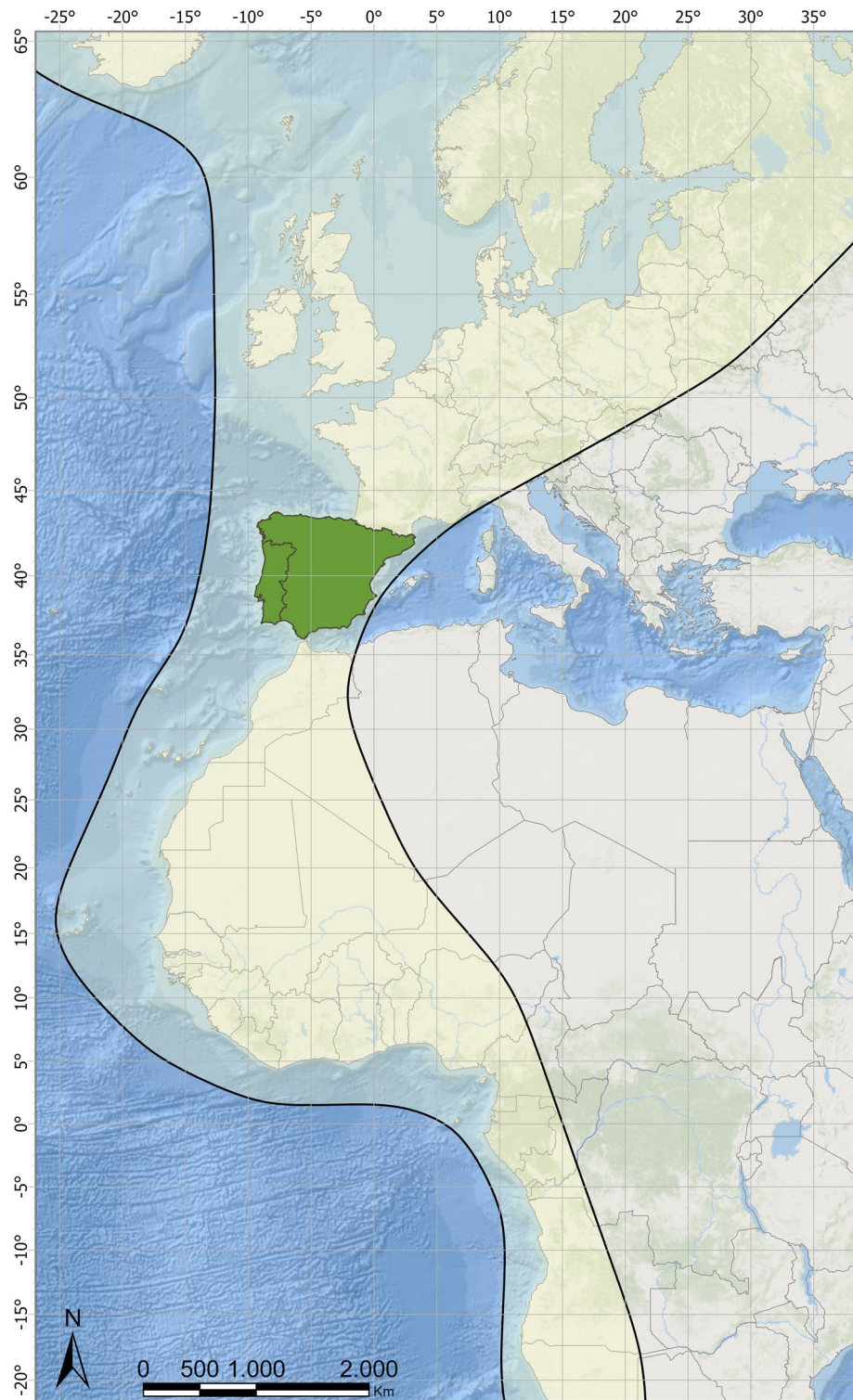


Fig. 1. Map depicting the location of the Iberian Peninsula (in green) in the European continent. The primary migratory flyway for birds in Western Europe is indicated as a yellow band³⁴.

woodlands, Southwest Iberian Mediterranean sclerophyllous and mixed forests)³⁵. These subdivisions underscore the significant environmental heterogeneity of the Iberian Peninsula. The habitats of interest within this region comprise Mediterranean forests, wetlands, hardwood forests, steppes (whether natural or agricultural pseudo-steppes), mountain ranges, and river systems, among others³⁶. This Peninsula boasts the highest contribution to the European Nature 2000 network, spanning over 222,000 km², and hosts 2,130 protected spaces designated by national or regional authorities³⁷.

Biodiversity data

Data for this study is sourced from three distinct platforms: two CS platforms, eBird (<http://www.ebird.org>)³⁸ and Observation.org (<https://observation.org/>)³⁹, along with GBIF (<https://www.gbif.org/>)⁴⁰, a biodiversity data repository that aggregates information from various sources, including the aforementioned platforms. GBIF updates data from original sources with variable time delays, usually from 3 to 18 months, influenced by factors such as data validation, standardization, technical capabilities, and formal agreements between institutions. The dataset was cleaned and filtered in several phases, following Escibano et al.¹⁹. Two types of records were initially eliminated: (1) records from islands and North African territories; and (2) records lacking a collection date or coordinates.

The combination of these three datasets could result in the creation of duplicate or triplicate records (the same observation uploaded more than once) for two main reasons: (1) an observer uploading the same observation to two platforms (eBird and Observation.org) and/or (2) both platforms (eBird and Observation.org) routinely uploading their records to GBIF. To mitigate potential duplications and reduce the effects of pseudoreplication in subsequent analyses, we applied a stringent criterion. Specifically, we retained only one record from datasets that shared identical dates and geographical coordinates with the same first two decimal places in decimal degrees (DD) notation. These decimals correspond to a circle of approximately 800 m radius, ensuring a conservative approach to data consolidation. Additionally, records lacking specific taxonomic assignment and hybrids were excluded. From the initially downloaded 42,453,700 records, a final set of 23,372,637 records was retained for use in this study. Each record was then assigned to its corresponding 10 x 10 km UTM grid cell.

We have compiled a comprehensive list of continental birds of the Iberian Peninsula by combining the species lists from Spain (SEO/BirdLife “Lista de las aves de España. Edición de 2019”²⁴) and Portugal (Matías et al.^{25,26}), encompassing a total of 595 species (Suppl. Materials 1, 2). We have followed to the taxonomic criteria outlined in the HBW & BirdLife International Illustrated Checklist of the Birds of the World⁴¹.

The dataset used for this study is publicly available. No living organism has been manipulated during the execution of this study, which has been entirely based on data obtained from the internet.

Taxonomic coverage

To assess taxonomic coverage, we conducted a comparison between the species documented in the CS data and our comprehensive list of species for the Iberian Peninsula. We exclusively considered native and exotic species present in this consolidated list.

Given that the platforms involved may exhibit slight variations in bird taxonomy nomenclature, we created a synonymous list to standardize the taxonomy across the entire dataset. This process ensured uniformity by aligning the taxonomy with our unique combined list for the Iberian Peninsula (Suppl. Material 2).

Inventory completeness

The assessment of inventory completeness was conducted across 10x10 km grid cells. The expected species richness for each cell was determined using the Chao2 metric (S_{Chao2})^{4,5}:

$$S_{Chao2} = S_{obs} + \left(\frac{N-1}{N} \right) + \left(\frac{a(a-1)}{2(b+2)} \right),$$

where S_{obs} is the number of species recorded in each cell; N indicates the number of unique combinations of records (species x date) available for each cell; a and b are the number of species recorded in exactly one day and two days in each cell, respectively. Subsequently, the bird inventory completeness (C) of each pixel is calculated as follows:

$$C = \frac{S_{obs}}{S_{Chao2}}.$$

To determine the threshold for considering a cell as well-sampled, various alternative sources were reviewed (Table 1). Cells were considered well inventoried under lax criteria when $N \geq 200$ and $C \geq 0.8$ and under strict criteria when $N \geq 1000$ and $C \geq 0.9$. Grid cells with fewer than 200 records were considered undersampled.

Geographical coverage

To evaluate the geographical coverage of well-inventoried cells, the count of unique coordinates per species and per grid cell was conducted. Moran's I was calculated using `esda.Moran` in Python (<https://pysal.org/esda/generated/esda.Moran.html>) to assess the geographical bias in the overall distribution of records, providing a quantitative measure of spatial autocorrelation⁴². A positive Moran's I value, along with significant spatial autocorrelation, indicates a biased geographical coverage. To compute this index, the spatial weight matrix was initially calculated using the K nearest neighbours (KNN) method, with $K=8$ (representing the eight nearest cells).

Temporal coverage

A time-based index (T), was employed to assess the temporal coverage of each species or grid cell⁴², aiding in the identification of well-surveyed cells over time. Within the dataset, the earliest available record was in January 1817, and the latest was on 31 December 2021, resulting in a total of 2172 months (Mo). Ideally, each species or grid cell should have at least one record per month. The variable M was computed as the number of months for which data was recorded for each species or grid cell in the dataset. Subsequently, the temporal coverage index (T) for each species or grid cell was calculated as follows:

Organisms	Records	Cell size	Criteria	N	C
Plants of Brazil ⁶	2,456,491	1/10°	Strict	> 5000	> 0.8
			Lax	≥ 1000	≥ 0.5
Birds of Mexico ⁴	277,249	0.25–2°	Strict	≥ 200	≥ 0.9
			Lax	≥ 200	0.8 ≤ C < 0.9
Amphibians, birds, freshwater bivalves, crayfish, freshwater fish, fungi, insects, mammals, plants, and reptiles of the United States ⁷	718–4,813,437	0.1°	High	≥ 50	≥ 0.8
			Low	≥ 10	≥ 0.6
			Moderate	≥ 25	≥ 0.7
Worldwide bees	137,809		Well surveyed		> 0.9
Scarabaeidae (Coleoptera) of the Iberian Peninsula and adjoining islands ¹⁸	15,142	10 × 10 km	Under-surveyed		< 0.5
Birds of India ⁵	901,665	0.25–2°	Strict	≥ 200	≥ 0.9
Terrestrial mammals of the Iberian Peninsula ¹⁹	179,767	10 × 10 km	Strict	≥ 50	≥ 0.8
			Lax	≥ 25	≥ 0.7
Birds of western Africa ¹⁷	438,877	0.1°–0.5°		≥ 200	≥ 0.8
Birds of the Iberian Peninsula (this work)	23,372,637	10 × 10 km	Strict	≥ 1000	≥ 0.9
			Lax	≥ 200	≥ 0.8

Table 1. Criteria used in different studies based on citizen science data to define well-inventoried cells, along with a proposal made in this study for Iberian birds. The organisms studied, the number of records for each case, the cell size, the rigor of the criteria, the number of organisms (N), and completeness (C) are specified.

$$T = \left(\frac{M}{M_o} \right) \times 100.$$

Additionally, considering the increase in data since the last decades of the 20th century, this index was computed for the period 1980–2021, encompassing a total of 492 months.

Moran's I was also calculated to assess geographical bias in temporal coverage. A positive Moran's I value with significant spatial autocorrelation indicates a notable aggregation of grid cells with high temporal coverage⁴³. Furthermore, Pearson correlations were calculated to examine the relationships between completeness, geographical coverage (localities per cell), and temporal coverage (months per cell).

Environmental and land use coverage

Environmental and land use factors can influence the way citizens collect bird data, potentially introducing some sampling bias^{44,45}. To evaluate differences (gaps among well-surveyed sites) in the environmental dimension, we obtained the annual mean temperature, annual precipitation, and altitude variables at 2.5' of arc (~ 5 km) spatial resolution from the WorldClim 2.1 dataset⁴⁶. Following the approach described by Freeman and Peterson¹⁷, this data was utilized to compute nearest neighbour Euclidean distances between cells in the environmental space. The focus was particularly on distances from each grid cell across the study area to well-surveyed grid cells.

For land cover categorization, we employed the first level of the Corine Land Cover classification. The assessment involved computing the dominant land use within each grid cell, considering instances where the coverage exceeded 50% of the occupied surface. Following this criterion, our cells were categorized as follows: forest, agricultural, wetlands, artificial surfaces (urban), water bodies (encompassing all marine cells), mixed (indicating the absence of a clear dominance), and data unavailable (NA).

Kolmogorov's test was applied to assess whether the frequency distribution of environmental and land use variables from well-sampled cells differed randomly from the global pool of cells. In cases where differences were identified, Tukey's post hoc test was employed to pinpoint the specific variables contributing significantly to these differences.

Results

Taxonomic coverage

Between 1817 and December 2021, a total of 553 bird species were recorded in the CS data from the Iberian Peninsula, representing 92.9% of the taxonomic coverage outlined in the unified checklist of birds for the region (Table 2). The primary source contributing to the bird species richness in the Iberian Peninsula within the CS dataset was GBIF (553 species), while observation.org documented the lowest count (511 species), and eBird fell in between with 523 species.

The taxonomic coverage of CS data for Iberian birds is notably high, with the majority of orders displaying values ranging between 90 and 100%. Only one order with more than 10 species exhibited a taxonomic coverage below 90%, Procellariiformes, which comprises exclusively pelagic seabirds with 23 species and a 73.9% coverage. Additionally, six other orders recorded values below this threshold, all characterized by a limited number of species, some of which are rare or exotic and present considerable difficulty in being covered by CS data (Table 2). For instance, two out of three species of Cuculiformes are accounted for in the CS dataset, with the sole exception being the rarity *Coccyzus americanus*.

Order	No. of species (checklist)	Covered species (CS data)	Taxonomic coverage (%)
Total	595	553	92.9
Accipitriformes	33	30	90.9
Anseriformes	62	56	90.3
Apodiformes	6	6	100
Bucerotiformes	1	1	100
Caprimulgiformes	3	3	100
Charadriiformes	124	115	92.7
Ciconiiformes	4	4	100
Columbiformes	9	8	88.9
Coraciiformes	5	5	100
Cuculiformes	3	2	66.7
Falconiformes	11	10	90.9
Galliformes	8	7	87.5
Gaviiformes	4	4	100
Gruiformes	18	17	94.4
Otidiformes	3	2	66.7
Passeriformes	217	208	95.9
Pelecaniformes	19	19	100
Phaethontiformes	1	1	100
Phoenicopteriformes	2	2	100
Piciformes	7	7	100
Podicipediformes	6	6	100
Procellariiformes	23	17	73.9
Psittaciformes	5	5	100
Pterocloriformes	3	2	66.7
Strigiformes	9	9	100
Suliformes	9	7	77.8

Table 2. Taxonomic coverage of citizen science data for birds in the Iberian Peninsula (1817–2021) at the order level. Citizen science data were sourced from eBird, Observation.org, and GBIF.

Inventory completeness

To assess the distribution and completeness of bird observations gathered from the CS platforms eBird, Observation.org, and GBIF, we analyzed a total of 23,372,637 records collected between 1817 and 2021, documented across 6185 10×10 km cells. The average completeness per cell for the entire Iberian Peninsula was 0.68, with interquartile ranges of 0.97–0.78 (Table 3).

A total of 2411 cells (38.8%) of the 10×10 km grid were considered well-sampled under strict criteria ($C \geq 0.9$), and 4418 cells (71.2%) met the criteria under lax conditions ($C \geq 0.8$). The remaining 1767 cells (28.6%) were considered not well-sampled. Completeness, particularly in well-sampled cells, exhibits a scattered pattern throughout the Iberian Peninsula, with two distinct areas, Portugal and eastern Spain, displaying well-defined maps characterized by cells with $C \geq 0.8$ (Fig. 2). Noteworthy concentrations of high completeness values are also observed in specific regions such as central, mid-western, north-eastern, or the southwest coastline of Spain,

	Dataset
Records	23,372,637
Time interval	1817–2021
Observed bird richness	553
Mean C	0.68
C quartiles (25–75%)	0.97–0.78
Sampled 10×10 km cells	6185
Well-sampled cells (lax criterion: $N \geq 200$, $C \geq 0.8$)	4418 (71.2%)
Well-sampled cells (strict criterion: $N \geq 1000$, $C \geq 0.9$)	2411 (38.8%)

Table 3. Summary of downloaded bird citizen science data from eBird, Observation.org and GBIF alongside bird inventory completeness of 10×10 km cells sampled across the Iberian Peninsula.

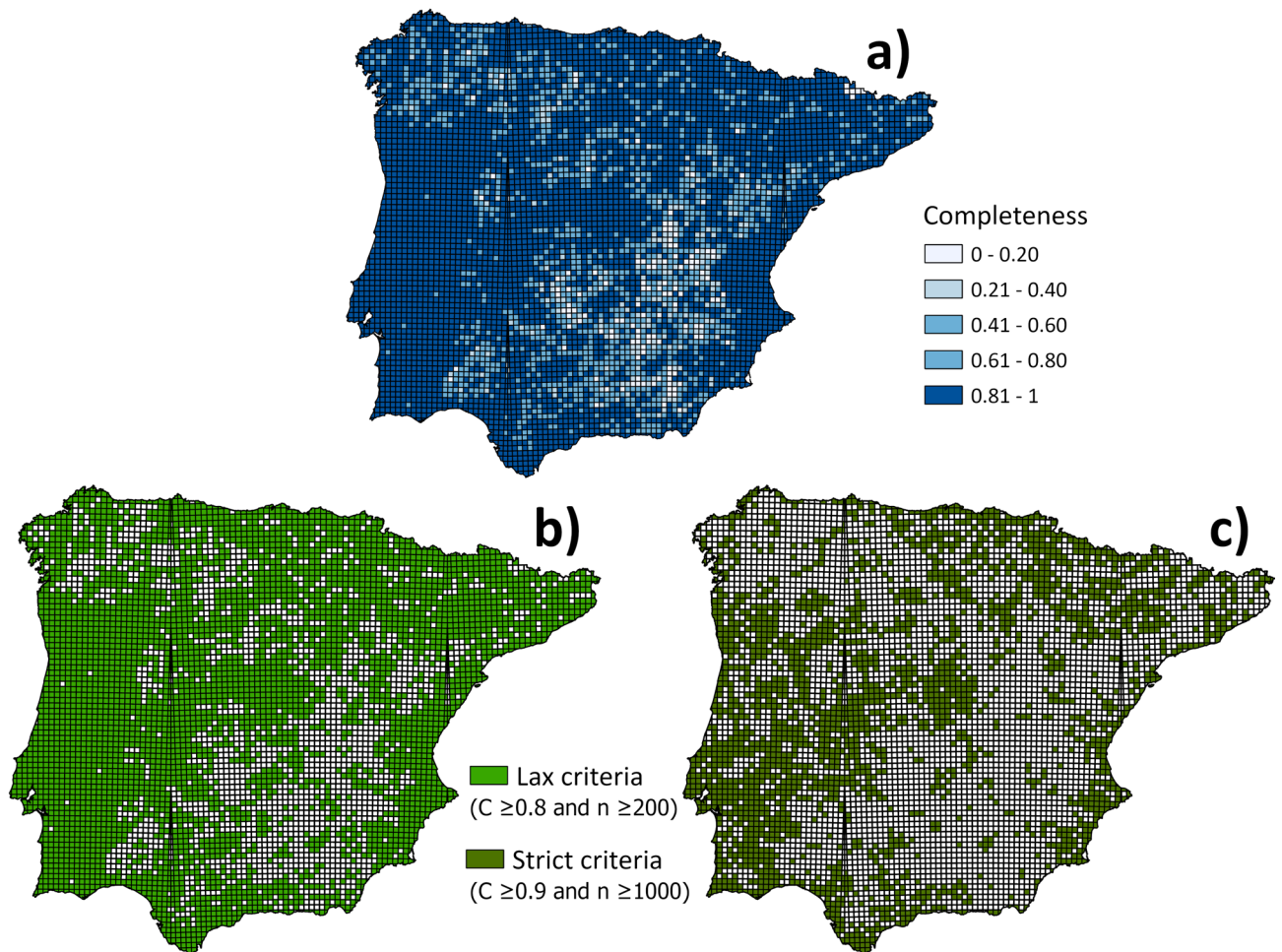


Fig. 2. Bird inventory completeness (C) for the Iberian Peninsula, based on citizen science data obtained from eBird, Observation.org, and GBIF. (a) Map summarizing the C value in 10×10 km cells. Colours indicate C range values. (b) Cells well-surveyed under lax criteria ($C \geq 0.8$ and $n \geq 200$). (c) Cells well-surveyed under strict criteria ($C \geq 0.9$ and $n \geq 1000$).

encompassing renowned birdwatching areas like Doñana National Park or the vicinity of the Strait of Gibraltar (Fig. 2).

Geographical coverage

All 6185 10×10 km cells within the Iberian Peninsula recorded at least one bird observation (Fig. 3). The maximum number of locations per cell reached 603. The CS dataset for the Iberian Peninsula exhibits a significant geographical bias (Moran's $I = 0.59$, $p < 0.01$). Regions with higher concentrations of locations per cell were primarily found along the coast, in the vicinity of major cities (such as Madrid, Lisbon, Barcelona, Valencia, Seville, Córdoba, Oporto or Bilbao), and within well-known protected areas (e.g. Doñana and Monfragüe National Parks, the Ebro delta, etc.) (Fig. 3). Notably, observations predating 1980 (comprising 56,267 records) still demonstrate a significant bias (Moran's $I = 0.32$, $p < 0.01$).

The regions with lower geographical coverage are dispersed across the south-eastern, central, and eastern quadrants of the Iberian Peninsula, typically corresponding to areas characterized by low human density, all within Spain.

Temporal coverage

The temporal distribution of bird records from the Iberian CS data reveals an exponential growth curve (Fig. 4a), attributed to the recent increase in the use of CS applications. The total number of months with data for all species is 793, with the majority of species having fewer than 100 months recorded (Fig. 4b). The temporal coverage for the entire period (1817–2021) was notably low (T index median: 2.7% of the months in the entire period) (Table 4). Focusing solely on the most recent forty years, the frequency of months with data is markedly higher (T index median 11.9%) (Table 4).

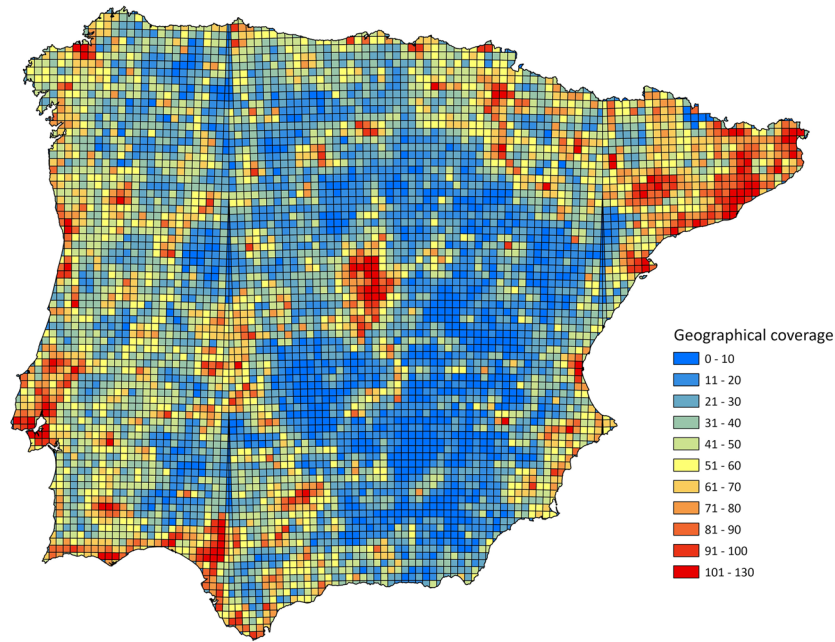


Fig. 3. Geographical coverage (locations per 10 × 10 km cell) of the Iberian bird’s citizen science data for the 1817–2021 period. Colours vary based on the number of locations per cell.

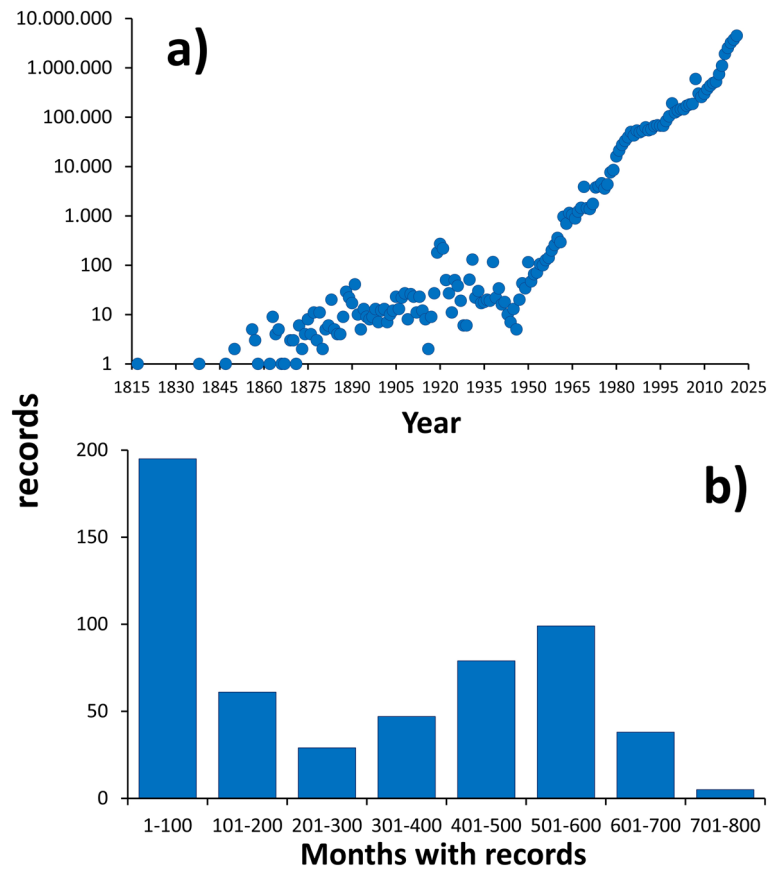


Fig. 4. Temporal coverage (records per year) of the Iberian bird species’ citizen science data between 1817 and 2021. (a) Temporal distribution of records. (b) Frequency distribution of months with records per number of species.

	Mean	Median	Max.	Min.	25% percentile	75% percentile
T (1817–2021)	3.8	2.7	27.8	0.0	1.3	5.0
T (1980–2021)	16.1	11.5	99.4	0.2	5.6	21.2
Months (1817–2021)	83.2	59	603	1	29	108
Months (1980–2021)	81.0	58	501	1	28	107

Table 4. Statistics for the bird’s citizen science temporal dataset from the Iberian Peninsula. Temporal coverage index (T) and total months.

The temporal coverage was generally low or very low for most species (Fig. 4b), with some exceptions such as *Carduelis carduelis* (742 months), *Erithacus rubecula* (716 months), *Alectoris rufa* (715 months), *Turdus merula* (715 months), or *Fringilla coelebs* (711 months) (Suppl. Material 3).

The spatial distribution of temporal coverage across the Iberian Peninsula, measured by the T index per 10 × 10 km cells, was generally low, except for specific locations around the largest cities, emblematic protected areas, and most of the coastal strip (Fig. 5). However, during the period 1980–2021, the T index reached significantly higher values compared to those for the entire period and the temporal Moran’s I indicated significant temporal data aggregation, both for the entire dataset ($I = 0.46$, $p < 0.01$) and the period after 1980 ($I = 0.19$, $p < 0.01$).

Common spatio-temporal patterns

All parameters, including inventory completeness, geographical coverage (locations per 10 × 10 km cell), and temporal coverage (months per 10 × 10 km cell), exhibited strong correlations (Spearman Rank correlations, $Rho > 0.52$, p -value < 0.001 , $n = 6185$) (see Figs. 2, 3, 5). The cells that were best sampled also had higher counts of locations and more recorded dates throughout the entire recording period (1817–2021).

Environmental and land use coverages

Environmental coverage was extensive for Iberian birds, with representatives from well-sampled cells found in more than 80% of the environmental range in all variables. The majority of well-sampled cells fell within the range of 11 to 17 °C for annual mean temperature, between 500 and 1000 mm for annual precipitation, and below 1000 m of elevation (Fig. 6). The distribution of well-sampled cells was narrower compared to that of incomplete cells, indicating a clear bias towards areas with less extreme environmental conditions, often coinciding with more densely populated areas (Kolmogorov’s test, p -value < 0.02 ; Tukey post hoc test identified wetlands and cells without information as the primary contributors to these differences).

The main land uses in the Iberian Peninsula, based on their dominance within each cell, are forestry and agriculture, with a small number of cells designated as artificial surfaces (17) and wetlands (3) (Fig. 7). Despite wetlands being the least represented cell type in the study area, they exhibit the highest coverage, with 100% of cells considered as well-sampled, while the rest of the land uses have proportions ranging between 40 and 60% of well-covered cells. The most dominant land uses, namely forest (2871 cells) and agriculture (2781 cells), display

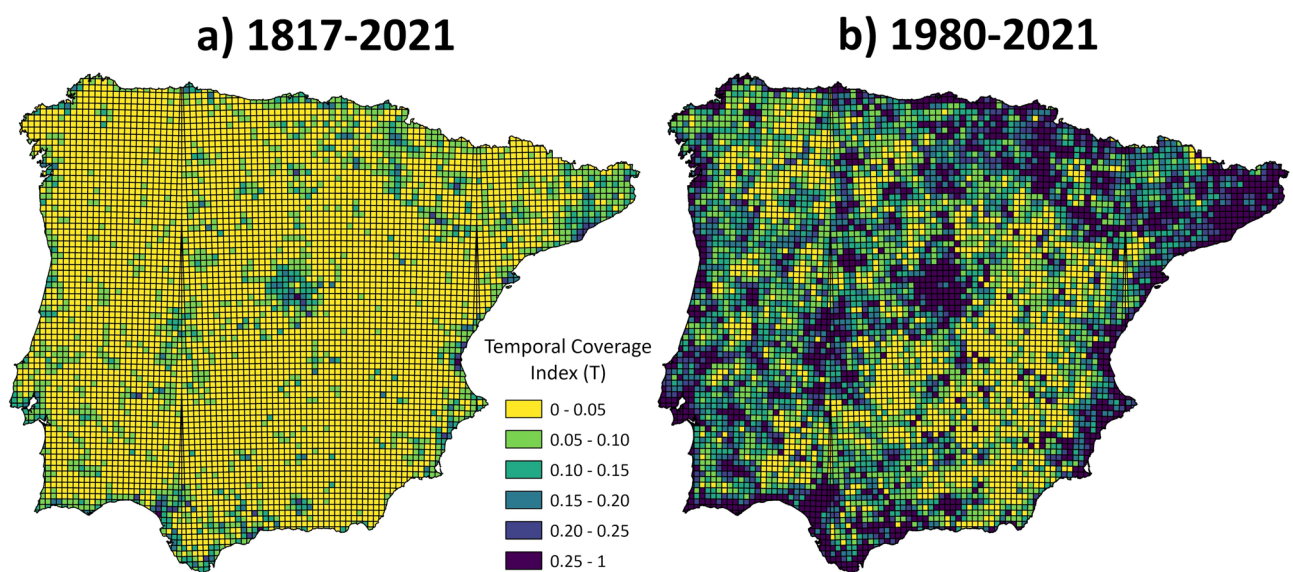


Fig. 5. Map of the temporal coverage Index (T) for the bird species’ citizen science dataset in the Iberian Peninsula between (a) 1817–2021 and (b) 1980–2021. The colours in the legend denote the temporal coverage index T in each grid cell.

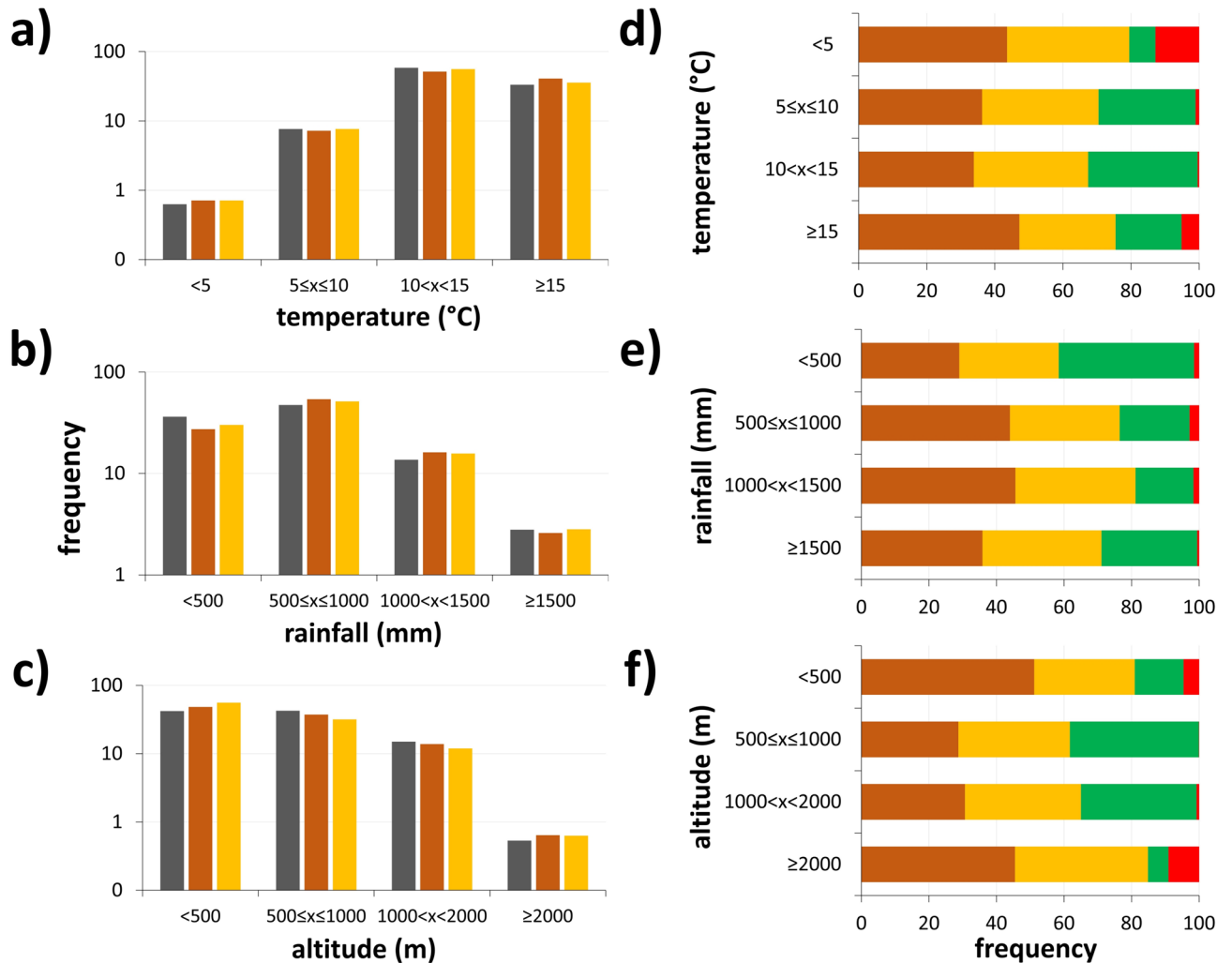


Fig. 6. 10×10 km cell completeness along three environmental variables for the bird species' citizen science dataset in the Iberian Peninsula. Frequency distribution of total (black bars) and well-sampled cells, considering strict ($N \geq 1000$ and $C \geq 0.9$) (brown bars) and lax ($N \geq 200$ and $C \geq 0.8$) (yellow bars) criteria by (a) annual mean temperature, (b) annual rainfall and (c) altitude (ordinates in log-scale). Frequency distribution of cell completeness along the same environmental variables: (d) temperature, (e) rainfall, (f) altitude. In (d–f) brown and yellow bars are well sampled cells according to strict and lax criteria, respectively. Green bars are cells not well-sampled and the red ones, cells with no data.

similar absolute values of cells represented and a comparable proportion of well-covered cells (40.5% and 44.1%, respectively, using strict criteria).

Global coverage of bird citizen science data

Open-access biodiversity datasets from CS platforms have become increasingly important in recent years. Currently, GBIF hosts nearly 2 billion records for the class Aves. Remarkably, over 95% of these records originate from just 40 countries (Table 5). The United States is the largest contributor, with almost 1 billion records. Spain and Portugal, the primary focus of this study, rank 9th and 21st, contributing 44.0 million and 12.3 million records, respectively. There is a significant correlation between the number of bird records and the number of avian orders present in a country (Pearson's correlation on log-transformed data, $r = 0.67$, $p < 0.0001$, $n = 40$). Additionally, the correlation between a country's area and the number of bird occurrences is also significant ($r = 0.32$, $p < 0.05$, $n = 40$). Table 5 arranges countries by the density of bird occurrences per square kilometre. In this ranking, Portugal and Spain are positioned 10th and 14th globally, with densities of 133 and 87 occurrences per square kilometre, respectively. These rankings indicate that the coverage and quality of CS data from the Iberian Peninsula are relatively high compared to the rest of the world.

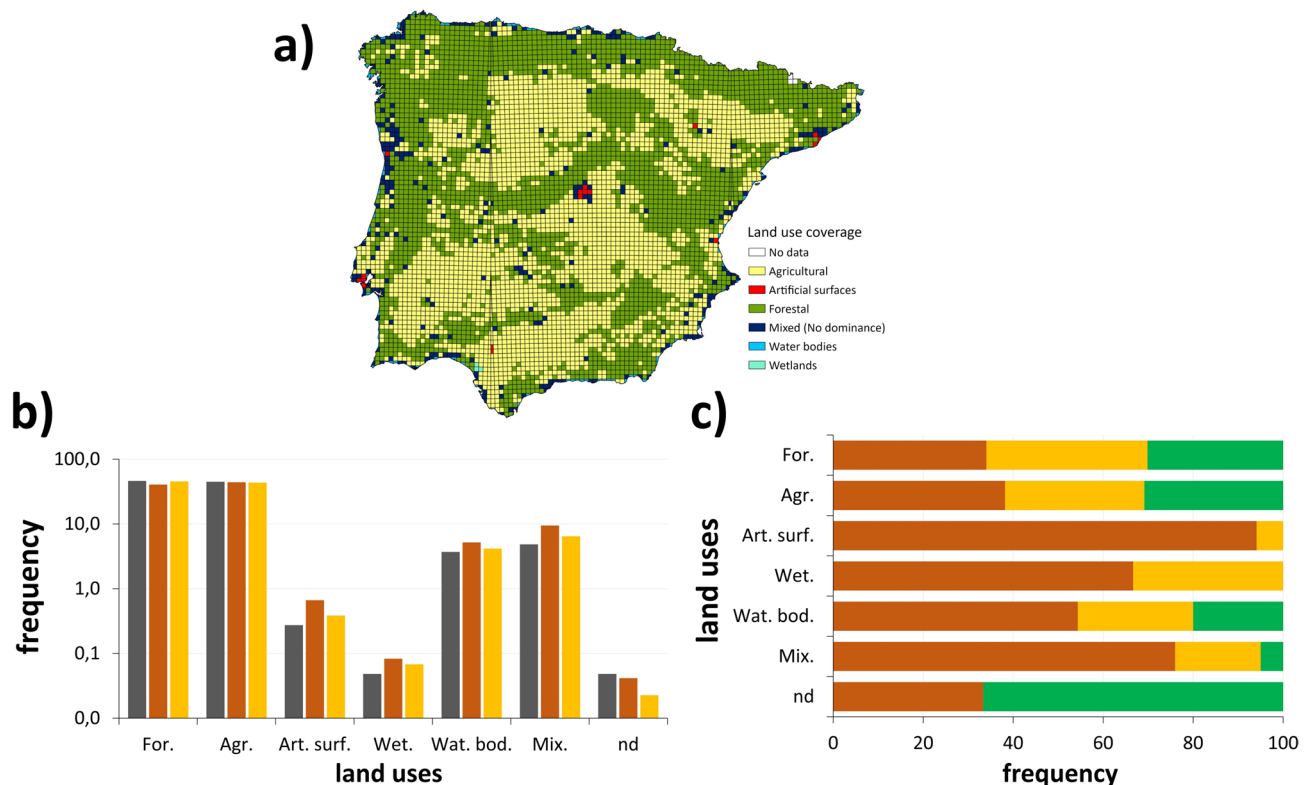


Fig. 7. Land use coverage of grid cells in the Iberian Peninsula bird's citizen science dataset. **(a)** Map illustrating land use distribution per cell. **(b)** Frequency distribution of total 10×10 km cells per land use type (*For.* Forests, *Agr.* Agricultural areas, *Art. surf.* Artificial surfaces, *Wet.* Wetlands, *Wat. bod.* Water bodies, *Mix.* Mixed, and *nd* no land use data), along with the frequency of well-sampled cells. **(c)** Frequency of well-sampled grid cells categorized by lax and strict criteria, and cells not thoroughly surveyed for each land use type. Black bars are total cells, brown and yellow bars are well sampled cells according to strict and lax criteria, respectively. Green bars are cells not well-sampled and the red ones, cells with no data.

Discussion

CS data are of significant interest for biodiversity conservation projects⁴⁷. They contribute to raising awareness of biodiversity and the challenges encountered in such studies, partly due to the growing public interest in biodiversity³. It is crucial for these inventories to maintain integrity concerning spatial, temporal, and ecological dimensions, thus preventing the transmission of uncertain data to future studies³. Insufficient data can result in an unreliable portrayal of species evolution within a territory, leading to incomplete insights into species' threats and other crucial knowledge⁴³. However, it's imperative to stress the importance of data quality to data collectors, as errors such as incorrect dates may be challenging to detect and could yield erroneous outcomes¹⁹. Notwithstanding, numerous successful studies based on CS have achieved significant conservation goals, including inventories, trend analyses, assessing biodiversity impacts, and selecting protected areas (e.g., Australian frogs inventory⁴⁸, fire effects on biodiversity at a large scale⁴⁹, spatiotemporal monitoring of coral reefs⁵⁰, and prioritizing conservation areas⁵¹).

In this study, we have examined the quality of CS data concerning birds in the Iberian Peninsula, driven by the increasing volume of data and user participation across various sources. Consequently, our research serves as an evaluation of the comprehensiveness of CS occurrence data, pinpointing geographic, temporal, and environmental gaps in coverage. It's essential to emphasize that this study should not be construed as an accurate representation of species diversity and richness in the Iberian Peninsula, nor should it be considered a predictor of bird species distribution concerning climate change⁵², owing to the many biases and lack of randomness in the records. Instead, at the moment, it should be utilized for comparative analyses with future inventories⁴.

With a dataset comprising 23,372,637 records, our study represents the most extensive data collection to date in CS assessments, both for the Iberian Peninsula^{18,19} and on a broader scale, including regions such as China, Mexico, or Brazil^{4-17, 43}. This significant volume of data can be attributed to the heightened level of volunteer surveillance in Europe compared to other regions (Table 5), coupled with the predominant focus on birds in the surveyed organisms (accounting for 54% of total records in GBIF)⁵³.

Taxonomic coverage

The taxonomic coverage across different platforms for the Iberian birds shows notable consistency, with slightly higher coverage observed in GBIF. While minor discrepancies in species representation among platforms may highlight the importance of utilizing all three sources, practical considerations suggest initially prioritizing the

	Country	Occurrences ($\times 10^6$)	Orders	Species	Area* (km ²) ($\times 10^3$)	Human population* ($\times 10^6$)	Occ./Order ($\times 10^5$)	Occ./area
1	Netherlands	52.7	32	971	42	17.7	16.5	1269
2	Denmark	34.5	33	978	43	5.9	10.5	804
3	Belgium	18.6	33	791	31	11.7	5.6	609
4	Taiwan	17.0	24	823	36	23.9	7.1	472
5	Costa Rica	21.1	31	1165	51	5.2	6.8	413
6	Belize	6.4	26	664	23	0.4	2.5	280
7	Israel	5.2	28	613	22	9.6	1.9	235
8	United Kingdom	48.2	33	1255	244	67.0	14.6	198
9	Sweden	87.2	31	765	529	10.5	28.1	165
10	Portugal	12.3	30	723	92	10.4	4.1	133
11	France	55.0	38	1128	549	68.0	14.5	100
12	USA	970.7	41	3241	9832	333.3	236.7	99
13	Panama	7.0	26	1179	75	4.4	2.7	92
14	Spain	44.0	37	983	506	47.8	11.9	87
15	Estonia	3.0	25	455	45	1.3	1.2	67
16	Finland	21.6	29	568	338	5.6	7.4	64
17	Norway	30.1	26	621	625	5.5	11.6	48
18	Germany	16.2	37	1753	358	83.8	4.4	45
19	Guatemala	4.2	25	911	109	17.4	1.7	39
20	Ecuador	9.4	30	1999	256	18.0	3.1	37
21	Czech Republic	2.9	27	466	79	10.7	1.1	36
22	Honduras	3.1	26	868	112	10.4	1.2	28
23	South Africa	32.3	31	1169	1219	59.9	10.4	26
24	New Zealand	6.7	25	548	268	5.1	2.7	25
25	Colombia	20.9	35	2315	1141	51.9	6.0	18
26	Poland	5.2	28	617	313	36.8	1.8	16
27	India	48.9	31	1727	3287	1417.2	15.8	15
28	Australia	82.5	33	1257	7741	26.0	25.0	11
29	Mexico	20.2	31	1504	1964	127.5	6.5	10
30	Thailand	5.1	24	1222	513	71.7	2.1	10
31	Canada	149.5	33	1148	15634	38.9	45.3	10
32	Chile	6.2	31	755	757	19.6	2.0	8
33	Japan	2.8	32	940	378	125.1	0.9	7
34	Kenya	3.8	29	1473	580	54.0	1.3	7
35	Peru	7.1	31	2230	1285	34.0	2.3	6
36	Argentina	12.3	30	1347	2780	46.2	4.1	4
37	Venezuela	3.4	29	1619	912	28.3	1.2	4
38	Brazil	18.1	36	2267	8516	215.3	5.0	2
39	China	3.2	29	1643	9563	1412.2	1.1	0
40	Russia	3.8	30	903	17098	144.2	1.3	0

Table 5. Statistics for the top 40 countries with the highest total bird occurrences in GBIF⁴⁰. This table includes data on the number of bird Orders and species recorded, the area of the country (in km²), and its human population. To control for the effects of taxonomic richness and the geographical area, the GBIF occurrences are also expressed as ratios: occurrences per Order and occurrences per km². *<https://data.worldbank.org/>. GBIF data accessed 7 July 2024.

most comprehensive database. GBIF aggregates data from eBird, Observation.org, museums, and others, but the lag time associated with data upload to GBIF implies the practicality of leveraging all available platforms. However, adopting this approach introduces additional challenges in data filtering and processing due to the heterogeneity of projects within GBIF, ranging from structured (strict protocols and methodologies, ensuring high data quality and consistency, ideal for precise scientific studies but requiring significant participant training and resource management; e.g., atlases and academic projects) to semi-structured (provide guidelines but allow flexibility, balancing between consistent data collection and participant autonomy, making them adaptable to diverse conditions while potentially introducing some data variability; e.g., eBird) or unstructured (grant complete freedom in data collection, encouraging broad participation and large data volumes, but often at the cost of data standardization and quality; e.g., iNaturalist) formats. As an efficient alternative, using only eBird data

may circumvent issues associated with mixed-origin data in GBIF or data scarcity in Observation.org. eBird pools the majority of bird data and offers the advantages of a semi-structured project.

Taxonomic coverage is a critical consideration as it can introduce bias when users prioritize certain taxa in their records (e.g., threatened species, rarities) or focus their surveys on areas where sought-after species are known to occur⁵⁴. This bias becomes evident when observing the concentration of well-sampled grid cells in popular natural spaces. However, our assessment reveals robust coverage across all taxonomic orders, with values exceeding 90% in most cases. For instance, Passeriformes (208 out of 217 listed species) and Charadriiformes (115 out of 124 species) demonstrate excellent coverage, with half of the total orders achieving 100% coverage of their species. Overall, the checklist for all bird species in the Iberian Peninsula also surpasses 90% coverage. Furthermore, orders with lower coverage typically comprise only a few species, with gaps often representing rarities rather than common species. For example, Otidiformes and Pterocliiformes exemplify this pattern, with the missing species, *Chlamydotis undulata* and *Syrhaptes paradoxus*, respectively, being notable rarities in the Iberian Peninsula. In other studies, low coverage within an order may also be attributed to the challenging observation of seabirds, given their tendency to inhabit pelagic areas for extended periods⁴³.

Geographical inventory completeness

Completeness, defined as the ratio between observed and expected species richness^{4,5}, is a crucial metric for assessing the utility of CS data^{6,7}. We have employed stringent criteria to identify cells with complete inventories (requiring values above 0.8 for lax criteria and 0.9 for strict criteria). This approach reveals that many areas initially deemed inadequately covered actually exhibit relatively high completeness (as illustrated by the extensive region with completeness values ranging between 0.6 and 0.8 in Fig. 3). Furthermore, instances of incompleteness may not necessarily stem from a lack of surveys but rather from data that has not yet been digitized, rendering it inaccessible for analysis⁶.

Our findings indicate a high level of completeness (mean completeness: 0.68), with a minimal number of cells categorized as not well surveyed, and all cells containing at least one bird record. We observed superior coverage particularly in Portugal and the eastern Spanish region. The areas with the highest sampling coincide with those having the greatest number of locations per cell. However, the distribution of sampled cells was non-random and exhibited significant bias, as indicated by the geographical Moran's I coefficient of 0.59⁴². This points out an uneven distribution of records, with a tendency to cluster in specific areas such as large metropolitan regions and protected areas. These findings align with similar studies, such as those conducted in India, which also report geographical completeness biases⁵.

The only prior study examining the quality of CS data coverage in the Iberian Peninsula focused on certain invertebrates¹⁸ and mammals¹⁹, identifying a few well-sampled regions, notably mid-western, north and north-east and eastern Spain. While this pattern is similar to our findings for avifauna, differences exist in the distribution of well-sampled cells. Whereas well-sampled areas for mammals appear to stem from specialized monitoring surveys, our analysis of bird data reveals a clear pattern linked to popular visitor destinations. Notably, well-sampled cells are frequently situated in major urban areas, densely populated coastal regions, and renowned natural habitats such as popular wetlands. Conversely, significant gaps in sampling occur in sparsely populated interior regions of central Spain. It's important to note that these regions are not necessarily devoid of bird biodiversity; rather, the records may not fully capture the true richness of these areas due to lower attractiveness to birdwatchers⁴³.

Other studies focusing on birds typically exhibit a proportionally smaller number of records and more pronounced gaps, as evidenced by research conducted in China, Mexico, and India. This phenomenon can be attributed to the relatively short history of birdwatching in these regions⁴³. Nonetheless, the exponential growth of CS data suggests that these gaps may only be temporary and could diminish in the coming years. In contrast, the study focused on mammals in the Iberian Peninsula revealed a low inventory completeness, with approximately 90% of cells failing to meet lax criteria, compared to around 30% in our avifauna study. This discrepancy may be attributed to differences in survey efforts between mammals and birds. Mammals, with their elusive behaviour and many being nocturnal, present greater challenges for observation compared to birds, which are often easily spotted flying or active during the day¹⁹. However, it's worth noting that the best-surveyed areas identified in the mammal study coincide with those highlighted in our avifauna analysis, although our study also emphasizes well-sampled urban areas and Portugal¹⁹.

Temporal coverage

Temporal coverage reveals a remarkable surge in data sharing following the widespread adoption of CS platforms, with exponential growth observed since around 2016, culminating in over 3 million records annually by 2021. This exponential trend extends back to approximately 1960 if we consider the period from 1817 to 1974. The exponential growth in data sharing raises the possibility that by the time of publication, the number of records will have increased substantially. This pattern prompts us to question whether the knowledge gap prior to the advent of CS platforms reflects a genuine absence of data or rather a problem stemming from limited data sharing.

The frequency distribution of months with available data exhibits discernible patterns, both natural and influenced by the activity of birdwatchers, leading to specific areas with better temporal coverage than expected⁴³. When considering the temporal index for the entire study period, we observe generally low values across the study area, with notable exceptions in major urban areas and certain coastal regions. However, focusing solely on the period from 1980 to 2021, the temporal index shows a considerable increase. This suggests that gaps in temporal coverage were more pronounced until the 1980s, with improvements in temporal completeness coinciding with the emergence of biodiversity platforms and CS initiatives². Regarding species-specific temporal coverage, it was generally low for most species, with some exceptions.

Common spatio-temporal patterns

It is logical to expect correlations between inventory completeness, geographical coverage, and temporal coverage, as they are often influenced by the same factors. Indeed, there is a strong correlation among them, as the most active observers tend to collect data from the same birdwatching sites (for amateurs) or study areas (for scientists). Consequently, these areas exhibit better coverage across all analysed aspects, as observed in previous studies^{4,6,7}. However, depending on the scale of the study, this phenomenon can introduce an autocorrelation bias that researchers must carefully consider before utilizing the data.

Environmental and land use coverages

The Iberian Peninsula is predominantly characterized by two main land uses: forestry, which encompasses natural or spontaneous vegetation, not limited to forests, and agricultural land. Areas well-surveyed under strict criteria slightly outnumber those in forestry use. This can be attributed to the higher human activity in agricultural regions, which provides increased opportunities for data collection. Additionally, forested areas in Spain tend to be sparsely populated and are less frequented by birdwatchers, although they may attract hikers who typically contribute less data. However, when lax criterion is considered, forested areas show moderate completeness compared to agricultural ones. The most comprehensively surveyed areas are artificial surfaces, primarily urban areas, followed by mixed land use areas with no dominance, and wetlands. This observation can be attributed to the intensive visitation of wetlands by birdwatchers, drawn by the high bird diversity in these habitats⁵⁵.

Environmental coverage appears to be high, with a more thorough surveying level observed in areas meeting specific environmental conditions. These conditions include an annual mean temperature either higher than 15 °C or lower than 5 °C, an annual mean precipitation ranging between 500 and 1500 mm, and altitudes either less than 500 m or exceeding 2000 m. This observation aligns with the concentration of human population in temperate zones⁵⁶. Interestingly, areas at the highest altitudes and with the lowest temperature also demonstrate a well-surveyed status. This phenomenon may be attributed to travellers in mountainous regions who document bird sightings during their journeys, even if these sightings primarily involve species they were actively seeking. However, the lax criterion in these areas tends to be higher compared to more moderate climates or densely populated regions. This difference may be due to the relatively smaller number of surveyors in these extreme environmental conditions.

Future prospects for bird's CS data in Iberia

CS is emerging as a crucial tool for biodiversity studies and has the potential to engage local communities in biodiversity conservation efforts²⁸. However, CS data exhibits several gaps and biases that necessitate addressing. First and foremost is the issue of data quality⁵⁷, which is heavily influenced by two main factors⁵⁸: (1) the preferences and skills of volunteers in identifying and counting birds, and (2) environmental variability that affects bird detectability (e.g., weather conditions, time of day, season, and disturbances such as background noise). Volunteer site preferences, inconsistent attention to different species, and varying time investment in sampling can introduce significant biases into the data. To mitigate these issues, some CS projects, such as eBird, encourage data collection through semi-structured protocols to ensure a minimum level of data quality. eBird allows participants to record the type of observation (e.g., point count, transect, incidental, historical, or other), the start time, and the duration of sampling. Such semi-structured approaches help standardize the data and improve its reliability. Moreover, depending on the taxa or complexities of data collection, CS practitioners may require training and mentoring²⁹. Various strategies exist to address low-quality data in CS projects, including spatiotemporal gaps and biases, both for existing datasets and future sampling efforts. Callaghan et al.⁵⁹ discuss several statistical techniques used to mitigate biases. These include accounting for differences in observer skills, pooling, filtering, or subsampling data to reduce errors and uneven sampling effort, and employing a 'big data' approach to enhance data quality (see⁵⁹ for detailed references). To fully harness the potential of CS platforms in filling gaps in our understanding of global biodiversity and driving research and conservation initiatives, it's necessary to expand survey efforts in under-sampled regions worldwide, particularly those with high conservation value^{60,61}.

The study conducted in China⁴³ provides valuable strategies to address significant information gaps for future CS projects: (1) Prioritize observation sessions in areas with low data coverage, guided by expert researchers. (2) Implement alternative monitoring techniques utilizing new technologies such as aerial vehicles and cameras. These innovative approaches can complement traditional field surveys and provide additional insights into species distribution and abundance. (3) Utilize indirect information from specimens, such as DNA samples. Incorporating genetic data can offer valuable knowledge into species diversity and relationships, particularly in cases where direct observation is challenging. And finally, (4) leverage existing literature on bird distribution⁴³. Incorporating information from scientific studies and field guides can enhance our understanding of species occurrence and habitat preferences. Another study underscores the importance of promoting less charismatic species to garner attention and conservation efforts. By highlighting the significance of these overlooked organisms, scientists can encourage society to prioritize their knowledge⁵⁴. Similarly, Callahan et al.⁶² summarize strategies to enhance CS sampling for significant gains in data quality through three key points: (1) shifting focus from taxon-specific approaches to spatial and temporal considerations, (2) recommending high-value sampling sites to collaborators, and (3) fostering meaningful participation in sampling via dynamic incentive systems. Implementing these proposals, while acknowledging real-world constraints, could substantially improve data quality and enhance the ability to measure bird trends over space and time, a crucial objective amid the ongoing Sixth Extinction.

In the case of the Iberian Peninsula, our data indicates that CS on birds is approaching completeness, making it suitable for scientific studies. However, it's crucial to thoroughly review the gaps identified in inventory,

geographical, temporal, land use, and environmental coverages to mitigate potential biases in research outcomes. The ongoing exponential growth in the number of records and users on CS platforms in the Iberian Peninsula is expected to partially alleviate some of the identified gaps. Nevertheless, we encourage the Iberian birder community to extend their focus to areas with deficient information. By actively contributing data from these underserved regions, birdwatchers can play a vital role in further enhancing the comprehensiveness and accuracy of biodiversity knowledge in the region.

Data availability

The dataset, “Birds of the Iberian Peninsula (1817–2021)”, is available on Mendeley Data: <https://data.mendeley.com/datasets/wrk27sjtc/2>.

Received: 7 March 2024; Accepted: 21 August 2024

Published online: 02 September 2024

References

1. Heberling, J. M. et al. Data integration enables global biodiversity synthesis. *PNAS* **118**, e2018093118 (2021).
2. Callaghan, C. T. et al. Three frontiers for the future of biodiversity research using citizen science data. *BioScience* **71**, 55–63 (2021).
3. Troudet, J., Vignes-Lebbe, R., Grandcolas, P. & Legendre, F. The increasing disconnection of primary biodiversity data from specimens: How does it happen and how to handle it? *Syst. Biol.* **67**, 1110–1119 (2018).
4. Peterson, A. T., Navarro-Sigüenza, A. G. & Martínez-Meyer, E. Digital accessible knowledge and well-inventoried sites for birds in Mexico: Baseline sites for measuring faunistic change. *PeerJ* **4**, e2362 (2016).
5. Peterson, A. T., Kumar, R. S., Nair, M. V. & Talukdar, G. Digital Accessible Knowledge of the birds of India: Characterizing gaps in time and space. *Curr. Sci.* **115**, 35–42 (2018).
6. Sousa-Baena, M. S., Garcia, L. C. & Peterson, A. T. Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Divers. Distrib.* **20**, 369–381 (2014).
7. Troia, M. J. & McManamay, R. A. Filling in the GAPS: Evaluating completeness and coverage of open-access biodiversity databases in the United States. *Ecol. Evol.* **6**, 4654–4669 (2016).
8. Gajji, S. et al. Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. *Biodivers. Inform.* **8**, 94–172 (2013).
9. Meyer, C., Krefl, H., Guralnick, R. & Jetz, W. Global priorities for an effective information basis of biodiversity distributions. *Nat. Commun.* **6**, 8221 (2015).
10. Geldmann, J. et al. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Divers. Distrib.* **22**, 1139–1149 (2016).
11. Ward, D. F. Understanding sampling and taxonomic biases recorded by citizen scientists. *J. Insect Conserv.* **18**, 753–756 (2014).
12. Nieto-Lugilde, D., Maguire, K. C., Blois, J. L., Williams, J. W. & Fitzpatrick, M. C. Multiresponse algorithms for community-level modelling: Review of theory, applications, and comparison to species distribution models. *Methods Ecol. Evol.* **9**, 834–848 (2018).
13. Beck, J., Böller, M., Erhardt, A. & Schwanhart, W. Spatial bias in the GBIF database and its effect on modelling species’ geographic distributions. *Ecol. Inform.* **19**, 10–15 (2014).
14. Feeley, K. J. & Silman, M. R. Keep collecting accurate species distribution modelling requires more collections than previously thought. *Divers. Distrib.* **17**, 1132–1140 (2011).
15. He, J. K., Krefl, H., Gao, E., Wang, Z. C. & Jiang, H. S. Patterns and drivers of zoogeographical regions of terrestrial vertebrates in China. *J. Biogeogr.* **44**, 1172–1184 (2017).
16. Tingley, M. W. & Beissinger, S. R. Detecting range shifts from historical species occurrences: New perspectives on old data. *Trends Ecol. Evol.* **24**, 625–633 (2009).
17. Freeman, B. & Peterson, A. T. Completeness of digital accessible knowledge of the birds of western Africa: Priorities for survey. *Condor* **121**, 035 (2019).
18. Lobo, J. M. et al. KnowBR: An application to map the geographical variation of survey effort and identify well-sampled areas from biodiversity databases. *Ecol. Ind.* **91**, 241–248 (2018).
19. Escribano, N., Galicia, D. & Ariño, A. H. Completeness of digital accessible knowledge (DAK) about terrestrial mammals in the Iberian Peninsula. *PLoS ONE* **14**, e0213542 (2019).
20. Myers, N. et al. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858 (2000).
21. Nieto-Feliner, G. Southern European glacial refugia: A tale of tales. *Taxon* **60**, 365–372 (2011).
22. Crespo-Blanc, A., Comas, M. & Balanyá, J. C. Clues for a Tortonian reconstruction of the Gibraltar Arc: Structural pattern, deformation diachronism and block rotations. *Tectonophysics* **683**, 308–324 (2016).
23. Booth-Rea, G., Ranero, C. & Grevemeyer, I. The Alboran volcanic-arc modulated the Messinian faunal exchange and salinity crisis. *Sci. Rep.* **8**, 1–14 (2018).
24. Rouco, M. et al. *Lista de las aves de España. Edición de 2019* (SEO/BirdLife, 2019).
25. Matias, R. et al. Lista sistemática das aves de Portugal Continental. *Anuário Ornitol.* **5**, 74–132 (2007).
26. Matias, R. et al. Adições e alterações à lista das aves de Portugal Continental: Primeira actualização. *Anuário Ornitol.* **8**, 105–117 (2011).
27. Guerrero, I. et al. Agricultural intensification affects birds’ trait diversity across Europe. *Basic Appl. Ecol.* **74**, 40–48 (2024).
28. Dickinson, J. L., Zuckerberg, B. & Bonter, D. N. Citizen science as an ecological research tool: Challenges and benefits. *Annu. Rev. Ecol. Syst.* **41**, 149–172 (2010).
29. Schmeller, D. S. et al. Building capacity in biodiversity monitoring at the global scale. *Biodivers. Conserv.* **26**, 2765–2790 (2017).
30. Gibert, J., Gibert, L. & Iglesias, A. The Gibraltar strait: A Pleistocene door of Europe? *Hum. Evol.* **18**, 147–160 (2003).
31. Gómez, A. & Lunt, D. H. Refugia within Refugia: Patterns of phylogeographic concordance in the Iberian Peninsula. In *Phylogeography of Southern European Refugia* (eds Weiss, S. & Ferrand, N.) 155–188 (Springer, 2007).
32. Garcia-Vázquez, A., Pinto Llona, A. C. & Grandal-d’Anglade, A. Post-glacial colonization of Western Europe brown bears from a cryptic Atlantic refugium out of the Iberian Peninsula. *Hist. Biol.* **31**, 618–630 (2019).
33. Finlayson, C. *Birds of the Strait of Gibraltar* (T. & A. D. Poyser, 1992).
34. BirdLife International. Spotlight on flyways. *State of the World’s Birds*. <http://datazone.birdlife.org/sowb/spotFlyway> (2010)
35. Goicolea, T. & Mateo-Sánchez, M. C. Static vs dynamic connectivity: How landscape changes affect connectivity predictions in the Iberian Peninsula. *Landsc. Ecol.* **37**, 1855–1870 (2022).
36. European Environment Agency. *The Natura 2000 Protected Areas Network*. <https://www.eea.europa.eu/themes/biodiversity/natura-2000> (2018).
37. Alcaraz, D., Paruelo, J. & Cabello, J. Identification of current ecosystem functional types in the Iberian Peninsula. *Glob. Ecol. Biogeogr.* **15**, 200–212 (2006).
38. eBird. *An Online Database of Bird Distribution and Abundance* (Cornell Lab of Ornithology, 2024).

39. Observation.org. <https://observation.org/> (2024).
40. GBIF.org. <https://www.gbif.org/> (2024).
41. del Hoyo, J. & Collar, N. *HBW and Birdlife International Illustrated Checklist of the Birds of the World, Vol 2: Passerines* (Lynx Edicions, 2016).
42. Legendre, P. & Legendre, L. *Numerical Ecology* (Elsevier, 2012).
43. Huang, X., Lin, C. & Ji, L. The persistent multi-dimensional biases of biodiversity digital accessible knowledge of birds in China. *Biodivers. Conserv.* **29**, 3287–3311 (2020).
44. Meyer, C., Weigelt, P. & Kreft, H. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* **19**, 992–1006 (2016).
45. Kosmala, M., Wiggins, A., Swanson, A. & Simmons, B. Assessing data quality in citizen science. *Front. Ecol. Environ.* **14**, 551–560 (2016).
46. Fick, S. E. & Hijmans, R. J. WorldClim 2: New 1 km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).
47. Soberon, J. & Peterson, T. Biodiversity informatics: Managing and applying primary biodiversity data. *Philos. Trans. R. Soc. B Biol. Sci.* **359**, 689–698 (2004).
48. Rowley, J. J. et al. FrogID: Citizen scientists provide validated biodiversity data on frogs of Australia. *Herpetol. Conserv. Biol.* **14**, 155–170 (2019).
49. Kirchhoff, C. et al. Rapidly mapping fire effects on biodiversity at a large-scale using citizen science. *Sci. Total Environ.* **755**, 142348 (2021).
50. Branchini, S. et al. Using a citizen science program to monitor coral reef biodiversity through space and time. *Biodivers. Conserv.* **24**, 319–336 (2015).
51. Chowdhury, S. et al. Insights from citizen science reveal priority areas for conserving biodiversity in Bangladesh. *One Earth* **6**, 1315–1325 (2023).
52. Triviño, M., Kujala, H., Araújo, M. B. & Cabeza, M. Planning for the future: Identifying conservation priority areas for Iberian birds under climate change. *Landsc. Ecol.* **33**, 659–673 (2018).
53. Chandler, M. et al. Contribution of citizen science towards international biodiversity monitoring. *Biol. Conserv.* **213**, 280–294 (2017).
54. Sastre, P. & Lobo, J. M. Taxonomist survey biases and the unveiling of biodiversity patterns. *Biol. Conserv.* **142**, 462–467 (2009).
55. Junk, W. J. et al. The comparative biodiversity of seven globally important wetlands: A synthesis. *Aquat. Sci.* **68**, 400–414 (2006).
56. Small, C. & Nicholls, R. J. A global analysis of human settlement in coastal zones. *J. Coast. Res.* **19**, 584–599 (2003).
57. Shen, F. Y., Ding, T. S. & Tsai, J. S. Comparing avian species richness estimates from structured and semi-structured citizen science data. *Sci. Rep.* **13**, 1214 (2023).
58. Johnston, A., Matechou, E. & Dennis, E. B. Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods Ecol. Evol.* **14**, 103–116 (2023).
59. Callaghan, C. T. et al. Improving big citizen science data: Moving beyond haphazard sampling. *PLoS Biol.* **17**, e3000357 (2019).
60. Pocock, M. J. et al. A vision for global biodiversity monitoring with citizen science. *Adv. Ecol. Res.* **59**, 169–223 (2018).
61. Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.* **7**, 9132 (2017).
62. Callaghan, C. T. et al. Conservation birding: A quantitative conceptual framework for prioritizing citizen science observations. *Biol. Conserv.* **253**, 108912 (2021).

Acknowledgements

This work was financially supported by the following three projects: (1) “Agricultura y conservación de la avifauna esteparia en Andalucía. Evaluación de la biodiversidad a partir de conocimiento accesible digital (DAK) y estudios in situ” (Andalusian Research Plan 2014–2020. Ref. P20_00354) (Operative FEDER Program); (2) “Procesos de extinción en paisajes agrarios mediterráneos: evaluación de biodiversidad al inicio de una nueva Política Agraria Común mediante conocimiento accesible digital (DAK) y estudios in situ” (Ref. UHU-202076) (Operative FEDER Program) and (3) “Influencia sobre la avifauna del uso y ocupación del territorio en el medio agrario andaluz por plantas fotovoltaicas y olivar intensivo y superintensivo” (BIOD22_00033_15_PPCB. Plan Complementario de I+D+i de Biodiversidad). Gross data for this study were provided by GBIF, eBird, and Observation.org. We extend special gratitude to Julio Rabadán from Observation.org for his invaluable support.

Author contributions

JP developed the concept and curated the initial bibliography, while also coordinating collaboration among the other authors. ELL compiled the results, produced the figures, and conducted certain analyses. JLD filtered the data, performed extensive calculations (including Completeness, Moran’s I, etc.), and handled data extraction. RFV provided expertise in GIS analysis, generated the 10x10 Km cell grid, and extracted environmental, ecoregional, and land use variables. JJN contributed to the discussion and provided feedback on the concept and main findings. All authors critically reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-70827-3>.

Correspondence and requests for materials should be addressed to J.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024