Gorleri Fabricio (Orcid ID: 0000-0002-7992-7831)
Areta Juan Ignacio (Orcid ID: 0000-0001-8588-3030)

Using photographic records to quantify accuracy of bird identifications in citizen science data

FABRICIO C. GORLERI,*[1,2] EMILIO A. JORDAN,[3] IGNACIO ROESLER,[2] DIEGO MONTELEONE[2] & JUAN I. ARETA[1]

[1]*Laboratorio de Ecología, Comportamiento y Sonidos Naturales, Instituto de Bio y Geociencias del Noroeste Argentino (IBIGEO-CONICET), Salta, Argentina*

[2]*Aves Argentinas/Asociación Ornitológica del Plata, Buenos Aires, Argentina*

[3]*Laboratorio de Ornitología, Centro de Investigaciones Científicas y Transferencia de Tecnología a la Producción (CICYTTP-CONICET), Diamante, Entre Ríos, Argentina*

*Corresponding author.

Email: fabriciogorleri@gmail.com

Twitter: @FabriGorleri

Citizen science data are increasingly used for biodiversity monitoring. However, concerns are often raised over the accuracy of species identifications in citizen science databases, as data are collected mostly by non-professionals. Misidentifications can simultaneously generate two error types: false positives (erroneous reports of a species) and false negatives (lack of reports of the misidentified species). Large-scale assessments of identification errors should bring insights into the strengths and weaknesses of citizen-science data. Here we show that citizen science photographic data for birds are trustworthy overall, although problems arise in hard-to-identify bird groups. We reviewed over 104 000 images of 377 passerine species from the southern Neotropics (Argentina) stored in eBird –a large citizen science platform– and quantified erroneous reports to calculate precision and recall metrics as measures for data accuracy. Precision increases with fewer false positives and recall increases with fewer false negatives, thus high values of precision and recall will mirror a higher data accuracy. We found that 97% of the photos of all species were correctly identified. Most species (77%; n = 291) showed high accuracy in their identifications (precision and recall > 95%), with 122 species showing no errors. A few hard-to-identify species (10%; n = 40) showed low levels of data quality (63-90% precision or recall). Similarly, few species (12%; n = 46) exhibited intermediate precision or recall scores (90-95%]. Further, we uncovered the existence of a complex network of cross-identifications composed of 272 species, with a predominance of tyrant-flycatchers and ovenbirds, reflecting the strong traffic of errors that occurs within these families. To our knowledge, our study provides the first large-scale quantification of identification errors in photos submitted by citizen-science contributors. We underscore the relevance of performing such assessments to understand how identification errors are distributed across a database before analyzing data and provide tools for

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/ibi.13137

citizen science stakeholders to direct more specific efforts toward species that need an improvement in data quality.

Keywords: Argentina, eBird, false negatives, false positives, misidentifications, Neotropics, network analysis, Passerines, precision, recall

Citizen science enterprises have become the fastest-growing contributors to bird occurrence data, with increasing application of these data in science and policy (Bonney *et al.* 2009, Cooper *et al.* 2014, Schubert *et al.* 2019). One well-known example is eBird (Sullivan *et al.* 2009), which to date concentrates millions of avian occurrence records of at least 10,500 bird species that are used to increase our knowledge on bird ecology and trends worldwide (Horns *et al.* 2018). Despite their emerging popularity and use, citizen science data are still frequently perceived as unreliable, as data are largely collected by non-professional users (Cohn 2008, Bonney *et al.* 2014, Brown & Williams 2019). Because of the existent variability in the sampling behaviour and expertise among observers (Tulloch & Szabo 2012, Johnston *et al.* 2018), the potential for error and biases in citizen science data is high.

Species misidentifications are a common source of bias in citizen science data. Misidentifications become a problem when they are systematically stored in databases, as they can bias estimates of species distributions derived from the data (Ensing *et al.* 2013, Costa *et al.* 2015). They can also alter our understanding of other relevant aspects of the ecology of a species, such as migratory patterns and phenology (Hull *et al.* 2010, Gorleri & Areta 2022). To avoid data misuse, users of citizen-science data need to be aware of these potential sources of bias (Aubry *et al.* 2017).

Surprisingly, few studies have assessed data quality with regard to species identification in citizen-collected data, even though projects devote large efforts to detecting and curating identification errors (Kosmala *et al.* 2016, Kelling *et al.* 2013, 2015a). Scattered evidence for non-avian taxa suggests that errors may be low or null for most species that are relatively easy to identify, but that errors increase as untrained observers find it harder to identify species (Swanson *et al.* 2016). For birds, information on the identification quality is practically nil, although there are studies highlighting problems caused when citizen scientists confuse hard-to-identify bird species that co-occur (Gorleri & Areta 2022, Rocha-López *et al.* 2021). Because of this lack of information on data quality, critical voices have often raised the question of whether citizen-science data are useful (Kosmala *et al.* 2016). A better understanding of the strengths and weaknesses of citizen-science databases in relation to the accuracy of species identifications is yet required.

Two error types occur when a species is misidentified: false-positives and false-negatives. A false-positive error is created for species A, when species B is erroneously reported as A (Figure

1); simultaneously, a false-negative error is created for species B, as it was misidentified and reported as A instead of as B (Figure 1). Because misidentifications involve different error rates for different species pairs, the proportion of false-positives and false-negative errors is expected to vary from species to species. For example, a given species would primarily suffer false-positive errors if the species is more commonly misreported rather than misidentified when detected. On the other hand, if the species is continuously misidentified when detected, but rarely misreported, it would suffer primarily false-negative errors. Species may suffer both error types if they are often misreported and misidentified.

While detecting and quantifying false-positive errors in databases are relatively easy with documented data, this is not the case for false-negative errors. False positives can be easily found by carefully examining the focal species database for identification errors. Conversely, finding false negatives is difficult because these errors are distributed across the full database and are potentially buried among several non-focal species. False-positive and false-negative errors contained in databases can significantly distort ecological knowledge if they are not corrected before data analysis (Royle & Link 2006, Ruiz-Gutierrez et al. 2016, Gorleri & Areta 2022). The relative weight of false-positive and false-negative errors can be analyzed through precision and recall metrics (see Figure 1 and Methods) that provide critical information on the error structure of a dataset.

On the other hand, misidentifications in databases can be conceptualized as a cross-identification network between species, where the error flow between the constituent species occurs in different directions and magnitudes. For example, errors can be unidirectional, where one species is continually misidentified as another but not the other way around, or mutual, where each of a pair of species is reciprocally confused in similar (symmetrical) or different (asymmetrical) proportions. In turn, there may also be certain species that concentrate errors or irradiate them to many other species, therefore, functioning as core species in the error network. In this context, knowing the degree of interconnection between the different species can be very important to help focus data curation efforts on those taxa (whether species or families) responsible for most misidentifications.

Here, we quantified and characterized false-positive and false-negative errors that stem from misidentifications in photographic reports submitted to eBird, for a broad group of nearly 400 passerine bird species from the southern Neotropics. We also created a network of cross-identifications among species to quantify the strength, reciprocity, and patterns of identification errors. To the best of our knowledge, this is the first thorough assessment of species-level identification data quality in a large citizen-collected avian dataset. The goals of this study are to (1) provide insights into the overall quality of publicly available photographic data gathered by

volunteer data colectors, (2) identify sets of species most likely to be erroneously identified in the database, and (3) propose measures to increase data transparency to encourage data reuse and applicability. We chose a group of Neotropical passerines as our study system, being good candidates for assessing the quality of a volunteer-collected database. Families, such as ovenbirds (Furnariidae), tyrant flycatchers (Tyrannidae), and pipits (Motacillidae) have numerous difficult-to-identify species that coexist in the region, and are particularly challenging for non-expert birdwatchers.

METHODS

Data compilation

We analyzed the photographic reports of 393 species of passerines from 26 families directly through the eBird media explorer tool (https://ebird.org/media). Therefore, the data that we assessed had passed through eBird's data validation systems. We only assessed photographs from Argentina. Analyses were limited to this country because we have a higher level of expertise on the identification and geographic and temporal distribution of species from our home country, all of which are relevant to building accurate networks of cross-identifications. Additionally, Argentina provides a good representation of passerine species that inhabit most of the ecoregions that are present in the Southern Cone of the Neotropics. We evaluated photographs spanning from 2010 to 2020. We reviewed these photographs asynchronously across a six-month period from June to December 2020. Two or more of the five authors identified the species in each photograph, with at least one being an expert on the species. For a list of photographs reviewed, the names of the reviewers, and the date of revision of each photograph, see Supporting Information Appendix S1.

We compiled and classified a total of 103,428 photographs representing 393 passerine species. We classified each photograph as 'correct', 'incorrect' or, when they did not allow us to reach species-level identification but had a possibility of being correctly identified, as 'uncertain'. Where possible, we identified the correct species in misidentified reports. If species-level identification was challenging, we conservatively assigned incorrect original identifications only to the correct genus, family, or order. We reported incorrect identifications through the eBird website or directly by contacting the corresponding data curator.

Data cleaning

Users can upload multiple photographs of the same bird in their eBird checklists. To ensure the independence of each observation, we removed duplicate reports that contained the same information on (1) species reported, (2) classification, (3) observer name, and (4) submission identifier. The column "classification" refers to the identity (species, genus, family, or order) that

we assigned to each report, and the submission identifier refers to the unique code that eBird assigns to each checklist. For species in a checklist having a mix of correctly identified and misidentified reports, we used the "classification" column to retain as separate reports the original designations. For example, the Yellow-winged Blackbird *Agelastiscus thilius* had four photographs uploaded to eBird checklist S80591607, three of which were correctly identified (apparently all photos of the same individual) while one was confused with the Brown-and-yellow Marshbird *Pseudoleistes virescens*. We therefore retained only one record as correctly identified (eliminating the others as potential duplicates) and another record as misidentified as Marshbird. We also excluded photographs showing (1) only a nest or fledgling of the reported species, and (2) no observable or identifiable bird (including habitat or other features not related to birds). Finally, to avoid biasing the quality estimates resulting from a low sample size, species containing five or fewer unique reports (including false-negative reports) were removed from analyses. After data cleaning, we obtained 69,699 unique photographic reports of 377 species.

Species identification quality analysis

We analyzed (1) the overall number of misidentified photographs in our sub-sample of eBird data, and (2) the number of misidentified photographs for each of the 377 study species. The overall accuracy of eBird data in the identification of photographic reports was measured as the number of correct photographs divided by the total and multiplied by 100. To assess identification accuracy for each species we used precision and recall indices that measure the two error types produced through misidentifications: false positives and false negatives (Figure 1). Precision was calculated as the proportion of true positives over the total number of identifiable photographic reports of a given species. Hence, the precision will be lower as the number of false positives for a given species increases in the database (Figure 1). Recall was calculated as the proportion of true positives over the sum of true positives and false negatives of a given species; hence, recall values will decrease as the number of false negatives in the database increases for a given species (Figure 1). If, for example, Chilean Elaenia *Elaenia chilensis* is often misidentified and reported as Small-billed Elaenia *Elaenia parvirostris* by contributors, but not vice-versa (as shown in Gorleri & Areta 2022), the recall score for Chilean Elaenia will be lower than the precision score (as false negatives outnumber false positives), with the opposite occurring for Small-billed Elaenia.

For practical purposes, we grouped species into three subjective quality categories based on the minimum values found for precision and/or recall (1) high-quality group: species with both precision or recall above 95%, (2) moderate-quality group: species with minimum precision or recall ranging from 90 to 95%, and (3) low-quality group: species with minimum precision or recall equal or below 90%. We chose these subjective thresholds because recent research has

demonstrated that accuracy metrics below 90% in citizen science databases may distort ecological estimates, such as the migratory phenology in certain species (Gorleri & Areta 2022; but see Discussion for further refinement).

Finally, to contextualize our findings, we highlighted the species that may present identification challenges to a non-expert birdwatcher. We tagged 85 out of the 377 species analyzed as hard to identify based on similarity in appearance, personal experience with the species, and bibliographical references stating that the species or genus is difficult to identify (see list of species tagged as difficult to identify in Supporting Information Appendix S2). This information aims to provide a better framework for the interpretation and discussion of our results.

Network analysis

We examined complex patterns of cross-identifications among species by performing network analysis. Network analysis allows visualization of how misidentifications behave by plotting the strength, reciprocity, and patterns of interconnections of misidentifications between species. To create the network, we first created a dataset indicating the number of misidentified reports between each possible pair of species and discarding misidentified reports to which we could not assign a correct species-level identification. The resulting dataset consisted of three columns: (1) reported species, (2) misidentified as, and (3) number of misidentifications. With the resulting information, we created the network using the R package *visNetwork* (v2.0.9; Almende *et al.* 2019). The network consisted of a series of nodes that represent each species, interconnected with arrows that represent the direction and magnitude of misidentifications. We used the Fruchterman-Reingold layout, which is a force-directed layout that orders nodes with more connections closer to each other, but repelling nodes when they get too close (Fruchterman & Reingold 1991). The resulting network places highly connected species and groups towards the centre and relegates species with few connections to the periphery. We also calculated attributes of the network, such as degree centrality, which is the number of links a node has to other nodes in the network, both incoming (indegree) and outgoing (outdegree). The R code is available in Supporting Information Appendix S1.

RESULTS

Of the total 69,699 unique photographic reports representing 377 species, 68,101 (97.7%) were correctly identified (true positives), while 1,002 (1.4%) were incorrect (false positives) and 596 (0.9%) were uncertain. There were 937 incorrect reports that we could assign to the correct species (assignable false negatives), while for the remaining 65 incorrect reports we could only identify genus, family, or order (unassignable false negatives). After grouping species into subjective quality groups, we found 291 (77%) species with a score in both precision and recall > 95% (high data quality), while 46 (12%) species had minimum values of either precision or recall ranging

from 90-95% (moderate data quality), and the remaining 40 (11%) species had either precision or recall ≤ 90% (low data quality) (Table 1, Figure 2). Of the 85 species that we initially tagged as difficult to identify, 26 had high, 27 had moderate, and 32 had low-quality data (Supporting Information Appendix S2).

As outlined, most species had high-quality data. In this group of 291 species, we found that 122 achieved a perfect score in both precision and recall, meaning that we did not find any false-positive or false-negative reports for these species. Leading this ranking were: Masked Gnatcatcher *Polioptila dumicola*, Chiguanco Thrush *Turdus chiguanco*, Scarlet-headed Blackbird *Amblyramphus holosericeus*, Lark-like Brushrunner *Coryphistera alaudina*, Many-colored Rush-Tyrant *Tachuris rubrigastra*, and Rufous-browed Peppershrike *Cyclarhis gujanensis* (for detailed full ranking see Supporting Information Appendix S2). As expected, most species that achieved high data quality were easy or relatively easy to identify, and often were very familiar birds, 'backyard' birds, with the exception of Shiny Cowbird *Molothrus bonariensis*. However, this group also included 25 (8.6%) species that we had considered *a priori* as difficult to identify (Supporting Information Appendix S2). In general, these were often rare or range-restricted birds (e.g., Rufous-breasted Leaftosser *Sclerurus scansor*, Olive Spinetail *Cranioleuca obsoleta*, Sharp-billed Treehunter *Heliobletus contaminatus*, Sooty Grassquit *Asemospiza fuliginosa*, Dull-coloured Grassquit *A. obscura*, Fuscous Flycatcher *Cnemotriccus fuscatus*, Mouse-colored Tyrannulet *Phaeomyias murina,* and Cordoba Cinclodes *Cinclodes comechingonus*), or widespread species with little geographic overlap with other similar-looking congeners (e.g., Scale-throated Earthcreeper *Upucerthia dumetaria*, Bar-winged Cinclodes *Cinclodes fuscus*, Tufted Tit-Spinetail *Leptasthenura platensis,* and Band-tailed Earthcreeper *Ochetorhynchus phoenicurus*).

The moderate-quality group contained 46 species, of which 27 (58%) were indicated *a priori* as difficult to identify (Table 1, Figure 2, Supporting Information Appendix S2). The tyrant-flycatchers (Tyrannidae) were the best-represented family with 22 species, with the presence of four species of elaenias (*Elaenia*), and species such as Southern Beardless Tyrannulet *Camptostoma obsoletum*, Southern Scrub Flycatcher *Sublegatus modestus*, and Suiriri Flycatcher *Suiriri suiriri* that were misidentified as several other species of tyrant-flycatchers (i.e., core species in the cross-identification network; see below in this section). The ovenbirds (Furnariidae) were also well-represented with 14 species, with the presence of three cinclodes (*Cinclodes*), three miners (*Geositta*), two canasteros (*Asthenes*), two thornbirds (*Phacellodomus*), two spinetails (*Synallaxis*), and two earthcreepers (*Upucerthia validirostris* and *Ochetorhynchus ruficaudus*).

The lowest quality group was the least numerous with 40 species, of which 32 (80%) had been previously classed as difficult to identify (Table 1, Figure 2, Supporting Information

Appendix S2). Most species belonged to the ovenbirds (Furnariidae; 16 sp.) and tyrant-flycatchers (Tyrannidae; 13 sp.) families. In relative terms, pipits (Motacillidae) was the worst-performing family, as all the three species of this family comprised the low-quality group: Correndera Pipit *Anthus correndera*, Hellmayr's Pipit *A. hellmayri,* and Short-billed Pipit *A. furcatus* (Figure 2). Regarding precision, 27 species had a score ≤ 90% (Figure 2), some with critically low values as in the case of Three-striped Flycatcher *Conopias trivirgatus* (64.2%), a species that was systematically and reciprocally confused with Social Flycatcher *Myiozetetes similis.* Other examples of a critically low precision score included White-winged Cinclodes *Cinclodes atacamensis* (74.2%) confused with various other sympatric *Cinclodes* sp. (Figure 3d), and Buff-fronted Foliage-gleaner *Dendroma rufa* (75%) reciprocally confused with the more common look-alike Ochre-breasted Foliage-Gleaner *Anabacerthia lichtensteini* (Figure 3a). In terms of recall, 17 species had a score equal to or lower than 90% (Figure 2), with the lowest scores reached by Greenish Tyrannulet *Phyllomyias virescens* (63.6%) and Tropical Pewee *Contopus cinereus* (76%), the latter largely reported as Smoke-colored Pewee *C. fumigatus*. However, note that recall indices may be overestimated, as we could not assign all misidentifications to the correct species. Only four species had low values in both precision and recall: Patagonian Forest Earthcreeper *Upucerthia saturatior*, Sclater's Tyrannulet *Phyllomyias sclateri*, Steinbach's Canastero *Pseudasthenes steinbachi,* and Greenish Tyrannulet (Figure 2), meaning that they were incorrectly reported and also reported as another species, thus, having high rates of both false positives and false negatives.

The general network of cross-identifications consisted of different clusters of species that varied from being isolated to poorly or highly interconnected (Figure 3; Supporting Information Appendix S3). Some species exhibited uni- or bidirectional interactions: a species confused only with another one, or pairs of species that were reciprocally confused at varying rates (Figure 3a). Some simple and relatively closed networks showed interactions among multiple species occurring mostly within a specific genus or family (e.g., Figure 3b for pipits [Motacillidae], and Figure 3c for woodcreepers [Dendrocolaptidae]). Other parts of the network became more complex and included a diversity of strengths of interactions among the constituting units, connecting species of different genera or even different families. Two "hotspots of misidentification" were observed: one formed mostly by tyrant flycatchers [Tyrannidae] and the other by ovenbirds [Furnariidae], reflecting the strong "traffic" of misidentifications occurring in these families. These hotspots had at their core a few geographically widespread species that were widely mistaken as several other look-alike species, obtaining high degree centrality scores (Table 2). Illustrated examples include Buff-winged Cinclodes *Cinclodes fuscus* (Figure 3d) in the ovenbirds group, and Southern Beardless Tyrannulet (Figure 3e) in the tyrant-flycatchers group.

DISCUSSION

The reliability of citizen science data in terms of species-identification quality remains a largely unexplored topic despite the increasing trend to use this data source for biodiversity monitoring, ecological and biogeographic studies, and conservation planning. In this work, we assessed the identification accuracy of photographic reports stored in the eBird database for 377 species of southern Neotropical passerines by assessing the effect of false-positive and false-negative reports on precision and recall values. We found that most species (77%; n = 291) had potentially high levels of identification accuracy of their photographic reports, while relatively few hard-to-identify or cryptic species (11%; n = 40) showed concerningly low levels of identification accuracy, and a similarly low number of species exhibited moderate accuracy scores (12%; n = 46). Further, we uncovered the existence of a complex network of cross-identifications, composed of different clusters of isolated to highly interconnected species, which underscores the relevance of large-scale assessments to understand how identification errors parse out across a database.

Strength and weaknesses of citizen-science data

Our large-scale assessment provides new insights into the strengths and weaknesses in the accuracy of species identification in photographic records in volunteer-collected databases. In this sense, we demonstrate that eBird photographic data are robust in general terms, highlighting their usefulness as a reliable source of information. Nonetheless, certain limitations exist in hard-to-identify species groups for which the data should be subject to validation and verification before use (see also Rocha-López *et al.* 2021). Our evaluation of photographic reports, however, is based on a snapshot of the current database of southern Neotropical species. Thus, the scores reported here may vary over time and for other geographic regions due to either change in the conditions under which identification errors occur (e.g., larger extent of range overlap in cryptic species groups or different identification challenges), intrinsic changes in the functioning of the platform (e.g. varying skills in the set of reviewers involved in data vetting; increased user expertise), or development of more refined identification criteria in difficult groups.

The largest limitation of eBird data was found in difficult-to-identify species belonging to cryptic genera, such as canasteros (*Asthenes* and *Pseudasthenes*), spinetails (*Synallaxis*), miners (*Geositta*), wood-pewees (*Contopus*), and pipits (*Anthus*), among others. These species represent a challenge for any database, including those largely compiled and curated by experts, such as museum collections. Of particular concern are those species that had precision or recall values equal to or below 90% (low-quality species group; 40 out of 377 species in our dataset; Figure 2), as the false-positive or false-negative errors contained in their databases can seriously affect analyses, and therefore products derived from their data (Royle & Link 2006, Miller *et al.* 2011, Ruiz-Gutierrez *et al.* 2016).

It is nonetheless important to highlight that moderate accuracies of 90-95% (46 out of 377 species in our dataset) could also threaten the usefulness of datasets. For example, Clare *et al.* (2021) demonstrated that even 5-10% of false-positive errors contained in ecological data could inflate the estimates of a species occurrence by 20-70%. In particular, a small number of errors can strongly affect data analyses if they are heavily clustered in certain geographic regions or seasons of the year which, for instance, may produce shifts towards the contaminant data biasing either spatial analyses (e.g., distribution maps; Costa *et al.* 2015, Aubry *et al.* 2017) or temporal analyses (e.g., phenological assessments; Gorleri & Areta 2022). For these reasons, assessing data quality demands specific and focused assessments to evaluate how errors are distributed through space and time, and thus, our subjective quality thresholds must be used only as rough rules-of-thumb in quality assessments.

Our analysis provides a partial picture of the error structure in the overall database as we quantified observable errors of individual photographs, and not at the checklist level. eBird contributors are encouraged to submit bird checklists with counts of individuals (Sullivan et al. 2009), rather than single photos (as would be the case in projects like iNaturalist: https://www.inaturalist.org), and such checklists often have photographs for only a fraction of all the individuals reported for each species. Thus, the finding of misidentified images of a species in a checklist does not mean that the undocumented records of that species in the same checklist were erroneous. However, assessing the data accuracy of undocumented records is challenging. Currently, reports of species without an accompanying photograph comprise more than 90% of all eBird records in Argentina. Because the repeated examination of photographs by users and data reviewers provides multiple opportunities for error correction, it is likely that photographic records will possess a higher accuracy than undocumented records. This suggests that values of precision and recall informed here may be optimistic in comparison to the real (unknown) error values for the complete dataset of a species, including photographically documented and non-documented records.

In the absence of evidence, it is virtually impossible to measure the error in undocumented ecological datasets. Hence, it is important to encourage citizen-science contributors to routinely add documentary evidence to their checklists whenever possible, complemented by notes describing how a hard-to-identify species was distinguished in the field. Many songbirds with problematic visual identification are easier to identify by their calls or songs and are highly vocal (e.g., *Elaenia* flycatchers, or pipits; Ridgely & Tudor 2009), therefore, sound recordings are also a feasible resource to use that could help decrease errors in datasets. On the other hand, model-based solutions exist to establish the statistical relevance of false-positive and false-negative errors contained in undocumented data (Miller et. al. 2011, Ruiz-Gutierrez *et al.* 2015, Clare *et al.* 2021). These models measure uncertainty in undocumented datasets for false-positive and

false-negative errors by using auxiliary data validated by experts with information on how errors behave in the studied taxa (but see Cruickshank *et al.* 2019). We strongly recommend researchers to explore these analytical methods to avoid misleading conclusions when analyzing undocumented data for hard-to-identify species.

Network analysis as a tool to characterize misidentifications

The network of cross-identifications that we have developed constitutes a useful tool to identify where more efforts are needed to increase data quality and showcase how errors are distributed in a large database by depicting how species interact. These networks can be used profitably to uncover misidentification hotspots, to pinpoint how the sets of easy-to-confuse species vary temporally and across space, and how symmetric or asymmetric are the confusion webs. In agreement with our previous expectations, our network revealed the existence of "core" species with a high degree of interconnections. These “core” species tend to have dull plumage (for example, brown or grey in ovenbirds or olive in tyrant flycatchers), small size, and a wide geographic range; attributes that may increase the chances of being misidentified as other species with overlapping distributions. Targeting “core” species may be fundamental for citizen-science projects as they can focus curation efforts (and contributors’ training) on these species, which would allow the most relevant misidentification links to be cut.

The observed links in the network of cross-identifications also revealed that the flow of errors may be asymmetrical between species. While "source" species have a larger number of outgoing errors (false positives of the focal species) than the number of incoming errors (false negatives of the focal species), in "sink" species the exact opposite occurs. In our network, for instance, White-winged Cinclodes functioned mostly as a source species, having a larger number of connections targeting other cinclodes, whereas Planalto Woodcreeper *Dendrocolaptes platyrostris* functioned as a sink species, since it only received errors from three other woodcreepers and did not distribute any (see Figure 3 D-B). This asymmetry in the flow of errors has direct implications for the recall and precision values. While source species would have lower precision than recall due to a larger number of false positives in relation to false negatives in the data, the opposite would occur in sink species. This complex asymmetry in the webs of confusion further highlights the usefulness of examining the misidentification problem through networking.

Finally, we only found two non-passerines in the passerine misidentification network: Picui Ground Dove *Columbina picui* reported as Saffron Finch *Sicalis flaveola* and Black-faced Ibis *Theristicus melanopis* reported as Scale-throated Earthcreeper. In both cases, these are clearly distinct pairs of species suggesting that the cause of error is data input rather than in-the-field misidentification. Quantifying such errors is complex as the intent of users during data entry is generally unknown. In any case, we believe that their overall impact is minimal compared to

actual identification errors, as we only detected a handful of errors involving species that are unlikely to be confused.

Recommendations to improve citizen-science data quality

The potentially high data quality in eBird data may stem from factors involving data validation systems adopted by the platform and the users participating in the project. eBird covers a broad front for error detection through the use of smart filters that detect unusual or outlier observations for a date and locality during data entry and the participation of expert reviewers (Wood *et al.* 2011). This combination of smart filters and expert reviewers generates an active feedback loop between humans and computers that demonstrated to improve the data quality of the eBird platform (Kelling *et al.* 2013). In Argentina, for example, 40 expert reviewers scour incoming data for accuracy and actively curate the database with the help of more than 145 filters set to flag unusual entries at a county level. Even so, current data validation systems have some flaws. In particular they fail to flag errors in several difficult-to-identify species, either because smart filters are not specifically designed to detect identification errors accurately (being better at flagging spatio-temporal anomalies during data entry; see Bonter & Cooper 2012), or because data reviewers are not adequately trained to detect errors in cryptic or hard-to-identify species (Gorleri & Areta 2022). On the other hand, recent evidence suggests that birders who are active in citizen science initiatives often have a high degree of specialization in bird identification (Kelling *et al.* 2015b, Randler 2021, Rosenblatt *et al.* 2022). It is, therefore, to be expected that, if most contributors are specialized birders, then the identification of species will be generally accurate when entered into eBird, in particular for those species that are not difficult to identify.

Unfortunately, because we were only able to analyze data from public eBird outputs, we were unable to quantify how much of the data was correct or incorrect when submitted by contributors. This is because eBird does not provide publicly available information about changes in identification or record removals following subsequent data revisions. We were therefore unable to distinguish whether the records that we analyzed were correctly entered, or whether the original identification was changed afterwards. We suggest that citizen science programs strive to include public metadata to allow tracking of changes in identifications (as, for instance, iNaturalist currently does). Having knowledge of the change history of each record will increase transparency and allow data users and citizen-science stakeholders to assess how erroneous data is entered into the database, and to identify where data-validation systems are either failing or are being more successful.

The need to reduce misidentification errors is important for future broad-scale citizen-science programs. Well-directed training can improve the identification skills of participants rapidly (Falk *et al.* 2019). This simple strategy can be used by citizen science stakeholders by

focusing training efforts on the core species responsible for most cross-identifications, or among species of cryptic genera. For large projects, various digital communication strategies can be used to reach a wide audience quickly, for example, through online bird identification workshops or the use of social media to spread targeted information. Another step that could be taken by citizen-science platforms is to highlight species or species groups in which data quality is known or suspected to be compromised by misidentifications. This would provide platform users and researchers with a useful reminder that these species demand more careful procedures when uploading and analyzing the data. In this context, Artificial Intelligence (AI) may also play a critical role for data quality in a timely and cost-effective manner, as these approaches become more sophisticated (Balázs *et al.* 2021, Sun *et al.* 2021). One recent example of AI applied to citizen science was outlined by Wessels *et al.* (2019) who combined expert knowledge and machine-induced models to identify unreliable observations automatically from a large volume of bird records in Europe, with an accuracy of ~85% in detecting erroneous data. Moreover, it is no longer far-fetched to imagine that models of smart photograph or audio recognition, for example, Merlin Bird ID (https://merlin.allaboutbirds.org) or BirdNET (Kahl *et al.* 2021), may soon be applied to data validation systems, resulting in a useful tool for automated error recognition during data entry.

In conclusion, researchers need to be cautious when analysing citizen science data (Areta & Juhant 2019, Gorleri & Areta 2022, Rocha-López *et al.* 2021), if their focal species are difficult to identify. We hope this work fosters a more critical use of citizen science data by researchers, besides providing tools for curators, reviewers, and managers of citizen-science data to direct more specific efforts toward species that need an improvement in data quality. We believe that citizen science has enormous potential as a trustworthy data source to serve science, policy, and conservation. Knowing and managing the inherent biases of each species will provide us with even more robust information.

We thank Emiliano Depino, Matías Juhant, Ingrid Holzmann, Juliana Benitez, Freddy Burgos, and Juan Amaya (all members of the ECOSON lab) for providing writing and statistical feedback on the manuscript. We thank the eBird Argentina reviewer team for their voluntary contribution to data quality. We also thank the reviewers W. Douglas Robinson and Judit Szabo, and the associate editors Dr. Alexandre Millon and Prof. Jeremy Wilson for the valuable suggestions and the improvement of the manuscript.

Data Availability Statement

The datasets used in this paper and the reproducible R code to perform all analyses are available at Zenodo: https://zenodo.org/record/6828335

REFERENCES

Almende, B.V, Thieurmel B. & Titouan, R. 2019. visNetwork: Network Visualization using 'vis.js' Library. R package version 2.0.9. https://CRAN.R-project.org/package=visNetwork

Areta, J.I. & Juhant, M.A. 2019. The Rufous-thighed Kite *Harpagus diodon* is not an endemic breeder of the Atlantic Forest: lessons to assess Wallacean shortfalls. *Ibis (Lond. 1859).* **161**: 337–345.

Aubry, K.B., Raley, C.M. & McKelvey, K.S. 2017. The importance of data quality for generating reliable distribution models for rare, elusive, and cryptic species. *PLoS One* **12**: 1–17.

Balázs, B., Mooney, P., Nováková, E., Bastin, L. & Arsanjani, J.J. 2021. Data Quality in Citizen Science. In: *The Science of Citizen Science* (K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, R. Samson, & K. Wagenknecht, eds), pp. 139–157. Springer, Cham.

Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S., Phillips, T., Kenneth, V., Shirk, J., Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S. & Phillips, T. 2009. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *Bio Sci.* **59**: 977–984.

Bonney, R., Shirk, J.L., Phillips, T.B., Wiggins, A., Ballard, H.L., Miller-Rushing, A.J. & Parrish, J.K. 2014. Next steps for citizen science. *Science* **343**: 1436–1437.

Bonter, D.N. & Cooper, C.B. 2012. Data validation in citizen science: a case study from Project FeederWatch. *Front. Ecol. Environ.* **10**: 305–307.

Brown, E.D. & Williams, B.K. 2019. The potential for citizen science to produce reliable and useful information in ecology. *Conserv. Biol.* **33**: 561–569.

Clare, J.D.J., Townsend, P.A. & Zuckerberg, B. 2021. Generalized model-based solutions to false-positive error in species detection/nondetection data. *Ecology* **102**: e03241.

Cohn, J. 2008. Citizen Science: Can Volunteers Do Real Research? *Bio Sci.* **58**: 192–197.

Cooper, C.B., Shirk, J. & Zuckerberg, B. 2014. The invisible prevalence of citizen science in global research: Migratory birds and climate change. *PLoS One* **9**: e106508.

Costa, H., Foody, G.M., Jiménez, S. & Silva, L. 2015. Impacts of Species Misidentification on Species Distribution Modeling with Presence-Only data. *ISPRS Int. J. Geo-Information* **4**: 2496–2518.

Cruickshank, S.S., Bühler, C. & Schmidt, B.R. 2019. Quantifying data quality in a citizen science monitoring program: False negatives, false positives and occupancy trends. *Conserv. Sci. Pract.* **1**: 1–14.

Ensing, D.J., Moffat, C.E. & Pither, J. 2013. Taxonomic identification errors generate misleading ecological niche model predictions of an invasive hawkweed. *Botany* **91**: 137–147.

Falk, S., Foster, G., Comont, R., Conroy, J., Bostock, H., Salisbury, A., Kilbey, D., Bennett, J. & Smith, B. 2019. Evaluating the ability of citizen scientists to identify bumblebee (*Bombus*) species. *PLoS One* **14**: 1–21.

Fruchterman, T.M.J. & Reingold, E.M. 1991. Graph Drawing by Force-directed Placement. *Software—Practice Exp.* **21**: 1129–1164.

Gorleri, F.C. & Areta, J.I. 2022. Misidentifications in citizen science bias the phenological estimates of two hard-to-identify *Elaenia* flycatchers. *Ibis (Lond. 1859)*. **164**: 13–26.

Horns, J.J., Adler, F.R. & Ş, H. 2018. Using opportunistic citizen science data to estimate avian population trends. *Biol. Conserv.* **221**: 151–159.

Hull, J.M., Fish, A.M., Keane, J.J., Mori, S.R., Sacks, B.N. & Hull, A.C. 2010. Estimation of Species Identification Error: Implications for Raptor Migration Counts and Trend Estimation. *J. Wildl. Manage.* **74**: 1326–1334.

Johnston, A., Fink, D., Hochachka, W.M. & Kelling, S. 2018. Estimates of observer expertise improve species distributions from citizen science data. *Methods Ecol. Evol.* **9**: 88–97.

Kahl, S., Wood, C.M., Eibl, M. & Klinck, H. 2021. BirdNET: A deep learning solution for avian diversity monitoring. *Ecol. Inform.* 61: 101236.

Kelling, S., Fink, D., La Sorte, F.A., Johnston, A., Bruns, N.E. & Hochachka, W.M. 2015a. Taking a 'Big Data' approach to data quality in a citizen science project. *Ambio* **44**: 601–611.

Kelling, S., Johnston, A., Hochachka, W.M., Iliff, M., Fink, D., Gerbracht, J., Lagoze, C., La Sorte, F.A., Moore, T., Wiggins, A., Wong, W.K., Wood, C. & Yu, J. 2015b. Can observation skills of citizen scientists be estimated using species accumulation curves? *PLoS One* **10**: 1–20.

Kelling, S., Lagoze, C., Wong, W., Yu, J., Damoulas, T., Gerbracht, J., Fink, D. & Gomes, C. 2013. eBird: A Human/Computer Conservation and Research. *AI Mag.* **34**: 10–20.

Kosmala, M., Wiggins, A., Swanson, A. & Simmons, B. 2016. Assessing data quality in citizen science. *Front. Ecol. Environ.* **14**: 551–560.

Miller, D., Nichols, J., McClintock, B., Campbell, E., Bailey, L. & Weir, L. 2011. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology* **92**: 1422–1428.

Pearman, M. & Areta, J.I. 2020. *Birds of Argentina and the South-west Atlantic*. Field Guide. Helm, London.

Randler, C. 2021. Users of a citizen science platform for bird data collection differ from other birdwatchers in knowledge and degree of specialization. *Glob. Ecol. Conserv.* **27**: e01580.

Ridgely, R.S. & Tudor, G. 2009. Field Guide to the Songbirds of South America. University of Texas Press, Austin.

Rocha-López, D., Quiñonez-Calle, M., Carantón-Ayala, D., Betancur-López, A. & Acevedo-Charry, O. 2021. La importancia de obtener evidencia multimedia: el caso de los semilleros piquigordos de Colombia, con registros de *Sporophila atrirostris* y un llamado a buscar *Sporophila maximiliani*. *Bol. SAO* **30**: 22–31.

Rosenblatt, C.J., Dayer, A.A., Duberstein, J.N., Phillips, T.B., Harshaw, H.W., Fulton, D.C., Cole, N.W., Raedeke, A.H., Rutter, J.D. & Wood, C.L. 2022. Highly specialized recreationists contribute the most to the citizen science project eBird. *Ornithol. Appl.* **124**: 1–16.

Royle, J.A. & Link, W.A. 2006. Generalized Site Occupancy Models Allowing for False Positives and False Negatives Errors. *Ecology* **87**: 835–841.

Ruiz-Gutierrez, V., Hooten, M.B. & Campbell Grant, E.H. 2016. Uncertainty in biological monitoring: a framework for data collection and analysis to account for multiple sources of sampling bias. *Methods Ecol. Evol.* **7**: 900–909.

Schubert, S.C., Manica, L.T. & Guaraldo, A.D.C. 2019. Revealing the potential of a huge citizen-science platform to study bird migration. *Emu* **119**: 364–373.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D. & Kelling, S. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biol. Conserv.* **142**: 2282–2292.

Sun, J., Futahashi, R. & Yamanaka, T. 2021. Improving the Accuracy of Species Identification by Combining Deep Learning With Field Occurrence Records. *Front. Ecol. Evol.* **9**: 1–10.

Swanson, A., Kosmala, M., Lintott, C. & Packer, C. 2016. A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conserv. Biol.* **30**: 520–531.

Tulloch, A.I.T. & Szabo, J.K. 2012. A behavioural ecology approach to understand volunteer surveying for citizen science datasets. *Emu* **112**: 313–325.

Wessels, P., Moran, N., Johnston, A. & Wang, W. 2019. Hybrid expert ensembles for identifying unreliable data in citizen science. *Eng. Appl. Artif. Intell.* **81**: 200–212.

Wood, C., Sullivan, B., Iliff, M., Fink, D. & Kelling, S. 2011. eBird: Engaging Birders in Science and Conservation. *PLoS Biol.* **9**: e1001220.

Figures

Figure 1. Schematic representation showing how false-positive and false-negative reports are simultaneously generated in databases when a species is misidentified. Note that a false-positive report of a species A simultaneously results in a false-negative report for another species B. Precision and recall metrics serve to evaluate the overall identification quality of a species database by considering the number of false positives (precision) and false negatives (recall) about the true positives. Illustrations from Pearman and Areta (2020) were reproduced with permission. This is a simplification of a more general problem; the flow of errors is better reflected in a network of cross-identifications (see Figure 3).

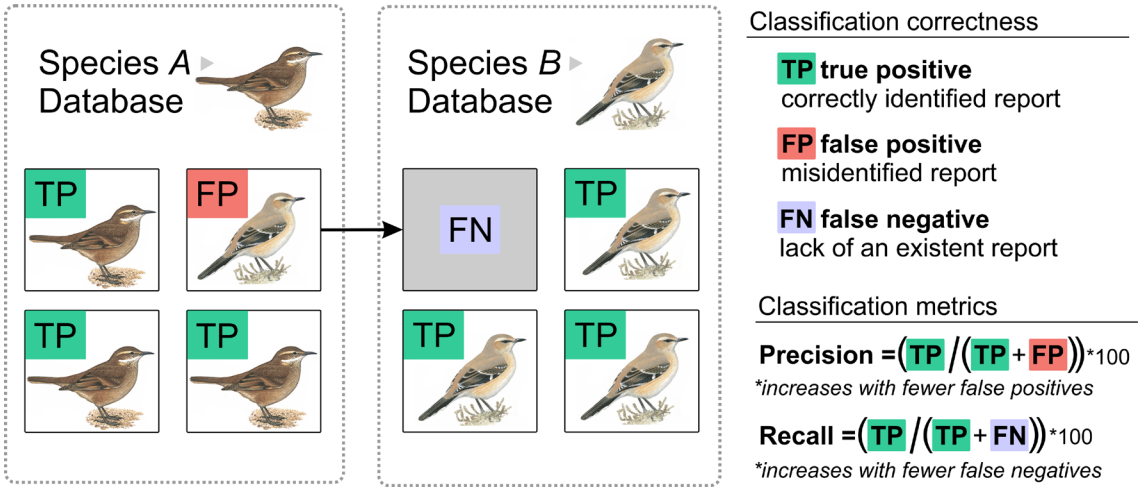


Table 1. Number of species of southern Neotropical passerines scoring high (green), moderate (orange), or low (red) in precision and recall metrics in the identification quality of photos submitted to eBird. High: species with scores > 95%, moderate: species with scores ranging from 90-95%, and low: species scoring ≤ 90%.

		Precision		
Recall		**High** (95-100%]	**Moderate** (90-95%]	**Low** [0-90%]
	High	291	15	18
	Moderate	20	11	5
	Low	4	9	4

Table 2. Top 20 species of southern Neotropical passerines with the highest degree-centrality scores obtained from the network of cross-identifications (see Figure 3). The degree centrality scores indicate the number of connections (links) a species has to other species in the network. While indegree refers to the number of incoming links of a species (number of arrows from other species pointing to the focal species), outdegree is the number of outgoing links (number of arrows from the focal species pointing to other species). Total links indicate the sum of indegree and outdegree.

Species	Family	Indegree	Outdegree	Total
Asthenes pyrrholeuca	Furnariidae	9	8	17
Molothrus bonariensis	Icteridae	8	9	17
Serpophaga subcristata/griseicapilla	Tyrannidae	9	8	17
Suiriri suiriri	Tyrannidae	5	10	15
Asthenes baeri	Furnariidae	6	8	14
Synallaxis frontalis	Furnariidae	5	9	14
Camptostoma obsoletum	Tyrannidae	8	4	12
Elaenia parvirostris	Tyrannidae	6	6	12
Sicalis flaveola	Thraupidae	4	7	11
Elaenia spectabilis	Tyrannidae	4	6	10
Myiophobus fasciatus	Tyrannidae	5	5	10
Rhopospina fruticeti	Thraupidae	6	4	10
Sublegatus modestus	Tyrannidae	5	5	10
Asthenes modesta	Furnariidae	4	5	9
Chrysomus ruficapillus	Icteridae	4	5	9
Cinclodes fuscus	Furnariidae	5	4	9
Molothrus rufoaxillaris	Icteridae	4	5	9
Phacellodomus sibilatrix	Furnariidae	4	5	9
Agelasticus cyanopus	Icteridae	1	7	8
Phacellodomus striaticollis	Furnariidae	4	4	8

Figure 2. Identification accuracy ranking in photographic reports submitted to eBird for 377 species of southern Neotropical passerines. The accuracy was measured using the precision and recall metrics (see Figure 1 and Table 1), and the ranking was based on the minimum value of either precision or recall for each species. Left panel: ranking including all species (see the ranking in detail in Supporting Information Appendix S2). Right panel: ranking showing only species comprising the "low-quality species group", i.e. species with either precision or recall scores ≤ 90%.

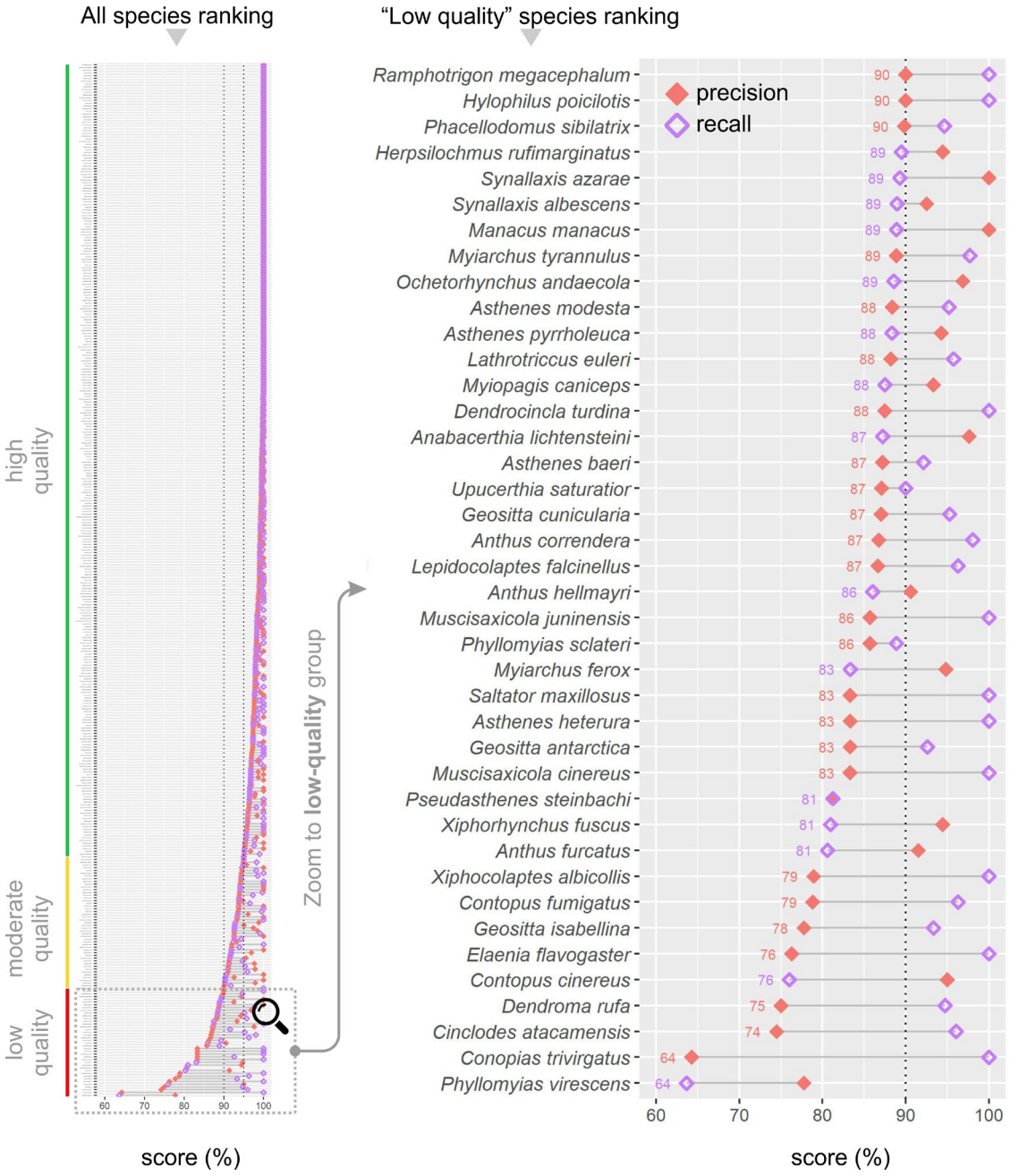
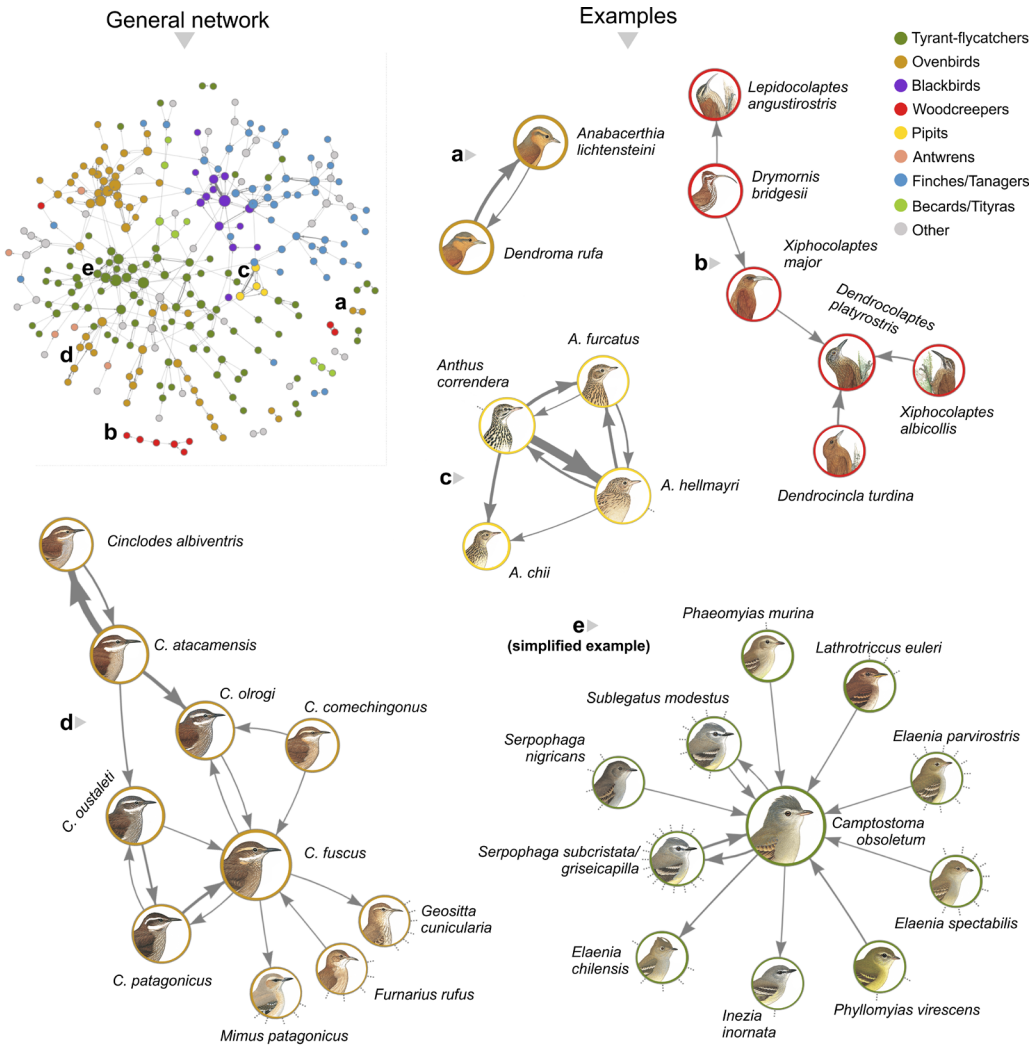


Figure 3. Network of cross-identifications in photo reports of 377 species of southern Neotropical passerines submitted to eBird. Each circle represents a species, coloured by family, with nearby circles representing species that are connected through misidentification. Arrows indicate the magnitude and direction of misidentifications, with thicker arrows corresponding to a higher number of misidentifications among the constituting units. Small, dashed lines indicate not-shown connections with species in or outside of the focal group. While examples (a-d) are accurate representations extracted from the general network, we simplified and re-arranged the example "e" relationships to represent only misidentifications involving the core species *Camptostoma obsoletum*. See full-resolution network in Supporting Information Appendix S3. Illustrations from Pearman and Areta (2020) were reproduced with permission.



Supporting Information

Appendix S1. The full dataset used in this study with a reproducible R code to perform data quality and network analyses. Available at: https://zenodo.org/record/6828335

Appendix S2. All species ranking of identification accuracy of photo reports submitted to eBird in Argentina. The ranking is first ordered by the minimum value found for either precision and recall scores, and second by the number of samples analyzed for each species. Species that were tagged as difficult to identify are indicated as 'TRUE' in column D named 'hard_to_id'. Available at: https://zenodo.org/record/6828335

Appendix S3. High-resolution network (Html file). Available at: https://zenodo.org/record/6828335