Biogeographia – The Journal of Integrative Biogeography 36 (2021): s002
SPECIAL SECTION: Citizen Science in Biogeography
https://doi.org/10.21426/B636049648

Scientists by chance: reliability of non-structured primary biodiversity data. Insights from Italian Forums of Natural Sciences

STEFANO DE FELICI[1,*], PAOLO MAZZEI[2,3], VALERIO SBORDONI[1,4], DONATELLA CESARONI[1]

[1] *Department of Biology, University of Rome "Tor Vergata", via della Ricerca Scientifica, 00133 Roma (Italy)*
[2] *Forum Natura Mediterraneo, www.naturamediterraneo.com/forum*
[3] *Forum Entomologi Italiani, www.entomologiitaliani.net/forum*
[4] *National Academy of the Sciences, called "Accademia dei XL", Via L. Spallanzani 5a/7, 00161 Roma (Italy)*
** corresponding author, email: stefano.de.felici@uniroma2.it*

Keywords: Biodiversity data, butterflies, citizen science, natural science forum, social networks.

SUMMARY

Forums and social networks store a big deal of data on flora and fauna, collected especially by amateurs. To what extent are these data useful to contribute to biodiversity data systems? In this paper, we addressed the question about the "suitability for use" of primary biodiversity data by exploring two popular and valued Italian Forums of Natural Science (Forum Natura Mediterraneo and Forum Entomologi Italiani) and tried to assess their scientific potential. The aim of our work was to evaluate and discuss taxonomic reliability of the identification of butterfly species and the accuracy of their geographic locations. For each forum thread, we examined the posted images of butterflies, checked the diagnoses and georeferenced the observations from the textual descriptions provided by the users. Then, we compared each final identification by users with an independent identification by expert taxonomists. Looking at species level identifications, users identified 3764 out of 4029 specimens (93.4%) and experts agreed with them in 3649 cases: a high percentage agreement (p_o = 96.9%). As for the geographic data, we were able to georeferenced 97.9% of the observations (70% with an estimated extent less than 2500m). Results of this study, although limited to butterflies, suggest that the final identifications from forums show a surprisingly small bias and that the 'democratic' approach to taxonomy ultimately produces few uncertainties. The selected forums contain large amounts of primary biodiversity data in digital format, correctly identified and georeferenced with satisfactory accuracy and this capital is too valuable to remain unused. The formalization of collaborations with scientific projects and institutions would bring the forums in the area of "official" citizen science initiatives, giving the forums a role of citizens' scientific training. The recognition of a scientific role makes forum

managers and users more deeply involved and data protection over time, currently entrusted to forum managers, would be greatly enhanced.

INTRODUCTION

Research in natural sciences has a long tradition of cooperation between professional researchers and amateurs (Dickinson and Bonney 2012, Miller-Rushing et al. 2012, Vermeulen et al. 2013). In recent decades, advances in Information and Communications Technologies (ICT) have multiplied the possibility to create new ways of networking between actors of a different background or large and dispersed research groups (Baker 2015, Wagner et al. 2015). Networking between natural scientists facilitates the global sharing of primary data on biodiversity, taxonomies, environmental data and general information (see e.g. GBIF, Catalogue of Life, LTER, LifeWatch).

One of the most significant developments of the impact of the Internet on environmental science is the "new dawn for Citizen Science" (Science Communication Unit 2013, Silvertown 2009), i.e. the growing involvement of non-scientist citizens in scientific projects. Citizen Science (CS) is a concept and a term highlighted both in the policy agenda of the European Commission and in the research community (European Union, 2020). In a relatively short time, the support of volunteers has become a pivotal method to approach large scale monitoring projects (Chandler et al. 2017, Devictor et al. 2010, Hallmann et al. 2017, Kullenberg and Kasperowski 2016). The CS approach is increasingly used in scientific projects to enlarge and improve the knowledge on biodiversity by collecting georeferenced data on the presence of rare or protected species or habitats (Kallimanis et al. 2017, Katsanevakis et al. 2015, Kullenberg and Kasperowski 2016, Martellos et al. 2016, Zapponi et al. 2017), setting up surveillance networks for alien species (Johnson et al. 2020, Maistrello et al. 2016, Zenetos et al. 2013) or digitalizing paper labels from museum specimens (e.g. Ellwood et al. 2015, 2018, Flemons and Berents 2012, Hill et al. 2012). More generally, a CS approach may be addressed to any problems in which human skills are irreplaceable for collecting and/or processing massive amount of data (Aceves-Bueno et al. 2017, Xue 2014).

As far as biodiversity is concerned, forums and social networks store a big deal of data on flora and fauna, collected especially by amateurs. Are these data potentially useful to contribute to biodiversity data systems?

In most cases, biodiversity data from generic social networks lack reliability, because they are not checked by expert taxonomists. These data, therefore, require a strenuous reconsideration by experts if they are being used. Instead, naturalists' forums are a particular way of networking, where users exchange knowledge and opinions through posts and related comments. Although these forums do not underlie a formal scientific CS project, they usually focus on a topic and both passionate non-scientist amateurs and professional researchers participate in discussions. As a consequence, these forums are potentially valuable sources of primary biodiversity data (Barve 2014, Lin et al. 2015, Morris et al. 2013). However, it should be considered that the aim of naturalistic forums of Natural Science is to promote the exchange of information between fans of specific taxonomic groups and not to maintain a rigorously scientific standard. Within forums, a thread usually starts when a user posts photo of animals or plants, asking for help in identifying them. Identifying an organism at the species level using only photographic images can be quite simple if the key features are clearly visible, yet not if the crucial characters are hidden. In the latter case, correct identification of the species is impossible and doubts about alternative species remain. As a rule, expert taxonomists manage and moderate forum

threads among more or less experienced users. Moderators also play an educational role and try to involve users in more in-depth discussions on the taxonomy, biology, distribution of the species of interest. Over the years, naturalists' forums stored a great deal of biodiversity data.

Now, the question is: do these data, produced by inexperienced citizens without following any scientific project guidelines, constitute a reliable source of primary data to feed big data systems on biodiversity?

In this paper, we addressed this basic question about the "suitability for use" of primary biodiversity data from the natural sciences forums and tried to assess their scientific potential. In particular, the aim of our work was to evaluate and discuss taxonomic reliability of the identification of butterfly species and their geographic location accuracy going through two popular and valued Italian naturalists' forums: "Forum Natura Mediterraneo" (FNM) and "Forum Entomologi Italiani" (FEI), having different stories, users and editorial policies. The pilot dataset only concerned butterflies (Lepidoptera Rhopalocera), for two main reasons: i) a limited number of taxa, included in a few Lepidoptera families; ii) a popular taxon among amateurs who like to observe, study and learn about these attractive insects.

Starting from forum threads devoted to identifying butterfly species through photos, we attempted to verify the correctness of user identifications and explored the advisability of obtaining accurate spatial and temporal data for each observation.

MATERIALS AND METHODS

Recording identifications and evaluation of their level of certainty

For each thread, our first steps were to examine the posted images, to carefully check the diagnoses and geo-reference the observations from the textual descriptions provided by the observers. Secondly, each final identification by users was compared with an independent identification carried out by expert taxonomists. The team of taxonomists included three forum moderators (in Acknowledgements) and the authors of this paper. The moderators only evaluated the threads in which they were not originally involved. We considered the identifications of the experts as "the gold standard", that is the best possible identification.

The identification of butterfly species from photos can sometimes be problematic, both for users and experts. In cases of unsolvable ambiguity in the identifications, alternative specific names were both written in the final diagnosis, for example *Erebia stiria / styx*, *Hipparchia blachieri / semele, Plebejus argyrognomon / argus.*

Reading the comments of users and moderators, different levels of certainty of the identifications can be recognized. Based on the wording of the various comments, we standardized the statements concerning the confidence of identification by scoring the level of certainty into five categories of s-value (Table 1). We assigned a level of certainty only to identifications at the species level; we did not assign any level of certainty to all other cases, namely: i) identifications at the genus or higher taxon level; ii) identifications with combined alternative names; iii) missing identifications. Similarly, we also scored the level of certainty of the identifications by experts.

In order to compare the levels of certainty of identification between users and experts we used a specific statistical analysis. The statistical adequacy of the inter-rater agreement coefficients is a controversial issue and any coefficients can suffer severe limitations according to the data nature and structure (Aceves-Bueno et al. 2017, Gwet 2014, Uebersax 2015). The percentage (or overall) agreement rate p_o, i.e. the sum of the agreement divided by the sample size is a

straightforward and widely used coefficient, based on empirical concept of probability, that can be considered a useful descriptive agreement statistic of the agreement (Uebersax, 2015).

$$p_o = \frac{\sum n_{ii}}{N}$$

In this case, p_o indicates the percentage of user identifications matching to identifications of the experts. Unlike what commonly happens in inter-rater evaluations, in our case the identifications of the experts are considered correct by default, so that the agreement between users and experts can be considered a correct rate of the identifications.

Table 1. Examples of concluding remarks in forum threads and levels of certainty assigned to each taxonomic identification. The levels of certainty were scored into five categories.

Types of remark on identification	**Level of certainty assigned**
Plain identification, doubtless "It is…"	Certain (95% < s ≤ 100%)
"I'm almost sure it is…" "Most likely it is…"	Probable (75% < s ≤ 95%)
"Could be…" "I think it is…" Genus cfr. species (e.g. *Pyrgus* cfr. *picenus*)	Possible (50% < s ≤ 75%)
"I think it is … but I'm not experienced with the genus"; "I'm not sure"; "I try to guess…" "Could be … but is better waiting for a more experienced friend…"	Doubtful (s ≤ 50%)
Identifications at the genus or higher taxon level (e.g. *Erebia* sp.; Hesperiidae) Identifications ascribed with alternative names (e.g. species1/species2) Missing identifications	Not Attributed

As a rule, the percentage agreement rate can overestimate the real agreement due to a random agreement between the raters, see e.g. (Jansen et al. 2003). For this reason, we used Cohen's kappa statistic that estimates and removes the expected percent of chance agreement between raters p_e. However, the chance agreement is generally related to the number of categories available for rater (Gwet 2014). Since the number of categories is very high (equal to the number of Italian butterfly species), the possibility of a random agreement should be very small and the difference between the percentage agreement and Cohen's kappa is expected to be negligible. Since the identifications are categorical data, we used Cohen's unweighted kappa on a contingency table in which we crossed identifications at species-level performed by the users and experts. To compute Cohen's kappa, we used the "psych" R package (Revelle 2018) in R environment (R core Team 2020). We used chi-square tests to point out possible differences in the ability to correctly identify observations at the species level i) between the two forum datasets, and ii) between identifications of adult butterflies and all other development stages. The identifications of experts and users were used as observed and expected data, respectively, after standardization with respect to the total of the observed values. To perform chi-square tests we used the "stats" package in R environment (R Core Team 2020), and Monte Carlo simulations were used to verify the performance of the statistical tests by using 1,000 replications.

Georeferencing and accuracy assessment for locality descriptions

For georeferencing localities, we used a technique designed for museum data (see

Tagliolato et al. 2017). Each locality was represented by a point, regardless of whether its textual description referred to a point spatial model (e.g. refuge, spring) or linear (e.g. rivers, transects, paths) or areal (e.g. protected areas, mountains etc.). To establish the geographical position of the representative points we used the following electronic gazetteers from authoritative sources:

- "Toponimi d'Italia IGM" (Italian National Geoportal, www.pcn.minambiente.it/mattm), was used for point localities, e.g.: "ME, Nebrodi, Monte Soro: Portella Femmina Morta (1524m)";
- "VI Elenco Ufficiale Aree Naturali Protette (EUAP)" " (Italian National Geoportal, www.pcn.minambiente.it/mattm), was used for localities referred to Protected Areas if no details were supplied, e.g.: "AQ, Parco Regionale Naturale del Sirente-Velino";
- Administrative boundaries for Regions, Provinces and Municipalities from ISTAT (Italian National Institute of Statistics), were used for administrative areas if no details were supplied, e.g. "Provincia di Torino";
- Getty Thesaurus of Geographic Names (www.getty.edu/research/tools/vocabularies/tgn), was used for geographic regions if no details were supplied, e.g.: "Monti Lepini";
- "Elementi idrici 10k" (Italian National Geoportal, www.pcn.minambiente.it/mattm), was used for river courses, e.g.: "Fiume Taro";
- Google Maps service was used for point localities not found in "Toponimi d'Italia IGM";

We defined the representative point of a locality as that provided in the gazetteers. In the absence of such references we used the centroid of the spatial model of the locality.

The spatial uncertainty of the named location, i.e. the "extent" of the observation, was estimated according to the point radius method (Liu et al. 2009, Wieczorek et al. 2004). All steps of the procedure were recorded for each observation. An example of the main fields recorded in georeferencing are showed in Figure 1.

verbatimlocality	localityname	localityqualifier	longitudeintopic	latitudeintopic
Pian di Gembro (SO) alt 1350 slm	PIAN DI GEMBRO	[null]	10.16015	46.167082
oasi Lipu di casacalenda	OASI LIPU DI CASACALENDA	[null]	[null]	[null]
BG, Treviglio	TREVIGLIO	[null]	[null]	[null]
Monti Sibillini	MONTI SIBILLINI	[null]	[null]	[null]
pressi Malga Cimana (Tn)	MALGA CIMANA	NEI PRESSI DI	[null]	[null]

refpoint_method	refpoint_name	refpoint_source	extent	extent_est_method
Su coordinate originali fornite dall'os...	[null]	[null]	350	Stima di errore quando sono for...
Centroide dell'area protetta	Oasi di Bosco Casale (Casacale...	Elaborazione su: http://wms.pcn...	1063	Distanza dal punto all'estremo p...
Punto su toponimo	Toponimo: TREVIGLIO	Toponimi IGMv (1:25.000)	2540	Distanza dal punto all'estremo p...
Centroide della regione geografica n...	[null]	Coordinate del punto da Getty T...	23823	Distanza dal punto all'estremo p...
Su punto Google Maps	[null]	Google Maps	1250	Stima di errore prefissata (= 250...

Figure 1. Some of the fields used to record the steps of the georeferencing procedure.

RESULTS

Starting from 3383 threads, we assembled the final data set including 4029 records of butterfly species identification (2091 from FNM, 1938 from FEI); 3810 records referred to the adult stage and 219 to other stages (eggs, larvae and pupae). The forum users achieved species-level identifications in 3764 records, whereas experts in 3783 records. The number of species listed was 250 for forum users and 251 for the experts. Therefore, the overall set of records included a taxonomic coverage higher than 86% of the 290 Italian butterfly species (Balletto et al. 2014). The species lists from users and experts shared 248 species; two species identified by the users (*Lycaeides abetonicus* and *Pyrgus folquieri*) were not validated by the experts and three species detected by the experts (*Erebia pronoe, Hipparchia blachieri* and *Leptidea reali*) were incorrectly or unidentified in the forum threads. Table 2 reports details on identifications from user and experts.

Table 2. Number and level of the identifications by users and experts.

Identification	Users	Experts
at the species level	3764	3783
ascribed with alternative names (e.g. species1 / species2)	12	10
at the genus or higher taxon level	229	236
missing	24	0
Total	**4029**	**4029**

Users and experts agreed in identifying 3812 cases out of 4029 (percentage agreement, p_o = 94.6%), but this is a result of little value because it puts together identifications at different taxonomic levels.

Comparisons between identifications by users and experts at the species level

Looking at species level identifications, users identified 3764 out of 4029 records (93.4%) and experts agreed with them in 3649 cases (percentage agreement p_o = 96.9%), i.e. the correct identification rate of the users was very high. As expected, the resulting values of Cohen's unweighted kappa Kc = 0.97 (confidence limits: min = 0.96 max = 0.97) was similar to the percentage agreement p_o. In addition, we checked for differences between data from the two forums and between identifications of adult butterflies and other stages, in order to assess homogeneity in the data structure (Table 3).

As regards the two forums, we tested the agreement between identifications by users and experts separately in the two forum datasets (Table 3-A) and we found no evidence of statistically significant differences between the levels of correctness of the identifications (χ^2=0.44714, d.f. =1, Monte Carlo p = 0.5064).

Since butterfly adults and pre-imaginal stages require different skills in taxonomic identification, we also explored the matching between identifications by users and experts separately for adults and all other stages of development (Table 3-B) and we revealed no statistically significant differences between them (χ^2=0.092161, d.f. =1, Monte Carlo p = 0.7971).

Table 3. The 3649 correct identifications at the species level were tested to assess homogeneity in the data structure and no statistically significant differences in correctness degree were found (A) between identifications from the two Forums (χ^2=0.44714, d.f.=1, Monte Carlo p = 0.5064) and (B) between identifications of adult butterflies and all other development stages (χ^2=0.092161, d.f.=1, Monte Carlo p = 0.7971.

A) Forum	Observed	Expected	B) Development stage	Observed	Expected
Forum Entomologi Italiani	1748	1719.8	adult	3443	3448.3
Forum Natura Mediterraneo	1901	1929.2	all other stages	206	200.7
Total	**3649**	**3649**	**Total**	**3649**	**3649**

Table 4 shows matches and mismatches of species-level identifications performed by users and experts, with their levels of certainty. Table 4-A reports all the species level identifications by users matching those by the experts, i.e. correct identifications, whatever the level of certainty; since all identifications by users are at the species level, "Not Attributed" row and column are empty. The entries in the main diagonal represent the "perfect match" of identifications by users and experts, i.e. 3330 cases (88.5% of the matching identifications at the species level) in which users and experts agreed on both the identification and its level of certainty. The cells above the main diagonal (sum = 62) gather information on higher levels of certainty in identification by user than by experts, while the cells below the main diagonal (sum = 257) lower levels of certainty by users than by experts. The comparison between these values indicates that the forum users tend to be more uncertain than experts in identifying butterfly species (χ^2=65.66, d.f. =1, Monte Carlo p = 0.0001).

Table 4. Matches (A) and mismatches (B) of species level identifications performed by forum users and by the team of experts split by levels of certainty assigned by users and experts respectively.

		Identifications by team of experts				
Identification by forum users	**(A)** 3649 matches	Certain	Probable	Possible	Doubtful	Not Attributed
	Certain	3207	30	15	6	
	Probable	63	76	3		
	Possible	144	41	44	8	
	Doubtful	7		2	3	
	Not Attributed					
	(B) 115 mismatches	Certain	Probable	Possible	Doubtful	Not Attributed
	Certain	14	3	1	1	4
	Probable	4	3		1	22
	Possible	12	4	4	1	30
	Doubtful		3		1	7
	Not Attributed					

Table 4-B shows identifications by users mismatching those by experts, that is misidentifications. The "Not Attributed" row is empty here too, since all identifications by users are at species-level, whereas the values in the "Not Attributed" column (sum = 63) are the species-level identifications by users not validated by experts because not identifiable at the species level, but only at genus or higher taxon level.

Filtering forum identifications by taking only "Certain" and "Probable" species level identifications into account, 311 records (9.1%) were discarded, percentage agreement between users and experts rose to 98.4% and misidentifications fall to 1.5%.

As it sometimes happens when the agreement level is high, analysing mismatching is important because it provides indications on "critical cases" to be used with special care (Uebersax, 2015). In the list of wrongly identified records, 51 out of 115 (44.3%) incorrect identifications concerned species of the family Lycaenidae; the experts usually disagreed with the species level identification by users, believing diagnosis cannot exceed the genus level. In addition, the subfamily Satyrinae, especially the genera *Hipparchia* and *Erebia*, including known critical species,

showed 17 out of 115 (14.8%) incorrect identification.

Comparisons between identifications by users and experts at the genus or higher taxon level

Forum users identified 265 subjects at a level higher than species level, but only 163 identifications were in agreement with identifications by experts (p_o = 61.5%); the unweighted Cohen's kappa Kc = 0.58 (confidence limits: min = 0.52 max = 0.64). Mismatches in identifications by users and experts were almost entirely due (71 out of 78 cases, 91.0%) to identifications at the species level by experts previously identified at the level of genus by users.

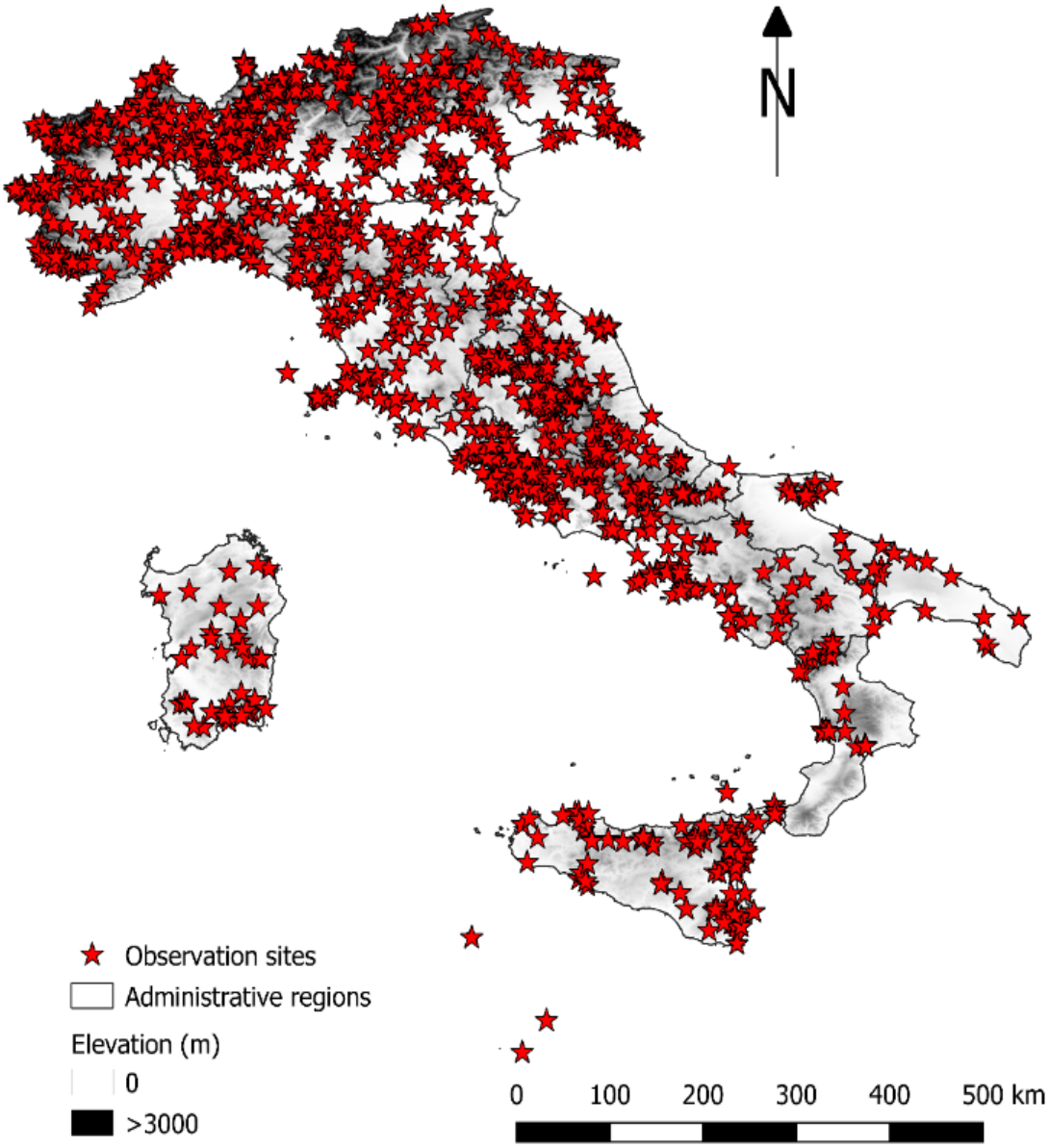

Figure 1. Map of the 1,402 georeferenced observation sites.

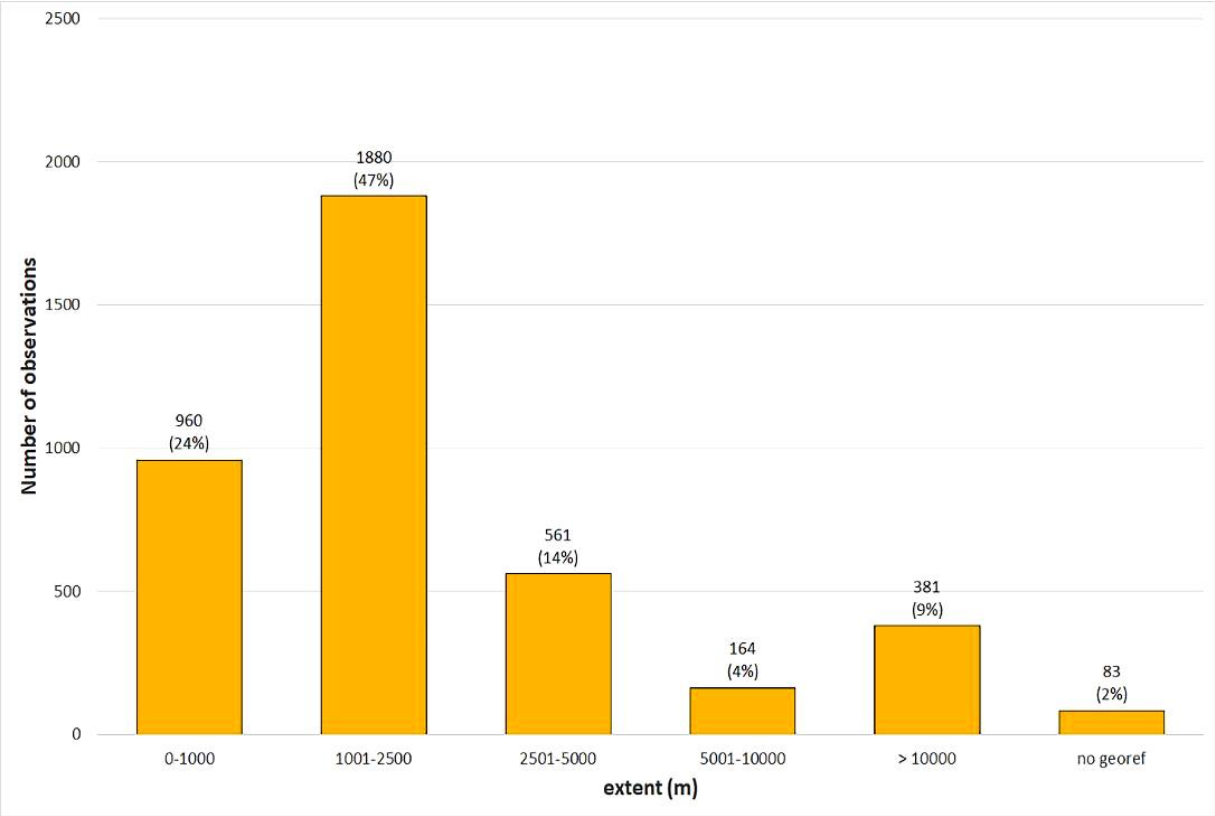

Figure 3. Class distribution of the estimated extent of georeferenced observations.

Geographic accuracy from sites descriptions

Although in recent years FNM e FEI forums were equipped with tools facilitating the geolocation of the observations, the spatial references found in the threads were mainly provided in textual form. Very different levels of accuracy characterized geographic data, ranging from an administrative region, to a river, to an accurate point-like area (e.g. a fountain or a mountain pass). We were able to georeferencing 3,946 out of 4,029 records in the dataset, the remaining 83 observations had no geographical indications or these were extremely vague (e.g. "Alps"). Recognized localities (distinct points) were 1,402 (Fig. 2), of which only 69 (4.8%) were directly georeferenced by the users (without accuracy of the reference points). Besides these few records including geographic coordinates, the estimated extent of the sites ranged from 100 m (e.g. “Roma, Porta Maggiore”) to 161,107 km (e.g. “Regione Piemonte”). Figure 3 illustrates the classes of the extent for the georeferenced observations; 2,840 out of 3,946 (70%) record localities have an estimated extent less than or equal to 2500 m).

DISCUSSION

In recent years, there has been a steadily increasing effort for biodiversity research and associated costs. The opportunity to spread awareness on biodiversity combined with the active involvement of citizens allowed the dissemination of citizen science projects, in which amateur naturalists largely participate (Devictor et al. 2010). However, scientific researchers often disagree on the reliability of data collected by amateurs. The need to validate these data gave rise to a literature as broad as the scientific initiatives involving citizens in the production of data are diversified (for a review see Aceves-Bueno et al. 2017).

Our case study analysed primary biodiversity data produced by untrained citizens in an unstructured context, where observational data are exposed. The identification attempts of butterfly specimens from photos, supervised by more or less expert users, were obtained comparing interpretations between users, and morphological characters were discussed along with ecological and geographical features. The percentage agreement on final identification between forum users and experts was very high, close to 97%. Because the inter-rater reliability is high, the users data from forum can be used interchangeably with expert identifications (Gwet 2014). An equally high agreement rate (98%) in the identification of taxa between experts and citizen scientists is reported by Kosmala et al. (2016). They surveyed a sample of about 4,000 randomly selected animal images within a huge archive of images from photographic traps positioned in the Serengeti National Park and 90% of the images were correctly classified. User interactions reduce the error rate, and quality control of data in qualified Citizen Science initiatives is based on this principle. In *iNaturalist* (www.inaturalist.org), a popular system of collecting citizens' observations, the quality level of the observations automatically upgrades from "casual" to “research grade" when 2-3 users agree on the identification (www.inaturalist.org/pages/help). In the forums, the specimen identification occurs in a non formalized way and the process is supervised by a moderator who ensures effective quality control.

An issue worth highlighting concerns the identification methods and the limitations associated with photos in nature, as opposed to museum specimens. The latter, of course, offers ample scope for taxonomic characters. In butterflies, but also in many other organisms, the potential ease of access to diagnostic characters, such as the morphology of the genitalia, androconial scales, chromatic reflection pattern at different wavelengths of the spectrum, DNA sequences, etc., is often crucial. Obviously, these insights are not possible on photographic images posted on the web.

Interestingly enough, the widespread use of apps for recognition in the field has also led to an evolution in the choice of diagnostic characters. These are often related to the insect behaviour in nature and unconventional if compared to the recognition standards traditionally based on mounted specimens. Thus, for example, a particularly subtle analysis of the colour pattern and design of the underside of butterfly wings, palpi, spinulation of legs, facilitates their diagnostic recognition, since most photos of butterfly species in the wild are in lateral view, enlarged and with closed wings.

Most scientific projects require mandatory geographical coordinates for species observation localities. In the forums, instead, precise geographical coordinates of posted observations are almost never supplied and have to be obtained starting from descriptions in the accompanying text. Using techniques for museum data, we obtained reliable estimates of the reference point positions and their extent, which were less than 2500 m in 70% of cases. We believe this order of magnitude is acceptable in most cases of data usage.

The authoritativeness of the sources and the transparency of the steps carried out in the extraction of geographic information from the textual description allow users to simply evaluate the suitability for use of each observation, giving these data high quality (Redmann 2001, Marta et al. 2019).

Lights, shadows and perspectives for the forum data

Results of this study, although limited to butterflies, indicate that the final identifications from forums show a surprisingly small margin of error and that the 'democratic' approach to taxonomy ultimately produces few uncertainties. The selected forums contain large amounts of primary biodiversity data in digital

format, correctly identified and georeferenced with satisfactory precision. Since Italian butterfly species are well known and well researched, a general agreement between users and experts on their identification was expected despite some difficulties in using images. However, the accuracy of the results provided by amateurs may vary on a case-by-case basis depending on the difficulty of the given tasks (Crall et al. 2011, Gardiner et al. 2012, Lewandowski and Specht 2015, Swanson et al. 2016, van der Velde et al. 2017). This means it will be advisable to apply our approach to citizens' data by analysing other taxonomic groups and other forums.

Biodiversity data stored in forums are too valuable to remain unused and if ignored they would be lost. However, their inclusion in organized systems is currently problematic. The formalization of collaborations with institutions and scientific projects would bring the forums in the area of "official" citizen science initiatives, giving the forums a role of citizens' scientific training. The recognition of a scientific role makes forum managers and users more deeply involved. Data protection over time, currently entrusted to forum managers, would be greatly enhanced. Cooperation between scientific institutions and forums is also essential for improving access to forum data. However, retrieving data from forums is a complex job, because the applications managing forum threads do not provide tools to effectively filter and export the thread contents, neither taxonomic nor geographic data.

The challenge of recovering data from millions of forum posts shares troubles with the digitization of museum samples that constitute the historical reservoir of primary biodiversity data. Traditionally, occurrence data on biodiversity come from museum collections. Museums house millions of specimens that represent a potential treasure to describe the space-time evolution of organisms in recent centuries. However, most of the potential offered by museums remains limited and unexpressed, pending adequate digitalization processes. The GBIF (Global Biodiversity Information Facility; www.gbif.org) database currently aggregates over 1.5 billion occurrence data (last accessed July 2020), and the data of museum specimens represent only 11.2%, that is, about 165 million records out of an estimate of 1.2-2.1 billion of subjects in the natural history collections (Ariño 2010). The CollMap project quantified the number of specimens preserved in Italian natural history museums in over 26,000,000 (http://www.anms.it/collmap/index.php?tipo=report).

The systematic digitization of natural science collections showed great difficulties, shared by most European museum institutions. Despite some increasing commendable efforts by the most active museums (BMNS, Harvard, Stockholm, Monaco, Leiden) much remains to be done. Several studies have addressed this issue and examined the organizational, procedural, practical and economic features to make museum collections virtual (Drinkwater et al. 2014, Heerlien et al. 2015, Hudson et al. 2015, Nelson et al. 2012, Tegelberg et al. 2014). In particular, a European research infrastructure for the digitization and sharing of data from museum collection was implemented: the Distributed System of Scientific Collections (DiSSCo). DiSSCo works for the digital unification of all European natural science assets under common curation and access policies and practices and it aims to make the data easily Findable, more Accessible, Interoperable and Reusable (FAIR) (Wilkinson et al. 2016).

The specimens observed by forum users are already in digital format, yet their identification and localization is obtained by interpreting the informal discussions. Today, many tools developed for automatic language processing allow designing a multilingual system capable of analysing data in a sophisticated way, extracting textual information, thus recognizing taxonomic names, time records and locations. The development of such a system will be crucial to recover large amounts of data from non-

institutional citizen science initiatives. It would be also functional in other cases where the digitization of primary biodiversity data is required and would significantly speed up data recovery processes, otherwise destined to remain forgotten forever.

ACKNOWLEDGEMENTS

We are deeply indebted to Tiziana Dinolfo, Federica Gioli and Claudio Labriola, forum moderators for butterfly threads of the Forum Natura Mediterraneo, for their precious support in the expert team for the identifications of butterflies.

This study was supported by the CSMON-Life Project (LIFE13 ENV/IT/000842), co-funded by the European Commission in the framework of the LIFE+ programme.

AUTHORS CONTRIBUTIONS

SDF, VS and DC conceived the study. DC supervised the project. PM managed forum moderators for butterfly identifications. SDF designed the archive structure, analysed the data and prepared the manuscript draft. All authors contributed to writing and revising the manuscript.

REFERENCES

Aceves-Bueno, E., Adeleye, A.S., Feraud, M., Huang, Y., Tao, M., Yang, Y., & Anderson, S.E. (2017) The Accuracy of Citizen Science Data: A Quantitative Review. The Bulletin of the Ecological Society of America, 98, 278–290. DOI: 10.1002/bes2.1336

Ariño, A.H. (2010) Approaches to estimating the universe of Natural History collections data. Biodiversity Informatics, 7, 81–92. DOI: 10.17161/bi.v7i2.3991

Baker, B. (2015) The Science of Team Science. BioScience, 65, 639–644. DOI: 10.1093/biosci/biv077

Balletto, E., Cassulo, L.A. & Bonelli, S. (2014) An annotated checklist of the Italian butterflies and skippers (Papilionoidea, Hesperiioidea) In Zootaxa (Vol. 3853) DOI: 10.11646/zootaxa.3853.1.1

Barve, V. (2014). Discovering and developing primary biodiversity data from social networking sites: A novel approach. Ecological Informatics, 24, 194–199. DOI: 10.1016/j.ecoinf.2014.08.008

Chandler, M., See, L., Copas, K., Bonde, A.M., López, B.C., Danielsen, F., Legind, J.K., Masinde, S., Miller-Rushing, A.J. & Newman, G. (2017) Contribution of citizen science towards international biodiversity monitoring. Biological Conservation, 213, 280–294. DOI: 10.1016/j.biocon.2016.09.004

Crall, A.W., Newman, G.J., Stohlgren, T.J., Holfelder, K.A., Graham, J. & Waller, D.M. (2011) Assessing citizen science data quality: An invasive species case study. Conservation Letters, 4(6), 433–442. DOI: 10.1111/j.1755-263X.2011.00196.x

Devictor, V., Whittaker, R.J. & Beltrame, C. (2010) Beyond scarcity: Citizen science programmes as useful tools for conservation biogeography. Diversity and Distributions, 16, 354–362. DOI: 10.1111/j.1472-4642.2009.00615.x

Dickinson, J.L. & Bonney, R. (2012) Overview of Citizen Science. In Citizen Science. Public partecipation in environmental science. (ed by J. L. Dickinson and R. Bonney), e-book edi, pp. 31–38 Comstock Publishing Associate. Ithaca, New York

Drinkwater, R.E., Cubey, R.W.N. & Haston, E.M. (2014) The use of Optical Character Recognition (OCR) in the digitisation of herbarium specimen labels. PhytoKeys, 30, 15–30. DOI: 10.3897/phytokeys.38.7168

Ellwood, E. R., Dunckel, B. A., Flemons, P., Guralnick, R., Nelson, G., Newman, G., Newman, S., Paul, D., Riccardi, G. & Rios, N.(2015) Accelerating the digitization of biodiversity research specimens through online public participation. BioScience, 65, 383–396. DOI: 10.1093/biosci/biv005

Ellwood, E.R., Kimberly, P., Guralnick, R., et al. (2018) Worldwide engagement for digitizing

biocollections (WeDigBio): the biocollections community's citizen-science space on the calendar. Bioscience 68:112-124. DOI: 10.1093/biosci/bix143

European Union (2020) Horizon 2020 Work Programme 2018-2020 16. Science with and for Society. (European Commission Decision C(2020)1862 of 25 March 2020)

Flemons, P. & Berents, P. (2012) Image based Digitisation of Entomology Collections: Leveraging volunteers to increase digitization capacity. ZooKeys, 217, 203–217. DOI: 10.3897/zookeys.209.3146

Gardiner, M. M., Allee, L. L., Brown, P. M., Losey, J. E., Roy, H. E. & Smyth, R. R. (2012) Lessons from lady beetles: Accuracy of monitoring data from US and UK citizen science programs. Frontiers in Ecology and the Environment, 10, 471–476. DOI: 10.1890/110185

Gwet, K. L. (2014) Handbook of inter-rater reliability (4th edition). Advanced Analytics.

Hallmann, C.A., Sorg, M., Jongejans, E., et al. (2017) More than 75 percent decline over 27 years in total flying insect biomass in protected areas. PLoS ONE, 2(10): e0185809. DOI: 10.1371/journal.pone.0185809

Heerlien, M., Van Leusen, J., Schnörr, S. & van Hülsen, K. (2015) The natural history production line: An industrial approach to the digitization of scientific collections. Journal on Computing and Cultural Heritage, 8(1), 289–294. DOI: 10.1109/DigitalHeritage.2013. 6744766

Hill, A., Guralnick, R., Smith, A., et al. (2012) The notes from nature tool for unlocking biodiversity records from museum records through citizen science. ZooKeys, 233, 219–233. DOI: 10.3897/zookeys.209.3472

Hudson, L.N., Blagoderov, V., Heaton, A., Holtzhausen, P., Livermore, L., Price, B.W., van der Walt, S. & Smith, V.S. (2015) Inselect: Automating the digitization of natural history collections. PLoS ONE, 10(11), 1–15. DOI: 10.1371/journal.pone.0143402

Jansen, R.G., Wiertz, L.F., Meyer, E.S. & Noldus, L.P. (2003) Reliability analysis of observational data: Problems, solutions, and software implementation. Behavior Research Methods, Instruments, and Computers, 35(3), 391–399. DOI: 10.3758/BF03195516

Johnson, B.A., Mader, A.D., Dasgupta, R. & Kumar, P. (2020) Citizen science and invasive alien species: An analysis of citizen science initiatives using information and communications technology (ICT) to collect invasive alien species observations. Global Ecology and Conservation, 21. DOI: 10.1016/j.gecco.2019.e00812

Kallimanis, A. S., Panitsa, M. & Dimopoulos, P. (2017) Quality of non-expert citizen science data collected for habitat type conservation status assessment in Natura 2000 protected areas. Scientific Reports, 7(1), 1–10. DOI: 10.1038/s41598-017-09316-9

Katsanevakis, S., Dreiu, I., Numes, A. L., et al. (2015) European Alien Species Information Network (EASIN)- supporting European policies and scientific research. Management of Biological Invasions, 6(2-Special Issue-Alien species related information systems and information management), 147–157.

Kosmala, M., Wiggins, A., Swanson, A. & Simmons, B.(2016) Assessing data quality in citizen science. Frontiers in Ecology and the Environment, 14(10), 551–560. DOI: 10.1002/fee.1436

Kullenberg, C. & Kasperowski, D. (2016) What is citizen science? - A scientometric meta-analysis. PLoS ONE, 11(1), 1–16. DOI: 10.1371/journal.pone.0147152

Lewandowski, E. & Specht, H. (2015) Influence of volunteer and project characteristics on data quality of biological surveys. Conservation Biology, 29(3), 713–723. DOI: 10.1111/cobi.12481

Lin, Y.P., Deng, D., Lin, W.C., Lemmens, R., Crossman, N.D., Henle, K. & Schmeller, D.S. (2015). Uncertainty analysis of crowd-sourced and professionally collected field data used in species distribution models of Taiwanese moths. Biological Conservation, 181, 102–110. DOI: 10.1016/j.biocon.2014.11.012

Liu, Y., Guo, Q., Wieczorek, J. & Goodchild, M. F. (2009) Positioning localities based on spatial assertions. International Journal of

Geographical Information Science, 23(11), 1471–1501. DOI: 10.1080/13658810802247114

Maistrello, L., Dioli, P., Bariselli, M., Mazzoli, G.L. & Giacalone-Forini, I. (2016) Citizen science and early detection of invasive species: phenology of first occurrences of Halyomorpha halys in Southern Europe. Biological Invasions, 18(11), 3109–3116. DOI: 10.1007/s10530-016-1217-z

Marta S., Brunetti M., Ficetola G.F., Stoch, F., Amori, G., Cesaroni, D., Sbordoni, V. & Provenzale, A. (2019) ClimCKmap: a spatially, temporally and climatically explicit distribution database for the Italian fauna. Sci Data 6, 195 (2019) DOI: 10.1038/s41597-019-0203-6

Martellos, S., Attorre, F., Cesaroni, D., Di Marco, S., Petruzzella, D., Spinelli, O., Tallone, G. & Mereu, A. (2016) CSMON-LIFE: data from the people, data for the people. First ECSA Conference 2016 Citizen Science – Innovation in Open Science, Society and Policy, (May 2016), 73–74.

Miller-Rushing, A., Primack, R. & Bonney, R. (2012) The history of public participation in ecological research. Frontiers in Ecology and the Environment, 10(6), 285–290. DOI: 10.1890/110278

Morris, R. A., Barve, V., Carausu, M., et al. (2013). Discovery and publishing of primary biodiversity data associated with multimedia resources: the audubon core strategies and approaches. Biodiversity Informatics, 8, 185–197. DOI: 10.17161/bi.v8i2.4117

Nelson, G., Paul, D., Riccardi, G., et al. (2012) Five task clusters that enable efficient and effective digitization of biological collections. ZooKeys, 45(209), 19–45. DOI: 10.3897/zookeys.209.3135

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Redmann, T.C. (2001) Data quality: the field guide. Digital Press, Boston.

Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.12.

Science Communication Unit, University of the West of England, Bristol (2013) Science for Environment Policy In-Depth Report: Environmental Citizen Science. Report Produced for the European Commission DG Environment. Available at: http://ec.europa.eu/science-environment-policy

Silvertown, J. (2009) A new dawn for citizen science. Trends Ecol. Evol., 24, 467–471. DOI: 10.1016/j.tree.2009.03.017

Swanson, A., Kosmala, M., Lintott, C & Packer, C. (2016) A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. Conservation Biology, 30, 520–531. DOI: 10.1111/cobi.12695

Tagliolato, P., Oggioni, A., Fugazza, C., Cianferoni, F. & De Felici, S. (2017) Georiferimento di campioni museali nell'infrastruttura LifeWatch Italia: le nuove prospettive dal web semantico. Museologia Scientifica, 11, 114–118.

Tegelberg, R., Mononen, T. & Saarenmaa, H. (2014) High-performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens. Taxon, 63, 1307–1313. DOI: 10.12705/636.13

Uebersax, J. (2015) Statistical Methods for Diagnostic Agreement. https://www.john-uebersax.com/stat/agree.htm

van der Velde, T., Milton, D. A., Lawson, T. J., Wilcox, C., Lansdell, M., Davis, G., Perkins, G. & Hardesty, B.D. (2017) Comparison of marine debris data collected by researchers and citizen scientists: Is citizen science data worth the effort? Biological Conservation, 208, 127–138. DOI: 10.1016/j.biocon.2016.05.025

Vermeulen, N., Parker, J. N. & Penders, B. (2013) Understanding life together: A brief history of collaboration in biology. Endeavour, 37(3), 162–171. DOI: 10.1016/j.endeavour.2013.03.001

Wagner, C.S., Park, H.W. & Leydesdorff, L. (2015) The continuing growth of global cooperation networks in research: A conundrum for national governments. PLoS ONE, 10, 1–15. DOI: 10.1371/journal.pone.0131816

Wieczorek, J., Guo, Q. & Hijmans, R. (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science, 18, 745–767. DOI: 10.1080/13658810412331280211

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., et al. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(March). https://doi.org/10.1038/sdata.2016.18

Wiemers, M., Balletto, E., Dinca, V., et al. (2018) An updated checklist of the European butterflies (Lepidoptera, Papilionoidea) ZooKeys, 811, 9–45. DOI: 10.3897/zookeys.811.28712

Xue, K. (2014) Popular Science. Harvard Magazine, 54–55. Retrieved from http://harvardmagazine.com/2014/01/popular-science

Zapponi, L., Cini, A., Bardiani, M., et al. (2017) Citizen science data as an efficient tool for mapping protected saproxylic beetles. Biological Conservation, 208, 139–145. DOI: 10.1016/j.biocon.2016.04.035

Zenetos, A., Koutsogiannopoulos, D., Ovalis, P. & Poursanidis, D. (2013) The role played by citizen scientists in monitoring marine alien species in Greece. Cahiers de Biologie Marine, 54, 419–426.

Submitted: 31 August 2020

First decision: 17 November 2020

Accepted: 1 March 2021

Guest editor for the special section Citizen Science in Biogeography: Stefano Martellos