The Xeno-canto collection and its relation to sound recognition and classification

Willem-Pier Vellinga[1] and Robert Planqué[1]

Stichting Xeno-canto voor natuurgeluiden (Xeno-canto Foundation), The Netherlands
{wp,bob}@xeno-canto.org

Abstract. This paper discusses distinguishing characteristics of the Xeno-canto bird sound collection. The main aim is to indicate the relation between automated recognition of bird sounds (or feature recognition in digital recordings more generally) and curating large bioacoustics collections. Not only do large collections make it easier to design robust algorithmic approaches to automated species classifiers, those same algorithms should also become useful in determining the actual content of the collections.

Keywords: LifeCLEF2015, BirdCLEF2015, Xeno-canto, bird sounds, automated recognition, citizen science, data mining.

1 Introduction

For the past two years the BirdCLEF challenge [1,2], part of the LifeCLEF workshops [3,4], has been based on sounds from Xeno-canto. Xeno-canto (XC) aims to popularise bird sound recording, to improve accessibility of bird sounds, and to increase knowledge of bird sounds. It tries to achieve these aims by facilitating and curating a collaborative, shared, global bird sound collection on www.xeno-canto.org. The collection was initiated by the authors in 2005 [5]. When XC started out it was mainly a project to aid identification of small collections of bird sounds made by the authors in tropical forests in Peru and Ecuador. Identifying species by sound using the means available at the time, mostly commercial cassette tapes or CDs with up to a hundred recordings, was cumbersome and many sounds were simply not available. (For a discussion see [6]).

Sjoerd Mayer's "Birds of Bolivia" CD-ROM's [7,8] were an inspiration. They increased the number of sounds available and species represented by an order of magnitude, made navigation of the sounds much easier, mapped locations, and identified background species on a recording. Mayer also engaged the birding community by welcoming and crediting contributions of sounds by birders and published corrections of errors on his website.

The authors essentially took these concepts a step further, and designed and constructed an interface to a non-commercial, open database situated on the world wide web. A number of guiding principles were formulated that distinguished XC from other sound collections at the time:

– Anyone with web access is invited to upload sounds. XC does not refuse recordings. Contributors can share any bird sound they find interesting, provided they are below a fixed maximum size (initially 1 MB, now 10 MB) and provided a required minimum set of metadata is given: species, recordist name, location name, country, recording date, time of day, elevation, and sound type(s). This system certainly has drawbacks: a considerable fraction of the recordings is short, of dodgy quality, or both. Still, such recordings may be useful. They may represent poorly known locations or vocalisations, or may simply contribute to the sample size of individual species. Also, in the context of automated species identification algorithms, it is clear that any real-life deployment of such an algorithm would have to deal with poor quality recordings as well.
– The recordings uploaded to XC are shared. Re-use of the recordings is intended, for purposes that are in line with the aims of XC, such as downloading to personal collections, embedding sounds in educational or personal web sites, use for scientific research, etcetera. The Creative Commons licenses (`http://creativecommons.org/`) offer a useful framework. After consultation with the community it was decided to settle for CC-BY-ND-NC (attribution, no-derivatives, non-commercial) licenses. Since this is in fact a rather restrictive license, nowadays one can also choose CC-BY-NC-SA (SA stands for share-alike) and CC-BY-SA licenses that allow more liberal re-use. In all cases attribution of the author/contributor on republication is mandatory. For discussion of the limits of the other terms, see the Creative Commons website. The XC website code is written in free, open source software. It is based on a standard LAMP (Linux, Apache, MySQL, PHP) set-up, with some additional software written to show sonograms, implement mapping, and so on.
– Anyone can contribute to the collection in some way. Apart from sharing recordings, people may contribute expertise on identification, set identification challenges, offer experience with equipment, write articles on-site, or just comment on recording achievements.
– Anyone can challenge an identification (ID) on the site. The vast majority of recordings have been identified correctly to species by the recordist, but errors are inevitable. When challenged, the recording is set aside and does not appear in search results until the ID is resolved by the community. This is usually done in an open discussion on the forum. If the ID is agreed upon, the recording is put back into the collection by the administrators. The administrators therefore have the role of arbiters, rather than authorities, and in fact there are no designated authorities that decide on species identification. This is one of the more uncommon features of Xeno-canto, and in this sense it differs from other well-known community projects on natural history, such as eBird (`ebird.org`), Observado (`waarneming.nl / observado.org`).

At present, May 2015, the XC collection contains some 243,000 recordings from over 9,300 bird species, shared by more than 2400 contributors from all over the world. In the rest of this paper, the development and current status of

the XC collection are illustrated and a few points relevant to its relation with automatic sound classification and recognition are discussed.

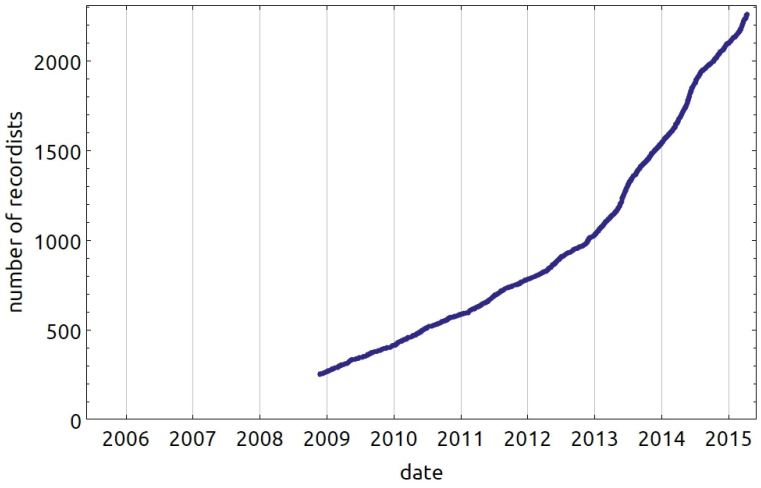

Fig. 1. Cumulative number of contributors over time.

2 Characteristics of the collection

The growth of XC is illustrated in Figures 1 and 2, by plotting the number of recordings and the number of contributors over time. Two things are noteworthy. Firstly, the data for the initial period is incomplete, since the uploading dates were initially not recorded. Secondly, there are pronounced seasonal effects, most obvious in the number of contributors. These points are remedied to some extent in subsequent figures by plotting versus the number of recordings instead of versus time.

2.1 Contributors

Both the number of recordings and the number of contributors grow at increasing rates. Remarkably, plotting the number of recordings versus the number of contributors shows that they have consistently increased at approximately the same rate. See Figure 3. This leads to a more or less constant average number of recordings per contributor, which turns out to be about 100. However, it should be noted that the distribution of recordings per contributor is very broad and skewed. At this moment 298 contributors contribute more than 100 recordings, and many more contributors, around 2100, contribute less than 100 recordings.

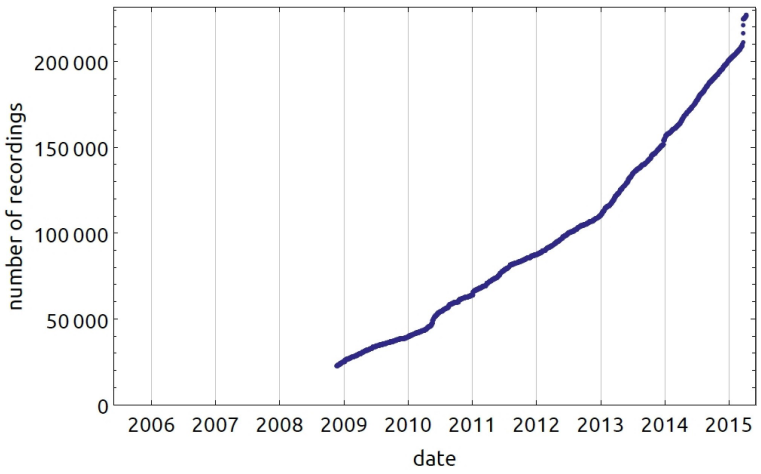

Fig. 2. Cumulative number of recordings over time.

The three largest contributions each comprise more than 10000 recordings, more than 100 times the average; 729 contributors contributed 1 recording, 100 times less than the average. The Zipf-like plots in Figure 4 serve to characterise the distribution at various stages during the development of XC.

2.2 Species

When a sound is uploaded a set of metadata is required, among which the name of the species. Specifying the subspecies is optional. The taxonomy of the site was initially based on the taxonomy in Neotropical Birds [11]. Other regions were added over time (North-America, Africa, Asia, Europe and Australasia) using other local taxonomies, which lead to problems with species occurring in several regions. In 2011 the global IOC (International Ornithological Council) taxonomy was adopted for all recordings and XC currently uses version 4.1 [12]. The constant revision of taxonomy at the species level means that the species assignment of the recordings needs to be updated frequently. This task falls to the team of administrators. Splits can be problematic, since the subspecific taxon to which a recording belongs may not be indicated (see below).

IOC 4.1 recognises 10,518 extant species and 150 extinct species; to this list XC has added 16 additional recently described or as yet undescribed species. About 9330 are represented in XC at this moment. To our best knowledge this constitutes the largest number of species in any public collection of bird sounds. (There is at least one private collection that has more species, but it includes sounds of all species from XC.)

The growth of the number of species may provide a clue about the moment of completion of the collection at the species level. Figure 5 shows the species

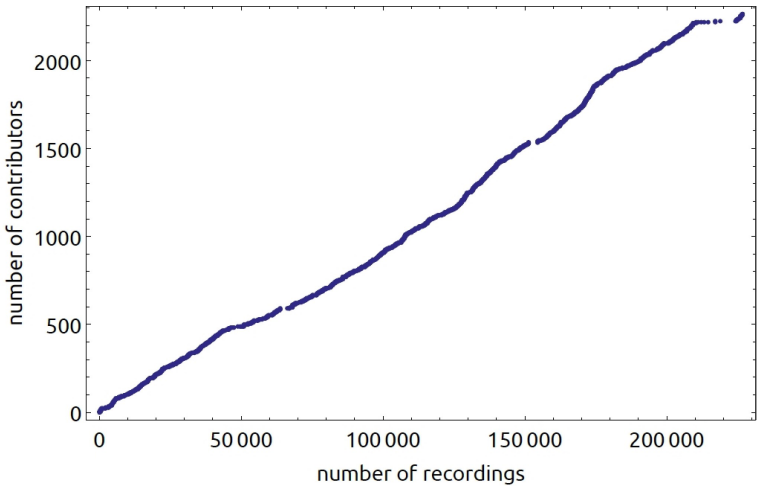

Fig. 3. Order of contributor versus number of first recording by that contributor. To a reasonable approximation the increase is linear the slope indicating that every contributor adds about 100 recordings on average.

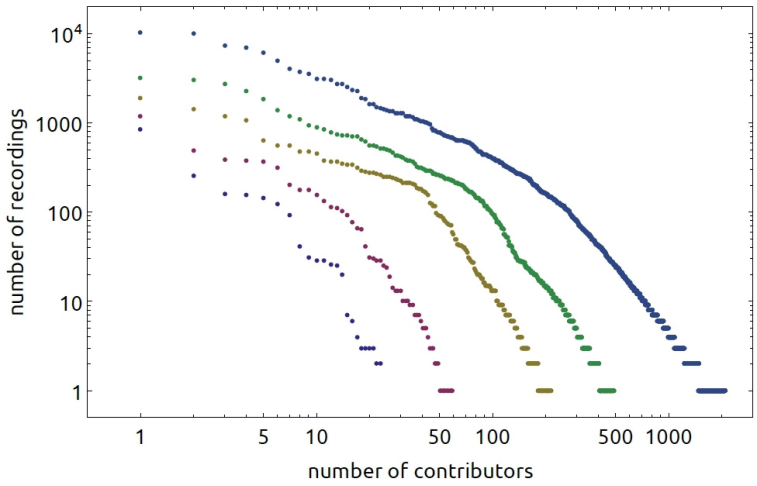

Fig. 4. Distribution of the number of recordings per recordist plotted in a Zipf plot after 2000, 5000, 20000, 50000 and 200000 recordings. These plots show that the distribution of the number of recordings per contributor is very wide.

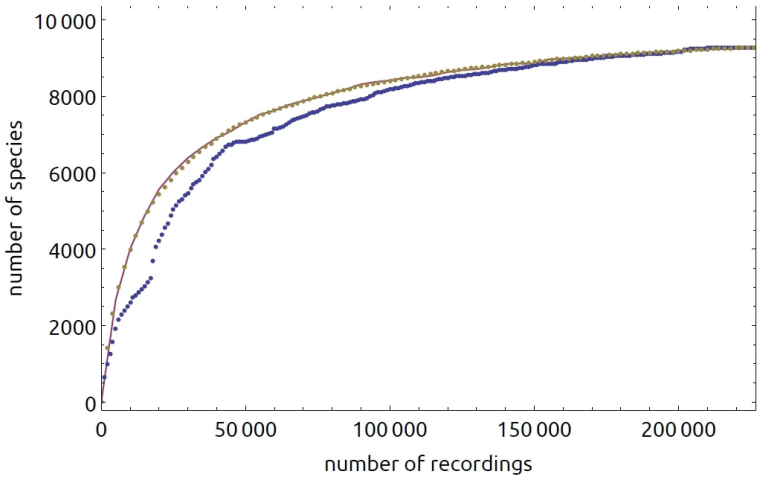

Fig. 5. Species accumulation curve (blue dots) and randomised species acccumulation curve (brown dots). The drawn line is an extrapolation shown further in figure 8.

accumulation curve up to may 2015, together with a randomised accumulation curve and a fit used for extrapolation. The randomised version is based on a random draw from all recordings present in XC. Clearly the two curves differ significantly. This is caused by the fact that XC started out with only Neotropical species, and that other world areas were added later. The randomised species accumulation curve does not take that into account. The two curves are seen to meet up after about 170000 recordings, well after XC went global.

For any number of reasons (abundance and size range, vocal (in)activity, accessibility of the range of the species, accessibility of the site to name just four) the recordings are not evenly distributed across the species. The current expectation value for the number of recordings per species is around 20. However some 20 species have attracted over 500 recordings, while around 1200 are still waiting to be uploaded. The distribution plotted in Zipf-fashion is shown in Figure 6, probability densities are shown in Figure 7.

The species abundance curves can be extrapolated into the future by making assumptions on the probability that species that are not represented at this time will be uploaded. A reasonable fit is achieved by assuming that the probability of a new species being uploaded is 1/3 of that of a species with 1 recording in XC, with the ratios between probabilities of species already represented remaining equal. An extrapolation based on that assumption is shown in Figure 8. Of course the extrapolation follows the randomised species abundance curve very well. The extrapolation is shown up to 900,000 recordings, at which point it is still about 600 species shy of the total number of species. The precise number will depend on the assumptions made, but it seems reasonable to assume that completion at the species level will take a multiple of the number of recordings present at this moment.

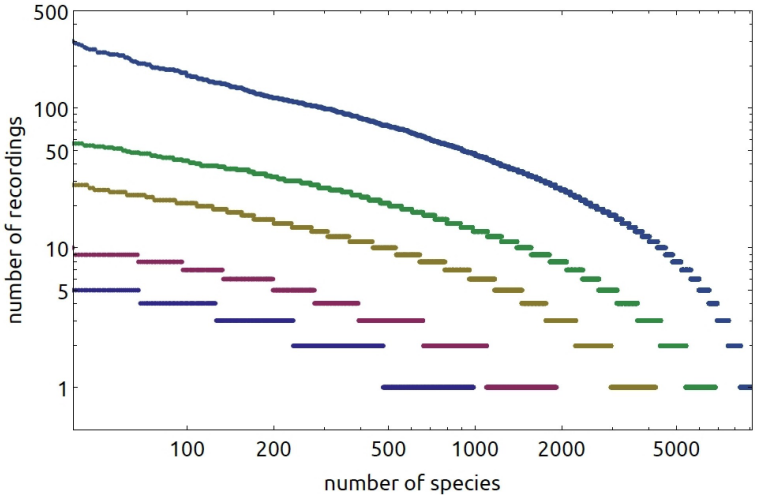

Fig. 6. Distribution of the number of recordings per species plotted in a Zipf plot after 2000, 5000, 20000, 50000 and 200000 recordings.

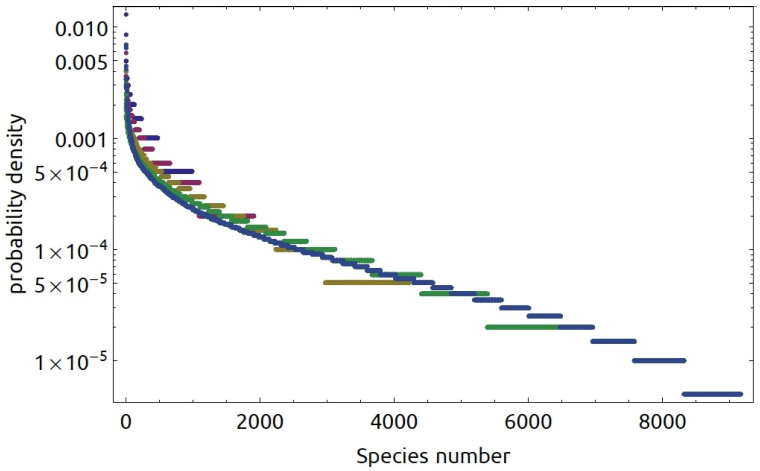

Fig. 7. Probability densities of species after 2000, 5000, 20000, 50000 and 200000 recordings.

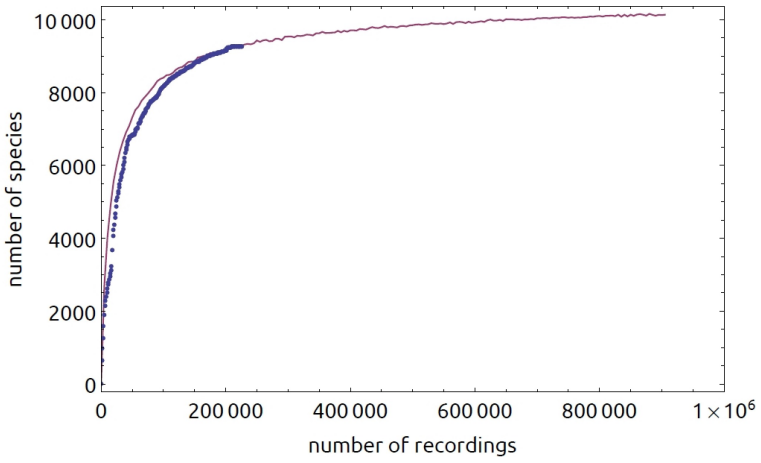

Fig. 8. Extrapolation of species accumulation curve assuming that as yet unrepresented species have a probability of being uploaded that is 1/3 that of a species represented with 1 recording. It is likely the collection needs to multiply in size before completion at the species level is reached.

3 Linking bird song databases and automated species recognition

The BirdCLEF workshop requires entrants to identify recordings from the XC collection to species level based on the species level identification provided by the XC community. It is worthwhile to have another look at the species level IDs in XC. For a number of reasons the ID to species level, even if correct, may be misleading.

- **Presence of unnamed background species** Although recordists are asked to mention the background species present in the recordings, not all recordists do so. On average 2 species are identified per recording, but it is certain that many more could be identified. Interestingly, the presence of named background species helps humans to identify a sound of interest (as the authors know from personal experience), but this does not seem to lead to a higher rate of identification in the algorithmic identifications [1,2].
- **Hidden diversity** The IOC 4.1 species list not only recognises 10668 species, but identifies another 20976 subspecies for 5093 of species, bringing the total number of taxa to 26551. On XC about 9330 species and 9140 additional subspecies have been identified. This does not mean that 18470 taxa are represented. It is likely that some recordings represent subspecies that have not been named now. This means that the currently recognised number is an underestimate. But it is also likely that in some cases the taxa represented by recordings without subspecific ID are in fact already named, which would lead to an overestimate. Of the 9300 species present on XC 4484 are monotypic. The 9100 subspecies therefore belong to about 4900 species adding

at least 4200 taxa. The maximum number of named taxa represented is therefore 18400 and the minimum number 13500. An estimate based on the number of species present (9300/10668)*26551 would lead to about 23000 taxa present at this time. Based on this estimate it seems likely that a considerable number of taxa remains to be named on XC. At the same time this also means that the species category may represent considerable taxonomic diversity. It is to be expected that such diversity hidden within species on XC is reflected in the sounds, since many subspecies are known to have distinct vocalizations [9,10].

Other contributions to the diversity of sounds which do not necessarily align with subspecies categories are geographical dialects, such as in Yellowhammer (*Emberiza citrinella*), and the size of the vocabulary of a species, such as in Common Nightingale (*Luscinia megarhynchos*). Little quantitative information is available on the extent of dialect formation and the size of the vocabulary across the range of the overwhelming majority of the 10518 species of birds. Apparently the effect of such diversity at the species level on the results of automatic recognition has not been quantified yet. Intuitively, given a set of training data, one would expect a species that shows little diversity to be recognised more faithfully than a species that shows a lot of variability.

In [1] it was concluded there that the recognition algorithms worked better on average for species with more recordings in the training set. It would be interesting to look for correlations with the number of subspecies recognised, or the known size of vocabulary.

4 Conclusion

The results from the 2014 and 2015 BirdCLEF challenges offer an interesting perspective on the use of automated algorithmic techniques on the one hand, and large accessible public archives of sound data on the other.

At present, the focus in the challenges lies squarely in the field of automated recognition, and understandably so. The large Xeno-canto database has been the basis of the challenges, and give the first general insights in automated feature extraction and classification to species level for general vocalizations. The species set included in the latest 2015 edition spans 1000 species with a huge range of different types of bird songs and calls. The BirdCLEF paper in this volume contributes to our understanding which techniques excel at this type of challenge.

We would welcome a second application of the algorithms, however, one that would allow a deeper insight into the variety of vocalizations actually represented in archives such as Xeno-canto. There is great potential for collaborative projects, in which estimates would be computed of a number of statistics. Example include (a) estimates of repertoire sizes in song birds (or other taxa); (b) discovery of subspecies with different vocal signatures; (c) the ability to extract

a small representative sample of different vocalizations for focal species, or focal localities. We hope that we may attract the computer science community to work with us to start to address these types of challenges.

References

1. Goëau, H., Glotin, H., Vellinga, W.P., Planqué, R, Joly, A.: LifeCLEF Bird Identification Task 2014. In: Proceedings of CLEF 2014 (2014)
2. Goëau, H., Glotin, H., Vellinga, W.P., Planqué, R., Rauber, A.,Joly, A.: LifeCLEF Bird Identification Task 2015. In: CLEF working notes 2015 (2015)
3. Joly, A., Müller, H., Goëau, H., Glotin, H., Spampinato, C., Rauber, A., Bonnet, P., Vellinga, W.P., Fisher, B., Planqué, R.: LifeCLEF 2014: multimedia life species identification. In: Proceedings of CLEF 2014 (2014)
4. Joly, A., Müller, H., Goëau, H., Glotin, H., Rauber, A., Bonnet, P., Vellinga, W.P., Fisher, B., Planqué, R.: LifeCLEF 2015: multimedia life species identification challenges. In: Cappellato, L., Ferro, N., Jones, G., and San Juan, E., editors (2015). CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/ Vol-1391/. (2015)
5. Vellinga, W.P., Planqué, R.: http://tinyurl.com/xcstart05 (2005)
6. Moore, J.V.: Ecuador's avifauna: the state of knowledge and availability of sound-recordings. Cotinga 29, 19–21 (2008)
7. Mayer, S.: Bird sounds of Bolivia / Sonidos de aves de Bolivia, 1.0. CD-ROM. Bird Songs International, Westernieland, The Netherlands. (1996)
8. Mayer, S.: Bird sounds of Bolivia / Sonidos de aves de Bolivia, 2.0. CD-ROM. Bird Songs International, Westernieland, The Netherlands. (2000)
9. Stotz, D.F., Fitzpatrick, J.W., III, T.A.P., Moskovits, D.K.: Neotropical Birds. University of Chicago Press (1996)
10. Gill, F., Donsker, D.: IOC World Bird Names v4.1. Available at www.worldbirdnames.org. CC-BY 3.0 (2015)
11. Kroodsma, D.E., Miller, E.H. (eds.): Ecology and evolution of acoustic communication in birds. Comstock Publishing Associates (1996)
12. Marler, P., Slabbekoorn, H.: Nature's Music. Elsevier Academic Press (2004)