eBird Data Challenge

Help spread the word. Share this image on Facebook and Whatsapp!

The eBird database for India is now very large, and is growing rapidly. But **what can be done with this information**? Are you interested in helping generate new ideas, patterns and results from the observations the database contains? If so, read on, or click on a heading below to jump directly to the section.

Results now available!

Background

Despite a long history of ornithology in India, many aspects of current bird distribution, abundance and movements are still poorly known. In early 2014, birdwatchers in India started using the eBird platform to record their sightings, and the information aggregated so far forms one of the largest online and publicly available databases on any aspect of Indian biodiversity (see The Database, below).

Collating information into one place is only the beginning. The next task is to convert the raw data into **meaningful information** which can lead to a better understanding of Indian birds, **for education, research and conservation**.

The Challenge

The challenge is simple:

1. Think of a **question** you wish to answer or a **problem** you would like to tackle (see the Ideas section below).
2. Download all **eBird data** from India, or only a subset (you can specify region, species, and date range if you wish).
3. Use the eBird data to **address this question or problem**. You can do so by creating one or more maps, graphs or other kind of output (e.g. animation).
4. Send your findings to us (see Entering the Challenge).

Who can get involved

Anyone interested in birds and/or data!

If you have an **idea** for a question/problem to apply to the data AND you have the **technical skills** to tackle this, you're all ready to go ahead!

If you are a birdwatcher or a conservation group and have a question/problem you are interested in, but don't have the training or skills to work with data, please submit your suggested question/problem to the Question Pool. Perhaps someone else who has the technical ability will take up your suggested question and tackle it.

If you have the technical skills, but don't know what question would be meaningful, take a look at the suggestions in the Question Pool – if one of these strikes your fancy, let us know in the comments below and we'll put you in touch with the suggester so that you can work together on this.

Remember, the challenge can be taken up by individuals as well as groups – so don't hesitate to join forces. Collaboration is good!

Entering the Challenge

To enter the challenge, first register your team. Your team can be just you, or a larger group, but there should be a single contact person specified. We ask that you register so that we can be sure to keep you informed about any updates to the challenge, and to let you know of ideas contributed into the Question Pool. The deadline for registering is extended to **20 December 2016**.

Once you have registered, download the data (see The Database below) and work on tackling the problem you would like to solve. Your output can be in any form: from one or more maps or charts, to an animation, all the way to an interactive website or app! The main idea is to state a clear question or purpose of your work, to describe in detail how you processed the raw data to arrive at your output, and then to reflect on what the output tells us about the original purpose.

Entries must be submitted by **31 December 2016** (extended from 15 December) and will consist of the following pieces of information:

1. About you and your group (names, contact details, etc.)
2. The question or problem that you set out to tackle
3. A detailed description of how you processed the data, accompanied by programming scripts, if you used them
4. A link to where we can see or download the output
5. Your thoughts on the answers to the question or problem stated in #2, based on the output you have generated.

Evaluation Process

Entries will be evaluated by a panel of jury members, whose composition is as follows:

Harini Nagendra, Azim Premji University
Taej Mundkur, Wetlands International
Ghazala Shahabuddin, Centre for Ecology, Development and Research
Vijay Barve, University of Florida
MD Madhusudan, Nature Conservation Foundation

Entries will be evaluated on the following criteria:

Novelty, creativity and importance of the question or problem being tackled.
Care and thought put into data processing and quality control.
Design of the output produced, such that it be understandable to and usable by a broad audience.
The connection made between the output and the original question or problem.

Prizes!

As a small incentive for helping understand Indian birds better, we are offering an overall prize for the best entry (to be announced, worth approx. Rs 10,000), plus two special mention prizes (each worth approx Rs 5,000). Details will be announced soon.

Ideas

A large variety of questions can potentially be tackled with the information in the eBird India database. Some brief examples are given below – they would have to be made more detailed and specific for this challenge – as well as some more detailed examples on the next page.

Gaps. Where and when are there gaps in information on birds? These can be gaps in both space and time (e.g. season). From this can we set some priorities to encourage birdwatchers to fill important gaps?

Birder behaviour. Where, when and with whom do birders go birdwatching? How far do they travel, and what are their eBird listing habits? Understanding how birdwatchers behave can lead to better design of citizen science projects like eBird.

Individual species. What is their distribution and seasonality? Can we detect local movements? Are there influences of habitat?* In flocking or colonial species, how does the number of birds counted vary by month or season? What is the breeding season of different species, and does this change geographically? Is there adequate information from different sites/regions to answer these sorts of questions?

Species diversity. How does the number of species change over time and space? What aspects of habitat or weather might influence this?* In a specific region, how many lists or (birds counted) does it take to get to the total number of bird species?

Locations or regions. What species are found in a specified location or region? Can we assess the adequacy of the data available to answer this question?

Detecting potential errors. In a large database like eBird, some errors are bound to creep in, and some (perhaps many) seeming errors are in fact not errors at all. Are there ways to flag possible errors such that observers can be requested to provide more information, to strengthen the database as a whole?

**For these questions, eBird data alone may not be sufficient, and other data sources may need to be brought in.*

Who knows, perhaps you'll come up with a great idea and find an interesting answer, and your work could be published in a formal outlet!

The Database

The eBird database currently contains just over **4 million records** of birds from India. However, the fundamental unit of observation in eBird is not a single record, but rather a list (and it's possible that a list contains only a single record). There are around **200,000 lists** from India, and many, if not most, useful analyses of eBird data are conducted at the level of a list, rather than that of a record. For example, to look at where eBird data come from, one would plot a map of list locations. Or, if we wanted to plot the distribution of a species, we would ask **not only which lists contain the species, but also which lists do not**. (For this, we would choose to analyse only **'complete' lists**, that is where all species seen or heard were reported.)

Most of the data in eBird from India come from 2014 and later, but a number of people have uploaded older records as well. Do consider exploring and using these historical records if they can help to answer your question.

When you download the eBird data, each row in the database corresponds to an observation, with all observations that come from a single list sharing a common ID code (called SAMPLING.EVENT.IDENTIFIER), so that you can collapse the data to the level of the list if you so wish. Various metadata about each list are also available in other columns, including the location (with State, District, latitude and longitude), start time, duration, distance covered, eBird protocol followed, whether the list is complete, and so on. These metadata fields are very important to consider when summarising and analysing eBird data.

Please also carefully read the terms of use document that comes together with the data download. In brief, eBird permits use of the data for research and education; any commercial purpose must receive written permission from eBird. Please **do not send the downloaded data to others**; anyone else interested should please **download the data for themselves directly** from eBird.

Getting Started

eBird data are made publicly available every quarter through the eBird Basic Dataset (EBD). You can

request access to the dataset, and when doing so, please state that this is **for use in the eBird-India data challenge**. When you are given access, you will receive an email. This may take 2-3 days at most; and do check your spam folder as well!

When downloading the data you can specify which region, species or date range you wish to download. The download format is a tab-separated text file. If you download the full India data, the file is some 110Mb in size (as of September 2016). Do consider if you want to download data for a single State instead, or for specific species. On the download form, there is also a checkbox to download “unvetted data”. These are records that have been flagged as unusual, but have not yet been verified. For most purposes, one should not use the unvetted data; and in general, any records marked with a ‘1’ in the ‘APPROVED’ column should be ignored/removed.

The download package contains a detailed description of each of the columns in the tab-separated file; please read these carefully.

Please note that the data is in spreadsheet format, but several standard spreadsheet software programs (including Excel) will not be able to open a file with 4 million rows. For this reason, if you want to work on the full India dataset, you will need to use more flexible software like R, Python, Matlab, etc. In the coming days we will post some tips for how to handle large volumes of eBird data. If you need specific help, please let us know in the comments below, or on the Facebook event page.

Some Cautions

When analysing data from eBird, do take some time to think about the quality and accuracy of data that you need. eBird has a detailed set of quality control processes, and over 70 volunteer reviewers help with eBird data quality in India. However, despite best efforts, in a database this size, there will be errors that creep in; and there are also other reasons to carefully think about which parts of the database you should use. You may have to subset the overall dataset to include only those lists/records that meet your needs and quality requirements. Some specific cautions:

Location accuracy. The geographical precision of lists in eBird varies from list to list. For example, even though each list is associated with a specific latitude and longitude, the birds recorded on that list may come from a very large area. This can usually be discovered by looking at the distance travelled field. If you need high location precision for your project, you may have to filter out ‘Travelling’ lists that cover a large distance, and/or those where distance is not specified. Similarly, lists are often tagged to a particular hotspot location, which means that the precise place where a list was created may be several kilometres from the lat-long of the hotspot. The geographical precision of such lists is likely to be relatively low.

Complete lists. If you intend to look not only at the presence of a species, but also absence (or more accurately non-detection), then you will want to use only those lists that are ‘Complete’ (ie, where all species seen have been reported). This would be needed if, for example, you wanted to examine the frequency of reporting of a species over space or time. ‘Incidental’ or ‘Casual’ lists are by definition incomplete, and should be removed from such analyses. For all other lists, you would want to use only those in which the observer has stated that all species have been reported (ALL.SPECIES.REPORTED=1 in the data file).

Detection probability. Even while using only ‘complete’ lists, as described above, please keep in mind that the absence of a species from a list doesn’t mean that it was necessarily absent from the area. Absence from the list could mean that the species was simply not detected and identified by the observer, even though it was actually present in the area. The degree to which this can affect your results depends on the probability of detection of the species, which in turn depends on a number of things, including how easy the species is to detect and the duration/distance of the list. Although interesting analyses can be done even when ignoring the complication of varying detection probability, please do keep in mind that true absences cannot be inferred from a list.

Possible errors in effort. Some lists will appear odd, for example being of very short duration (e.g. 5 min) but reporting very many species (e.g. 60 species). One can also see the opposite – long duration (e.g. 2 hrs) or distance lists (e.g. 5 km) with very few species (e.g. 3 species), even though the list is marked ‘complete’. This is most likely a mistake made while uploading the list, and you may want to look for and exclude such lists, depending on your purpose.

Resources

eBird data products. These may give you some ideas for how to get started: maps (eg for Red Junglefowl), seasonality charts (eg for Kerala), species lists for a region (eg for Goa). More data products.

GBIF best practice guide for data gap analysis.

Useful software. R | more to come…

Examples

Here are some example analyses carried out in the past by the Bird Count India team to explore some of the data in eBird. In each case, we provide step-by-step procedures so that you can get a sense of the kind of data processing that was done to arrive at the result.

First things first!

The very first thing to consider, even before downloading the data, is whether you want to use the entire dataset for India. If you do, be warned that there are 4 million rows in the dataset, and you will not be able to open this file in Excel. If your software has a limit on the number of rows it can read (Excel’s limit is around 1 million), then it might be best to download a subset of the data, rather than all India. For example, you can download the data for only a single State or for a single species, or for a restricted date range. (Note that the download page gives the option to ‘include unvetted data’. For most purposes, it’s best to exclude such data.)

If you do want the entire India data, you will need to use a suitable software program to subset what you want. For example, one commonly used software platform for analysis and graphics is R. In R, you would open and subset the data using commands similar to those below.

```
dat <- read.delim("ebd_IN_prv_relAug-2016.txt", na.strings = c("NA", "", "null"), quote="")

## check the column names

names(dat)

## subset only data from Kerala

kl <- subset(dat, STATE_PROVINCE == "Kerala")

## check if number of rows are OK

nrow(kl)

## gives 889,898, just about OK for Excel!
```

If you have trouble opening the downloaded data, or are stuck at any stage, please let us know in the comments below, or on the Facebook event page.

Example question 1: What is the country-wide distribution of birding effort represented in eBird?

The motivation behind this question might be to assess which areas (let's say Districts) have the most active eBirders, and also to identify gaps in bird information such that efforts can be made to fill them.

To answer this question, it is necessary to count up the number of lists per District. We might first want to create a new spreadsheet (or 'dataframe', in R) in which each list is represented by one row (in contrast to the raw data, in which each record is represented by one row, and therefore each list may have several rows). In Excel, this can be done using pivot tables.

Once this is done, we can tabulate the number of rows per District and then arrange in descending order, perhaps. If we want to display the results on a map, then we need to download map data for India showing administrative boundaries, match the District names from eBird to the map data; and then use software like QGIS to display a map of effort.

Examining this map could tell us where eBirders are most active, and conversely, where the major gaps in information are. See an example of this kind of analysis in an earlier post on birding gaps.

Example question 2: How much birding is required in a single spot in order to find all or at least most species?

In other words, **how much birding effort is required to adequately document the birds** of a place? For this, we might take a single location that has substantial birding effort, and look at how species numbers accumulate with effort. To put it another way, as the total amount of time spent birding in that location increases, what is the pattern of increase in the total number of species seen? We would expect there to be a rapid increase at first, and then gradually to stop increasing. The point (amount of time, number of lists) where new species stop being found might be considered adequate effort for that location.

To examine this, after choosing a particular location, we would want to order the birding lists in sequence (from earliest to latest, the checklist ID gives this), calculate the accumulated species seen and plot that versus the accumulated number of lists or the accumulated time spent birding.

To see an example of this done in the past, please look at an earlier analysis of repeated lists at a location.

Example question 3. How well are Important Bird and Biodiversity Areas (IBAs) in India covered in eBird?

IBAs are areas of particularly rich and/or threatened bird diversity. Clearly it is important to document and monitor the birdlife within them. Very few Indian IBAs have regular monitoring programmes; can birdwatchers uploading their birdlists to eBird contribute useful information?

For this, we first need a digital map of IBAs from India. Once we have the boundaries of different IBAs, we can check which eBird lists have been contributed from within those boundaries by looking up the latitudes and longitudes of the lists. From those eBird lists from within IBA boundaries, we can then calculate the amount of eBirding (in terms of lists and/or duration) from each IBA, and sort them to see which are heavily and which are poorly eBirded.

For a single IBA (or a set of them), we could also calculate the reporting frequency of each species. The reporting frequency for a particular species is the proportion of ‘complete’ lists that contain that species. The reporting frequency can be compared (with caution!) across IBAs and across seasons/years.

A brief analysis looked at these sorts of questions for IBAs before.

From across the World

eBird data have been used in a large number of research and conservation applications across the world, and cited in various scientific publications. Here we highlight two examples of interesting work that uses eBird data; both from North America.

In the first piece of work, an animated migration map of a large number of species was created by carrying out species distribution modelling on data submitted to eBird by participants. The map shows the movement of transcontinental migrants based on the aggregation of large amounts of information on observation of birds in space and time.

The second piece of work we highlight here is an assessment of how observation skills of individual birders increases over time. It looks at how individual birders at a particular place discover new species and whether this can be used as a measure of differences and changes in observation and identification skills. The conclusion is unsurprising: the more you watch birds, the better a birder (in terms of observation and identification skills) you can become!

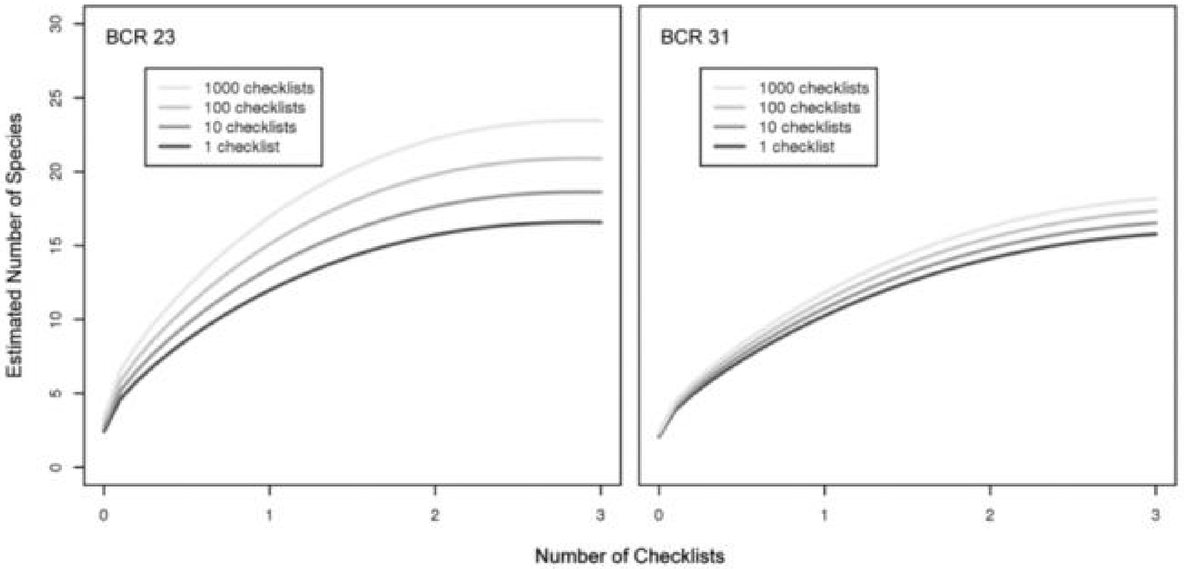


Fig 8. The change in SACs as a function of the cumulative participation in eBird. We estimated average changes in shapes of species accumulation curves with increasing number of checklists submitted to eBird from our BCR-specific models of species accumulation curves, to visualize whether observers report more species after they have submitted more eBird checklists. Note that while increased participation leads to a higher rate of accumulation of species, this effect is highest for beginning participants and slows with increased participation.

doi:10.1371/journal.pone.0139600.g008

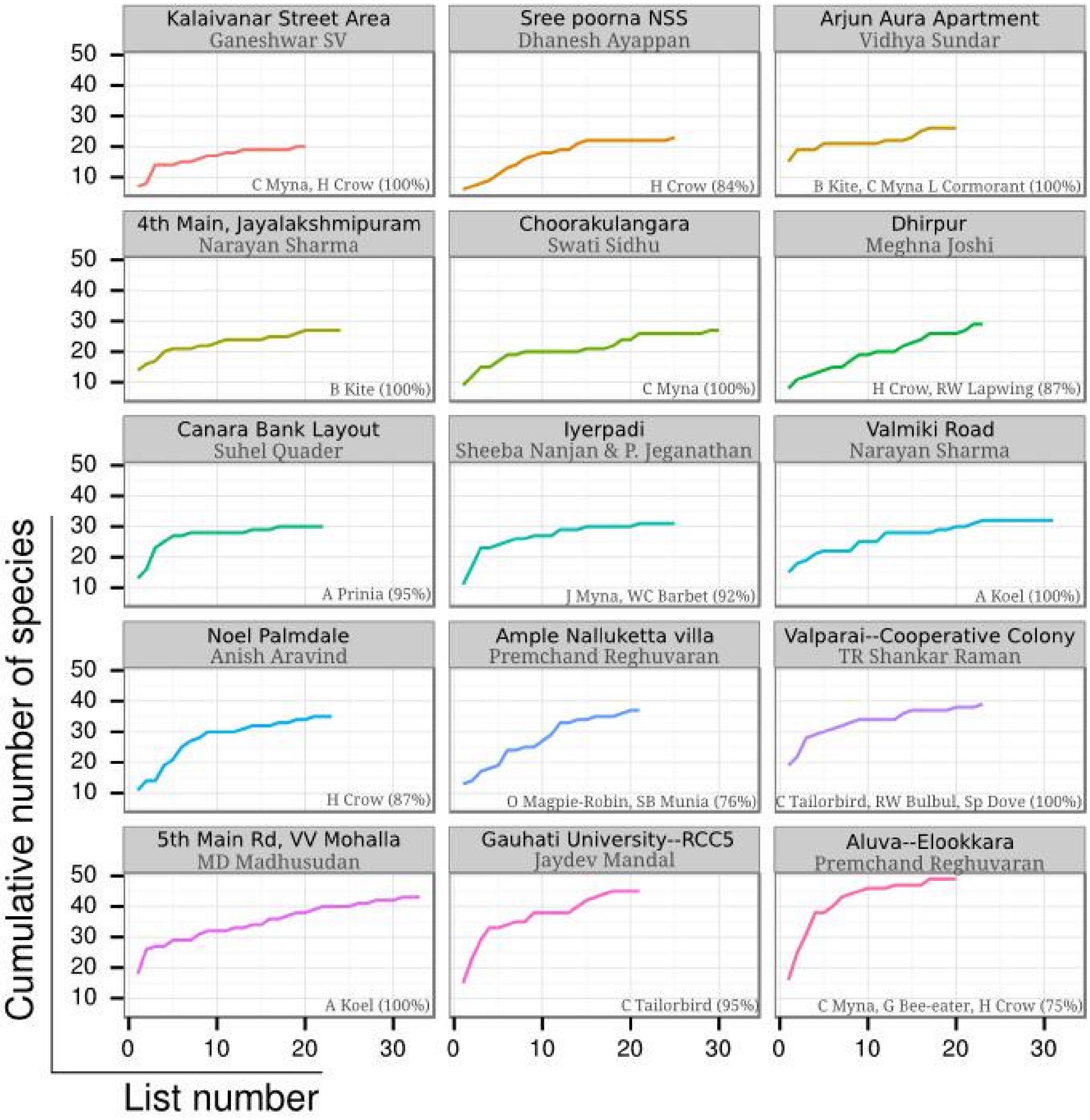
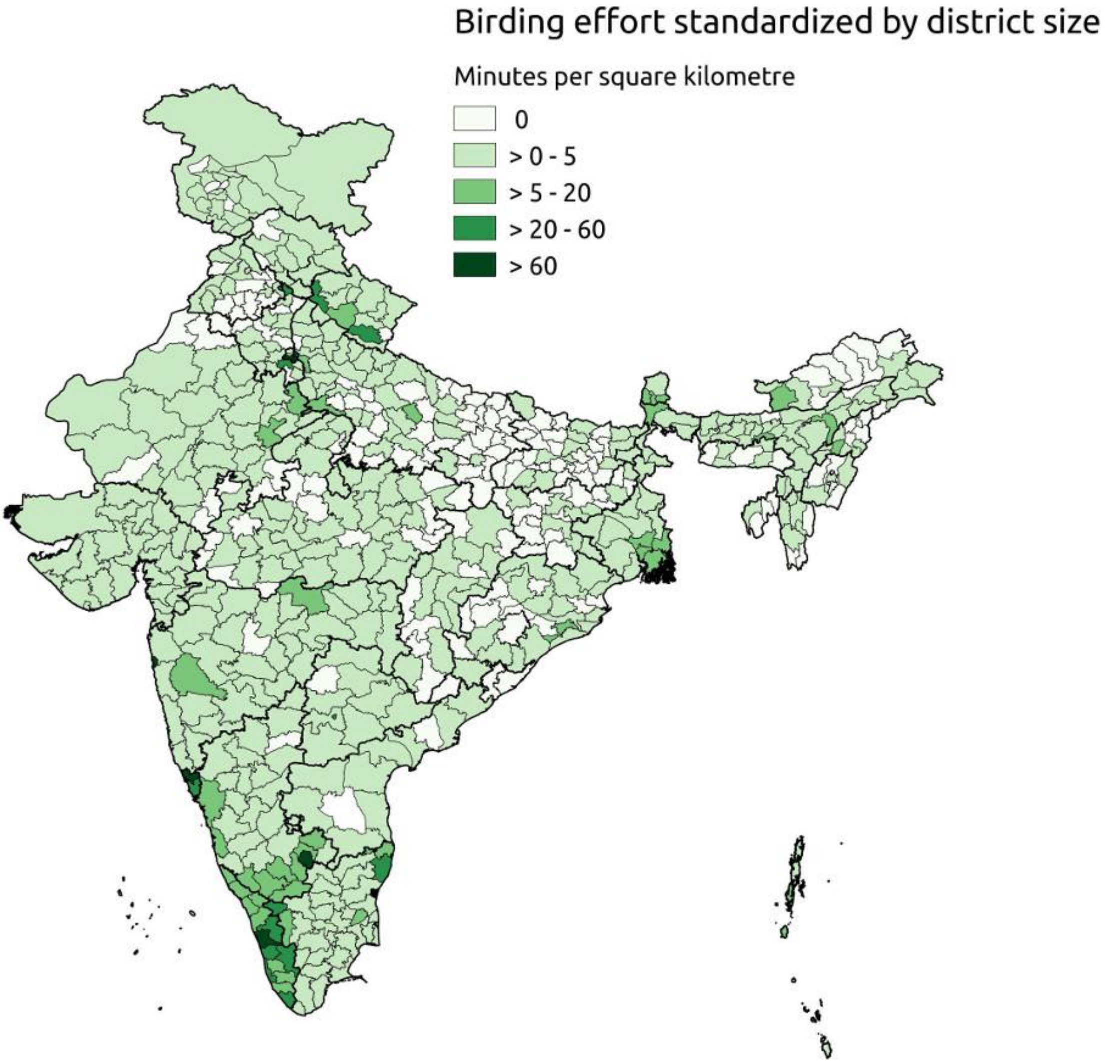

Kalaivanar Street Area
Ganeshwar SV
C Myna, H Crow (100%)
Sree poorna NSS
Dhanesh Ayappan
H Crow (84%)
Arjun Aura Apartment
Vidhya Sundar
B Kite, C Myna L Cormorant (100%)
4th Main, Jayalakshmipuram
Narayan Sharma
B Kite (100%)
Choorakulangara
Swati Sidhu
C Myna (100%)
Dhirpur
Meghna Joshi
H Crow, RW Lapwing (87%)
Canara Bank Layout
Suhel Quader
A Prinia (95%)
Iyerpadi
Sheeba Nanjan & P. Jeganathan
J Myna, WC Barbet (92%)
Valmiki Road
Narayan Sharma
A Koel (100%)
Noel Palmdale
Anish Aravind
H Crow (87%)
Ample Nalluketta villa
Premchand Reghuvaran
O Magpie-Robin, SB Munia (76%)
Valparai--Cooperative Colony
TR Shankar Raman
C Tailorbird, RW Bulbul, Sp Dove (100%)
5th Main Rd, VV Mohalla
MD Madhusudan
A Koel (100%)
Gauhati University--RCC5
Jaydev Mandal
C Tailorbird (95%)
Aluva--Elookkara
Premchand Reghuvaran
C Myna, G Bee-eater, H Crow (75%)
50
40
30
20
10
0
10
20
30
Cumulative number of species
List number

Table 1. *The ten Indian IBAs with the most birding effort (as measured by number of birding hours) represented in the eBird database. The rightmost column is the total number of species endemic (or near-endemic) to South Asia.*

IBA	Birding hours	Total lists	Total observations	Total species	South Asia endemics
Keoladeo National Park	1503	486	29076	427	66
Corbett Tiger Reserve	926	395	15419	532	75
Thattekkad Wildlife Sanctuary	696	292	9741	254	76
Periyar Wildlife Sanctuary	674	353	9707	275	76
Kedarnath Musk Deer Sanctuary	584	116	5685	272	38
Ranthambore NP and Tiger Reserve	543	191	9025	343	56
Kaziranga National Park	470	222	10149	557	46
Wild Ass Wildlife Sanctuary	374	105	5604	300	39
Kanha National Park	367	76	4523	325	65
Eaglenest and Sessa Sanctuaries	352	107	4801	362	17


Birding effort standardized by district size
Minutes per square kilometre
0
> 0 - 5
> 5 - 20
> 20 - 60
> 60