

Discovering and Developing Primary Biodiversity Data from Social Networking Sites

By
Vijay Vasant Barve

Submitted to the graduate degree program in Geography and the Graduate Faculty of the
University of Kansas in partial fulfilment of the requirements for the degree of Doctor of
Philosophy.

Chairperson J. Christopher Brown

A. Townsend Peterson

Xingong Li

Terry Slocum

Jorge Soberón

Date Defended: May 5, 2015

The Dissertation Committee for Vijay Vasant Barve
certifies that this is the approved version of the following dissertation:

Discovering and developing
primary biodiversity data
from social networking sites

Chairperson J. Christopher Brown

Co-Chairperson A. Townsend Peterson

Date approved: May 5, 2015

Abstract

An ever-increasing need exists for fine-scale biodiversity occurrence records for a broad variety of research applications in biodiversity and science more generally. Even though large-scale data aggregators like GBIF serve such data in large quantities, major gaps and biases still exist, both in taxonomic coverage and in spatial coverage. To address these gaps, in this dissertation, I explored social networking sites (SNS) as a rich potential source of additional biodiversity occurrence records.

In my first chapter, I explored the idea of discovering, extracting, and organizing massive numbers of biodiversity occurrence records now available on SNSs. I presented a proof-of-concept with Flickr as the SNS and Snowy Owls (*Bubo scandiacus*) and Monarch Butterflies (*Danaus plexippus*) as target species. The methods presented in this chapter can easily be used for any other SNS, region, or species group. These approaches are broadly applicable to animal and plant groups that are photographed, and that can be identified from photographs with some degree of confidence (e.g., birds, butterflies, cetaceans, orchids, dragonflies, amphibians, and plants). SNS thus offer a rich new source of biodiversity data.

To understand the strengths and weaknesses of biodiversity data, we need effective tools by which to explore and visualize these data. I developed a suite of such tools in an R package called `bdvis`, which is described in chapter two. The package allows

users to explore spatial, temporal, and taxonomic dimensions of biodiversity data sets to highlight gaps and identify strengths.

In the third chapter, I explored Flickr further as a source of biodiversity data for the birds of the world, to assess the potential of augmenting the largest portal to biodiversity occurrence data, i.e., the Global Biodiversity Information Facility (GBIF). GBIF provides access to $\sim 190 \times 10^6$ bird records, compared to $\sim 7 \times 10^6$ that I could discover from Flickr, out of which only $\sim 1.3 \times 10^6$ were geotagged. However, the Flickr data showed the potential to add to knowledge about birds in terms of geographic, taxonomic, and temporal dimensions, as Flickr data tended to be complementary to the GBIF-derived information.

Finally, I developed a case study to investigate the quantity of records existing, and the quality of identifications by users on Flickr. I developed a detailed case study of Indian swallowtail butterflies, and implemented a crowd-sourcing platform to recruit identification expertise and apply it to butterfly photographs from the SNS. Results were encouraging, with $>93\%$ correct identities for records of this family of butterflies from across India.

Acknowledgments

First and foremost I would like to thank my research committee members, each of whom helped me in shaping my research. Dr. J. Christopher Brown and Dr. A. Townsend Peterson for guiding me throughout each and every step when required and letting me explore on my own when necessary. This dissertation would not have been possible without them. Dr. Terry Slocum and Dr. Xingong Li for guiding me through GIS, Cartography, Visualization Techniques and Volunteer Geographic Information. Dr. Jorge Soberón was always there for me to discuss biodiversity informatics related topics.

I would like to thank the faculty, staff, and graduate students of the Department of Geography, the Biodiversity Institute, the Kansas Biological Survey, and the Department of Ecology and Evolutionary Biology of the University of Kansas for their support during my tenure as a PhD student. Special thanks to the KU ecological niche modelling group for testing my code and giving very useful suggestions to improve it.

I wish to thank my family, starting with my grandmother and parents who always have supported and encouraged my biodiversity explorations. The rest of my family in United States and back home for believing me in my potential to complete this. My wife, Narayani, who brought me to KU with her which led to the opportunity of enrolling as a graduate student and my daughter, Toshita, who is always an inspiration and support in my work.

Thanks to the people who have helped me along the way. My sincere gratitude goes to all of you who are not named here.

Table of contents

INTRODUCTION.....	1
DISCOVERING AND DEVELOPING PRIMARY BIODIVERSITY DATA FROM SOCIAL NETWORKING SITES: A NOVEL APPROACH.....	5
Abstract.....	7
Introduction.....	8
Historical Context	10
Need and Potential	11
Why Flickr	13
Extraction of Data from Flickr	14
Data gathering and organization.....	15
Case Studies and results	17
Discussion.....	18
Acknowledgements	20
References.....	20
VISUALIZING BIODIVERSITY DATA IN R USING THE BDVIS PACKAGE.....	27
Abstract.....	29
Introduction.....	29
Obtaining biodiversity data in R.....	30
Datasets.....	31
Visualization functions.....	31
Using bdvis	32

Visualizations.....	33
Future Plans.....	36
Obtaining bdivs.....	36
Acknowledgements	37
References.....	37
EXPLORING FLICKR AS A NOVEL SOURCE OF PRIMARY, VOUCHERED, OCCURRENCE DATA FOR BIRDS OF THE WORLD.....	54
ABSTRACT	56
INTRODUCTION	56
METHODS	58
Results.....	62
Discussion.....	63
Acknowledgments	65
References.....	65
FLICKR BIODIVERSITY DATA QUALITY: A CASE STUDY WITH SWALLOWTAIL BUTTERFLIES FROM INDIA.....	74
Abstract.....	76
Introduction.....	76
Methods	77
Results.....	80
Discussion.....	80
Acknowledgements	82
References.....	82
CONCLUSION	88

Introduction

The availability of high-quality biodiversity information is vital to all aspects of biodiversity research, and in particular to efforts focused on biodiversity conservation and its sustainable use. Researchers are using biodiversity information for an increasingly wide range of studies in areas including biogeography, ecology, invasive species biology, and climate change; this information is also being applied in more applied studies related to food security, control of disease vectors, and marine productivity. One of the most important components of biodiversity information is what has been termed “primary biodiversity data”: records that document occurrences of particular species at particular points in time and space. Associated metadata may include who recorded the species, who identified the species or verified the identification, climatic parameters, microhabitat information, number of individuals, sex, size, etc. Although these records are compiled in different formats for different kinds of studies (e.g., for assessing threats to a particular species, one might need all records of the species from both historical and current periods), a common format, Darwin Core has been developed that communicates the essence of primary biodiversity records, allowing broad data integration and extensive re-use of data, such that single data records may see numerous and diverse applications. When such data records are cast in such universally accepted formats, published openly, and integrated with other such data streams, they have been called “Digitally Accessible Knowledge” or DAK.

Sources of biodiversity occurrence records are diverse, but they may be grouped into

three broad classes: directed surveys, broad-scale surveys, and biological collections. Directed surveys focus on a particular organism or set of organisms (e.g., a species) across a particular area, at times involving specimen collections. Broad-scale surveys are more general assessments of a major taxon across a region, where observations are recorded by individuals ranging from trained scientists to interested citizens (e.g., Breeding Bird Surveys and Christmas Bird Counts in North America). Finally, biological collections are sets of specimens assembled to document the phenotypes and genotypes of biotas worldwide; biological collections are typically high-quality, but relatively low-volume sources of data on biological diversity.

More and more biodiversity occurrence records are being made available through aggregators like the Global Biodiversity Information Facility or GBIF (<http://www.gbif.org/>), VertNet (<http://www.vertnet.org/>), and eBird (<http://ebird.org/>) at global scales; on regional scales, portals like speciesLink (<http://splink.cria.org.br/>), SABIF (<http://www.sabif.ac.za/>), BioCASE (biocase.org) and Indian Biodiversity Portal (indiabiodiversity.org), are actively serving biodiversity occurrence records. GBIF currently serves $>5 \times 10^8$ records, and is growing rapidly with the help of data publishing institutions and networks. Major citizen science initiatives like eBird and iNaturalist have joined the venture and have fueled the growth of data served by GBIF massively.

This study focuses on a novel source of photo-vouchered primary biodiversity data records: social networking sites (hereafter referred to as SNS). SNS like Flickr, Facebook, Google+, and Picasaweb provide large numbers of users with the capability to share

images and associated metadata with other users. Many users post high-quality photographs that include identifications and geographic references, thereby including the basic ingredients of primary biodiversity data records. SNS focused on photographs are generally geo-aware: that is the user has ability to specify the location of a picture taken via coordinates or by choosing a location on a map. Indeed, increasing numbers of devices used to take photographs have built-in GPS units that record the precise location at which the image was captured. For photos taken without this technology, Internet-based mapping websites like Open Street Map and Google Earth offer improved graphical user interfaces and assist users in obtaining geographic coordinates easily and with greater precision. All of these components enable users to create primary biodiversity occurrence records out of photographs, and share them with others for viewing, appreciation, and comment.

The focus of this research is to discover (search and collect), organize (store and assess for quality and quantity) primary biodiversity occurrence records from SNS. In the first chapter, I have provided a proof-of-concept, showing that useful data are out there on SNS, and can be searched and tabulated for further use. In the second chapter, I developed the means by which to explore these data in spatial, temporal, and taxonomic dimensions, to understand the gaps and strengths of biodiversity data sets. This platform was developed to allow users access to interesting visualization tools developed in R as part of the package `bdvis`. In the third chapter, I gathered data for birds of the world from Flickr (an example SNS), and compared the resulting information with GBIF-mediated data to assess the potential of SNS-derived information to augment and improve the GBIF

data. In the final chapter, I assessed taxonomic identification quality issues to validate the fitness for use of SNS-harvested data.

Chapter 1

Discovering and Developing Primary Biodiversity Data from Social Networking Sites: A Novel Approach¹

¹ Barve, V. 2014. Discovering and developing primary biodiversity data from social networking sites: A novel approach. *Ecological Informatics* 24:194–199.

Title: Discovering and Developing Primary Biodiversity Data from Social Networking
Sites: A Novel Approach

Author: Vijay Barve^{1,2}

1. Present Address: Department of Geography, 1475 Jayhawk Blvd, 213 Lindley Hall, University of Kansas, Lawrence, KS, 66045 USA. vijaybarve@ku.edu
2. Permanent Address: No 9, Sneha Nagar, Amruthahalli, Byatarayanapura, Bangalore 560092 India. Vijay.barve@gmail.com

Abstract

Detailed, authoritative Digital Accessible Knowledge (DAK) about biodiversity is crucial to any biodiversity informatics or conservation project. In most developing nations, significant DAK gaps exist both geographically and taxonomically. This paper explores a novel source of photo-vouchered biodiversity occurrence data, in the form of records associated with photos posted on social networking sites (SNSs). SNSs like Flickr, Facebook, and Picasaweb allow naturalists to share images and associated metadata with other users. I explore the idea of discovering and organizing massive numbers of biodiversity occurrence records now available on SNSs. I present a proof-of-concept with Flickr as the SNS and the Snowy Owl (*Bubo scandiacus*) and the Monarch Butterfly (*Danaus plexippus*) as target species, but methods presented here can easily be used for any other SNS, region, or species group, particularly in developing, mega-diverse countries where the need for biodiversity DAK is particularly acute. These approaches are broadly applicable to animal and plant groups that are photographed and that can be identified from photographs with some degree of confidence (e.g., birds, butterflies, cetaceans, orchids, dragonflies, amphibians, and plants), and thus offer a rich new source of biodiversity data.

Keywords: social networking sites; biodiversity informatics; primary biodiversity records; data discovery;

Introduction

The availability of high-quality biodiversity information is vital to all aspects of biodiversity science and in particular to efforts focused on its conservation and use.

Researchers are using biodiversity information for an increasingly wide range of studies in areas including biogeography, ecology, invasive species biology, and climate change; this information is also being applied in studies related to food security, control of disease vectors, and marine productivity (Chavan, Sood, and Arino 2010). One of the most important components of biodiversity information is what has been termed “primary biodiversity data”: records that document species’ occurrences in time and space. More specifically, these observations place an identified organism in a particular context in which time, date, and location are recorded. Associated metadata may include who recorded the species, who identified the species or verified the identification, climatic parameters, microhabitat information, number of individuals, sex, size, etc. (Kelling 2008). Although these records are compiled in different formats for different kinds of studies (e.g. for assessing threats to a particular species, one might need all records of the species from both historical and current records), a common format has been developed that communicates the essence of primary biodiversity records, allowing for broad data integration (Darwin Core Task Group 2009). When such data records are cast in such universally accepted formats, published openly and integrated with other such data streams, they have been called “Digitally Accessible Knowledge” or DAK (Sousa-Baena, Garcia, and Peterson 2014).

Sources of biodiversity occurrence records are diverse, but they may be grouped into three broad classes: directed surveys, broad-scale surveys, and biological collections.

Directed surveys focus on a particular organism or set of organisms (e.g., a species) across a particular area, at times involving specimen collections. Broad-scale surveys are more general assessments of a major taxon across a region, where observations are recorded by individuals ranging from trained scientists to interested citizens (e.g. Christmas Bird Counts in North America). Finally, biological collections are sets of specimens assembled to document the phenotypes and genotypes of the biotas worldwide; biological collections are typically high-quality, low-volume sources of data on biological diversity (Kelling 2008).

This study focuses on a novel source of photo-vouchered primary biodiversity data records: social networking sites (hereafter, SNSs). SNSs like Flickr, Facebook, Google+, and Picasaweb provide large numbers of users with the capability to share images and associated metadata with other users. Many users post high-quality photographs that include identifications and geographic references, thereby constituting primary biodiversity data records. The focus of this research is to discover (search and collect), organize (store and assess for quality), and utilize (develop ways to make use of) primary biodiversity occurrence records from SNSs. Figure 1 shows a diagrammatic representation of the proposed scheme.

Historical Context

Directed surveys have been carried out for centuries. These surveys require careful planning and are expensive to execute. As a consequence, such surveys are generally conducted on limited spatial and temporal scales. Over time, many such surveys generate a lot of data, but their diverse nature and foci can introduce heterogeneity in reporting formats, making it difficult to integrate data and use these data in a meaningful way. Such data are highly susceptible to loss, as they are often considered personal research materials, such that they require a great deal of care and maintenance (Kelling 2008). Typical examples of this type of data collection might include Ph.D. dissertation projects or other efforts financed by small grants, which in most cases do not require data sharing or deposition of data in a central repository.

Broad-scale surveys generally involve engaging large numbers of citizen scientists to collect data. Protocols are usually standardized, at least to some degree, but resulting datasets may show biases toward accessible sites near population centers. Data from such surveys tend to concentrate on the most charismatic and visible species (e.g., birds); this situation is changing as more projects are developed involving citizen scientists that focus on less showy creatures (Cohn 2008). In recent years, most of the data from such efforts are available in Internet-accessible databases (Silvertown 2009), although some glaring exceptions remain.

Scientific collection of specimens has occurred across the globe over at least the last three centuries. An estimated 3 billion specimens exist in museums worldwide (Beaman and Conn 2003; Ariño 2010). Perhaps less than 40% of the information associated with

these specimens is accessible as DAK (Faith et al. 2013). In the coming years, we might expect to see almost 100% of collections data available via the Internet. However, across the world's scientific collections, gaps and biases in taxonomic and geographic coverage are significant (Boakes et al. 2010; Yesson et al. 2007; Ballesteros-Mejia et al. 2013; Sousa-Baena, Garcia, and Peterson 2014; Peterson, Ball, and Cohoon 2002).

SNSs are web-based services allowing individuals to create personal profiles, articulate connections with other users, share digital materials with other users, and browse through other user's profiles. Since 2003, SNSs have built a massive user base (Boyd and Ellison 2007), estimated at more than 1.5×10^9 users. SNSs allow users to share digital objects and tag content with information detailing attributes. Images are being uploaded on SNSs at a rate of billions of images per month, many showing elements of biodiversity. Indeed, some citizen science projects are now utilizing SNSs as a platform to upload occurrence records in the form of photographs for a broad spectrum of taxa (e.g., Encyclopedia of Life <http://eol.org/>, Arkive <http://www.arkive.org/>;(Kirkhope and Williams 2010).

Need and Potential

More than 435 million primary biodiversity data records are being made available through GBIF and other biodiversity data portals, but these records still provide an incomplete picture of the magnitude of global biodiversity (Yesson et al. 2007; Ballesteros-Mejia et al. 2013). To fill gaps in this knowledge, building scientific data repositories is a priority. These developments will focus on data available in museums

(Blagoderov et al. 2012) and through research, but new sources of such data should also be explored.

Photographs are increasingly accepted as vouchers for primary biodiversity records if they are diagnostic and accompanied by high-quality metadata. An important recent call is to mobilize such data into DAK (Morris et al. 2013). Data generated by citizen scientists are increasingly appreciated as a key input into mainstream biodiversity data sources. GBIF already incorporates data sources like eBird and iNaturalist, and technologies like smart phones are contributing to this push with their ability to run applications that capture and contribute records, complete with geo-located photographs (Bacher 2012).

SNSs have become an increasingly popular way for people to share all kinds of information. Naturalists have been especially inclined to use SNSs to share photographs of organisms with others in their communities. Indeed, numerous communities have been formed to share photographs and information about particular taxonomic groups or regions (Kumar et al. 1999). These communities generate appreciation for biodiversity, in addition to helping in confirmation of taxonomic identifications and building overall knowledge bases. Owing to the interactive and engaging nature of SNSs, large quantities of data are being generated on these sites. The most popular sites among naturalists include Flickr, Facebook, and Picasa / Google+.

Why Flickr

Flickr was chosen as a target SNS for the initial study, allowing a proof-of-concept assessment. Flickr is a photo-sharing website that was launched in 2004; it is popular with photographers, thanks to its rich set of features and open architecture that allows development of applications using the site's contents. Flickr presently serves (as of March 2013) 8×10^9 photos, growing at a rate of 3.5×10^6 photos daily (Jeffries 2013).

Since late 2004, Flickr has supported Application Programmers Interfaces (APIs), having one of the most comprehensive and mature set of APIs in the industry. APIs are nothing but hooks provided by the web service that allow developers to build on the features provided by the service. Flickr encourages API-based applications, and it actively promotes their use. Of special interest to naturalists, Flickr supports features like geo-coding of photographs, the ability to tag photos with keywords, and even a machine-tagging feature, which can be used for storing taxonomic hierarchies within the system.

Flickr searches are powerful, and users can explore data effectively using searches based on tags or free text. Several projects, like Encyclopedia of Life (<http://eol.org/>) and Arkive (<http://www.arkive.org/>), use Flickr to get user contents. Users may make photos available to those sites simply by posting them in particular Flickr communities or tagging them with particular tags suggested by these initiatives.

The strength of Flickr lies in allowing users to assign different copyright licenses while uploading photographs. When users upload photos, they can choose from a range of

licenses, from strict copyright protection (viewable only) to liberal Creative Commons licenses, which permit other users to download and use the photos for profit or nonprofit purposes and with or without attribution. Flickr also provides fine-grained control over who can see photographs posted by the users, from friends and family only to any interested user. All of these features make Flickr an ideal candidate for exploration as a novel source of primary biodiversity records.

Extraction of Data from Flickr

Flickr provides an API through which specific queries can be passed to the website and data returned to the software (Flickr Development Team 2014). These APIs are similar to manual searches that one might run on any website, but the data returned are machine-readable. These data can then be used either for display in human-readable formats (web page) or for storage in a database. One peculiarity of the Flickr API is that blank spaces in the search string (like the space between genus and species in scientific names) needed to be replaced with a plus sign (“+”); with spaces, the API calls produced unpredictable results.

Typical APIs return data in data exchange formats like JavaScript Object Notation (JSON; <http://www.json.org/>) or eXtensible Markup Language (XML; <http://www.w3.org/XML/>), very similar to Hypertext Markup Language (HTML) popularly used for web pages. JSON and XML are plain text data formats used to share structured data that can be interpreted by machines. XML was used for this project.

Once XML data are received from Flickr, the data need to be parsed to extract relevant data in the form of a table. All essential elements are tabulated in relevant fields and stored. Metadata are added to the table to document query terms and date of download for later reference.

R (R Development Core Team 2012) is a language and environment for statistical computing and graphics. I used R in this project for its ability to use APIs and store resulting data in a simple and lightweight database for further analysis. The XML package (Temple Lang 2012) was used to parse the XML data received from the Flickr website, and the sqldf package (Grothendieck 2012) was used to store and retrieve data for analysis.

Data gathering and organization

The first step in data gathering is to define the focus and scope of the research. In short, what species are to be studied? An exhaustive list of scientific names, and common names if possible, must be compiled; typically, this list comprises a species group and a particular geographic area (e.g., birds of Togo). If no authority list exists *a priori* for that region, such a list needs to be procured or created from reliable sources and curated. This step may take the form of tabulating available species names from the data themselves (e.g., from existing biodiversity data), then reducing this list by combining variants and synonyms and checking for errors.

The next step is to include all synonyms for the species included in the list. Often, recent nomenclature changes are not reflected well in literature and data documenting biodiversity. Consequently, citizen-scientist recording and reporting of organisms on SNSs often produces lists in which one species has several different names. These synonyms must be enumerated and included among search terms, or valuable data will be lost. In R, a package called `taxize` (Chamberlain, Szoecs, and Boettiger 2012) has some useful functions for this exercise; it is important to list all synonyms and maintain their links to accepted names so that data for species can be pooled in later analyses. Common names in multiple languages must also be used in searching SNSs, as citizen naturalists often refer to the organisms they observe only by common names; this custom is understandable, because communication via SNSs is more casual and informal than it is formal and scientific. Websites like ITIS (<http://www.itis.gov/>) and EOL (<http://eol.org/>) have proven useful in this regard; one can develop scripts to extract common names from these websites.

Depending on the project and number of species under study, quantities of data may grow rapidly, such that it is very important to manage data effectively. For smaller numbers of species, data can be managed in simple comma separated value (CSV) text files. As projects grow larger, however, storing and managing data in flat file formats may prove difficult. For moderate-sized projects, platform independence and manageability can be achieved with a simple database like `sqlite`, which is very simple to use and lightweight, does not require installation of software, and can be handled using R-based tools like the `sqldf` package.

Case Studies and results

The Monarch Butterfly (*Danaus plexippus*)

A well-known species, *Danaus plexippus*, was selected as a first test to see how much data are available on SNSs. The Monarch Butterfly is a flagship species in North America owing to its size and coloration, and particularly its spectacular long-distance migration. The migration phenomenon has long been a very interesting phenomenon for citizen scientists to track (Monarch Watch 2014; Prysby and Oberhauser 2004; Howard and Davis 2004). I queried Flickr using the scientific name, and it returned 16,474 records of which 4,799 were geo-tagged. When queried with the common name “Monarch Butterfly,” I acquired 46,684 records, of which 5,318 were geo-tagged.

Mapping these records (figure 2) shows the expected concentrations of Monarch Butterflies across their distributional area in North America and introduced distribution areas in Australia. Records geo-tagged in Asia are either photographs misidentified as Monarchs or just use Monarch Butterfly as points of comparison for the similar species. The common name African Monarch is used for the species *Danaus chrysippus*, so the search sometimes returns photographs of African Monarch butterflies.

The Snowy Owl (*Bubo scandiacus*)

In light of a recent population irruption that placed many individuals of this spectacular Arctic bird much farther south than usual, I queried Flickr on the scientific name (*Bubo scandiacus*), Flickr returned 5,701 records of which 1,488 were geo-tagged. When

queried with the common name “Snowy Owl,” I acquired 51,598 records of which only 1,743 were geo-tagged. Alternative scientific names yielded some more records:, “*Bubo scandiaca*” 116 (22 geo-tagged), “*Nyctea scandiaca*” 2,007 records (516 geo-tagged) and “*Nyctea nyctea*” 234 records (55 geo-tagged).

Mapping these records (figure 3) shows the expected Holarctic distribution of records. Upon inspection of metadata of the few records geo-tagged in Asia and Africa, many were found to be from zoo or special exhibitions..

Discussion

Both of the examples highlight not only the richness of data available, but also the need to use common names in queries on Flickr and other SNSs to compliment scientific names. Indeed, searches yielded many more records using common names than with scientific names: 2.8 times more for Monarch Butterflies, and 6.4 times more for Snowy Owls. The large numbers of records for Snowy Owls under misspellings of scientific names and synonyms reflect recent taxonomic changes, highlighting the need to use these alternative terms in queries.

About 16% of the butterfly records and 6% of the bird records downloaded in queries had associated geo-tags, making it quick and easy to map the records. These tags offer immediate opportunities for quality control of records (Chapman 2005), and they also

lend themselves readily to incorporation in many analyses (Peterson et al. 2011). As with other biodiversity records, additional records could be geo-tagged (geo-referenced) *post hoc*, following detailed protocols now available (Guo, Liu, and Wieczorek 2008).

Scientists frequently note that confidence in results of biodiversity studies would be enhanced by greater access to much larger quantities of Digitally Accessible Knowledge (DAK) regarding elements of biodiversity. In today's world, if the data are not digital and not freely accessible (open access via the Internet), they are very hard to use. DAK is crucial in biodiversity studies in light of the large numbers of species and records used in many applications of analysis and modeling. Biodiversity DAK is especially critical for efforts towards conservation in mega-diverse developing countries (Soberón and Peterson 2009)

To address the issue of biodiversity DAK, several major initiatives have dedicated major efforts to enabling and making accessible large amounts of data. The Global Biodiversity Information Facility (GBIF) has been working with biodiversity scientists worldwide to publish their data; GBIF has also incorporated large amounts of data based on observations from projects like eBird. GBIF is now also working with communities, such as amateur divers, to share the biodiversity data that they capture through the GBIF data portal (GBIF 2013). With proper discovery and documentation of biodiversity data on Social Networking Sites, all of these efforts can be extended, thereby augmenting the biodiversity DAK currently available.

Questions still remain concerning how many of the SNS-derived records will pass data quality and fitness-for-use tests (Chapman 2005). Assessment of the accuracy of taxonomic identifications provided by the citizen scientists is critical, as is determination of the accuracy levels of the geo-tagging. Further work will be needed as well to assess the degree to which this data source is useful regarding less-well-known elements of biodiversity beyond birds and butterflies. Finally, it will be very useful to explore more SNSs beyond Flickr, which might have promise for additional regions and taxonomic groups.

Acknowledgements

Thanks to J. C. Brown and A. T. Peterson for helpful comments on drafts of the manuscript.

References

- Ariño, Arturo H. 2010. "Approaches to Estimating the Universe of Natural History Collections Data." *Biodiversity Informatics* 7: 81–92.
- Bacher, Sven. 2012. "Still Not Enough Taxonomists: Reply to Joppa et Al." *Trends in Ecology & Evolution* 27 (2) (February): 65–6; author reply 66. doi:10.1016/j.tree.2011.11.003.
- Ballesteros-Mejia, Liliana, Ian J. Kitching, Walter Jetz, Peter Nagel, and Jan Beck. 2013. "Mapping the Biodiversity of Tropical Insects: Species Richness and Inventory Completeness of African Sphingid Moths." *Global Ecology and Biogeography* (February 22). doi:10.1111/geb.12039.
- Beaman, Reed S., and Barry J. Conn. 2003. "Automated Geoparsing and Georeferencing of Malesian Collection Locality Data." *Telopea* 10 (1): 43–52.

- Blagoderov, Vladimir, Ian J Kitching, Laurence Livermore, Thomas J Simonsen, and Vincent S Smith. 2012. "No Specimen Left behind: Industrial Scale Digitization of Natural History Collections." *ZooKeys* 146 (209) (January): 133–46. doi:10.3897/zookeys.209.3178.
- Boakes, Elizabeth H., Philip J. K. McGowan, Richard A. Fuller, Ding Chang-qing, Natalie E. Clark, Kim O'Connor, and Georgina M. Mace. 2010. "Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data." *PLoS Biology* 8 (6) (June). doi:doi:10.1371/journal.pbio.1000385.
- Boyd, Danah M., and Nicole B. Ellison. 2007. "Social Network Sites: Definition, History, and Scholarship." *Journal of Computer-Mediated Communication* 13 (1) (October 17): 210–230. doi:10.1111/j.1083-6101.2007.00393.x.
- Chamberlain, Scott, Eduard Szoecs, and Carl Boettiger. 2012. "Taxize: Taxonomic Search and Phylogeny Retrieval."
- Chapman, AD. 2005. "Principles and Methods of Data Cleaning - Primary Species and Species-Occurrence Data." *Global Biodiversity*. Copenhagen.
- Chavan, V. S., R. K. Sood, and A. H. Arino. 2010. "Best Practice Guide for Data Discovery & Publishing Strategy and Action Plans 2010." *Global Biodiversity*. Copenhagen: Global Biodiversity Information Facility.
- Cohn, Jeffrey P. 2008. "Citizen Science: Can Volunteers Do Real Research?" *BioScience* 58 (3): 192–197. doi:10.1641/B580303.
- Darwin Core Task Group. 2009. "Darwin Core." *Taxonomic Databases Working Group*. <http://rs.tdwg.org/dwc/>.
- Faith, Dan, Ben Collen, Arturo Ariño, Patricia Koleff, John Guinotte, Jeremy Kerr, and Vishwas Chavan. 2013. "Bridging the Biodiversity Data Gaps: Recommendations to Meet Users' Data Needs." *Biodiversity Informatics* 8 (1): 41–58.
- Flickr Development Team. 2014. "Flickr API Documentation." *Yahoo! Inc.* <http://www.flickr.com/services/api/>.
- GBIF. 2013. "Diveboard: Diveboard - Scuba Diving Citizen Science Observations." <http://www.gbif.org/dataset/66f6192f-6cc0-45fd-a2d1-e76f5ae3eab2>.
- Grothendieck, G. 2012. "Sqlf: Perform SQL Selects on R Data Frames."
- Guo, Q., Y. Liu, and J. Wiecek. 2008. "Georeferencing Locality Descriptions and Computing Associated Uncertainty Using a Probabilistic Approach." *International Journal of Geographical Information Science* 22 (10) (October): 1067–1090. doi:10.1080/13658810701851420.

- Howard, Elizabeth, and AK Davis. 2004. "Documenting the Spring Movements of Monarch Butterflies with Journey North, a Citizen Science Program." *Monarch Butterfly Biology and Conservation*: 105–116.
- Jeffries, Adrienne. 2013. "The Man behind Flickr on Making the Service 'Awesome Again.'" *The Verge*. <http://www.theverge.com/2013/3/20/4121574/flickr-chief-markus-spiering-talks-photos-and-marissa-mayer>.
- Kelling, S. 2008. "Significance of Organism Observations: Data Discovery and Access in Biodiversity Research". Copenhagen.
- Kirkhope, CL, and RL Williams. 2010. "Social Networking for Biodiversity: The BeelD Project." In *International Conference on Information Society (i-Society)*, 625–626.
- Kumar, Ravi, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. 1999. "Trawling the Web for Emerging Cyber-Communities." *Computer Networks* 31 (11-16) (May): 1481–1493. doi:10.1016/S1389-1286(99)00040-7.
- Monarch Watch. 2014. "Monarch Watch." <http://www.monarchwatch.org/>.
- Morris, Robert A., Vijay Barve, Mihail Carausu, Vishwas Chavan, José Cuadra, Chris Freeland, Gregor Hagedorn, et al. 2013. "Discovery and Publishing of Primary Biodiversity Data Associated with Multimedia Resources : The Audubon Core Strategies and Approaches." *Biodiversity Informatics* 8 (1): 185–197.
- Peterson, A. Townsend, Lisa G. Ball, and Kevin P. Cohoon. 2002. "Predicting Distributions of Mexican Birds Using Ecological Niche Modelling Methods." *Ibis* 144 (1) (February 27): E27–E32. doi:10.1046/j.0019-1019.2001.00031.x.
- Peterson, A. Townsend, Jorge Soberón, Richard G. Pearson, Robert P. Anderson, Enrique Martínez-Meyer, Miguel Nakamura, and Miguel B. Araújo. 2011. *Ecological Niches and Geographic Distributions (MPB-49)*. Princeton University Press.
- Prysbly, MD, and K Oberhauser. 2004. "Temporal and Geographic Variation in Monarch Densities: Citizen Scientists Document Monarch Population Patterns." In *Monarch Butterfly Biology and Conservation*, 9–20.
- R Development Core Team. 2012. "R: A Language and Environment for Statistical Computing". Vienna, Austria: R Foundation for Statistical Computing.
- Silvertown, Jonathan. 2009. "A New Dawn for Citizen Science." *Trends in Ecology & Evolution* 24 (9) (September): 467–71. doi:10.1016/j.tree.2009.03.017.
- Soberón, Jorge, and A Townsend Peterson. 2009. "Monitoring Biodiversity Loss with Primary Species-Occurrence Data: Toward National-Level Indicators for the 2010 Target of the Convention on Biological Diversity." *Ambio* 38 (1) (February): 29–34.

- Sousa-Baena, Mariane Silveira, Leticia Couto Garcia, and Andrew Townsend Peterson. 2014. "Completeness of Digital Accessible Knowledge of the Plants of Brazil and Priorities for Survey and Inventory." Edited by Lluís Brotons. *Diversity and Distributions* 20 (14) (October 24): 369–381. doi:10.1111/ddi.12136.
- Temple Lang, Duncan. 2012. "XML: Tools for Parsing and Generating XML within R and S-Plus."
- Yesson, C., P.W. Brewer, T. Sutton, N. Caithness, J.S. Pahwa, M. Burgess, W.A. Gray, et al. 2007. "How Global Is the Global Biodiversity Information Facility?" *PLoS ONE* 2 (11). doi:doi:10.1371/journal.pone.0001124.

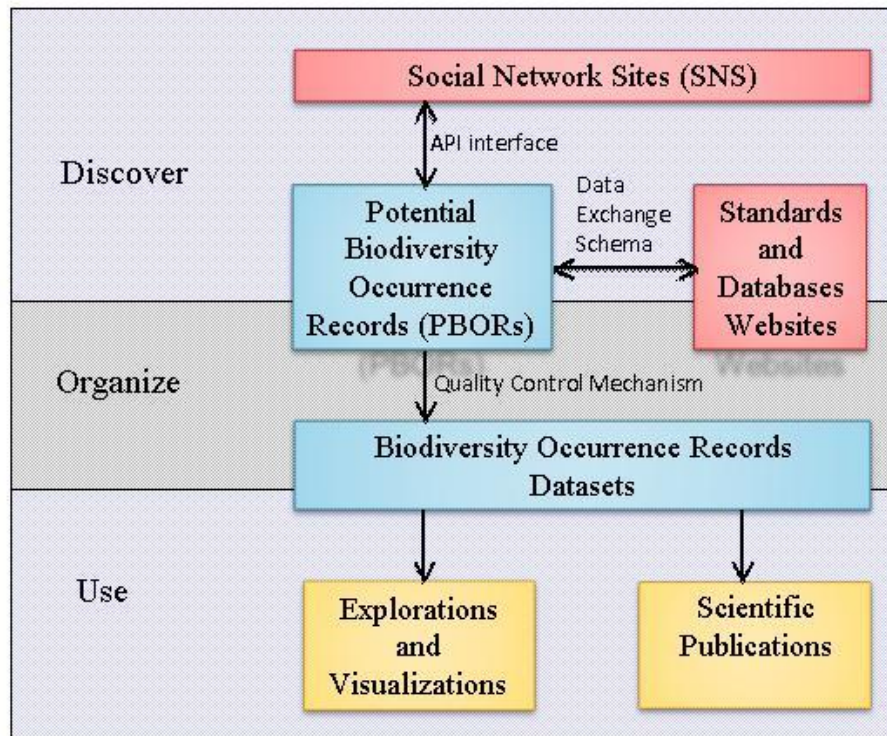


Figure 1: Proposed scheme of discovery, organization and use of biodiversity occurrence records held in SNS data repositories

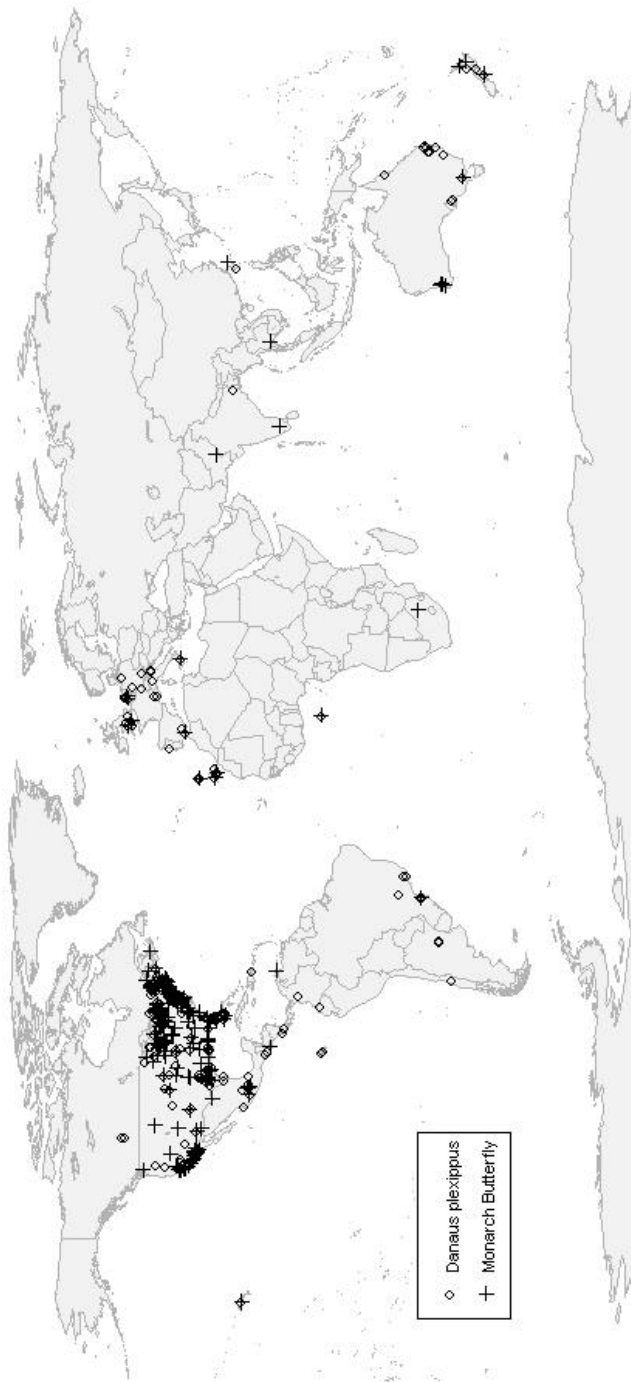


Figure 2: Map of geo-tagged records of Monarch Butterfly (*Danaus plexippus*) obtained via automated queries on Flickr

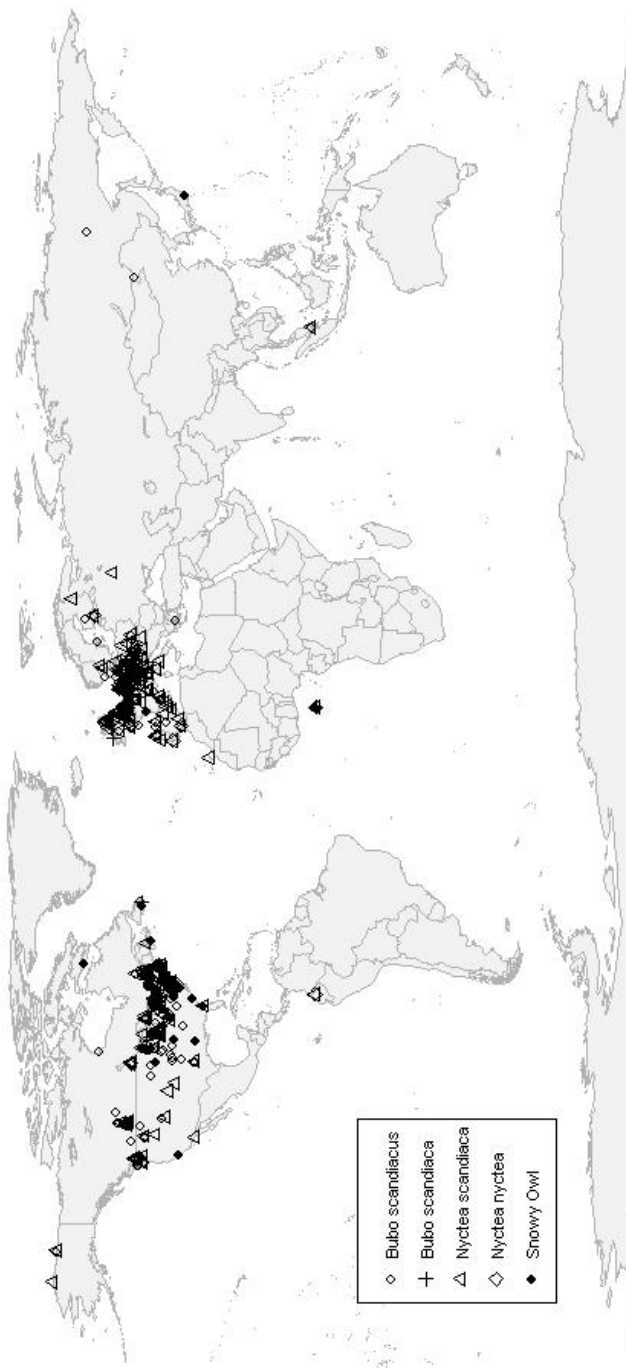


Figure 3: Map of geo-tagged records of Snowy Owl (*Bubo scandiacus*) obtained via automated queries on Flickr

Chapter 2

Visualizing biodiversity data in R
using the `bdvis` package

Visualizing biodiversity data in R using the bdivis package

Vijay Barve^{1,2}, Javier Otegui³

1. Present Address: Department of Geography, 1475 Jayhawk Blvd, 213 Lindley Hall, University of Kansas, Lawrence, KS, 66045 USA. vijaybarve@ku.edu
2. Permanent Address: No 9, Sneha Nagar, Amruthahalli, Byatarayanapura, Bangalore 560092 India. Vijay.barve@gmail.com
3. Postdoctoral Research Scholar at University of Colorado, Boulder, CO, USA. javier.oteguitellechea@colorado.edu

Abstract

Biodiversity studies are relying increasingly on primary biodiversity records (PBRs) for modeling and analysis, which are compiled by the researchers themselves, or obtained from data aggregators such as the Global Biodiversity Information Facility. Because biodiversity data are frequently ‘found’—i.e., not collected by the researcher for that particular study—researchers need to be aware of strengths and weaknesses of their data before they venture into further analysis. R is becoming a *lingua franca* of data exploration and analysis. We describe here an R package “`bdvis`” that facilitates efforts to understand the gaps and strengths of PBR data with quick and interesting visualization functions.

Introduction

Biodiversity studies are critical, because of the perceived risk of mass extinction due to rapid environmental changes in recent years. Most of these studies rely on primary biodiversity records (PBR) (Andrew et al. 2012, de la Torre et al. 2012, Ramírez-Bastida et al. 2008), which are simply records of species’ occurrence in certain places at certain times. PBR are relevant to almost every aspect of human endeavor, from basic needs like food and shelter to science and politics (Chapman and Speers 2005). Publications citing data served by the Global Biodiversity Information Facility (GBIF), which is a rich source of PBR, cover diverse areas like invasive alien species, climate change effects, conservation, human health, agriculture, etc. (GBIF 2015), which illustrates their broad relevance.

Informatics tools are seeing increased use in biodiversity science for improved management, presentation, discovery, exploration, and analysis (Soberón and Peterson 2004), challenges that are collectively referred to as biodiversity informatics. It is a relatively young, but rapidly growing, field. Visualizing data is a powerful technique by which to identify quickly the gaps and

strengths of the data in terms of geo-spatial, temporal, and taxonomic scales (Otegui et al. 2013). These assessments help data holders either to invest in improvement of the data or use the data with a better understanding of the gaps (Otegui and Ariño 2012).

More and more PBRs are being made available through aggregators like GBIF, VertNet, eBird at a global scale; on regional scales portals like BioCASE (biocase.org) and Indian Biodiversity Portal (indiabiodiversity.org) are actively serving PBR. GBIF currently serves $>5 \times 10^8$ PBRs, and is growing rapidly with the help of data publishing institutions and participating networks. Major citizen science initiatives like eBird (ebird.org) and iNaturalist (inaturalist.org) have joined the venture, and have fueled the growth of data served by GBIF in recent years.

R, the language and environment for statistical computing and graphics (The R Development Core Team 2012) is rapidly becoming the preferred tool for all kinds of data analysis. The package ecosystem supported by R is very effective in making reusable functions available to users. R has numerous packages (CRAN repository has more than 6000) for a wide multitude of tasks, several of which are useful for various biodiversity informatics-related tasks. This paper presents and explains the functionality of package `bdvis` with illustrations from a few sample datasets.

Obtaining biodiversity data in R

Several packages are available that allow users to access, manage, and visualize biodiversity data. The package `rvertnet` allows users to download PBRs for vertebrates. Similarly, the package `rgbif` allows user to download data from the GBIF data portal. Data portals such as VertNet and GBIF have their own gateways to download data, but packages like `rgbif` and `rvertnet` allow users to automate data download for multiple species, or download data under

advanced criteria. The package `spocc` allows users to access data from multiple sources.

Table 1 lists some of the functions available to access PBRs from various sources in R.

Datasets

Datasets selected for demonstrating the features of `bdvis` are summarized in Table 2. Our attempt was to select diverse datasets with different spatial and taxonomic coverage. The first dataset iNaturalist derived “research grade records” from GBIF, which are global in nature with broad taxonomic coverage, in the sense that iNaturalist includes data on all organisms that can be photographed and identified using photographs. The second dataset, India, consists of data served by GBIF for a specific country, India. Here again, the taxonomic coverage is all organisms, but spatial coverage is only India. The third dataset, the genus *Icterus*, is for a specific genus of birds, corresponding to the New World orioles. The dataset is taxonomically restricted to a single genus, and spatially to the Americas.

Data were downloaded from the GBIF website using the data portal directly for the iNaturalist and India datasets. For the *Icterus* data set, data were downloaded using the function `occ_search` in the `rgbif` (Chamberlain et al. 2014a) package. Since this data set is fairly large (almost 1,000,000 records), it took several hours to download the data using R.

Visualization functions

Table 3 lists functions currently available in `bdvis`. The package’s functions may be classified broadly as follows: (1) Helper functions to convert data to the correct format to be used in `bdvis`, and enriching an initial dataset with additional data like higher taxonomy and grid identifiers; (2) geographic visualizations; (3) temporal visualizations; (4) taxonomic

visualizations; and (5) assorted other graphs and charts. Many of these functions are available for the first time in R, and the seamless interface to access them from within a single R package makes them still more accessible and useful. Once the data are in appropriate formats, all of the package's functions work as single line commands in R.

Using `bdvis`

The data need to be in a format that the package understands for it to function. A couple of functions that help to achieve this are `fixstr` and `format_bdvis`. Function `fixstr` helps change required field names like scientific name, date collected, latitude, and longitude so that visualization functions work seamlessly. The function `format_bdvis` does the same task automatically for frequently-used data formats like the GBIF website download, `rgbif` package and the iNaturalist data with `rinat` package (Barve and Hart 2014).

Once field names are standardized, a next step is to calculate grid cell numbers as per the GBIF geographic scheme, as well as centi-grid cell numbers that function in parallel to provide finer spectral resolution. These data are used by functions like `mapgrid` to calculate statistics and plot maps. The function for this step is `getcellid`, which creates and calculates the fields `cellid` and `centicellid`.

For taxonomic visualization, higher taxonomy information is required to be associated with all records. Many times, biodiversity data have only scientific names, and fields like family, order, class, and phylum are missing. To add this information, the `gettaxo` function may be used. This function uses the `classification` function from the package `taxize` (Chamberlain et al. 2012) to retrieve higher ranks of taxonomic classification from the Encyclopedia of Life (Parr

et al. 2014) backbone classifications.

The function `datasubset` can be used to extract smaller subsets of data for comparative analysis, or for separating unwanted data from needed data. This step could be achieved by passing parameters like scientific name, or minimum and maximum year, or by writing a SQL statement to apply complex filters.

Once data are prepared for use in `bdvis` using the above-mentioned functions, they can be checked quickly using the `bdsummary` function. Referring to Figure 1, this function lists summary details like total number of records, range of dates in the data set, geographic range covered in the form of a bounding box, actual cells covered and taxonomic coverage. This function serves dual purposes: (1) validation that the data are prepared correctly for further exploration with the package, and (2) a summary of data gaps or richness as the case may be.

Visualizations

The function `mapgrid` creates a map of the data points in grid format. The data points are aggregated in grid cells and different colors are used to represent the density of points in respective grid cells. The maps may show just presence or absence of data; figure 2 illustrates presence-only data in the iNaturalist data set. Since iNaturalist data are contributed by citizen scientists, the red areas on the map are either regions where naturalists are based (e.g., North America and Europe) or places that they visit for excursions (e.g., East Africa). This function can also be used to display the number of species or number of records in each grid cell, depending on the `p_type` parameter (Figure 3 shows uses of `p_type`: the map in the left pane shows species richness, whereas the map in the right pane shows density of observations) Maps displayed can be global, or can be restricted by a bounding box; when using bounding boxes,

country names need to be specified in `regions` parameters to display borders.

Another exploratory approach is to plot records on a web map that can be zoomed and panned. In the background, Open Street Map is displayed, and occurrences are shown as pins on that map. Clicking on a pin displays the metadata about that record (see, e.g., Figure 4 shows a map of all Lepidoptera records in the GBIF India dataset). This approach is useful to verify the geocoding of records. The function `bdwebmap` creates this map, and lets the user explore it with functions like `zoom` and `pan`. In the background, this function creates a geojson file and uses the `leafletR` (Graul 2015) package to generate this map.

For temporal explorations, three functions are available. The function `tempolar` provides a polar plot of temporal data, which can be plotted using three different time scales (daily, weekly, monthly). The advantage of a polar plot is that the temporal continuity is maintained, in the sense that December connects to January, unlike the typical linear plots. This function is useful in biodiversity data to understand seasonality of data. The graph can be plotted with points (*s*), lines (*l*), or polygons (*p*), or a combination of these types, using the `plottype` parameter. Records can be averaged over years, rather than plotting raw values, using the `avg` parameter. Figure 5 shows a comparison of *Icterus* and iNaturalist data. The *Icterus* data, have a bias towards the month of May, whereas iNaturalist data are biased towards the warmer months of the Northern Hemisphere, from where most of the data are posted on the site.

The function `chronohorogram` creates another polar plot representation wherein each day is represented by a color dot, and each year, as a concentric ring in the plot, with 365 dots for each day of that year. The color of the dot summarizes the number of records on that particular day. The color scale is from blue to red, i.e. blue indicated few records and red indicated high

volume of records. This function is useful in highlighting the seasonality of the data collection or of the occurrence of the taxa in question. This function is also useful in identifying temporal gaps in data. A chronohorogram of iNaturalist data, from years 1980 to 2014, is shown in Figure 6.

Another interesting temporal visualization is to plot records on a calendar, using colors to indicate numbers of occurrence records. This visualization is a great reference diagram by which to explore temporal trends over a few years of data in a diagrammatic form. Figure 7 shows a calendar heat map of records from India over the past 5 years. Except for 2012 and 2013, the data are quite sparse; data for 2014 may still be getting processed, but earlier years certainly hold gaps in temporal coverage.

Among taxonomic visualizations is `taxotree`, which is basically a treemap representation of taxonomic data. In essence, treemaps display hierarchical data as a set of nested rectangles. The size of the rectangle typically denotes the number of records and the color of the rectangle indicates numbers of genera. The summary can be customized using the `sum1` and `sum2` parameters. Numbers of boxes can be controlled using the `n` parameter to avoid clutter. Figure 8 shows the top 30 families in the India dataset in a treemap. The darker color for family Orchidaceae indicates that close to 200 genera are covered in the data, though the number of records may not be high, as indicated by the size of the box. This result is well-justified, given that this family is one of the largest plant families with close to 900 genera worldwide. Note also the high number of genera with family unassigned, shown as dark green box in the top left corner.

The function `distrigraph` helps in visualizing the data distribution in terms of species, records per degree cell (geographic area), and sampling effort. A frequency graph of the

number of records per cell gives an idea of the evenness of geographic coverage. Figure 9 shows the distribution of records per cell in the India dataset, with records on a log scale so most of the cells can be seen to hold 10–100 records. A frequency graph of the number of records available in the dataset for each species gives an idea of the taxonomic coverage distribution, as illustrated in Figure 10: for most species of *Icterus* fewer than 50,000 records are present, whereas only four species have more, the highest being >400,000. The sampling effort graph summarizes the temporal coverage of the data: Figure 11 shows how the number of records in the iNaturalist dataset has increased exponentially after 2010 thanks to more and more people participating in citizen science initiatives.

Future Plans

Since July 2013, the package `bdvis` has been available on github, and users have been using the package. Several improvements in the aesthetics, as well as better control to the user, have been suggested by users: e.g., control over color selection of maps and graphs, ease of positioning of legend, etc. Seamless support for accessing more data sources, like Bison and eBird is also being requested. Support of additional parameters to add functionality, for functions like `taxotree` is also being suggested. Improved integration with packages like `rgbif` and `spocc` would also help users of these packages.

Obtaining `bdvis`

The `bdvis` package requires R installation (freely available from <http://cran.r-project.org/>), and can be downloaded from CRAN at: <http://cran.r-project.org/web/packages/bdvis/index.html>. The package is under development, and developmental releases can be downloaded from

<https://github.com/vijaybarve/bdvis>. We welcome bug reports and feedback, including suggestions for features to be included in future versions.

Acknowledgements

We are thankful to Google Inc. for the Google Summer of Code initiative, which brought the authors together to work on this package. We also thank the R Project for Statistical Computing for their support. For comments on early guidance on package development, we thank Scott Chamberlain, Carl Boettiger, Karthik Ram and Handley Wickham. We are also thankful to A. Townsend Peterson, Jorge Soberón, and Robert P. Guralnick, for ideas and guidance during development of the package, and Narayani Barve and Andrés Lira-Noriega for testing the package and offering suggestions on the user interface. Toshita Barve offered helpful suggestions on the manuscript.

References

- Andrew, M. E., M. a. Wulder, N. C. Coops, and G. Baillargeon. 2012. Beta-diversity gradients of butterflies along productivity axes. *Global Ecology and Biogeography* 21:352–364.
- Barve, V., and E. Hart. 2014. rinat: Access iNaturalist data through APIs. R package. Available at <http://cran.r-project.org/package=rinat>.
- Chamberlain, S., and V. Barve. 2012. rvertnet: Search VertNet database from R. Available at <http://cran.r-project.org/package=rvertnet>.
- Chamberlain, S., K. Ram, V. Barve, and D. Mcglinn. 2014a. rgbif: Interface to the Global Biodiversity Information Facility API. Available at <http://cran.r-project.org/package=rgbif>.
- Chamberlain, S., K. Ram, and T. Hart. 2014b. spocc: R interface to many species occurrence data sources. Available at <http://cran.r-project.org/package=spocc>.
- Chamberlain, S., E. Szoecs, and C. Boettiger. 2012. taxize: Taxonomic search and phylogeny retrieval. Available at <http://cran.r-project.org/package=taxize>.

- Chapman, A. D., and L. Speers. 2005. Uses of primary species- occurrence data, version 1.0. Copenhagen.
- GBIF. 2015. Peer-reviewed publications using GBIF data. Available at <http://www.gbif.org/mendeley>.
- Graul, C. 2015. leafletR: Interactive web-maps based on the Leaflet JavaScript library. Available at <http://cran.r-project.org/package=leafletR>.
- Hijmans, R. J., S. Phillips, J. Leathwick, and J. Elith. 2012. dismo: Species distribution modeling. Available at <http://cran.r-project.org/package=dismo>.
- De la Torre, L., C. E. Cerón, H. Balslev, and F. Borchsenius. 2012. A biodiversity informatics approach to ethnobotany: Meta-analysis of plant use patterns in Ecuador. *Ecology and Society* 17.
- Otegui, J., and A. H. Ariño. 2012. BIDDSAT: Visualizing the content of biodiversity data publishers in the Global Biodiversity Information Facility network. *Bioinformatics (Oxford, England)* 28:2207–8.
- Otegui, J., A. H. Ariño, M. a. Encinas, and F. Pando. 2013. Assessing the primary data hosted by the Spanish node of the Global Biodiversity Information Facility (GBIF). *PLoS ONE* 8:e55144.
- Parr, C. S., N. Wilson, P. Leary, K. Schulz, K. Lans, L. Walley, J. Hammock, A. Goddard, J. Rice, M. Studer, et al. 2014. The Encyclopedia of Life v2: Providing global access to knowledge about life on earth. *Biodiversity Data Journal* 2:e1079. Pensoft Publishers.
- Ramírez-Bastida, P., A. G. Navarro-Sigüenza, and A. T. Peterson. 2008. Aquatic bird distributions in Mexico: Designing conservation approaches quantitatively. *Biodiversity and Conservation* 17:2525–2558.
- Soberón, J., and a T. Peterson. 2004. Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 359:689–98.
- The R Development Core Team. 2012. R: A language and environment for statistical computing. Available at <http://www.r-project.org/>.

Figure 1 - Output of the function `bdsummary` in the package `bdvis`

```
Total no of records = 335
Date range of the records from 0010-07-07 to 2014-11-17
Bounding box of records 8.088306, 73.757774 - 27.491903, 96.212478
```

Taxonomic summary...

```
No of Families : 82
No of Genus : 124
No of Species : 217
```

Spatial coverage ...

```
Degree cells covered : 13
% degree cells covered : 2.974828
```

Figure 2 – Example of a mapgrid map of presence data in iNaturalist

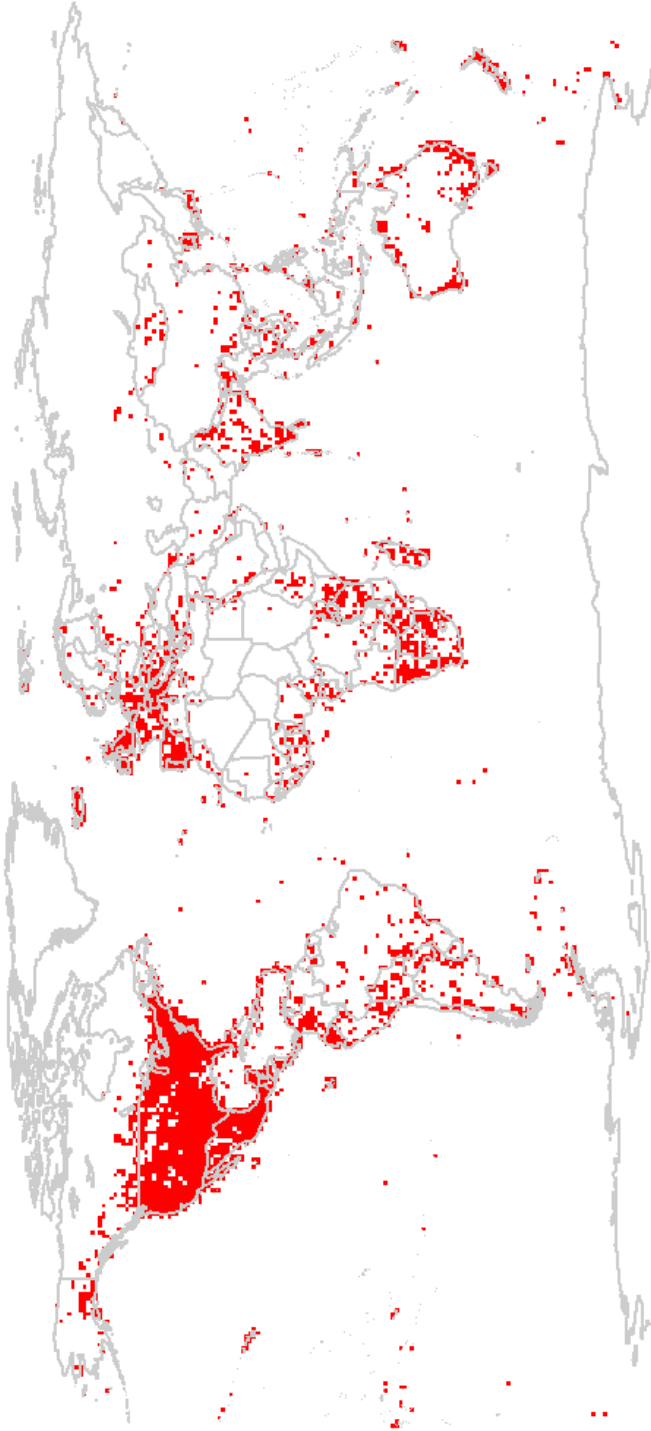


Figure 3 – Illustration of the function mapgrid to compare *Icterus* species richness versus record density across North America in the *Icterus* data set.

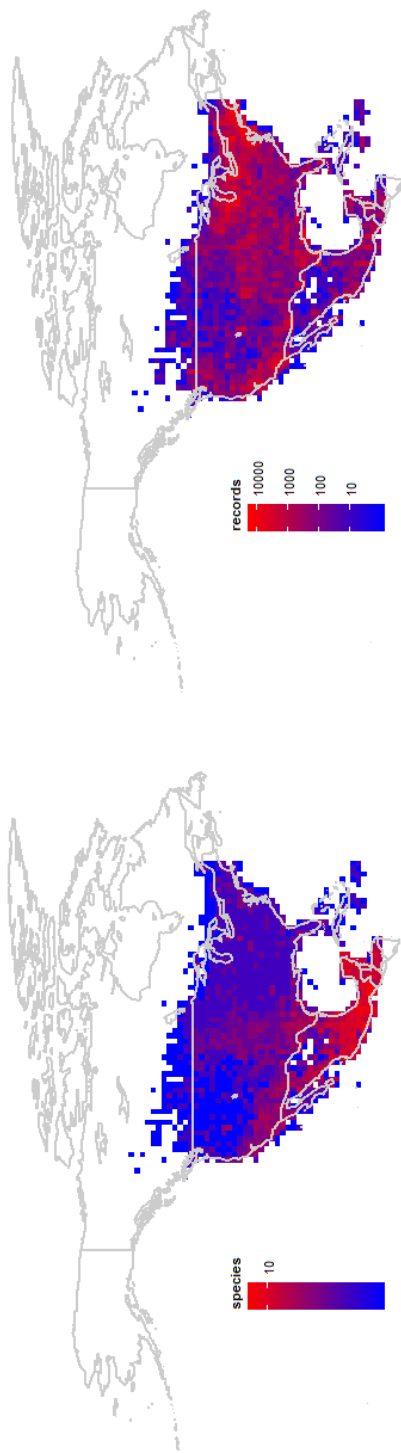


Figure 4 – Illustration of output `bdwebmap` in showing the geographic distribution of Indian Lepidoptera data served by the GBIF data portal

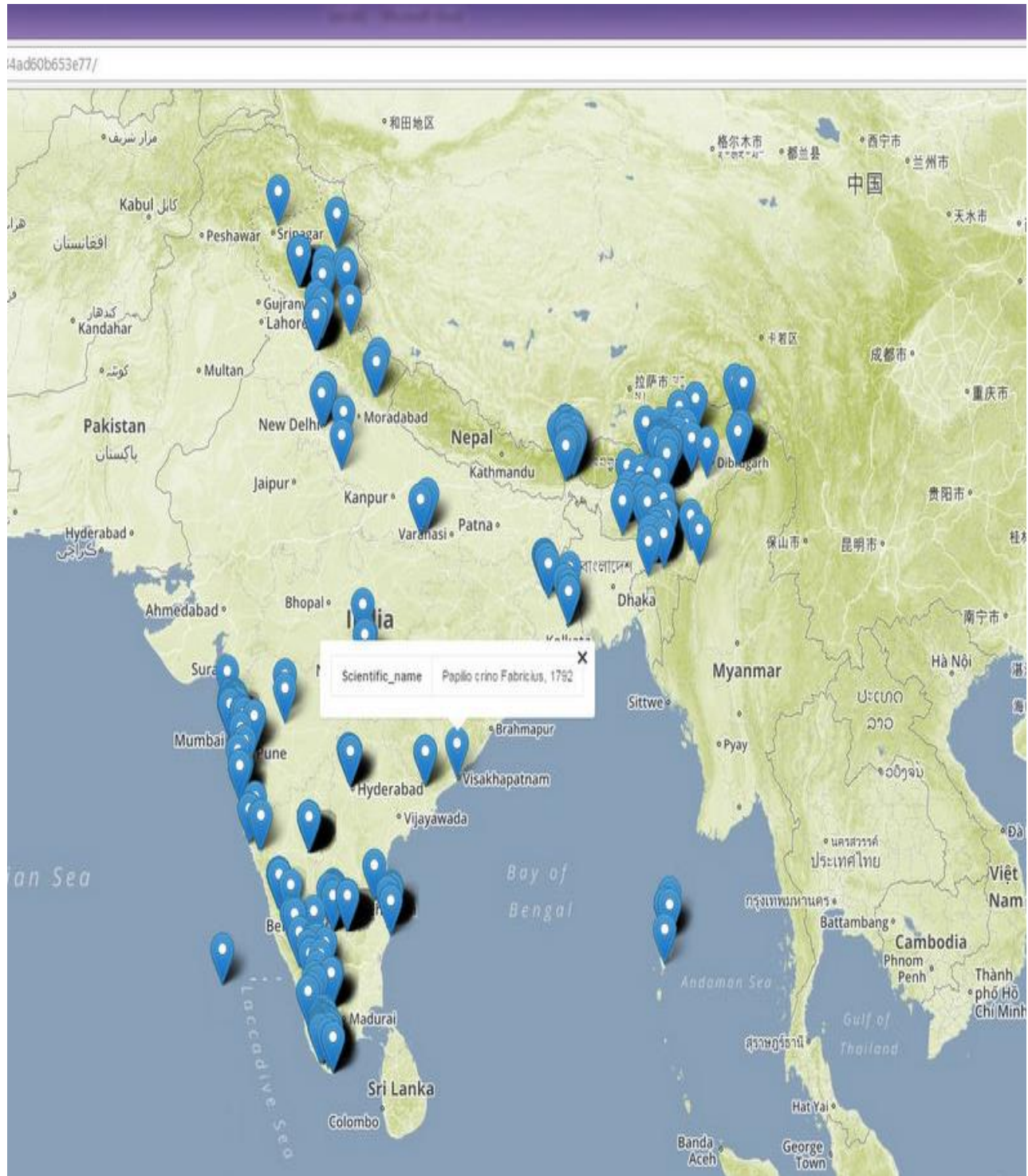


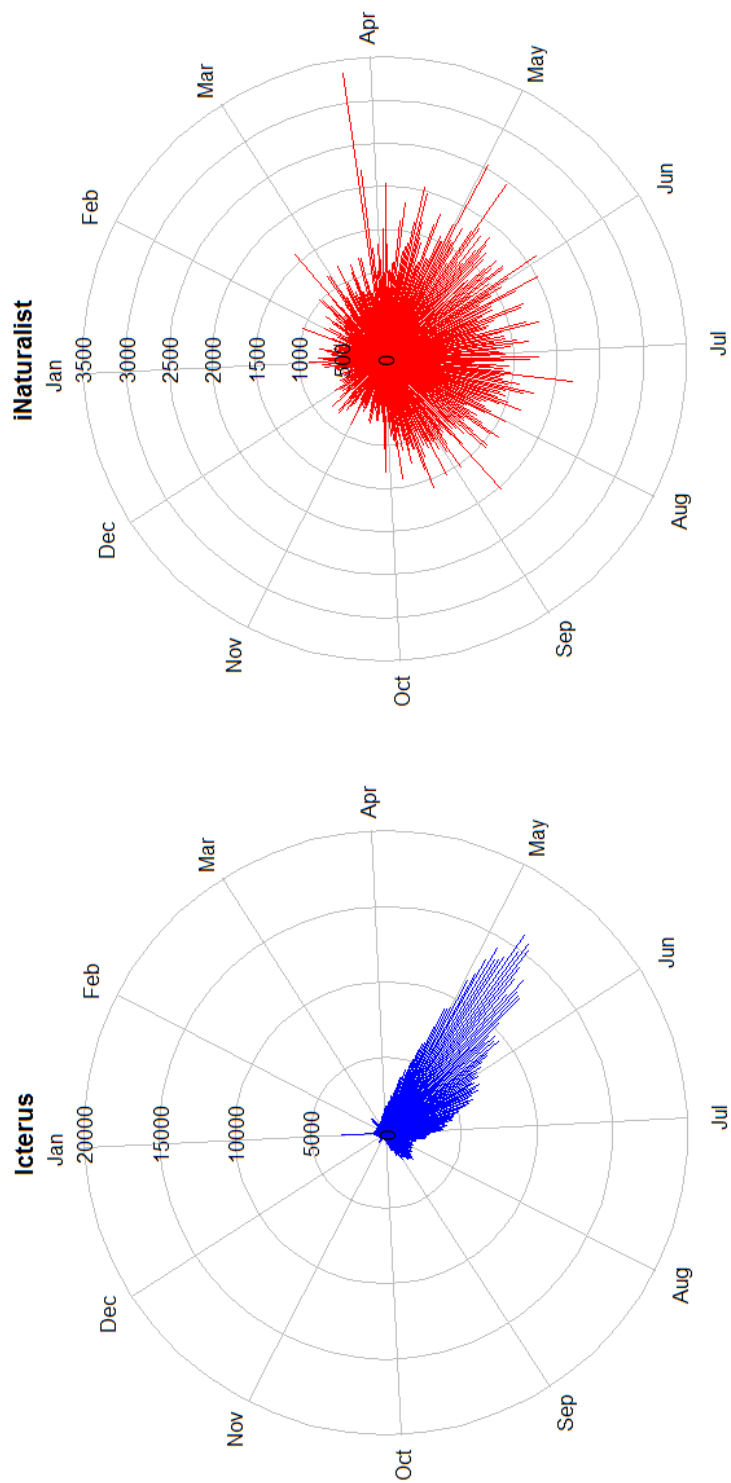
Figure 5 – Temporal data coverage of *Icterus* and iNaturalist using function tempolar

Figure 6 - Chronohorogram summarizing iNaturalist records from 1980 to 2014. Year 1980 in the center.

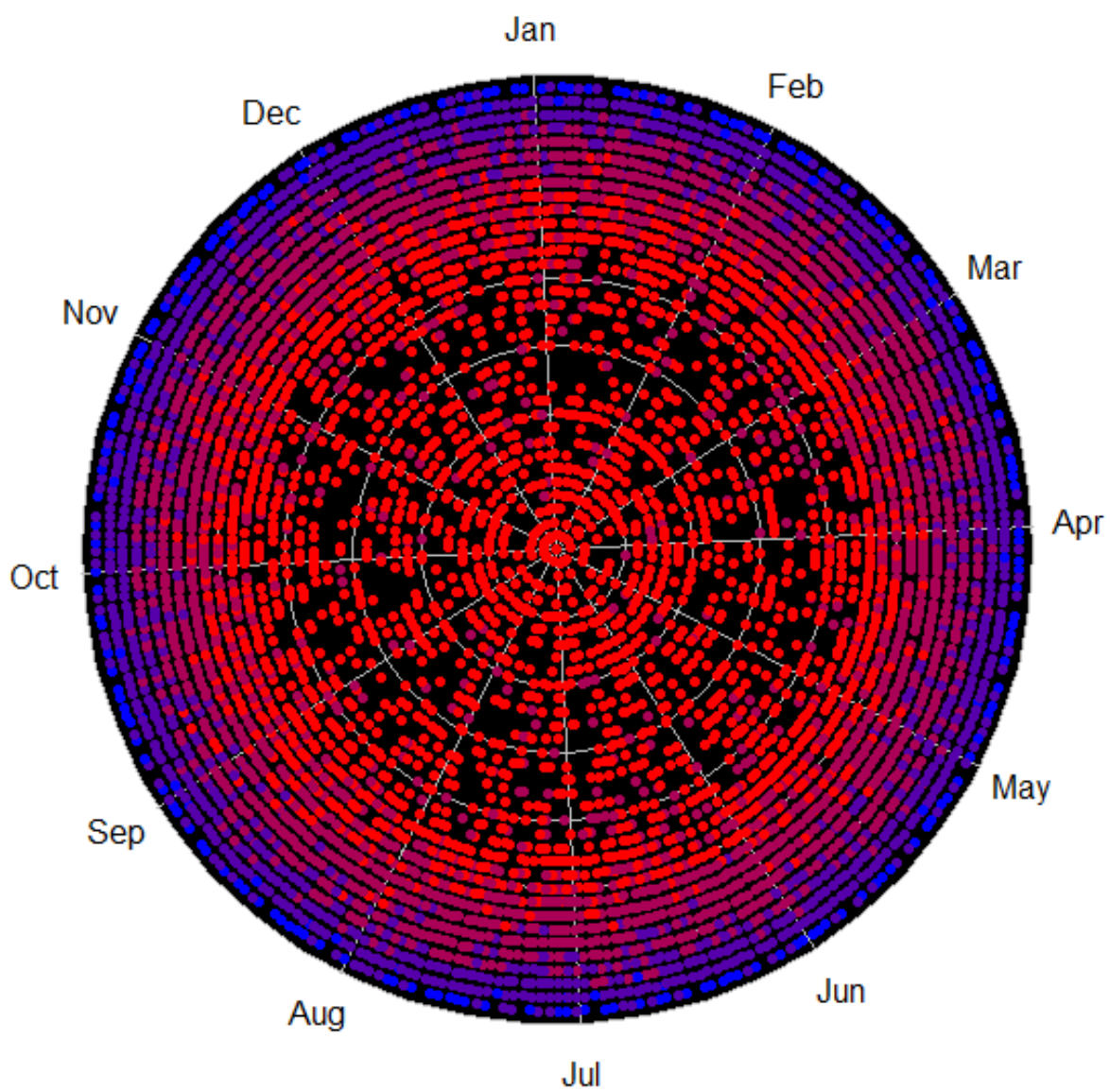


Figure 7 – Example of calendar heat map summarizing numbers of records from India in the GBIF mediated data store.

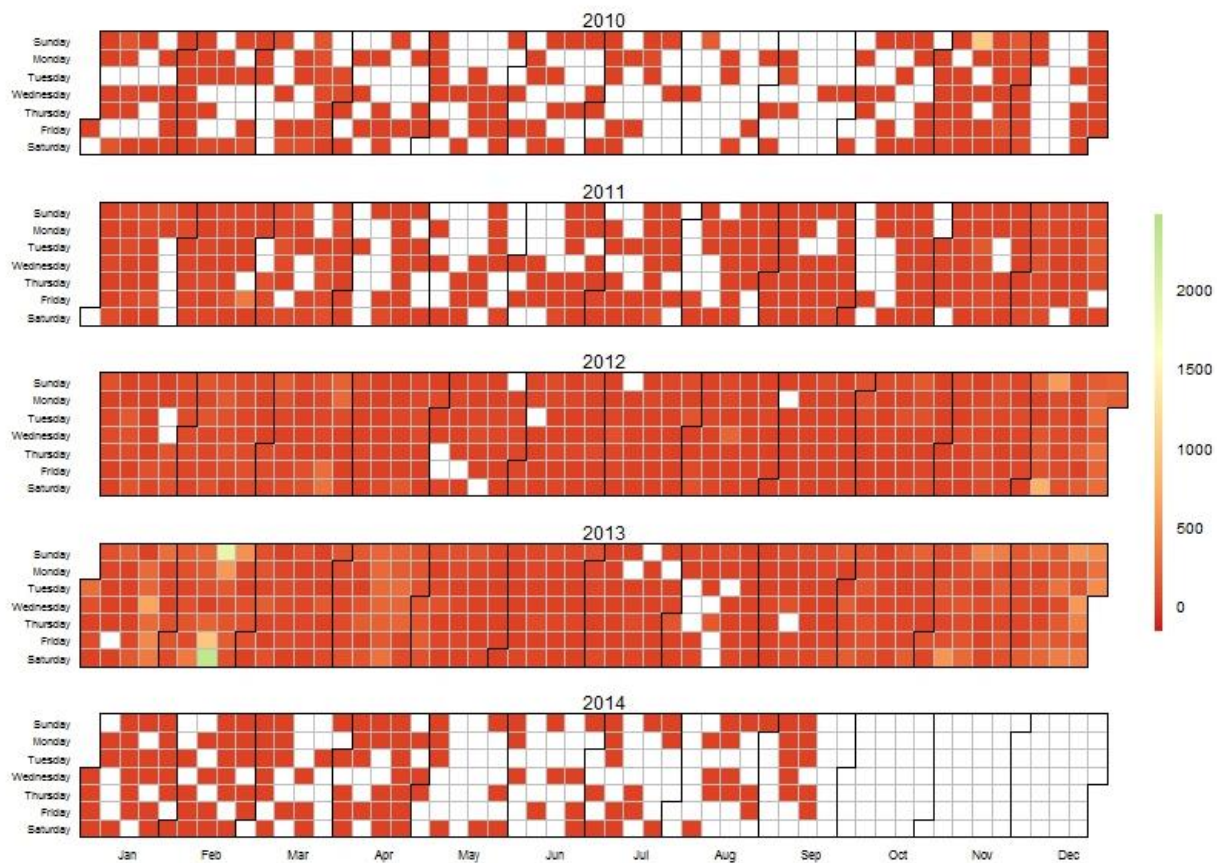


Figure 8 – Example of taxotree function output displaying number of records for top 30 families represented in the GBIF India data set.

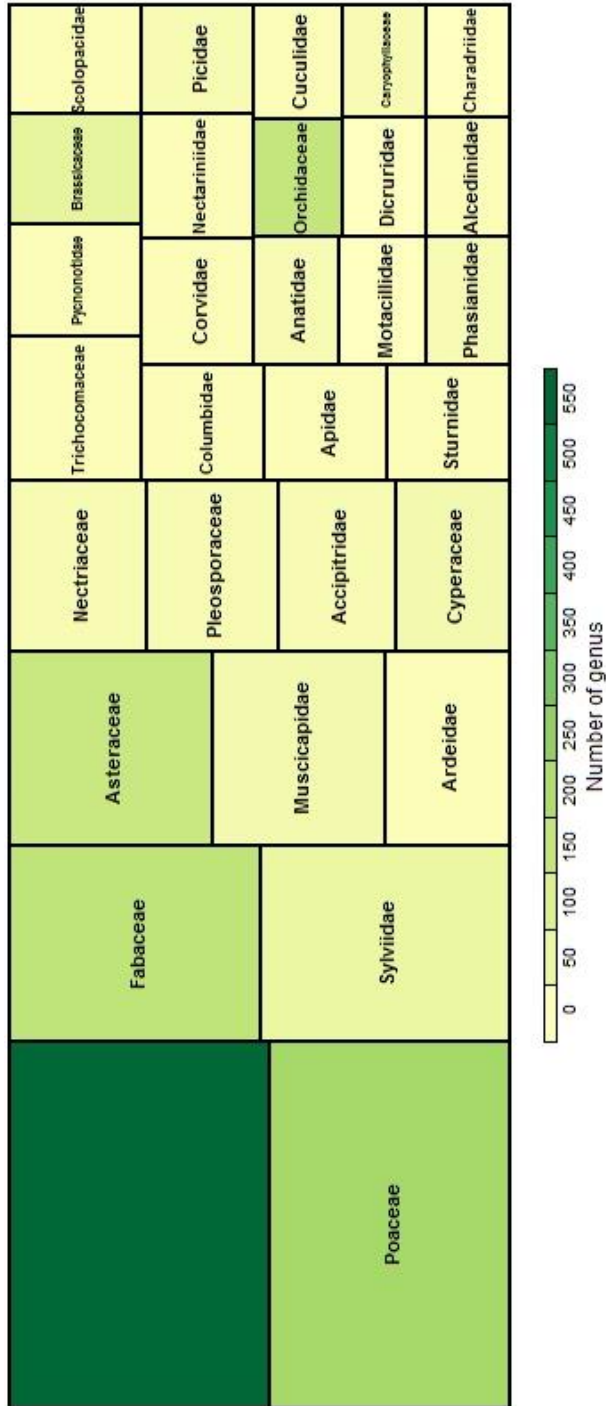


Figure 9 – Illustration of distrigraph function output for India dataset species per cell

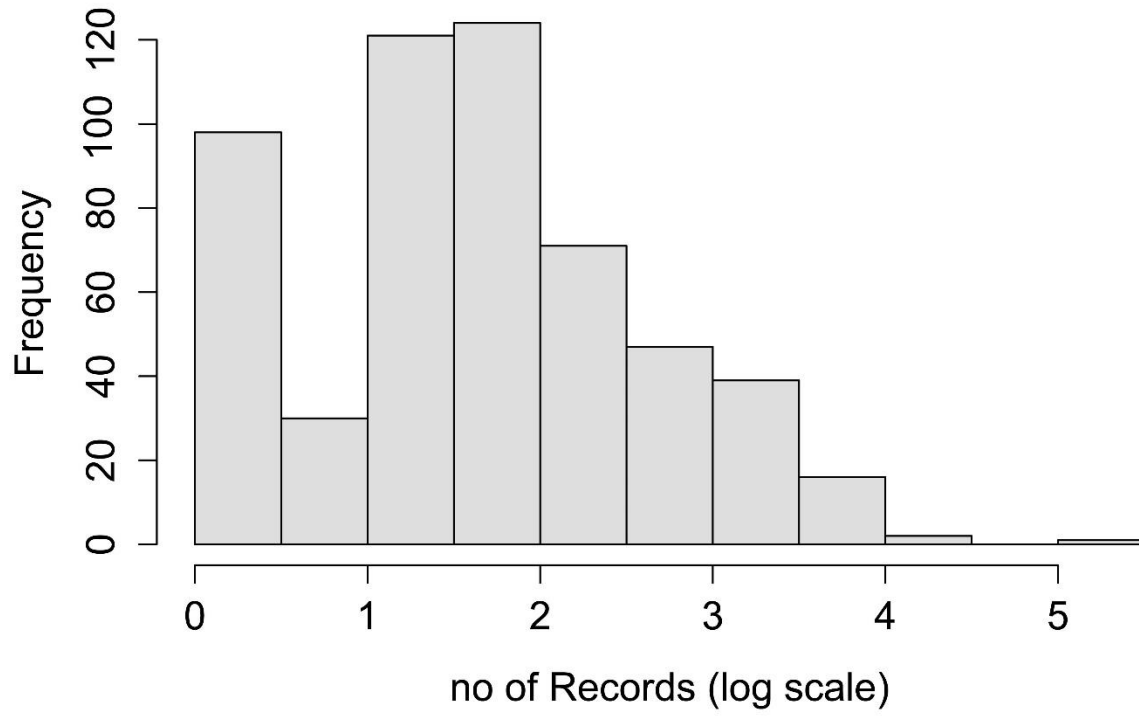


Figure 10 – Illustration of distrigraph function output for *Icterus* dataset records per species

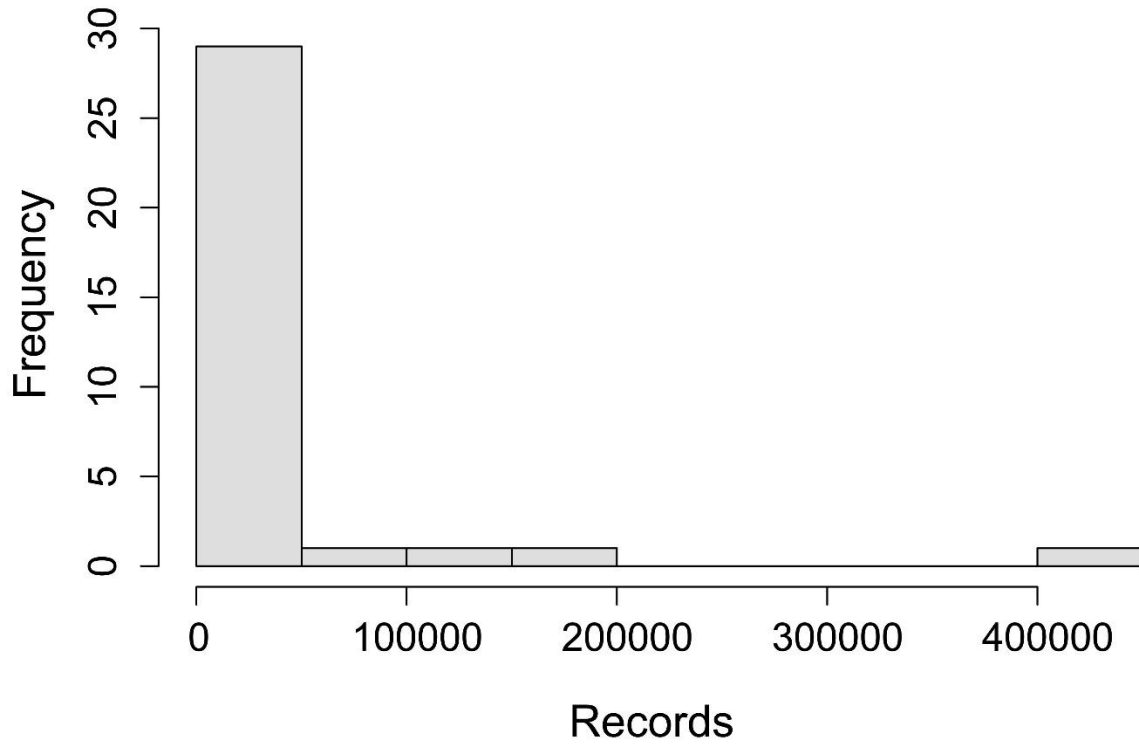


Figure 11 – Illustration of distrigraph function output for the iNaturalist dataset accumulation of records from 2000-2014

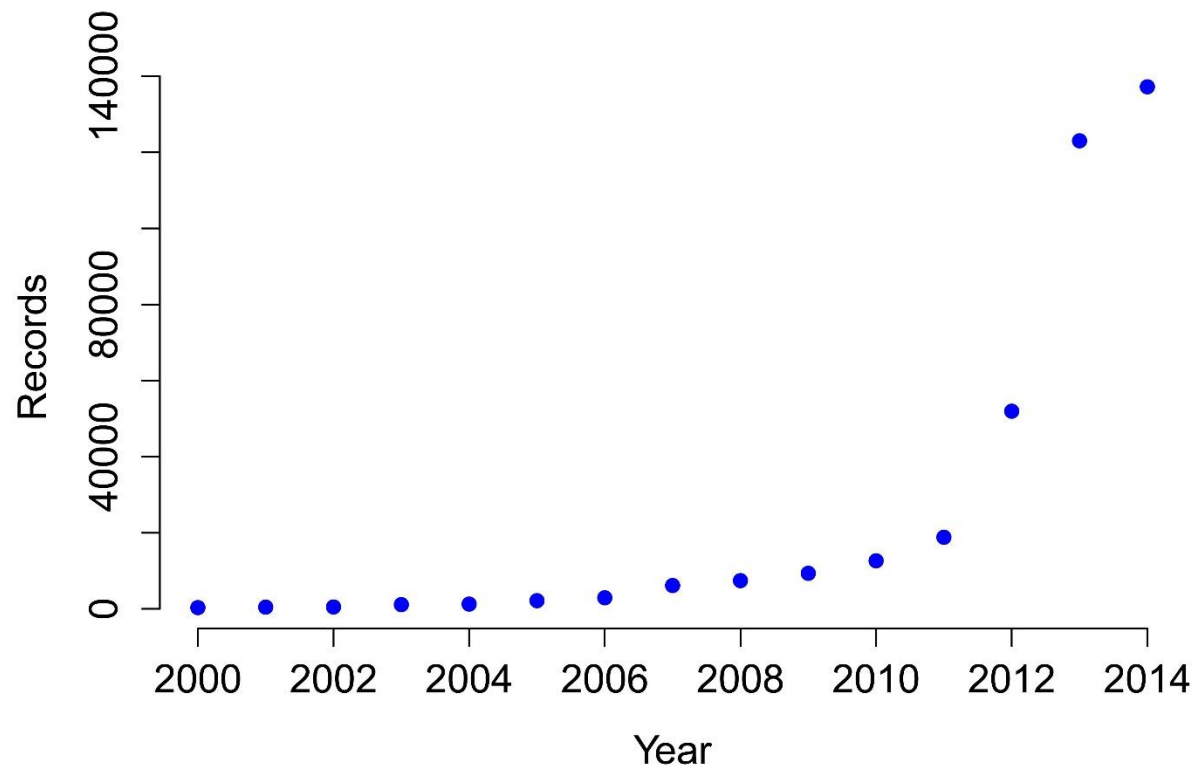


Table 1 R functions serving PBRs

Package	Functions	Data sources	Description
dismo (Hijmans et al. 2012)	gbif	GBIF	This function downloads species occurrence records
rgbif (Chamberlain et al. 2014a)	occ_search	GBIF	Searches for occurrence data on GBIF
rvertnet (Chamberlain and Barve 2012)	vertoccurrence	VertNet	Retrieves occurrence records from VertNet v2 portals.
rinat (Barve and Hart 2014)	get_inat_obs	iNaturalist	Retrieves observations from iNaturalist
spocc (Chamberlain et al. 2014b)	occ	GBIF, VertNet, Bison, eBird	Searches across multiple data sources and species

Table 2 Datasets used to illustrate the functions of bdvis

Dataset	Geographic coverage	Taxonomic Coverage	Number of records
iNaturalist (research grade)	World	All	378,432
India	India	All	791,951
Icterus	New World	Single genus	925,194

Table 3. Summary of functions in `bdvis` that enable visualization of biodiversity data

Function Name	Description
<code>bdcalendarheat</code>	Calendar heat map of biodiversity data
<code>bdcomplete</code>	Computes completeness values for each 1° cell; currently returns Chao2 index
<code>bdsummary</code>	Provides summary of data like number of records, families, genus, species, bounding box of locations, date range, and spatial coverage
<code>bdwebmap</code>	Creates interactive web page based map of records
<code>chronohorogram</code>	Draws a chronohorogram of records
<code>datasubset</code>	Subsets data for analysis in <code>bdvis</code>
<code>distrigraph</code>	Creates distribution graphs
<code>fixstr</code>	Fixes structure of the data frame to match the key fields to GBIF-style data field names
<code>format_bdvis</code>	Converts data to <code>bdvis</code> native format
<code>getcellid</code>	Assigns GBIF-style degree cell ids and centi-degree cell ids for each record
<code>gettaxo</code>	Retrieves higher taxonomy fields data from EOL
<code>mapgrid</code>	Maps the data points in grid format
<code>taxotree</code>	Draws a treemap based on taxonomic hierarchy of records

tempolar

Draws a polar plot of temporal data

Chapter 3

Exploring Flickr as a Novel Source of
Primary, Vouchered, Occurrence Data for
Birds of the World

Exploring Flickr as a Novel Source of Primary, Vouchered, Occurrence Data for Birds of the World

Vijay Barve, A. Townsend Peterson

Biodiversity Institute, University of Kansas, Lawrence, KS, USA

ABSTRACT

Interest is increasing in exploring Social Networking Sites (SNS) for crowdsourcing knowledge, including biodiversity. Showcasing interesting biodiversity elements through photographs is common among naturalists, which in turn generates masses of biodiversity occurrence data on SNS. We explored Flickr, one of the top ten popular SNS, for birds of the world, to assess the potential of augmenting the largest portal to biodiversity occurrence data, i.e., the Global Biodiversity Information Facility (GBIF). GBIF provides access to $\sim 190 \times 10^6$ bird records compared to $\sim 7 \times 10^6$ that we could discover from Flickr, out of which only $\sim 1.3 \times 10^6$ were geotagged. However Flickr data showed potential to add to knowledge about birds in terms of geographic, taxonomic, and temporal dimensions, as Flickr data tend to be complementary to the GBIF-derived information.

INTRODUCTION

Social Networking Sites (SNS) have gained massive popularity, and are increasingly being used by naturalists to share interesting photographs of biodiversity elements. SNS focused on photographs are generally geo-aware: that is the user has ability to specify the location of a picture taken via coordinates or by choosing a location on a map. Indeed, increasing numbers of devices used to take photographs have built-in GPS units that record the precise location at which the image was captured. For photos taken without this technology, Internet-based mapping websites like Open Street Map and Google Earth offer improved graphical user interfaces and assist users in obtaining geographic coordinates easily and with greater precision. All of these components enable users to create primary biodiversity occurrence records out of photographs, and share them with others for viewing, appreciation, and comment (Deng et al. 2014,

Stafford et al. 2010).

The idea of “communities” or “groups” on SNS is helping users to learn about biodiversity from others in a peer-to-peer manner. Most such communities include a mixture of amateur naturalists and scientists. This partnership helps naturalists obtain identifications for their photographs, and also acquire more information about the organisms they are photographing, including taxonomy, identification characters, biotic associations, biogeography, etc. On the other hand, scientists get to see a large number of interesting records of biodiversity from sites that they are not able to visit, and generally benefit from a much larger community of observers who may detect and document important records (Aravind 2013, Deng et al. 2012).

As an example of this potential, the possibility of harvesting primary biodiversity data from SNS was demonstrated in a previous contribution using a popular photo sharing website, Flickr (Barve 2014). This resource was shown to be a rich store of photo-documented, geo-tagged GPS tagged records, but assessing its full potential requires further exploration of its potential in terms of number of records, and coverage in terms of taxa, geography and seasonality. We thus present here a case study of SNS-derived records of birds of the world, developed in comparison with digital records available via the Global Biodiversity Information Facility (GBIF) portal.

METHODS

Data Collection

To explore the contents of Flickr for all bird species of the world, we started with the checklist of birds of the world compiled and maintained by the International Ornithologists' Union, formerly called the International Ornithological Committee (IOC; Gill and Donsker 2013). The checklist that we used was IOC version 3.5, which includes 10,650 species.

A script was developed in R (R Development Core Team 2014), to query and download data for all the bird species via the Flickr site (see Barve 2014 for details). The R script exploits the Application Programming Interfaces (APIs) provided by Flickr to download data. APIs are 'hooks' provided by the website to third-party developers to build applications that can improve website usage in innovative ways. Since the number of queries is substantial (> 10,000), a second script was developed to serialize data downloads for all species. The data were aggregated and stored in a SQLite database (Hipp 2014) for further analysis. At times, the data downloading script would get interrupted owing to data transfer losses between the client machine and Flickr website; this problem required a manual restart to continue the download process.

As described in Barve (2014), to get complete data for each species, it was important to query Flickr for common names and any synonyms of scientific names of each species. The IOC checklist provides a set of common names for all bird species, so the same script and query process was repeated for common names, and data downloaded. To distinguish between records downloaded via the scientific name and common name

queries, a tag was added to each record to indicate the query type by which it was downloaded.

For synonyms, another script was developed in R, which utilized the APIs of the Encyclopedia of Life (EOL) website (Parr et al. 2014). EOL provides automated access to nomenclatural databases like Catalogue of Life, the GBIF backbone taxonomy, Integrated Taxonomic Information System, etc. (Anonymous 2015). The “synonyms” function in the R package *taxize* (Chamberlain et al. 2012) was used for this task. The script fetched and stored all canonical synonyms for each accepted scientific name in the IOC list. This list of synonyms was again used to download all data from Flickr corresponding to these names, and “synonym” tag was appended to these records.

These three datasets (i.e. records derived based on scientific names, synonyms, and common names) were merged to produce a complete Flickr dataset. This dataset included numerous duplicate records, because the same Flickr photo may often be tagged with both scientific and common names, and sometimes even synonyms, and would be returned identically in all these iterations. These duplicate records were eliminated to avoid inflating the number of records.

For comparison, a dataset comprising all bird records was obtained from the GBIF data portal, consisting of $\sim 190 \times 10^6$ records. Data fields included GBIF identifier, scientific name, higher taxonomy fields (kingdom, class, order, and family), date of record, latitude, longitude, and taxon identifier as provided by GBIF. This data set was received as a ~ 32 GB text file, and was analysed using functions in R and SQLite.

To compare the GBIF and Flickr datasets, a first step was to separate GBIF-mediated records that did not match with names on the IOC list so that the two datasets would apply to same set of species names. The number of records remaining after this elimination step was ~181M. This data loss could have been avoided by investing efforts into resolving synonyms in GBIF-mediated data; we decided against this effort, since the effort would have added less than 5% additional records to the already large GBIF dataset.

Data Preparation

We explored the data to see how well the Flickr-derived data coverage compares to that of GBIF-derived data in spatial, temporal, and taxonomic dimensions. For spatial coverage, we generated simple gridded heatmaps. Since our analysis was at a global extents, we used a spatial resolution of 1° for this map. Each geocoded record in the Flickr dataset was assigned a grid cell identifier according to its geographic coordinates using the `getcellid` function from the R package `bdvis` (Barve and Otegui 2014). The `mapgrid` function of `bdvis` was used to create a record-level map, as well as maps of species richness. We compared the two datasets using maps that summarize representation in one, none or both of the data sets.

To explore temporal coverage for completeness and biases in the dataset, we generated polar plots of numbers of records based on the date of observation or the date on which the photograph was taken. Dates needed to be in standard formats, for

which we used the function `fixstr` in the R package `bdvis`. The `tempolar` function in `bdvis` was used to generate polar plots with daily, weekly, or monthly aggregation using the `timescale` parameter. Further temporal exploration was via `chronohorograms`, a multidimensional version of polar plots, in which yearly information is added as concentric circles (Otegui and Ariño 2012).

In terms of representation of species, to make a fair comparison of the two datasets, we generated subsets of GBIF data of the same size as the Flickr geocoded records dataset (~1.3M records). From the total pool of ~180M GBIF records, 138 mutually exclusive subsets were formed at random and stored. These subsets were then compared with the Flickr dataset, but now with sample size controlled.

Initially, we compared the 139 datasets for uniqueness in terms of species present. For this comparison, a list of unique species was developed for each subset. These lists were then compared with all the remaining GBIF subsets and the Flickr set to establish how many species were unique to each of the subsets.

A similar analysis was conducted with the rarest 25% species of the overall dataset (GBIF + Flickr). For this step, all occurrences were counted for all species, and the rarest 25% of species were stored as a rare species list. Each GBIF subset was then matched to see how many rare species that subset contained, and numbers were stored in a table. The same steps were performed for the Flickr dataset to find out how many rare species it included.

The geographic coverage of each subset was tested with a similar methodology. The same 1° grid was used to find out how many grid cells were covered by each of the 139 GBIF subsets

and the Flickr data set. Then the grid cells covered by each subset that were not covered by any of the remaining subsets were tabulated. The process was repeated for poorly represented 25% of grid cells. Numbers of these least represented grid cells were tabulated.

Results

Figure 1 shows the geographic coverage of the GBIF-derived and Flickr datasets. Please note that the colour scale is not same for these two maps. Both datasets have good coverage in regions like North America, Europe, and Australia. Direct comparisons of GBIF and Flickr data records in a spatial sense are shown in Figure 2. The areas in blue, particularly in Brazil, parts of Africa, and the Indian Subcontinent, highlight the potential of Flickr data to improve the spatial coverage of GBIF data: each blue pixel is a grid cell from where no GBIF-mediated records are present, but Flickr records are present, and this result is aggregated across all species. More than 1024 such Flickr-only pixels are present, as compared with 11,309 GBIF-only pixels; it should be borne in mind that this comparison is rather unbalanced, as the GBIF data are more than 139 fold more numerous than the Flickr data.

In terms of seasonality of the data, GBIF and Flickr records are compared in Figure 3. The GBIF-mediated data show clearly a positive bias in March-May, perhaps owing to a tendency by observers to record migratory birds, or to increased bird activity in temperate regions in these months. Although the Flickr data do show a bias towards the first half of the year, they are more balanced in terms of seasonality. The chronohorogram (Figure 4), however, shows the GBIF data as fairly balanced, with steady increase in numbers of records per year. Since Flickr has a more recent origin than GBIF-mediated data sources, few historical data are available, but numbers of records increase among Flickr data over time as well.

In the sample-size controlled analysis of species representation, 401 species were represented

only in Flickr and not in any GBIF subset. Among the GBIF subsets (Figure 5), however, >60 subsets had no species unique to them, and >50 had just one unique species. The maximum number of unique species in any GBIF subset was three (Table 1). Looking at coverage of rare 25% species, the GBIF subsets had 266-369 species (Figure 6), Flickr had 1336 species covered among the rare 25% of the species.

Distribution of unique geographic grid cells covered by each GBIF subset (Figure 7), had a bell-shaped distribution with slightly longer tail. The number of unique cells was between 7 and 26 in the GBIF subsets, whereas Flickr had 1024. Looking at the rare 25% of grid cells, the GBIF subsets showed bell-shaped distribution with values between 33 and 68 as compared with 1273 such cells in Flickr (Figure 8).

Discussion

In overview, we found that Flickr data have considerable potential to augment and improve existing biodiversity data, when compared to GBIF, the largest aggregator of such data. Our results also indicate that social networking sites such as Flickr may be more effective in biodiversity data collection than traditional approaches (i.e. dataset-level, institutional-based sharing) in developing countries, where efforts towards sharing and enabling biodiversity data have been less effective as compared to developed countries. This point is illustrated in Figure 2, where Flickr-only data are seen in Brazil, parts of Africa, and on the Indian subcontinent. In our analysis of rare species of birds, we found that Flickr had better representation of rare species data when compared with GBIF, probably as photos of such species were picked out particularly for sharing.

This analysis is very much scale-dependent, but we believe that the 1° resolution that we used is appropriate for representation of the whole world. If we move to a very fine resolution, digital knowledge represented in GBIF and Flickr will appear to be very sparse. We used 1° resolution to be able to analyse digital knowledge at the level of the whole world, which yields more than 64,000 grid cells. This resolution could be refined to 0.1° cells (about 10 km x 10 km) to perform similar analyses which might give interesting results on regional scales.

GBIF-mediated data include museum specimens as well as observations from sources like eBird and iNaturalist, which are citizen science efforts curated by experts; hence, the quality of the GBIF data is expected to be better than of Flickr. Flickr data comes mostly from amateur naturalist and hobbyists, and are not curated systematically except via peer comments. One caution we have in mind while analysing these data is that the quality of geo-tagging of the photographs and identifications may be dubious. This point needs much deeper exploration and possibly crowdsourcing to correct, since the dataset we have obtained is perhaps too large to be handled by a single team. Involvement of experts to verify identifications and assess the overall quality of records is being attempted in a separate study.

More generally though, Flickr data have the potential to augment and improve the digital accessible knowledge of birds of the world. Flickr data improves on the temporal coverage of the GBIF dataset (Figure 3), being more balanced in terms of seasonality. Flickr data also offers some information about the rarest species and least-represented places on Earth, that GBIF-enabled data do not. A mechanism needs to be put in place for periodic extraction and addition to the global store of biodiversity data.

Acknowledgments

We thank Laura Russell for help in procuring GBIF data. VB was supported by the GBIF-Young Researcher Award 2014 for this study.

References

- Anonymous. 2015. EOL API: Provider Hierarchies. *Encyclopedia of Life*. Available at http://www.eol.org/api/docs/provider_hierarchies.
- Aravind, N. A. 2013. Potential of social network and internet media for biodiversity mapping and conservation. *Current Science* 105:291–293.
- Barve, V. 2014. Discovering and developing primary biodiversity data from social networking sites: A novel approach. *Ecological Informatics* 24:194–199.
- Barve, V., and J. Otegui. 2014. *bdvis: Biodiversity data visualizations*. Available at <https://github.com/vijaybarve/bdvis>.
- Chamberlain, S., E. Szoecs, and C. Boettiger. 2012. *taxize: Taxonomic search and phylogeny retrieval*. Available at <http://cran.r-project.org/package=taxize>.
- Deng, D., G. Mai, T. Chuang, R. Lemmens, and K. Shao. 2014. Social web meets sensor web: From user-generated content to linked crowdsourced observation data. Pages 1–10 *in* *Linked Data on the Web (LDOW2014)* (C. Bizer, T. Heath, S. Auer and T. Berners-Lee, Eds.).
- Deng, D., G. Mai, C. Hsu, T. Chuang, T. Lin, H. Lin, K. Shao, R. Lemmens, and M. Kraak. 2012. Using social media for collaborative species identification and occurrence: Issues, methods, and tools. Pages 22–29 *in* *GEOCROWD '12 Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information* (M. Goodchild, D. Pfoser and D. Sui, Eds.). ACM, New York.
- Gill, F. B., and D. B. Donsker. 2013. IOC world bird list (v 3.5). Available at http://www.worldbirdnames.org/DOI-3/master_ioc_list_v3.5.xls.
- Hipp, D. R. 2014. SQLite database. Available at <https://www.sqlite.org/>.
- Otegui, J., and A. H. Ariño. 2012. BIDDSAT: Visualizing the content of biodiversity data publishers in the Global Biodiversity Information Facility network. *Bioinformatics (Oxford, England)* 28:2207–8.
- Parr, C. S., N. Wilson, P. Leary, K. Schulz, K. Lans, L. Walley, J. Hammock, A. Goddard, J. Rice, M. Studer, et al. 2014. The Encyclopedia of Life v2: Providing global access to knowledge about life on earth. *Biodiversity Data Journal* 2:e1079. Pensoft Publishers.

R Development Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at <http://www.r-project.org/>.

Stafford, R., A. G. Hart, L. Collins, C. L. Kirkhope, R. L. Williams, S. G. Rees, J. R. Lloyd, and A. E. Goodenough. 2010. Eu-social science: The role of Internet social networks in the collection of bee biodiversity data. PLoS ONE 5:e14381.

Table 1.

	GBIF min	GBIF Max	Flickr
Unique species	0	3	411
Rare 25% of species	266	369	1336
Unique cells	7	26	1024
Rare 25% of cells	33	68	1273

Figure 1. Heat maps of numbers of records of birds in GBIF and Flickr.

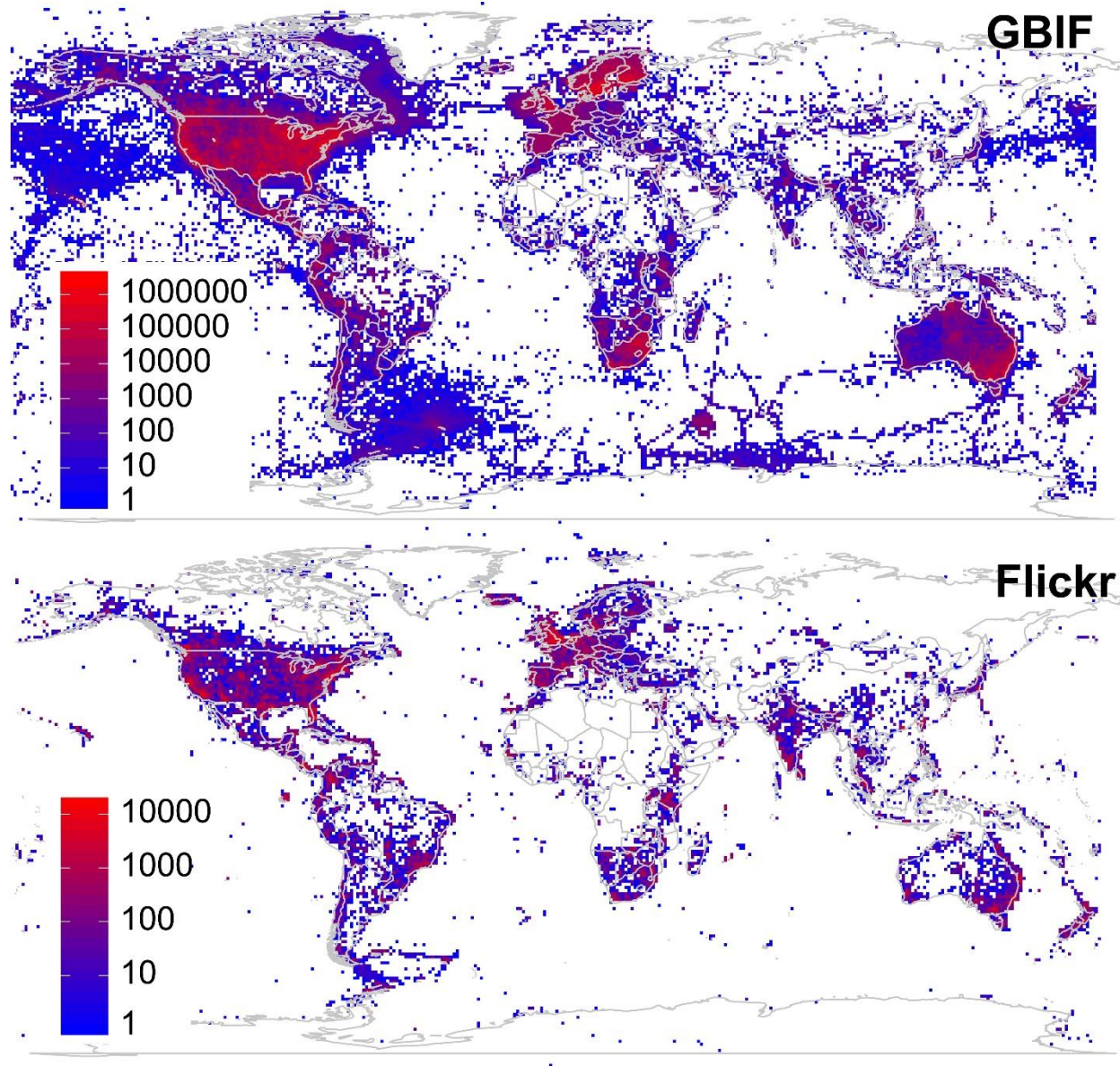


Figure 2 Comparison of GBIF and Flickr records in terms of spatial coverage.

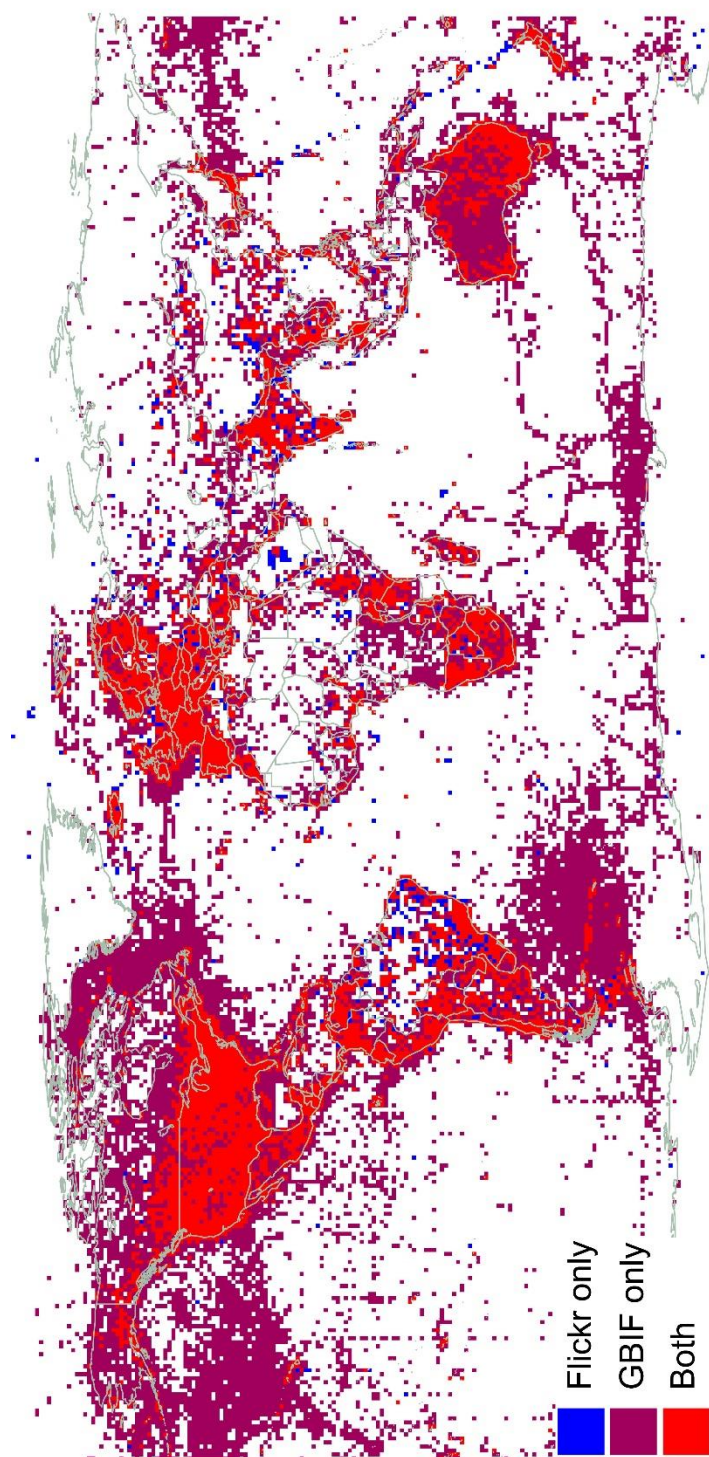


Figure 3. Comparison of seasonal coverage of GBIF and Flickr data.

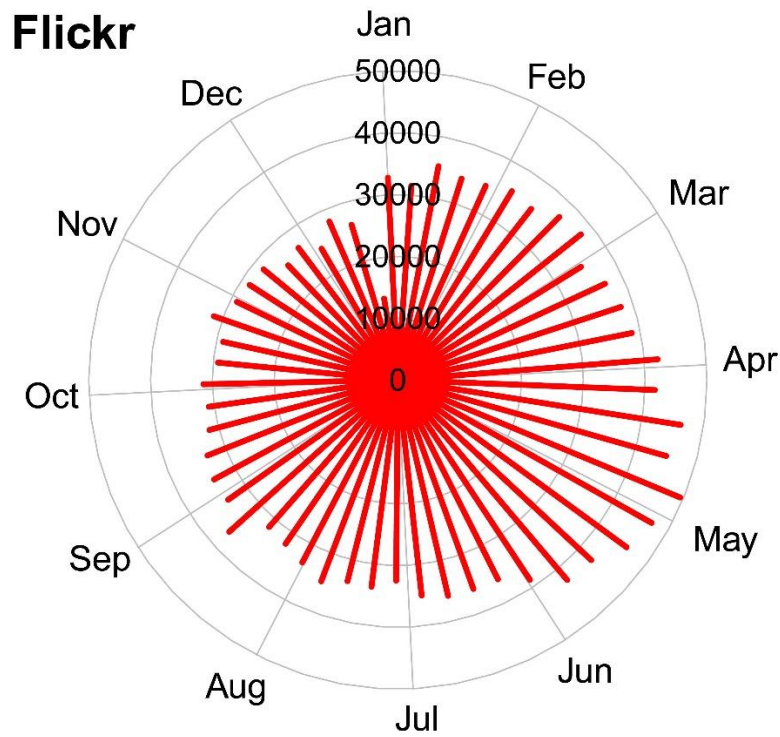
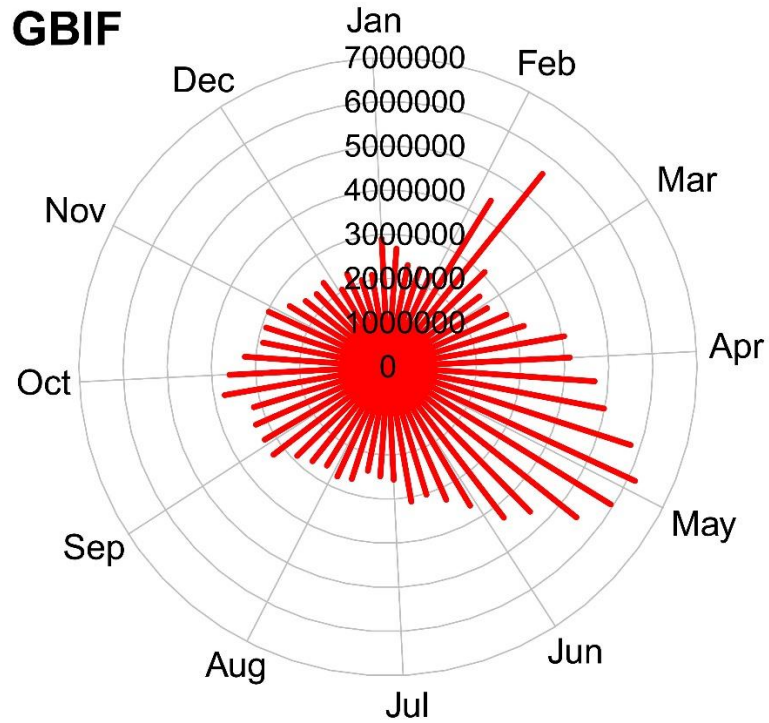


Figure 4. Chronohorogram showing seasonal and temporal coverage of GBIF and Flickr data over the period 1980-2014.

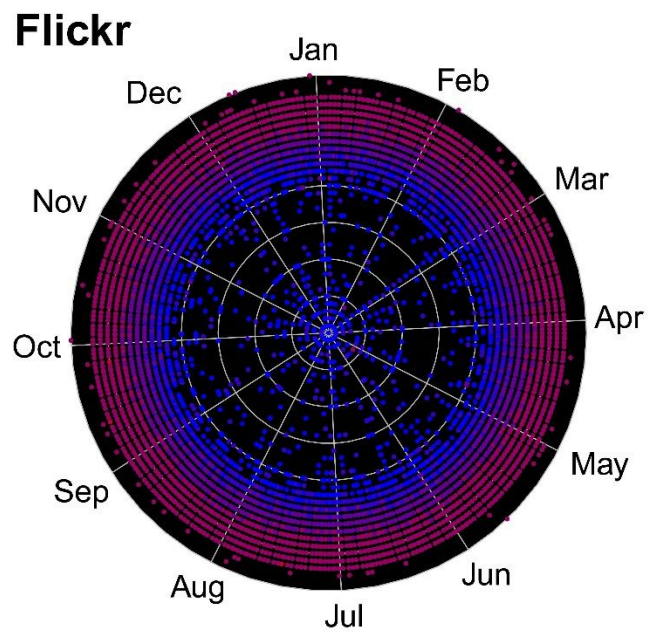
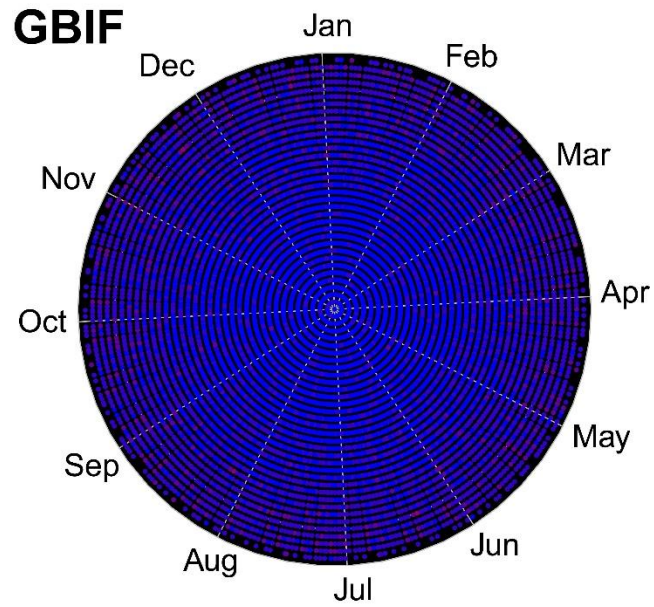


Figure 5. Numbers of species unique and rare in each of 139 GBIF data subsets. Flickr has 411 unique species (not shown).

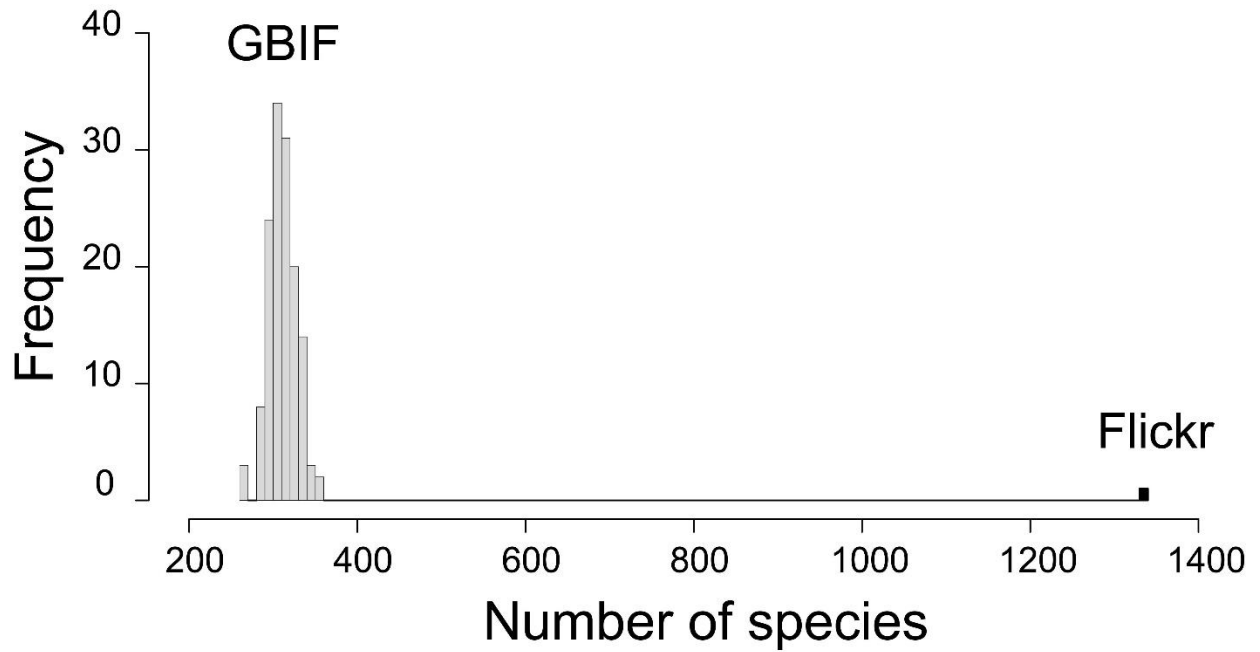
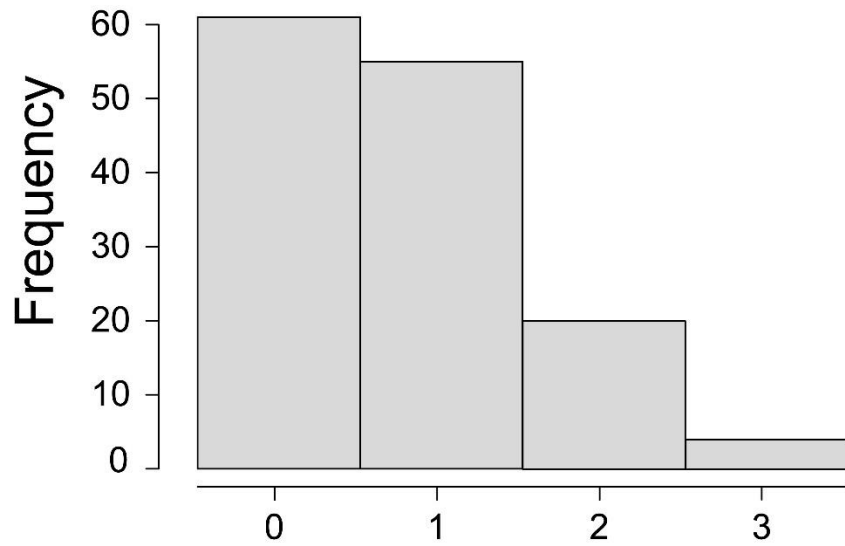
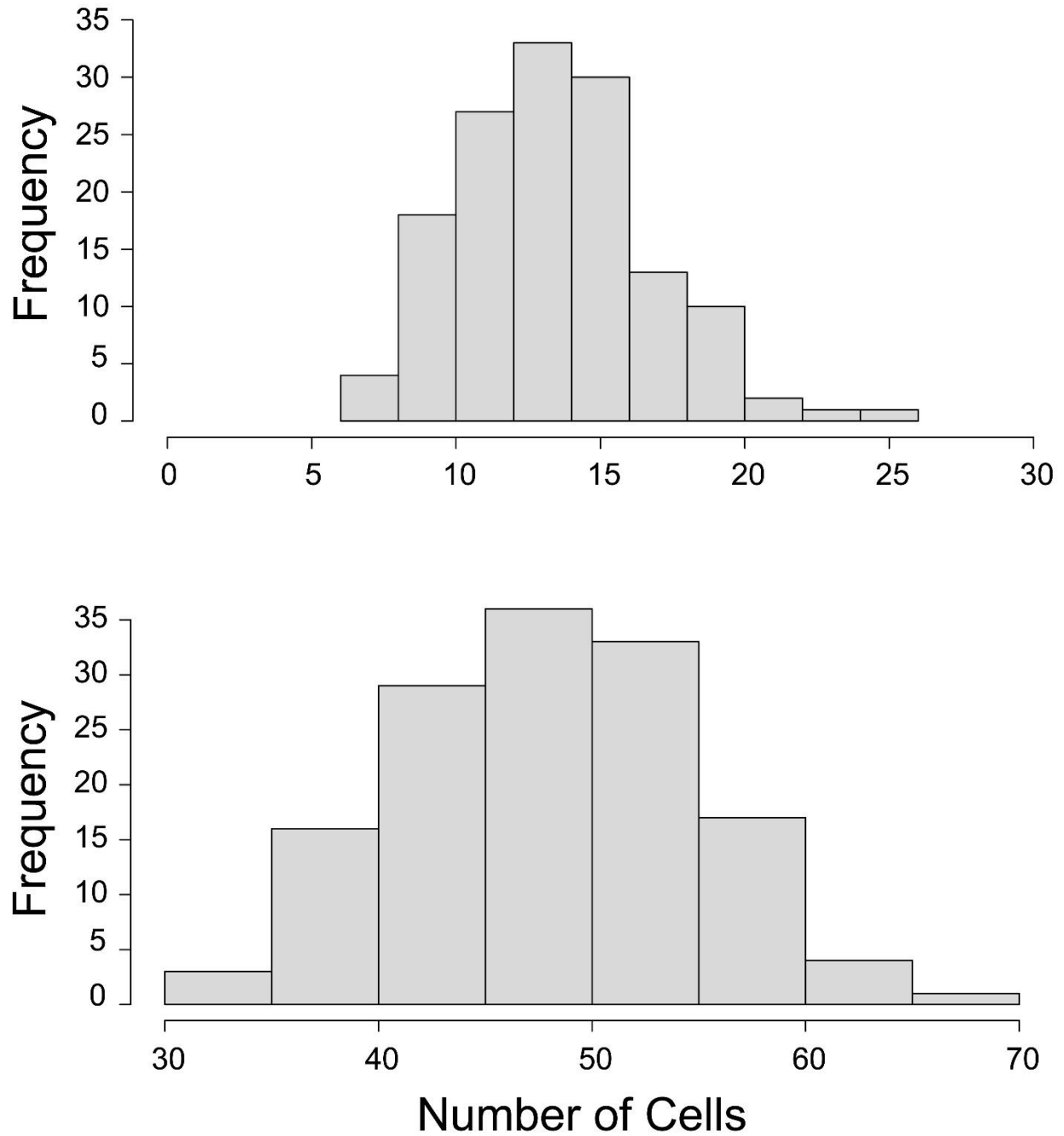


Figure 6. Number of unique and rare cells in each of the 139 GBIF data subsets. Flickr data include 1024 unique and 1273 rare cells (not shown).



Chapter 4

Flickr biodiversity data quality:
A case study with Swallowtail butterflies
from India

Flickr biodiversity data quality: A case study with Swallowtail butterflies
from India

Vijay Barve*, J. Christopher Brown

Department of Geography, 1475 Jayhawk Blvd, 213 Lindley Hall, University of Kansas,
Lawrence, KS, 66045 USA.

Abstract

An ever-increasing need exists for fine-scale biodiversity occurrence records for a variety of research applications in biodiversity and science. Even though large aggregators like GBIF serve such data in large quantities, major gaps and biases exist. To address these gaps, here we explore social networking sites (SNS) as a rich potential source of additional biodiversity occurrence records. We present a case study to investigate the quantity of records existing, and the quality of identifications of Indian swallowtail butterflies by users on Flickr. We explored these data by developing a website and presenting photographs with associated metadata to select ButterflyIndia Yahoo Group members, to check the accuracy of the identifications provided on Flickr. Results were encouraging; with >93% correct identities for records of this family of butterflies from across India.

Introduction

Biodiversity occurrence records have been accumulated for centuries now, in recent years in digital format and served by aggregators such as the Global Biodiversity Information Facility (GBIF) in large quantities. GBIF serves $>5.2 \times 10^8$ occurrence records, but major gaps and biases still exist in these data (Meyer et al. 2015) and completeness of the data for sites and local communities is a concern (Sousa-Baena et al. 2014). GBIF offers reasonable representation of vertebrates and plants, but data records for insects comprise <8% of the total set of records, even though insects have many more species than any other major taxon.

Interest is increasing in augmenting available species occurrence data by discovering it from social networking sites (SNS) (Morris et al. 2013). The advantage of acquiring such data from SNS is that they are vouchered with the photograph, and even peer reviewed in some sense by members of the SNS by way of appreciation and comments. The method of harvesting

occurrence data is demonstrated in Barve (2014), and its potential to augment and complement GBIF data in Barve & Peterson (in prep.)

The swallowtail butterflies of the family Papilionidae are generally large and colourful butterflies, which are simple to identify for amateurs in most cases. Even though they are mainly tropical, they are found almost all over the world except Antarctica (Reed and Sperling 2006).

Worldwide, about 550 species are found, and in India about 85 species have been recorded, some of which are rare or known from single records only. This family was chosen for this study in light of their attractive nature, which appeals to citizen scientists. India was selected based on availability of an organized community of butterfly lovers, combined with unavailability of substantial data served via the GBIF data portal for this family of insects.

Methods

The list of Indian swallowtail butterflies was created by referring to the website Butterflies of India (Kunte et al. 2015). All the butterflies from the family were listed with scientific and common names, and saved in a comma-separated values file. Higher classification within the family was obtained from the Encyclopaedia of Life (Parr et al. 2014) website through its Application Programmers Interface (API) in R (The R Development Core Team 2012) using a classification function in the package *taxize* (Chamberlain et al. 2012).

Flickr data records were harvested using the methodology described in Barve (2014) for all species on the list. Records were tagged based on whether they had scientific names, common names, or both. Initially, data were stored in a SQLite database using the R package *sqldf* (Grothendieck 2012). For this study, data records with restrictive photo-sharing (e.g, friends and family only) were separated, since such photos may not be displayed on other websites. All data downloaded were not necessarily from India since not all the photographs were geotagged.

A web-based module was developed, as illustrated in Figure 1, to display the photographs to be verified with the metadata provided on the Flickr website. The user is presented with a menu of taxa from which to choose, and is requested to select the lowest-level taxon for the photograph that can be verified with confidence. A comment can also be left with the verification. An option to skip the record exists if the user wants to come back to the record with more information or reference material. Metadata like the title and description, user-defined tags, date taken, and location are also displayed. This metadata is useful at times to provide additional information that might aid in verifying identity. A web-link to visit the original photo page on the Flickr site is also provided. The user may visit the page to refer to more details, inspect related photographs, and check comments from other Flickr users, before verifying the identity. For early life stages of butterflies, this function was found particularly useful, because more photos of the same insect are frequently available. The common name of the species was displayed to aid citizen scientists in understanding the species under consideration. A link to detailed species accounts on the Butterflies of India website is provided as initial reference, in case the user wants to compare the image in question to images on the website for verification of certain characters.

Users were provided with login credentials and responses were stored in the database with information on user name and time of data entry. Along with the response, the unique image identification number, the search text used for the image, and comments if any were all stored. Each image was presented to every user only once. To collect a maximum of data each time a user logged in, the image with the least number of responses were presented, so that responses for all images were collected in a balanced manner.

For the web module, the scripts were developed in PHP, and the backend database was stored in a MySQL database. The module is hosted at <http://diversityindia.org/snsrec/> and can be

accessed only by invited members. Metadata for a total of 1882 photographs were uploaded on the website for this study. Data uploaded were for 43 different species with number of photographs ranging from 1 for rare species to about 1200 for the most common species. 1106 of the 1882 photographs were geotagged, leaving 776 not geotagged. Records containing scientific names as a search string totalled 1646, as compared to only 236 found with common names.

Butterfly enthusiasts from India were contacted through personal emails, and were invited to pilot test the module and to provide identity verification data for the photographs. A good mix of scientists and seasoned citizen scientists were invited to participate, based on perceived reputation of the members in the ButterflyIndia Yahoo Group (Anonymous 2015).

Users were requested to confirm identities to the lowest level of taxonomy with which they were comfortable. If they were sure that the species listed in the search and the photograph matched, they were to select the species; if the identity was not correct at the species level, then the user was to select the genus, and so on.

The data were downloaded from the server using the RMySQL package (Ooms et al. 2015) in R. Responses were classified according to data availability classes. Based on a search using scientific names versus common names and geotagged versus non-geotagged, the records were classified in four classes. Statistics on accuracy of identification were tabulated in each of the four classes. A treemap (Tennekes 2014) was created using numbers of records for each species and percentage error in identifications.

Results

At the time of this analysis, 3706 responses had been collected on the website within less than a month, from a total of 11 members. The overall accuracy of the identifications on Flickr for this family was found to be more than 93%. Comparing identification accuracies for geotagged images versus non-geotagged photographs accuracies were 92.9% and 93.8%, respectively (Figure 2), which was not significant in a contingency table test ($P=0.0255$).

However, comparing between photographs retrieved based on scientific names versus common names, the difference was almost 11%, with accuracies of 94.7% and 83.4%, respectively (Figure 2). The contingency test revealed a significant interaction ($P \leq 0.00001$): records tagged only with common names had triple the chance of being wrongly identified. Finally, looking Figure 2(e) and 2(f), dividing the data by both geotagging and search type (scientific name versus common name), the least accuracy was with the records which are not geotagged and with only common names.

Assessing the distribution of identification errors across species (scientific names or common names, figure 3), 27 of 56 entities had no misidentifications. Of course most of the species without any misidentifications have very limited records, with an exception of a couple of species (Figure 4).

Discussion

The website for expert review of Flickr data that was developed for this project proved to be very useful, collecting large amounts of data (more than 3000 responses) over a short time (less than three weeks). Personal rapport with members of ButterflyIndia Yahoo Group was crucial to success of this data collection, which accumulated >3000 records in less than one month. This website has considerable potential for use with groups of organisms, and needs further

investment for improvement towards more flexibility in terms of user management, setting up projects for groups of organisms and facility to provide correct identifications rather than just marking them wrong.

Early stages of butterflies were included among the Flickr photographs. At times, identifying eggs or larvae with certainty based on a single photograph is difficult. This task often required more investigation by the experts on the Flickr website, using the “Visit original Image page” link provided in the web interface. In many cases, the Flickr user has more images of previous and later stages of the insect, and the identity could be confirmed with this further information.

Lower accuracy levels for records which were tagged with a common name only probably indicates that those records were posted by more casual users, who are either not competent enough to identify the butterflies or not connected to competent peers who could offer them help in correcting identifications. At times, these photographs are posted by naturalists interested in other taxa who chanced upon an interesting butterfly, and photographed and posted the image out of simple curiosity.

Accuracy assessment according “search term” highlighted the fact that if a record had only a common name specified, it had a greater probability of wrong identification. This idea can be explored further with the case of a group of three species, the Common Mormon (*Papilio polytes*), the female of which mimics two species, the Crimson Rose (*Pachliopta aristolochiae*) and the Common Rose (*P. hector*). For all three species, accuracy levels for scientific name-tagged records were higher than 95%, but common name-tagged records were 91%, 59%, and 80% accurate, respectively. For detailed species-wise explorations, much more data in terms of

species as well as records will be required, which will allow us to detect species that are cryptic or challenging to identify.

More generally, the results of this study were encouraging with greater than 93% accuracy in identifications attached to photographs as a rich source of biodiversity occurrence data. Of course, this approach will need more exploration with more species, and results for cryptic species will clearly be much lower. Nevertheless, this study shows that approaches that partner citizen naturalists with crowdsourced expertise have considerable potential, especially for taxa that are photographed frequently.

Acknowledgements

We would like to thank Tukaram Dokhale, who contributed to development of the website. We are also grateful to Amol Patwardhan, Aravind N. A., Arjan Basu Roy, Ashok Sengupta, Milind Bhakare, Ravi Vaidyanathan, Nithin R, Subramanyam Kalluri, Sarika Baidya, Toshita Barve, and Tarun Karmakar, for participating in the study by providing confirmations of identifications on the website. We are thankful to Narayani Barve for providing illuminating ideas on the development of the website. VB received support from a GBIF-Young Researcher Award 2014.

References

- Anonymous. 2015. ButterflyIndia Yahoo group. *DiversityIndia*. Available at <https://groups.yahoo.com/neo/groups/ButterflyIndia>.
- Barve, V. 2014. Discovering and developing primary biodiversity data from social networking sites: A novel approach. *Ecological Informatics* 24:194–199.
- Chamberlain, S., E. Szoeacs, and C. Boettiger. 2012. taxize: Taxonomic search and phylogeny retrieval. Available at <http://cran.r-project.org/package=taxize>.
- Grothendieck, G. 2012. sqldf: Perform SQL selects on R data frames. Available at <http://cran.r-project.org/package=sqldf>.

- Kunte, K., P. Roy, S. Kalesh, and U. Kodandaramaiah. 2015. Butterflies of India, v. 2.10. *Indian Foundation for Butterflies*. Available at <http://www.ifoundbutterflies.org/>.
- Meye, C., H. Kreft, R. Guralnick, and W. Jetz. 2015. Global priorities for an effective information basis of biodiversity distributions. *PeerJ PrePrints*:3:e1057 Available at <https://dx.doi.org/10.7287/peerj.preprints.856v1>.
- Morris, R. A., V. Barve, M. Carausu, V. Chavan, J. Cuadra, C. Freeland, G. Hagedorn, P. Leary, D. Mozzherin, A. Olson, 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.
- Ooms, J., D. James, S. DebRoy, H. Wickham, and J. Horner. 2015. RMySQL: Database interface and MySQL driver for R. Available at <http://cran.r-project.org/package=RMySQL>.
- Parr, C. S., N. Wilson, P. Leary, K. Schulz, K. Lans, L. Walley, J. Hammock, A. Goddard, J. Rice, M. Studer, et al. 2014. The Encyclopedia of Life v2: Providing global access to knowledge about life on earth. *Biodiversity Data Journal* 2:e1079. Pensoft Publishers.
- Reed, R. D., and F. A. H. Sperling. 2006. Papilionidae. The swallowtail butterflies. Version 07 July 2006. *The Tree of Life Web Project*, <http://tolweb.org/>. Available at <http://tolweb.org/Papilionidae/12177/2006.07.07>.
- Sousa-Baena, M. S., L. C. Garcia, and A. T. Peterson. 2014. Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions* 20:369–381.
- Tennekes, M. 2014. treemap: Treemap visualization. Available at <http://cran.r-project.org/web/packages/treemap/index.html>.
- The R Development Core Team. 2012. R: A language and environment for statistical computing. Available at <http://www.r-project.org/>.

Figure 1. Web interface developed to capture expert opinions on butterfly identifications.

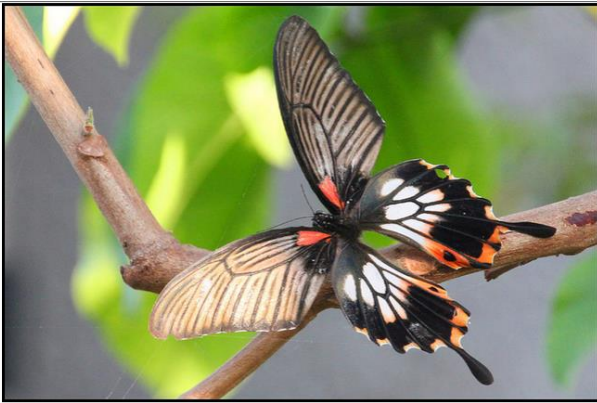
Biodiversity SNS

diversityindia.org/snsrec/fnlhd.php?image_id=6968787738

Search

Verification of Biodiversity Data from Social Networking Sites

vijaybarve : [Log Out](#) Last Refreshed on 2015-04-13 13:00:06. **Papilio polytes**



[Visit original Image Page](#)

Title Common Mormon (*Papilio polytes romulus*) at the Royal Flora butterfly exhibition, Chiang Mai

Description

Date Taken 07-01-2012 Owner Name 26822473@N06

Latitude 18.730639 Longitude 98.933944

Tags butterfly commonmormon papiliopolytesromulus totallythailand

Previous Comments:

Scientific Classification

[Common Mormon](#)

0 Kingdom : Metazoa

0 Phylum : Arthropoda

0 Class : Insecta

0 Order : Lepidoptera

0 Family : Papilionidae

0 Genus : Papilio

0 Species : Papilio polytes

Comment:

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

For feedback / comments email: vijay@diversityindia.org

Figure 2: Percent accuracy of identifications across different classes of images. Sci = Scientific name, Com = Common name, and +Geo and -Geo indicate with and without geotagging, respectively.

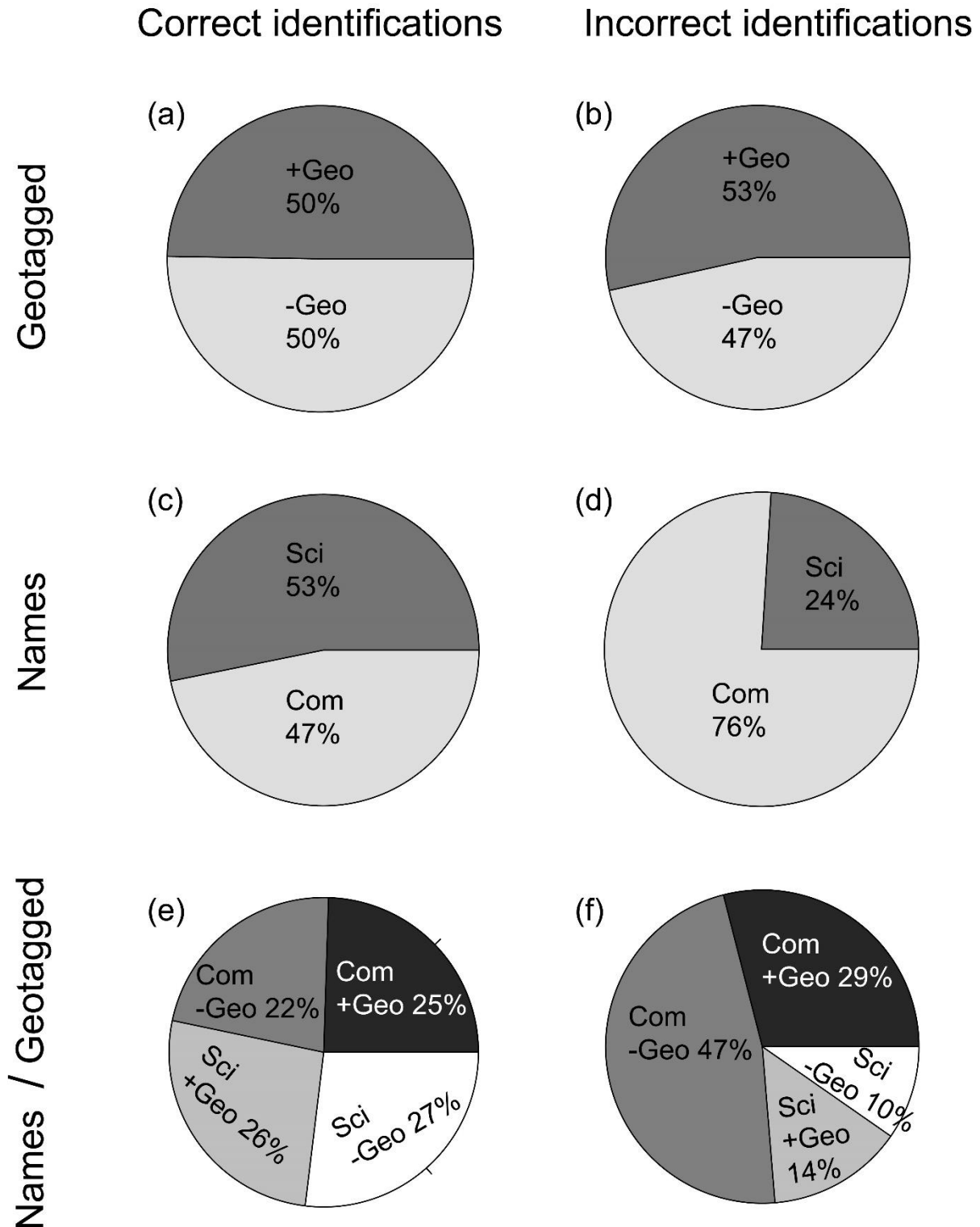


Figure 3. Accuracy assessment of each search entity. Grey colour indicated correct identification and black colour indicates incorrect identification.

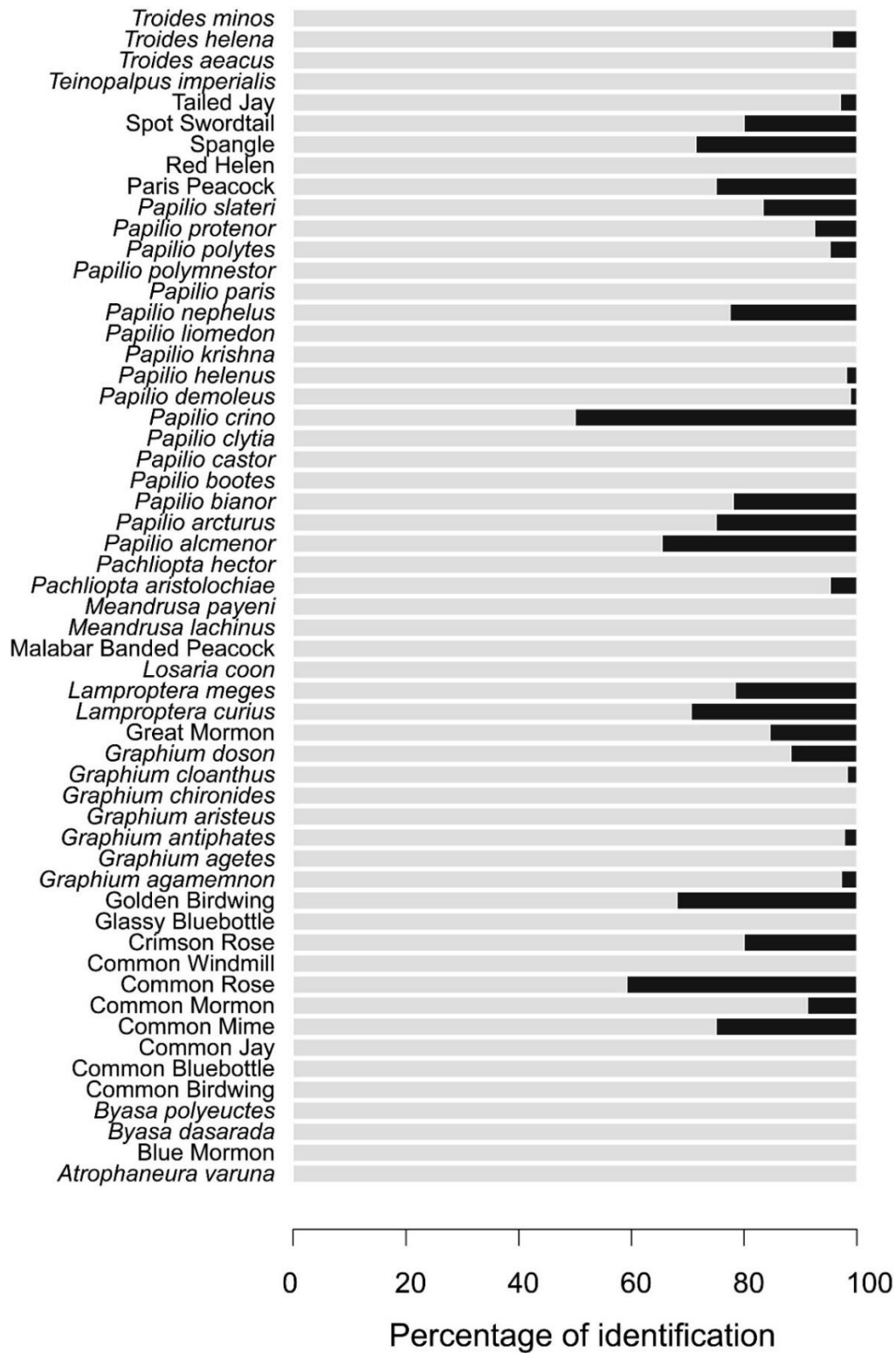
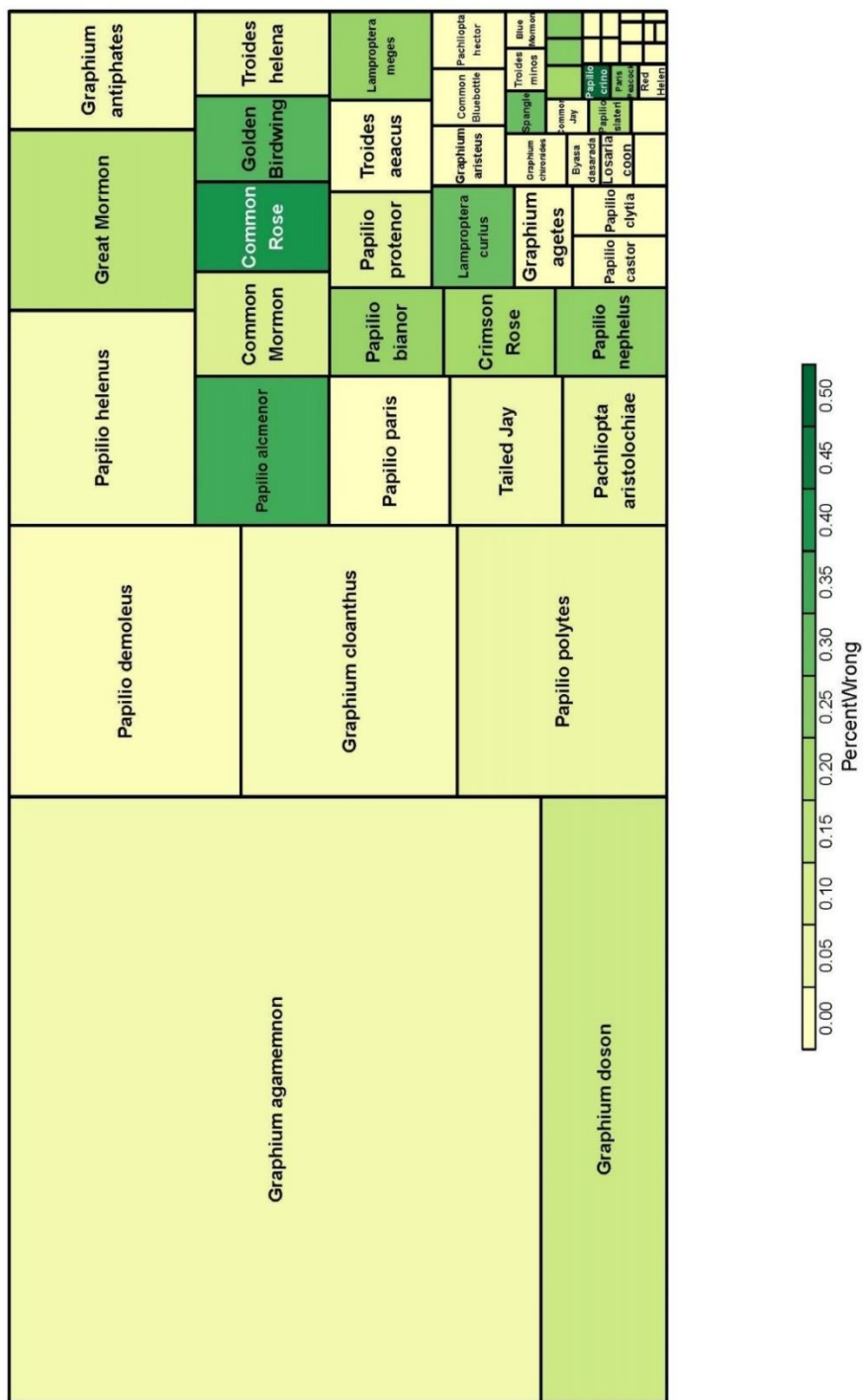


Figure 4. Treemap of search entities and percent error in identification. Size of the box indicates number of records and colour indicates percent error.



Conclusion

The first chapter demonstrated that the Flickr website has a good amount of biodiversity data worthy of exploration and highlighted not only the richness of data available, but also the need to use common names in queries on Flickr and other Social Networking Sites (SNS) to complement scientific names. The large numbers of records for Snowy Owls under misspellings of scientific names and synonyms reflect recent taxonomic changes, highlighting the need to use these alternative terms in queries. About 16% of the butterfly records and 6% of the bird records downloaded in queries had associated geo-tags, making it quick and easy to map the records.

In chapter three, the case study of birds of the world, we found that Flickr data have considerable potential to augment and improve existing biodiversity data, when compared to GBIF, the largest aggregator of such data. Our results also indicate that social networking sites such as Flickr may be more effective in biodiversity data collection than traditional approaches (i.e. dataset-level, institutional-based sharing) in developing countries, where efforts towards sharing and enabling biodiversity data have been less effective as compared to developed countries.

We also found that Flickr had better representation of rare species data when compared with GBIF, probably because photos of such species were picked out, particularly for sharing. Better representation within the degree grid cells in terms of uniqueness as well as poorly represented cells illustrates the potential to cover DAK poor regions.

As GBIF-mediated data include museum specimens as well as observations from sources like eBird and iNaturalist, the quality of GBIF data is expected to be better than that of Flickr. Flickr data come mostly from amateur naturalists and hobbyists, and are not curated systematically except via peer comments. One caution we have in mind while analyzing these data is that the quality of geo-tagging of the photographs and identifications may be dubious. This point needs much deeper exploration and possibly crowdsourcing to correct, since the dataset we have obtained is perhaps too large to be handled by a single team. Involvement of experts to verify identifications and assess the overall quality of records is being attempted in a separate study.

More generally though, Flickr data have the potential to augment and improve the digital accessible knowledge of birds of the world. Flickr data improves on the temporal coverage of the GBIF dataset, being more balanced in terms of seasonality. Flickr data also offers some information about the rarest species and least-represented places on Earth, something that GBIF-enabled data do not provide. A mechanism needs to be put in place for periodic extraction and addition to the global store of biodiversity data.

In the last chapter, the website for expert review of Flickr data that was developed for this project proved to be very useful, collecting large amounts of data over a short time. This website has considerable potential for use with groups of organisms, but further investment would be required to make this website flexible in terms of user

management, setting up projects for groups of organisms and facility to provide correct identifications rather than just marking them wrong.

Lower accuracy levels for records which were tagged with only common names likely indicate that those records were posted by more casual users, who are either not competent enough to identify the butterflies or not connected to competent peers who could offer them help in correcting identifications. At times, these photographs are posted by naturalists interested in other taxa who chanced upon an interesting butterfly, and photographed and posted the image out of simple curiosity.

More generally, the results of this study were encouraging with greater than 93% accuracy in identifications attached to photographs as a rich source of biodiversity occurrence data. Of course, this approach will need more exploration with more species, and results for cryptic species will clearly be much lower. Nevertheless, this study shows that approaches that partner citizen naturalists with crowdsourced expertise have considerable potential, especially for taxa that are photographed frequently.