Do you believe your (social media) data? A personal story on location data biases, errors, and plausibility as well as their visualization

Tobias Isenberg, Zujany Salazar, Rafael Blanco, Catherine Plaisant

Do you believe your (social media) data? A personal story on location data biases, errors, and plausibility as well as their visualization

Tobias Isenberg, Zujany Salazar, Rafael Blanco, and Catherine Plaisant

Abstract—We present a case study on a journey about a personal data collection of carnivorous plant species habitats, and the resulting scientific exploration of location data biases, data errors, location hiding, and data plausibility. While initially driven by personal interest, our work led to the analysis and development of various means for visualizing threats to insight from geo-tagged social media data. In the course of this endeavor we analyzed local and global geographic distributions and their inaccuracies. We also contribute Motion Plausibility Profiles—a new means for visualizing how believable a specific contributor's location data is or if it was likely manipulated. We then compared our own *repurposed* social media dataset with data from a dedicated citizen science project. Compared to biases and errors in the literature on traditional citizen science data, with our visualizations we could also identify some new types or show new aspects for known ones. Moreover, we demonstrate several types of errors and biases for *repurposed* social media data.

Index Terms—Social media data; Flickr; Panoramio; iNaturalist; data bias; data error; data plausibility; data obfuscation; citizen science.

1 Introduction

Social media is one of today's main forms for sharing thoughts and messages among people in developed countries. Its development has not only dramatically changed the way we communicate and share information, it also has led to huge datasets that may be useful for scientific research. Many citizen science projects already benefit from such data sources, usually in the form of dedicated projects for data collection by the public [63]. Scientific data, however, can also be extracted from social media sites even though posters did not have a specific scientific question in mind when posting (e. g., [4], [18], [19], [39]).

The geo-tagged and time-stamped pictures that many people share on social media today provide the opportunity to study the location of depicted items such as plants, in particular the distribution of endangered species. While this status unfortunately applies to many species today, personal interest led us to study specifically the distribution of carnivorous plants. Past studies have quantified the conservation threats of different genera and species [35], providing information to prioritize certain areas of conservation. Yet accurate geospatial distributions of the different species is difficult to obtain. Traditionally, researchers use environment data and current habitat sightings, to then apply species distribution modeling (SDM). SDM uses statistical models to improve the data quality of distribution estimates. The use of geo-tagged social media images has the potential of improving or, at least, contributing to this process and can also allow scientists to track specific populations over time.

A fundamental question is, however, to what degree the geo-tagged data from social media is plausible. What are the inherent data biases, identifiable data errors, and potential data hiding in such data collections? Based on a personal data collection we identify and visualize the biases and errors relevant to carnivorous plants, and present our exploration as a case study. We report our work using a structure that is rather unusual in today's scientific literature, one that may be more in line of early forms of scientific writing akin to those, e. g., by Darwin [20] and his contemporaries: We follow our actual path of exploration, which was not initially driven by a scientific question but by personal interest and the resulting dataset (Sec. 3). From this resource came successes and failures of locating habitats, which in turn led us to analyze the biases in our data (Sec. 4) and attempt to verify entries (Sec. 5). The specific verification attempt of one of a small set of entries then led us to create a novel visual representation we call "Motion Plausibility Profiles" (Sec. 6). These allowed us to analyze the data from specific individual contributors. After developing this representation, our team expanded and we became aware of the data collected by the citizen science project *iNaturalist*, which we then compared to our own and also analyzed with our motion plausibility profiles (Sec. 7). Based on this process we contribute:

- a detailed analysis of a geographic data sample extracted from image sharing sites; we describe many sources of bias and error that affect the global and local distribution of observations as extracted from social media posts, which can guide future social media data extraction for citizen science,
- a demonstration that the data contains intentionally introduced errors, likely due to people trying to protect the species' habitats,
- Motion Plausibility Profiles as a new, integrated time+space visual representation to analyze whether location data from specific contributors is plausible, useful for both individual data collations and citizen science projects such as *iNaturalist*, as it allows users to identify intentionally introduced errors,
- a comparison of two datasets on the same topic—one whose contributors intentionally contributed habitat data to citizen science, the other based on "repurposed" social media data, and
- a comparison of the errors and biases we found with the citizen science literature, highlighting new findings, in particular, for our repurposed social media data.

- *T. Isenberg is with Université Paris-Saclay, CNRS, Inria, LISN, France. E-mail: tobias.isenberg@inria.fr.*
- *Zujany Salazar, Rafael Blanco are with Télécom SudParis, France. E-mail: {zujany | rblancog25}@gmail.com.*
- *Catherine Plaisant is with the University of Maryland, USA, and with Inria, France. E-mail: plaisant@cs.umd.edu.*
- *People with color impairments may consider our alternative paper version.*

Manuscript received Apr. 2, 2021; revised Nov. 29, 2021; accepted Dec. 20, 2021. Author version. DOI: 10.1109/TVCG.2022.3141605

2 Related Work

The analysis of social media data and, in particular, geo-located posts has been a focus of visualization research in the past. Crandall et al. [19], for instance, studied people's visits of popular places by means of the pictures they took there. Andrienko et al. [4] and Kisilevich et al. [39] also used photos shared on social media sites to support the spatio-temporal analysis of people's points of interest, behavioral patterns, and events. Later, Chen et al. [18] and Krueger et al. [41] focused on tracking people's movements over time, while Wang et al. [72] extracted salient regions of people's interest. Anthony [5] recently developed Fireant, a tool for the collection and analysis of social media data. In contrast to all these approaches (and many more similar examples [17], [74]), we do not treat a geo-tagged picture as evidence for a person being at a location at a given time, but rather as evidence for the depicted subject matter—a plant in our case—being present at the location, thus as evidence for the plant's habitat. While our data also contains time stamps, we examine primarily the pictures' locations, and analyze potential geographic data biases, errors, and obfuscation.

We are also inspired by past discussions on the quality of social media data. For instance, Preece [63] and Kosmala et al. [40] noted that citizen science data in general—with dedicated data collection for a given topic—is subject to various data quality threats, yet they are not unlike those in professional data collection [40]. The situation is different for social media data as in our case where people did not intentionally contribute to the creation of a dataset but where the data was extracted based on a given set of criteria. Olteanu [54], e. g., demonstrated the inherent bias of social media discussions on current events compared to reports in the news. Later, Morstatter and Liu [51] studied Twitter data and showed that bias arose due to the API limiting access to 1% of the data. Olteanu et al. [55] defined several bias types: general data bias, population biases, behavioral biases, content production biases, linking biases, temporal biases, and redundancy. While we also show some of these bias types in our work, we provide a more in-depth description of the geographic data quality, in contrast to previous work's focus on the topic or contents of posts.

The issue of data quality is not new [24], [25], [29], [30], [37]. Several authors (e. g., [15], [16], [48], [65]) mentioned, in particular, the quality of geographic location data on observations in citizen science projects. Zizka et al. [76], e. g., cite reasons for incorrect geo-locations such as automated geo-referencing based on vague descriptions, switching latitude and longitude, data entry errors, records from cultivation (e. g., botanical gardens), rasterized locations, and rounding. To address such problems, some citizen science projects use volunteer training and testing as well as standardized and calibrated equipment [40], [66]. Such methods are not applicable, however, to geo-locations extracted from image sharing sites. Yet, data errors can still automatically be analyzed. Zizka et al.'s CoordinateCleaner [76], e. g., automatically filters out observations for botanical gardens and in urban centers, locations in the oceans, obvious coordinate entry errors (zeroed coordinates as also shown by Fisher [27]), and more. In our initial data acquisition we used some of these techniques, but we extend the analysis to within- and between-contributor data comparison as well as to our motion plausibility profiles, which allow us to effectively analyze the data quality of active contributors. Later in Sec. 8, after having discussed the data issues we found, we pick up this topic again, categorize our classes of errors and biases, and compare them to those discussed in the literature.

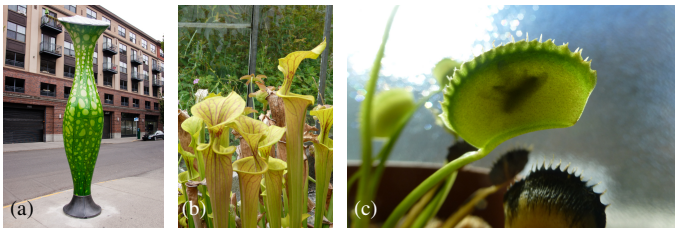


Fig. 1. Examples for images that we manually excluded from the search results because they do not show plausible plant habitats: (a) fake/artificial plants (Flickr image 9479368060 by Craig Moe; CC BY-NC 2.0), (b) botanical gardens or similar (Flickr image 5086658928 by Jane Nearing; CC BY-NC 2.0), and (c) plants kept at home (Flickr image 10753181585 by Mike Linksvayer; CC0 1.0).

Fig. 2. Screenshot of our initial data exploration interface, with markers colored by genus and image thumbnails shown at the bottom.

Our work also relates to the topic of uncertainty visualization. While numerous approaches have been described and surveyed in this context [10], [31], [36], [57], [62], we are mostly interested in geographic and geospatial uncertainty visualization [38], [45], [46]. In contrast to this past work, however, we focus on the analysis of the plausibility of citizen-contributed data [26] and identifying biases and errors by means of visualization, rather than the representation of this uncertainty in a final visualization. Our work also relates to the notion of *implicit error* proposed by McCurdy et al. [50] as we discuss in our conclusion.

3 Initial Data Acquisition and Analysis

As we stated in Sec. 1, we did not begin our work with a scientific question in mind but with the interest in the data itself, to plan visits to carnivorous plant habitat locations during future travel. Our work thus started with a data collection [56]. Originally unaware that specialized citizen science projects (like iNaturalist, see Sec. 7 below) may include data on carnivorous plants, our initial analysis (by the first three authors) focused on general online geolocation collections. POI (points of interest) sites such as `waymarking.com` contain categories about the subject, but largely focus on botanical gardens or similar places. One of the main reasons for this lack of habitat data is the protection status of many of the species, many of which are endangered (e. g., see the IUCN Red List [33]).

We thus began by collecting a social media dataset on carnivorous plants by searching Panoramio[1] and Flickr for geo-tagged images whose label or description included at least one from a series of search terms.[2] These search terms included Latin genus and family names, the term "carnivorous plant(s)," as well as common plant names for different species in Chinese, Danish, Dutch, English, French, German, Italian, Japanese, Norwegian, Polish, Portuguese, Russian, Spanish, and Swedish. This resulted in more than 49,000 candidate images, most of which did not

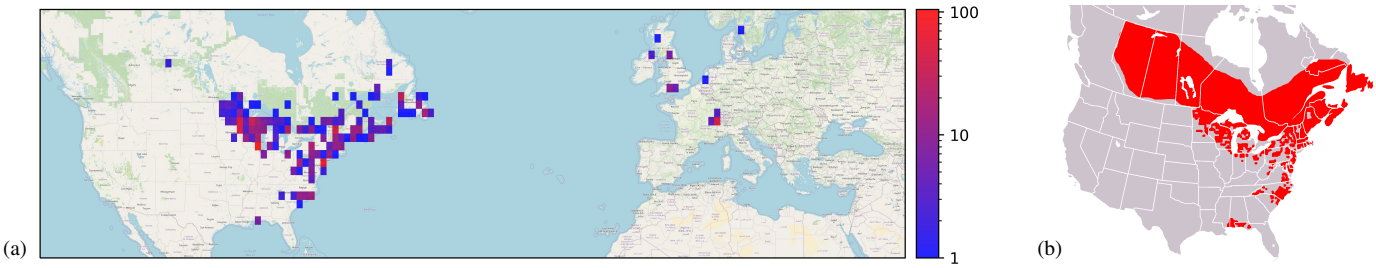


Fig. 3. Comparison of distribution maps for *Sarracenia purpurea*: (a) based on our social media data, (b) map from Wikipedia (public domain).

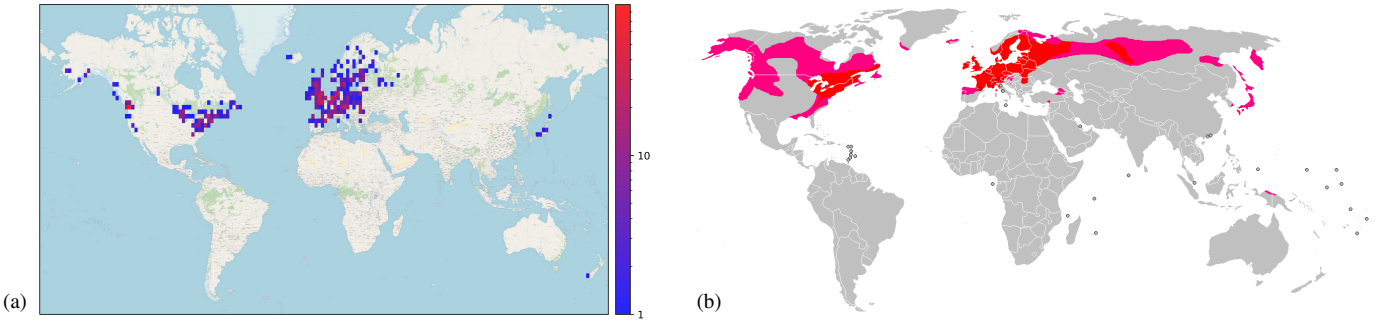


Fig. 4. Comparison of distribution maps for *Drosera rotundifolia*: (a) based on our social media data, (b) map from Wikipedia (public domain).

actually show plants or habitats. To remove false positives (e. g., Fig. 1) we used methods that we later-on found to frequently be used in citizen science projects [73]: For each of the found images, we manually looked at both the image itself and its location on the map using an online satellite map, to determine whether the images showed a true habitat and whether the location data was believable. For example, we removed image locations in the middle of urban areas. Some people posted images of plants kept at home or used locations that were not plausible, so we removed certain user IDs and locations to speed-up the inspection process. The remaining plausible images amounted to a list of more than 9,700 observations (images with location). For each plant location, we recorded its scientific name,[3] its geographic position on the map as reported by the service (extracted from the photo itself or reported by the contributor), its elevation (based on an elevation look-up for the geographic position), the time the picture was taken (again extracted from the photo or reported by the poster), the social media service and the respective image ID, text description of the geographic location such as area name and country, the name and ID of the person who uploaded the images, and the image itself.

With this data we designed an initial map-based data exploration tool (using Google's Maps API, Flask for Python, HTML5, CSS3, and JavaScript) that indicates the location of each observation in the dataset [9]. With the tool we could explore the recorded locations (e. g., Fig. 2) and filter for plant characteristics (genus, species) and/or image data (location, date it was taken, social network). Based on the data we also visualized the geographic distribution of different species. For example, we can see that our data supports the distribution of *Sarracenia purpurea* in North America (Fig. 3)—with the exception of the less populated areas of central and northern Canada. Yet we can also see populations of the species in Europe—including well-known places where the plant was introduced in Switzerland [60] or places in the UK, Scandinavia, and the Netherlands [1]. Data like ours thus has the potential to close gaps in biodiversity databases [3]. In another example, we find the distribution of *Drosera rotundifolia* in our data also largely on Wikipedia's distribution map (Fig. 4), yet again it does not support the full established distribution.

[1]The Panoramio service has since been retired by Google.

[2]None of the authors has contributed to this dataset through image posts.

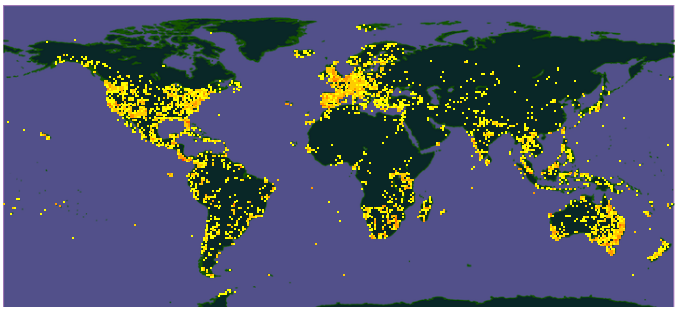
[3]We generally used the Latin names provided in the descriptions. If no species name was provided or if it was wrong, we either classified the plant ourselves if we knew the species well, or we only recorded the plant's genus.

4 Sources of Data Biases and Errors

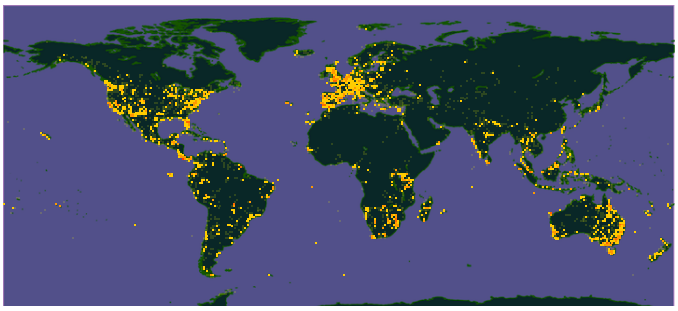
While our initial data collection (approx. up to 2019, continued later) was driven by personal interest (and served its initial purpose), we were immediately intrigued by the data artifacts we saw in the data. The distribution differences of *Drosera rotundifolia*, e. g., are likely due to a strong bias in our data. We thus decided to further investigate this data bias and other data errors more systematically, which then became the focus of our work. We specifically targeted the spatial data as extracted from the social media posts as we were mainly interested in the geographic location and distribution of specific species, genuses, and the plant group in general. We first analyze sources for global and local location bias as well as sources of non-geographic contributor bias that can affect the spatial data distribution. We then discuss and extract geographic location errors, before mentioning other relevant error sources.

4.1 Global location of contributors

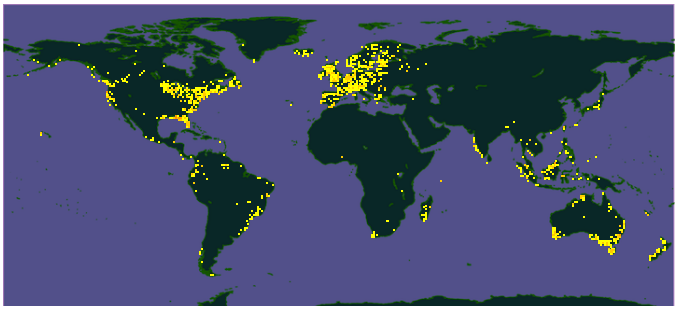
We saw that our data is subject to population biases: posters are not representative of the population at large. Most come from a few countries and include only people with access to digital cameras or camera phones as well to online social media. In our data we also saw that, in addition, contributors are also not evenly distributed over the globe. They concentrate in urban centers, and favor traveling to popular places. To illustrate this global bias we looked for a related but more general dataset. We thus used posts in Flickr's *Encyclopedia of Life* (EOL) group, which focuses on pictures of any form of life. We used Page's `flickreolmap` tool [58] to create a global heatmap of all geo-tagged image posts of this group (Fig. 5(a)). Fig. 5(c) provides the same heatmap generated for our dataset. Because the number of entries for Flickr's EOL group is much larger than our data, in Fig. 5(b) we show Fig. 5(a) again with the smallest class (< 10 images per spatial unit) transparent, to allow us to better focus on the overall distribution.



(a) Heatmap of all 172,083 geo-located EOL group posts, up to March 2020.



(b) EOL group posts, with the smallest class (<10 images; yellow in (a)) shown transparently to better emphasize the locations with multiple image posts.



(c) Heatmap for the 9,720 entries of our own dataset (also up to March 2020).

Fig. 5. Geographic distribution analysis based on posts in the *Encyclopedia of Life* (EOL) Flickr group which collects images and videos of organisms: Due to the similar subject matter (for all species in general) we use the maps as an indication of Flickr's geographic bias for habitat pictures [59]. Generated with Page's [58] tool, colors indicate image count in a given region: yellow: 1–9; light orange: 10–99; medium orange: 100–999; dark orange: 1,000–9,999; red: ≥ 10,000. Regions where the background map is visible do not have a single image. Note that the used Mercator projection has limitations, in particular, close to the poles.

The heatmaps show that our data on carnivorous plants, on a global level, roughly matches the distribution of all Flickr contributors interested in habitat pictures (i. e., entries in the EOL group, Fig. 5(a)), suggesing that they are subject to the same global bias. For example, few if any images exists for large stretches of Asia (e. g., Siberia), Africa, Southern America, inner parts of Australia, and northern parts of North America. This fact then also explains the lack of evidence of the distribution of *Drosera rotundifolia* in, e. g., Siberia that we saw in Fig. 4, as well as the lack of sites of *Sarracenia purpurea* in central Canada in Fig. 3.

Naturally, the maps in Fig. 5 cannot show the population bias of the user base of Flickr in general. If we assume that, among Flickr posters, those interested in images of living species are evenly distributed, then a comparison of Fig. 5 with a map of the world's population density (Fig. 6) can indicate the population bias in Flickr posters overall. We see, for example, that the population centers in India and China are heavily under-represented, while Europe and North America are heavily over-represented.

The first author of this article also personally visited several

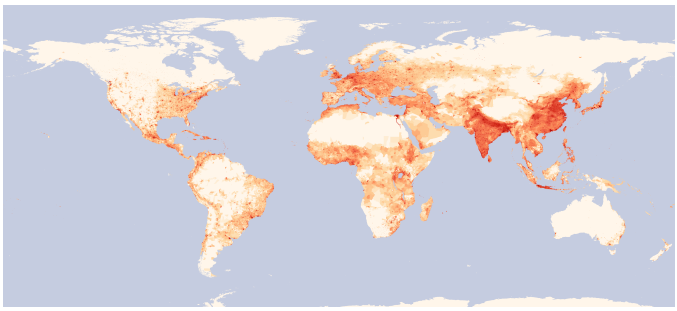


Fig. 6. World population density in 2000. Image by Robert Simmon, NASA Earth Observatory, based on data by the Socioeconomic Data and Applications Center (SEDAC), Columbia University (© public domain).

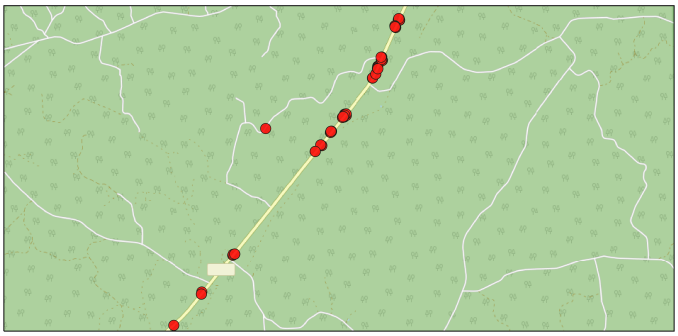


Fig. 7. Illustration of a local geographic bias due to ease of access: many sites are recorded near roads/paths, yet virtually none far from them.

habitat locations in Europe and North America and saw, for example, *Drosera* and *Pinguicula* species that, despite these locations generally being well visited by hikers, were not represented even by a single entry in our dataset. This observation illustrates another global aspect: Our dataset is relatively small (less than 10,000 observations for the whole planet). Even for regions with relatively frequent observations such as in Europe and parts of North America not all existing habitats are represented.

4.2 Local observation focus points

In addition to these global geographic biases, however, our dataset also exhibits several local geographic biases. In particular, people tend to only take pictures at locations that they can easily access such as near roads or on hiking paths as illustrated in Fig. 7. Moreover, some locations are well known and popular such as natural parks and similar places. Such sites are places that many people visit during vacations or trips. They thus have a high motivation to post images from such visits compared to, e. g., their everyday environment. Fig. 8 illustrates this bias for two sites: one well-known State Natural Site in the US where one can observe *Darlingtonia californica* (Fig. 8(a)) and a national park in Sarawak, Malaysia, where one can find several *Nepenthes* species (Fig. 8(b)).

4.3 Uneven observation counts

In addition to these geographic biases our dataset also has important biases based on the people that contribute the observations. In particular, the number of observations is heavily skewed as shown in Fig. 9(a). For example, only 3.9% of people (i. e., 88 out of 2233 people) reported 50% of the images. Many people only report very few images, i. e., 56.1% people only post a single image, 72.0% post ≤ 2, 80.1% post ≤ 3, and 88.0% post ≤ 5. We fitted a discrete power law distribution to the observation counts per person using Python's `powerlaw` package [2], which yielded a fit with parameter

Fig. 8. Local geographic bias due to the certain spots such as natural parks being more popular for visitors than others: (a) Darlingtonia State Natural Site in the United States and (b) Bako National Park in Malaysia. In both cases observations are clustered at marked locations in the parks.

$\alpha = 2.07$ and standard error $\sigma = 0.043$, for $x_{min} = 3$. The power law distribution is also evident in the almost linear log-transformed post count over rank plot in Fig. 9(a).

This power law distribution of people's contributions further amplifies the discussed geographic biases: The data locations in the dataset are heavily biased by where the few major contributors live or where they travel to. Moreover, it demonstrates the different interest of people: Most of them do not care particularly about carnivorous plants and thus only contribute "by accident," while a few have a core interest and contribute a lot of observations. In fact, the non-linear "bulge" for the contributors of rank 4 and 5 in Fig. 9(a) may be an artifact of the special interest: these 5 people are likely enthusiasts and thus contributed more images than what one would expect from a plain power law distribution.

4.4 Erroneous data locations

So far we only discussed biases that arise from population and sample distribution. However, we also need to consider error sources, in particular with respect to the reported geographic locations. Errors in our case may arise, e. g., from the reporting of cultivated plants (i. e., false positives for habitats). We manually filtered these out in the data acquisition process and are confident that only few, if any, of such cases remain in the data. Other errors, however, cannot easily be filtered out as we discuss next.

One main source of location error arises due to the inaccuracy of position tracking. GPS accuracy highly depends on the situation, location, as well as used equipment and numerous ways exist to measure its accuracy [64]. We can generally assume, however, that most GPS-enabled smartphones are, on average, accurate such that measured positions lie within a 4.9 m radius of the true location, in open-sky conditions [70]. While this situation probably matches many of those in which pictures in our dataset were taken, there is no way to correct for this error as the ground truth is not known. Nonetheless, this relatively small error is typically irrelevant in our case. The location error can grow, however, in less than ideal conditions. These include sky blockage such as from trees or equipment conditions. Location errors for images taken immediately after turning on a camera or phone, e. g., before a GPS lock is fully acquired, can be a lot larger than the normal inaccuracy. Also, modern phones do not always rely on GPS for positioning but also use the cell network's signal strength. The accuracy of this form of geo-location is lower, in the order of 90 m or worse [42] and there are techniques to analyze such errors (e. g., [43], [67]).

More important in our case, however, are situations when a person manually specifies the coordinates. For example, an image's coordinates may have been rounded at some stage. Or a picture file may not contain location data (e. g., older, scanned pictures), in which case the poster on Flickr or Panoramio may have retroactively

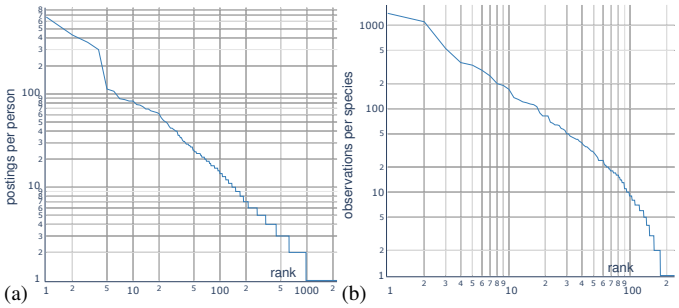


Fig. 9. Logarithmic plots of the number of (a) the contributors' postings over their rank and (b) the observations per species over the species' rank, showing their power-law-like distribution (tailed for (b)).

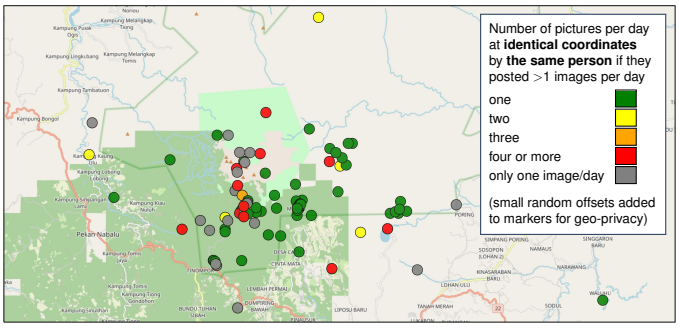


Fig. 10. Within-poster verification: Pictures taken in and close to Mount Kinabalu National Park, Malaysia, color-coded by how many pictures **by the same person** are at exactly the same geo-coordinates in a day.

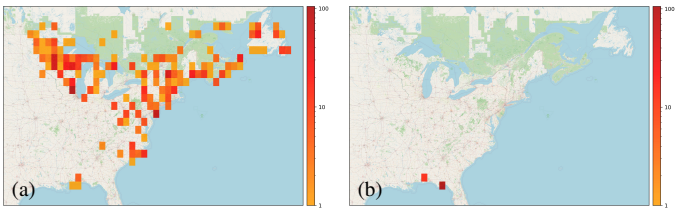


Fig. 11. Distribution of (a) *Sarracenia purpurea* and (b) *Sarracenia rosea*. The cases of *S. purpurea* in the same area of *S. rosea* are likely, in fact, *S. rosea*, which was only elevated to species status in 1999.

located the pictures when uploading them—with the best intentions. Then data entry errors are possible such as those discussed by Zizka et al. [76] and shown in Fig. 37, yet we found none of them (Fig. 38). This process, however, may lead to several pictures by the same person to be placed at exactly the same GPS coordinates as illustrated in Fig. 10. The same effect, however, can also be caused by people who are aware of and care deeply about the plants' protection status: They may actively enter false coordinates to prevent others from using the posted locations for poaching. For example, Flickr user `biodivinf` posted a habitat image of the rare *Cephalotus follicularis* and stated "coordinates changed" in the comments (https://www.flickr.com/photo.gne?id=3020563046). Another reason may also "burst shots" of many images in a second or two, but we saw none of those cases in our data.

4.5 Species popularity

The list of species, in general, is also not equally covered. Similar to posts per person (Fig. 9(a)), we also see a heavily skewed post count per species (Fig. 9(b)). A discrete power-law fit for the observation counts per species (excluding the genus-only observations) yielded the parameter $\alpha = 1.95$ with a standard error of $\sigma = 0.123$, for $x_{min} = 24$. The heavily tailed behavior likely arises from rare species being much less well known to the general public.

This popularity aspect also affected the classification. Naturally, most Flickr or Panoramio contributors are not botanists and, thus, are no nomenclature experts. In our data collection we only included images on which we could clearly identify the plants' genus, which is thus unlikely to contain errors. For the specific species, however, we relied on people's classification in the image title, description, or tags. For a small subset of species (including, e. g., *Drosera rotundifolia* and several *Sarracenia* species) that we personally knew well we sometimes corrected the classification. For images without species that we could not classify ourselves we only used the genus name (11.3% of the observations). To avoid spelling mistakes, we ensured the correctness of names using a list of known species. Nonetheless, we expect our current species classification to contain numerous errors.

In addition to "simple" incorrect classifications, incorrect species names can also arise due to the introduction of new species for plants that were previously considered part of another species. *Sarracenia purpurea* subsp. *venosa* var. *burkii*, e. g., which is native to a small area of the southern USA, was re-classified in 1999 to *Sarracenia rosea* [53]. Some people may not know about this fact and still classify these plants as *S. purpurea*. We see the effect in Fig. 11—in the *S. rosea* range we still observe many observations of *S. purpurea*, most or all of which are likely *S. rosea*.

4.6 Other sources for errors or biases

Other biases may arise from our search terms, despite using Latin and common names as well as multiple languages. Also, some posters do not label their images, so we likely missed some relevant pictures. In addition, we noticed that Flickr's API does not seem to return all posts that would match a search term: On repeated searches with exactly the same search terms we saw results within a previously covered ID range that we did not get in the first search (but this effect is small, see Fig. 28 in the additional material).

5 Means of Data Verification

In addition to the general discussion of biases and errors it is, of course, desirable to identify data points that are likely to be incorrect. Due to the nature of social media, posters who contributed to our dataset did not intend to do so and thus largely did not actively provide means to verify their data. Below we discuss ways to infer plausibility measures to indicate how much we can rely on the data. These measures are most relevant for a local data analysis, but they also provide means to verify the data globally.

5.1 Within- and between-poster data point comparison

We had already hinted at a first measure in Sec. 4.4: verification by within-poster comparison. As we show in Fig. 10, we can count the observations for precisely the same location data from the same poster, independent of the species they reported. Multiple posts at exactly the same place can indicate some form of intentional data hiding or at least manual entry, as we had discussed. So we can use this measure to assign a data quality grade to each data point.

We can, however, derive another quality measure by comparing data from different people, for the same species, genus, or any other genus. We base this second measure on the assumption that plants of the same species grow in the same habitats, plants from the same genus also like similar habitats, and even most carnivorous plants share the preference for a nutrient-poor environment. We can thus search, within a given radius, for other observations of the same species, genus, or any other plant by ***different*** people in the dataset, and color the observations accordingly. Fig. 12 shows this visualization for a site for two radii, 1 km and 4 km, showing that some of the observations are, indeed, confirmed by others.

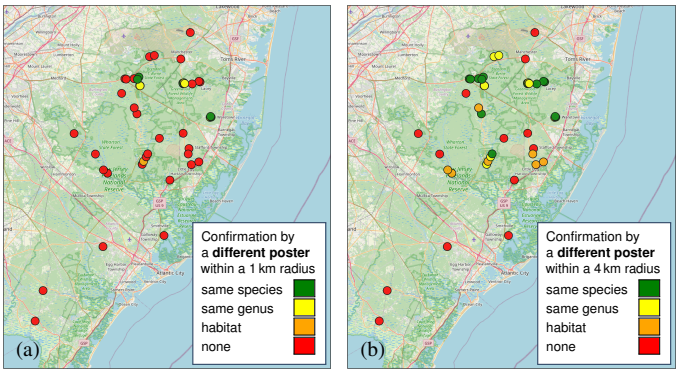


Fig. 12. Between-poster verification with radii (a) 1 km and (b) 4 km.

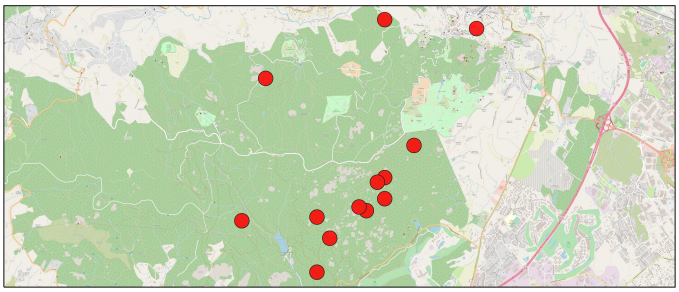


Fig. 13. Observations in our dataset of *Pinguicula lusitanica* near Lisbon, Portugal, that we visited but were unable to confirm.

Fig. 14. Observations in our dataset of *Pinguicula vallisneriifolia* in Spain. We confirmed the central two sites, yet the top site (>5 km away) appears manually moved: its Flickr picture shows the same site as the center ones. The incorrect coordinates exhibit no rounding artifacts and differ to a different degree in latitude and longitude (distance > 0.04 resp. 0.01).

Both measures are independent and can also be combined to a single grade. They are not, however, absolute values but depend on the number of analyzed observations. They may thus change as more observations are added to a data collection.

5.2 Personal inspection

In addition to such verification methods that rely exclusively on the collected data, we can naturally also compare the data with personal observations that we did prior to or after the data collection. Due to the personal interest of one of the co-authors in the subject we could verify several sites, e. g., in the Netherlands (*D. rotundifolia*, *D. intermedia*), in Canada (*D. rotundifolia*, *D. linearis*, *S. purpurea*, *Utricularia*), in the USA (*S. purpurea*, *Darl. californica*), and Spain (*P. vallisneriifolia*). We visited other sites (e. g., in the UK and Germany) based on the collected data, yet were unable to find

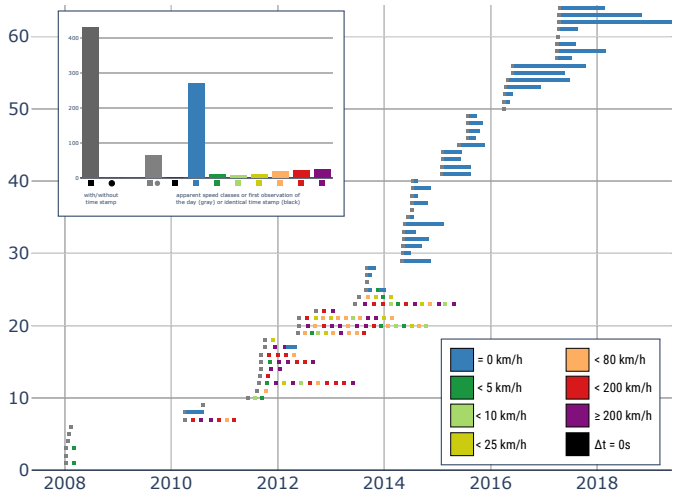


Fig. 15. Motion plausibility profile showing actively adjusted geo-locations. Each row is one day worth of images, its date indicated by the x-position of the day's **first** picture (each row's left-most gray dot). Each dot represents one image, the dots are spaced apart by the binary logarithm of the distance of the consecutive picture locations, colored by the apparent speeds from the previous location to the current one. Each daily dot sequence thus uses a different time scale than the x-axis. Top-left: histogram of entries; bottom-right: color scale of the speed classes.

any plants. Nonetheless, we still consider some of such sites as plausible due to their habitat matching that preferred by the plants.

Yet, we also visited sites such as the one depicted in Fig. 13 in which the local habitat did not match. All of the observations in this case were from the same person, so we used within-poster analysis (Sec. 5.1). Yet all of the depicted locations were single images, so within-poster analysis cannot indicate any data adjustment.

In addition, we also noticed, for some sites that we personally confirmed, that our data not only contained observations from others within a few hundred meters but also some several kilometers away. For one particular site in a national park in Spain (Fig. 14), for instance, we noticed another observation more than 5 km away whose Flickr image was depicting the exact same site as the one we saw, and four pictures were posted at exactly the same coordinates. Even more interestingly, this apparently manually adjusted observation—whose inaccuracy was much greater than typical GPS imprecision—was reported by the same person who also had posted the previously mentioned locations in Portugal: the second-most active person in our dataset. We thus asked us whether it would be possible to analyze potential data adjustment of, in particular, active people more holistically.

6 Individual Motion Plausibility Profiles

To do so, we propose a new design to visualize an individual poster's behavior to reveal possible intentional errors. For this purpose we analyze the whole sequence of a person's observations over the course of a day and for all of their observation days. We then use this tool to analyze the frequent posters in our dataset.

We begin with each image's location and the date and time ***when it was taken*** (i. e., not when it was uploaded). We can thus compute, for each image except the first in a sequence of a day's images, the ***minimum apparent*** speed at which the photographer would have had to travel, using geodesic distances between the locations. A speed of up to 5 km/h, e. g., is plausible for walking, while a speed of over 200 km/h would imply flying an airplane.

To understand the behavior of a person based on a sequence of images, we now arrange colored marks that represent the images and their speeds sequentially (see Fig. 15 where each horizontal line of marks represents a day of observation). We space the marks apart depending on their geographic distance, log_2-transformed so that we can both accommodate and see small and large distances. We assign colors based on meaningful speed intervals (no motion, walking, speed walking, bicycles, cars, sports cars without a speed limit, and airplanes) as shown in the color legend of Fig. 15. In the visualization, however, we do not only want to see short time spans such as a day but also the evolution of the posting behavior over the whole time a person contributed (typically several years). We thus anchor each day sequence with respect to the x-axis based on the day's date, and follow the sequence from that location horizontally as just described. We use the y-axis to stack several days, and each line's y-coordinate simply reveals the number in the sequence of days on which a person contributed to the dataset. Notice that in this mapping the x-axis represents two notions of time. First, on a time scale of years it shows the day on which pictures were taken—whose date is indicated by each ***first*** gray box in a row. Second, it visualizes a notion of sequence during a single day, independent of a correct timestamp: the sequence of colored boxes in a row whose date is marked by the first box in each row.

The resulting *Motion Plausibility Profile* allows us to gauge to what degree a frequent poster's geo-data is plausible. Looking at the person who we hypothesized in Sec. 5.2 had adjusted their data (Fig. 15), we can indeed confirm that they moved (or manually specified) the geographic locations of their images. In their first phase, in 2008, they recorded only one or two images per day (we mark the first image of a day in gray as we cannot compute a speed for it). Next, between 2010 and 2013, the locations of each image was changed individually (e. g., with a random offset to the real position), leading to a variety of apparent speeds—including some that would imply multiple airplane trips a day. This period also includes the sites we showed in Fig. 13 with pictures from May 2012. The EXIF data provided by Flickr reports the used camera model, which provided GPS capture only as an optional accessory so the poster may or may not have used automatic coordinate capture. Finally, starting in late 2013, the poster located all their pictures from a day at a single spot (including the site in Fig. 14), as indicated by the blue markers (i. e., 0 km/h). Still, even at that time this person continued to use the same camera model. Starting in 2016, however, they used a different camera with integrated GPS coordinate recording. Yet this new camera does not seem to have changed how they treated picture locations according to Fig. 15.

Fig. 16 shows the motion plausibility profiles of several additional active contributors ($\geq$80 images each). The most active poster in Fig. 16(a) with 669 images shows largely plausible speeds. The black markers in this profile indicate images with an identical time stamp, which is plausible for this person as they contributed to both Panoramio and Flickr and we merged the profiles from both services for those posters we could identify as identical people. The profiles in Fig. 16(c) and (f) also seem largely plausible, while Fig. 16(b) shows some adjustment and Fig. 16(d),(e) has a lot of evidence for data changes. Notice that we use circular markers for those observations that do not have a valid time stamp such as in the example in Fig. 16(c). Some motion plausibility profiles also show multiple entries with identical time stamps, such as Fig. 16(e). These may arise from uploading multiple pictures for which no time stamps were recorded, similar to several images initially without coordinates being located at a single position.

The profile also shows different contribution rhythms. The posters in Fig. 16(a),(c), e. g., contributed consistently throughout

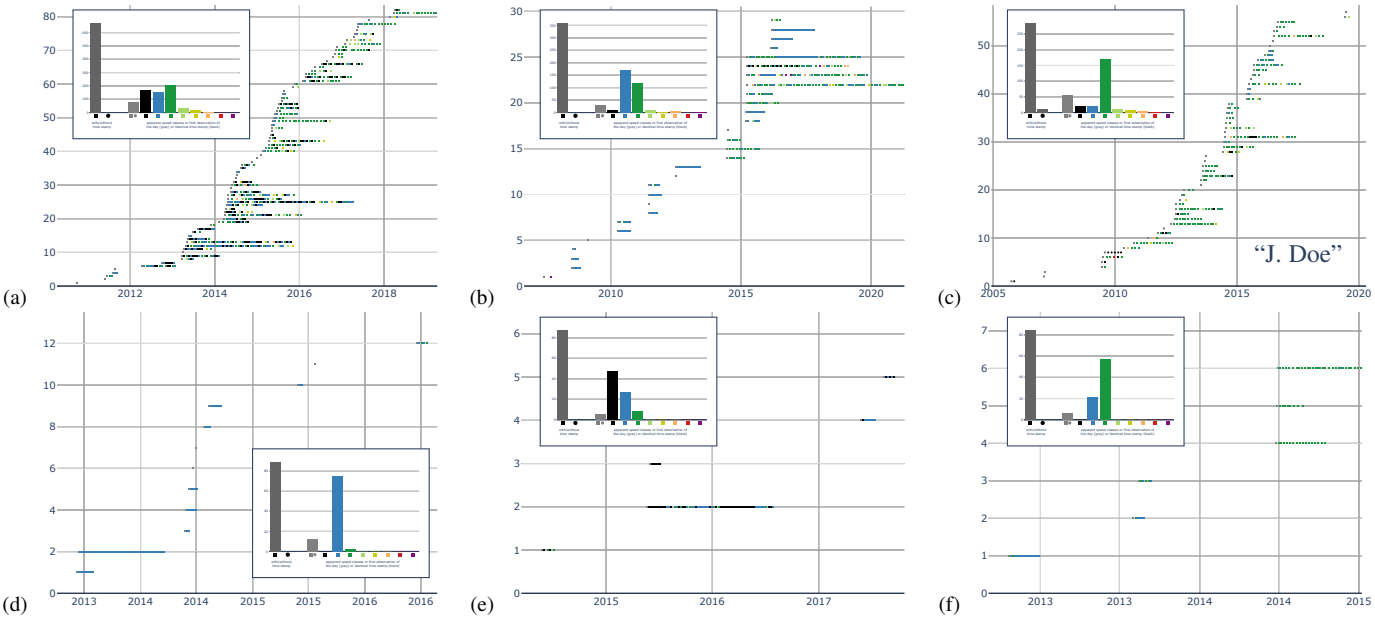


Fig. 16. Motion plausibility profiles of some of the other top contributors in the Panoramio/Flickr dataset. Same principle and color scale as in Fig. 15.

the years, while others (Fig. 15, Fig. 16(b)) mainly during quite specific times (vacations?). Still others (Fig. 16(d),(e),(f)) only contributed on a few days overall, but then many pictures per day.

A limitation of the motion plausibility profiles is that they require a certain minimum number of observations per person, i. e., they are not useful for most contributors. Yet they allow us to review the behavior of the most active posters and mark certain locations reported by them as suspicious. This process could be combined with the within-poster analysis (Sec. 5.1), which does not require visual inspection and also works for a lot fewer observations per person. Together with the positive confirmation from the between-poster analysis we could derive a single plausibility factor, parametrized to a given application case. We purposefully do not specify specific weights here as any mapping has pros and cons.

A negative ranking from any of the measures, however, does not mean that a specific data point would be useless. Even those social media posters for whom we could demonstrate data adjustments are not likely to move the locations of observations, e. g., to other countries or continents. We believe (but cannot prove) that the manually specified locations are within a few kilometers of the actual site, such that the data can still be used for a general habitat analysis as we did in Fig. 3 and 4. Evidence for this hypothesis is that, to the best of our knowledge, the locations in the dataset correspond to the natural habitats of the respective plants or to places where the plants have been introduced (as previously discussed for *S. purpurea*. The only exception we found were several locations of *Nepenthes* plants in Central America, which we excluded during data collection. As mentioned before, a main reason for data hiding for our specific type of data is to protect the plants. For example, the person examined in more detail in Fig. 15 works for a botanical garden, according to their Flickr profile.

7 Citizen Science Data: iNaturalist

After our analysis of habitat locations extracted from social media posts had progressed as we have described so far, we learned about the existence of the citizen science biodiversity site iNaturalist—of which we had previously been unaware. It allows users to post their observations of any living species including pictures and geo-locations. By design, the site thus works fundamentally different

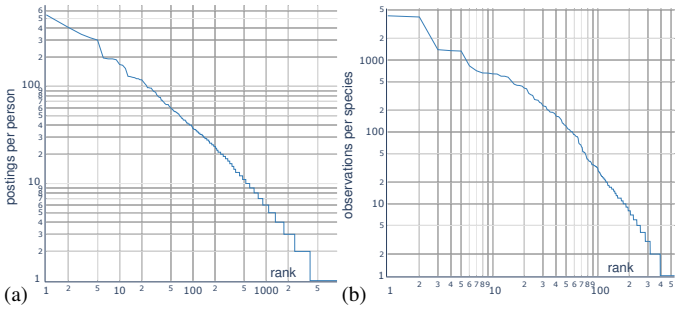


Fig. 17. Logarithmic plots, for the iNaturalist carnivorous plant data, of the number of (a) the contributors' postings over their rank and (b) the observations per species over the species' rank, showing their power-law-like distribution—comparable to the Panoramio/Flickr dataset (Fig. 9).

from the ‘scraping’ of habitat information from social media such as Panoramio or Flickr that we had analyzed so far: any contributor is aware that their data is recorded and potentially used to track down the species. Moreover, users are encouraged to confirm or correct the observed species from posts of others, leading to less classification errors. Advantageous for us, iNaturalist makes their data easily accessible to interested parties and also contains data about the subject matter of our interest. This fact thus provided us with a unique opportunity to compare the two different approaches to socially collected data, which we describe next.

7.1 Analysis of iNaturalist

iNaturalist consciously takes geo-privacy seriously and offers controls to hide the true geolocation of data points [68, Box 10.1 on pp. 355–356] (also called “fuzzing” of locations [11]). For each observation, the reporting person can choose to set a `geoprivacy` tag to either obscure the precise location of a data point[4] or to completely hide it from public display. Moreover, the geo-locations of some taxa are automatically obscured using a `taxon_geoprivacy` tag. The latter, however, is location-specific: for instance, we found obscured entries for *Drosera rotundifolia* only in North America, while un-obscured entries for the species

[4]The publicly shown geo-locations are random points within a $0.2° \times 0.2°$ area that contains the true coordinates (i. e., a 22 km × 22 km area at the equator).

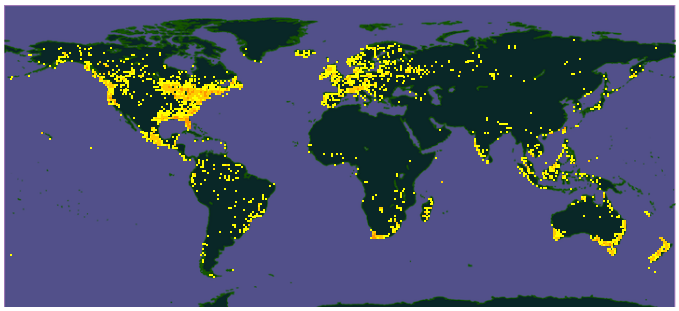


Fig. 18. Geographic distribution of the iNaturalist carnivorous plant dataset, created using the same means and mapping as in Fig. 5.

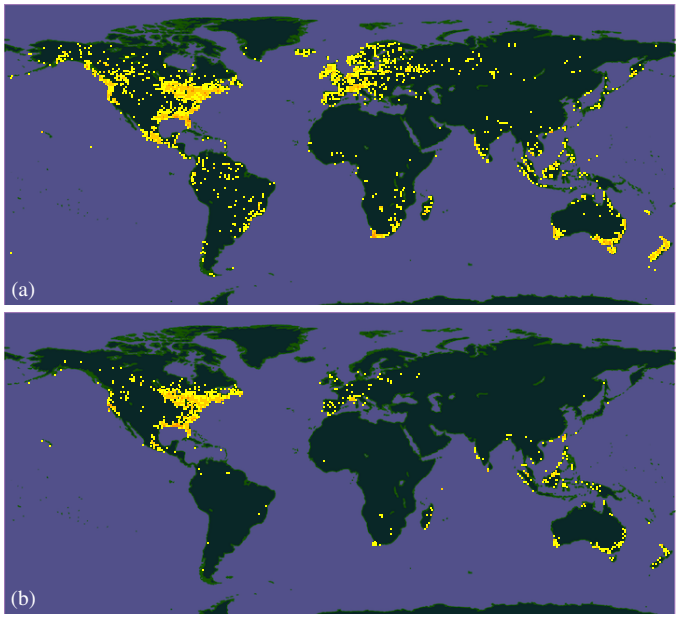


Fig. 19. Comparison of the geographic distibution of (a) the precise and (b) the obscured observations in the iNaturalist carnivorous plant dataset.

exist in North America, Europe, and Asia. Overall, for our chosen carnivorous plant subset, approx. 26.7% (8,035 out of 30,118) of the observations are obscured (as of March 2020).[5]

In total we thus have about 3× as many data points compared to those from Panoramio and Flickr. Nonetheless, we can observe similar overall data properties. As before, we find a power-law-like distribution for people's participation (Fig. 17(a)): 6.3% of people (533 out of 8,431 total people posting observations) have posted half of the observations. We fitted a discrete power law distribution fit with parameter $\alpha = 2.08$ and standard error $\sigma = 0.022$, for $x_{min} = 3$. We can also find a power-law-like distribution for species (Fig. 17(b)); a discrete power law fit yielded the parameter $\alpha = 1.40$ and a standard error of $\sigma = 0.017$, for $x_{min} = 1$. The lack of a clear tail compared to the Panoramio/Flickr dataset may be due to the increased emphasis on classification (we excluded the non-classified genus-only entries in Fig. 9(b) and 17(b)).

The fundamental difference to our previous dataset, in fact, is iNaturalist's goal of "research-grade" data, which includes not only precise positions but also reliable species classification. They achieve this goal by encouraging all contributors to verify or correct other poster's classifications. In our data subset, 89.4% of the entries were classified as research grade and it only included 5.3% genus-only entries, compared to 11.8% for our own dataset.

[5]Only a single observation in this dataset was independently contributed by a co-author of this article, before becoming involved in the work.

Moreover, iNaturalist focuses on organisms in the wild, tagging cultivated animals or plants (if included) specifically so that we did not have to identify and filter them out.

Yet iNaturalist data is not without bias either. It focuses, in particular, much more on North America and some regions of South Africa, Australia, New Zealand, and the Alps in Europe (Fig. 18), compared to the (also heavily biased, as we showed) global distribution of our Flickr data (Fig. 5). This bias also shows in the mentioned regional differences in the use of the `taxon_geoprivacy` tag, as well as in the posters' use of iNaturalist's data obscuring (Fig. 19: entries are primarily obscured in North America). As the site only started in 2008, this may thus be an artifact of local interest/publicity and also a language barrier.

Particularly interesting, however, is the comparison through our motion plausibility profiles. Here, we need to extend the visualization to also take the precision of an observation into account because it does not make sense to include imprecise locations into distance and speed computations. We thus represent those observations that we know are imprecise (either through the `taxon_geoprivacy` tag or through contributors themselves marking the `geoprivacy` tag) as outlined shapes, and we exclude them from the apparent speed computations. We show the result for some of the top iNaturalist contributors in Fig. 20 (we show all top 12 Panoramio/Flickr and top 30 iNaturalist contributors' profiles in Fig. 66–72 in the additional material).

We can observe that many observations, in fact, have plausible motion patterns. Overall, we still observed several cases where a set of images was assigned to the exact same location, yet overall this seems to be less frequent than in the Panoramio/Flickr dataset. Moreover, it does not mean that the data is actively hidden, it could be manually entered pictures without (or with erroneous) location information. What is especially interesting is that we were able to identify people who (based on matching names and descriptions on the different services) contributed to both data sources. For example, we matched the motion plausibility profile in Fig. 16(c) (299 observations) to the one in Fig. 20(f) (192 observations). This person ("J. Doe") made active use of geoprivacy, with 50% of locations obscured. While only three images with location information were posted on Flickr after 2016, new observations continue to be added in iNaturalist. We identified two more likely matches for people with more than 70 contributions to the Panoramio/Flickr data, yet they had only few entries in the iNaturalist data so far (5 and 41 observations, respectively) so creating a motion plausibility profile for them does not yet make much sense.

7.2 Dataset combination

We merged both sources as shown in Fig. 21. The simple combination seems unproblematic because, with only the mentioned single exception, the top contributors do not seem to have been active in both Panoramio/Flickr and iNaturalist. For the combined visualization we treat precise and obscured data points as two separate data sources, showing the latter ones in a different color to visually indicate that their locations are only approximations. We can thus now use the precise data from iNaturalist specifically as references to validate social media points as done with between-poster verification in Fig. 10. We can treat the precise iNaturalist as validated (due to their verification within iNaturalist), while we know that the obscured ones should not be used to verify locations of others within any radius smaller than 25 km. Nonetheless, all data points (including obscured ones as well as the ones in the Panoramio/Flickr dataset we considered to have been adjusted) can

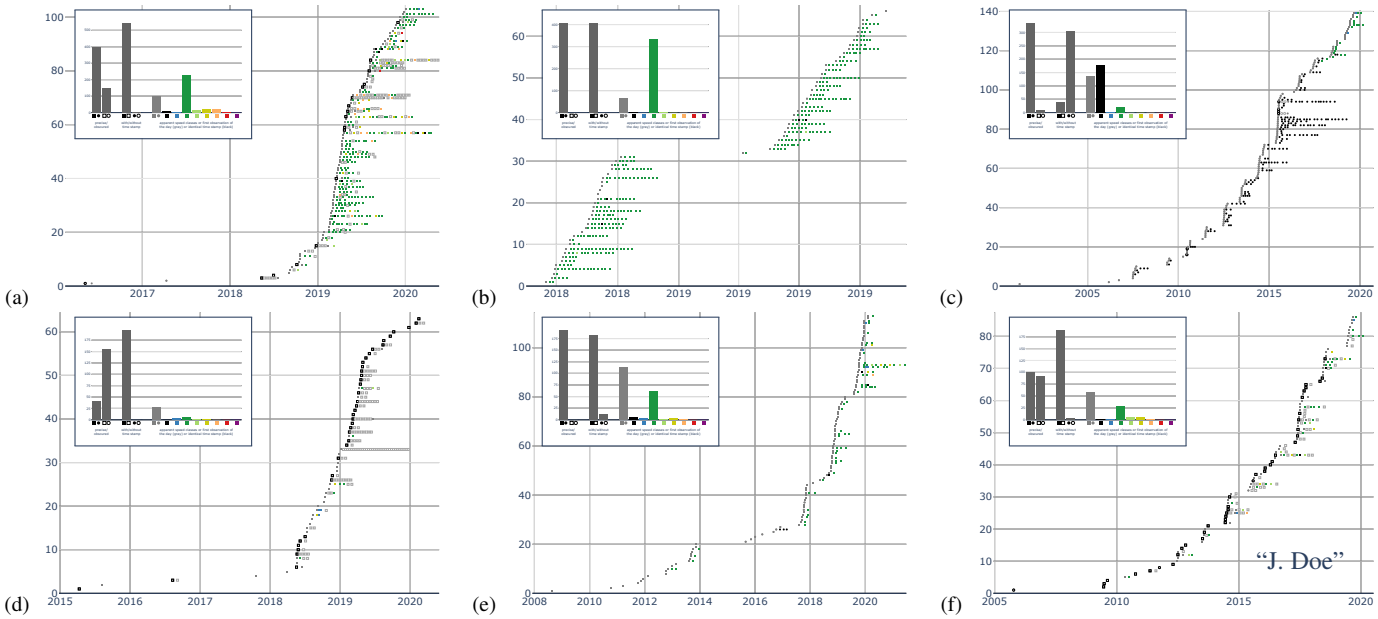


Fig. 20. Selection of motion plausibility profiles from the top contributors in the iNaturalist dataset. Same principle and color scale as in Fig. 15, with the addition of entries without full time stamp marked using circles. These date-only entries are also not included in the color scheme considerations.

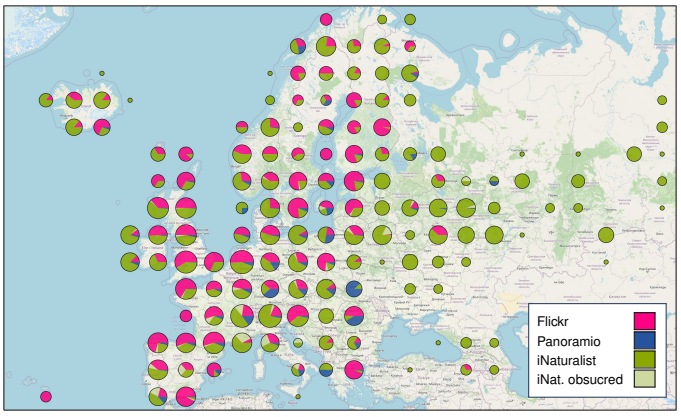


Fig. 21. Entries from both datasets shown via graduated [6] pie charts, scaled by the logarithm (base 1.2) of the entry count in the respective grid cell, for the example of Europe and parts of Asia (other examples in Fig. 50–64). Interestingly, the iNaturalist data also extends into central Russia, an area largely missing from the Panoramio/Flickr data.

be used for global habitat maps as we had explained. And as one can see in Fig. 21, for some sectors the Panoramio/Flickr dataset provides the majority or even the only set of observations.

It would certainly be useful to develop a form of joined scale of plausibility for the observations we have in our data. Finding a generally applicable measure, however, is difficult because it is unclear how to weigh the different verification levels for individual data points as well as data contributors. We thus leave this as an open problem and only provide the individual measures.

8 Summary and Comparison to the Literature

To assess the contribution of our case study we can now ask to what degree the discussed biases and errors have been documented in the literature in the past. In fact, we can find that some have been discussed previously, in particular with respect to citizen science, as we also outlined in Sec. 2. For example, the problem of inappropriate entries is well established in citizen science, which can be addressed through participant training [40], [66] and data cleaning [76]. Also the limitations of GPS-derived location data is well known (e. g., [16], [42], [70]). Finally, the fact that the number of contributions from people to social media follows a power law distribution is also known. For example, it has previously been shown for OpenStreetMap data [44].

To approach this comparison with past work more systematically, after our described data analysis we first collected all errors and biases that we found, both for the Panoramio/Flickr and for the iNaturalist dataset. We then subdivided categories, e. g., for population and popularity biases to account for the different effects we found in our analysis, in particular to distinguish global and local effects. We also took the above-mentioned literature into account and systematically searched for references for each of the identified bias or error class. In fact, the structure of our description in Sec. 4 already reflects this systematic approach, in that it discusses the different types in a structured way. Here we now list the summary of the results of our analysis in Table 1,[6] which describes each bias or error briefly, provides links to our evidence throughout the paper and our additional material, and lists past discussions in the literature. For example, we provide visual evidence for identical GPS coordinates for multiple observations in iNaturalist data and for cultural differences of the use of the geo-privacy feature in iNaturalist. Furthermore, several of the biases or errors are also typical problems in data analysis such as the power laws we showed, yet we demonstrate them for citizen-contributed habitat data. Finally, for some biases we show new aspects, such as our visual demonstration of the influence of local site access to observations of plant habitats.

More importantly, however, we provide evidence for many of the data biases and errors for habitat data derived from social media, i. e., data that was not collected in a citizen science context—we had extracted our own species habitat data from online image

[6] In Table 1 we only list biases and errors we found, expected to find, or discussed above, but others exist in citizen science as noted, e. g., by Kandel et al. [37] and Waller [71]. There are also biases in general social media contribution that we did not identify in our data. For example, gender participation disparities exist in projects such as OpenStreetMap [21], [28], yet—while a similar bias may exist in our data—it is unlikely, due to our focus on a quite specific subject matter, that this would lead to "gender content disparities, in which content of interest to men is better represented than content of interest to women" [21].

TABLE 1
Biases and errors as discussed in Sec. 4–7: our evidence (visualizations, examples), and links past literature where these biases/errors have been discussed before (incl. those newly demonstrated for either data source, known problems in data analysis but here demonstrated for species data collected by citizen scientists, and known problems with new aspects shown). Figures beyond Fig. 21 can be found in the additional material.

name and description	type	shown for Panoramio/Flickr	shown for iNaturalist	past discussion for regular citizen science data	past discussion for "repurposed" social media data
general population bias, people are not evenly distributed on Earth	bias	Fig. 6	Fig. 6	e. g., [13]	e. g., [61]
poster-population bias, posters do not represent the overall population	bias	Fig. 5(a) vs. 6	Fig. 18 vs. 6	e. g., [12], [13], [40]	e. g., [17], [18], [55]
cultural tool bias, social media posters adopt tools or services differently based on their region of origin	bias	comparison of both datsets: Fig. 21, Fig. 49–64		typical problem of social media (e. g., [34])	
global site popularity, popular regions where people live, travel to, or spend vacations get more reports (e. g., national parks)	bias	Fig. 5	Fig. 18	e. g., [22]	none *tbk.**
local site popularity, specifically marked places in natural/state/etc. parks (e. g., walks, view points) get many reports in very small places	bias	Fig. 8, 44	Fig. 45	none *tbk.**	none *tbk.**
local site access, more reports near or close to roads or paths	bias	Fig. 7, 46	Fig. 46	abstract ("convenient locations") [13], coarse (road/path density) [8], [23], [47]; not detailed as Fig. 7	none *tbk.**
individual poster contribution skew, power law of number of contributions by individual posters	bias	Fig. 9(a), Fig. 31	Fig. 17(a), Fig. 34	e. g., [44] (typical problem; e. g., [52])	none *tbk.** (typical problem; e. g., [52])
inapplicable entries, some entries should not have been included; e. g., cultivated plants at home, in stores, or in botanical gardens	error	Fig. 1, but data is "involuntary"	no; rare due to *cultivation* tag	participant training [40], [66], data cleaning [76]	n/a but typical problem, data cleaning [76] possible
inherent geographic location offset, natural limitations of GPS location measurement	error	inherent, Fig. 44	inherent, Fig. 45	e. g., [16], [42], [70]	only in passing: [18]
small offset through (well-intentioned) manual location entry w/o GPS data from memory (e. g., prior to public GPS start in 1980s)	error	inherent; e. g., F 7826985010	inherent; e. g., iN 48182055	none *tbk.**	none *tbk.**
coordinate errors through (well-intentioned) manual location entry w/o GPS data, coordinate flips, zeroed coordinates (i. e., major errors)	error	none found, Fig. 5(c) vs. 37	none found	e. g., [71], [76]	e. g., [27], [37]; strangely, our data lacks such errors
species popularity, popular plants species get more reports, power law	bias	Fig. 9(b), 32	Fig. 17(b), 35	e. g., [13]	typical problem (e. g., [13])
only genus classification, lack of poster expertise	error	e. g., F 35188184	rare	e. g., [16]	typical problem (e. g., [16])
species misclassification, lack of poster expertise	error	e. g., F 16332954669	rare	e. g., [13], [15], [16]	none *tbk.**
species classification change, errors that appear over time because recognized species classification has changed	error	Fig. 11, 47	Fig. 48	e. g., [14]	typical problem (e. g., [75])
lack of species classification, images are unlabeled so less likely to be used because difficult to find	bias	e. g., F 24857758530	n/a, actively prevented	usually n/a	none *tbk.**
search tool limitations, search returns incomplete/inconsistent results or encourages posters to remove their contributions	bias	Fig. 27, Fig. 28	none found or inherent	none *tbk.*, also not found by us, probably n/a	e. g., [51]; but specific aspects (Fig. 27, 28) new
single item offset through intentionally false location entry, location offset because posters do not want to reveal exact location	error	Sec. 5, Fig. 15, F 3020563046	none found explicitly	usually n/a	only in passing: [17]
multiple items on single, offset location through intentionally false location entry, exact same location for multiple items, possibly to avoid the extra effort for entry or due to deliberate location obfuscation	error	Sec. 5, Fig. 10, Fig. 15	Fig. 70(e), 71(a)	locations recorded at region centroid or on a raster [76]; intentional obfuscation more likely for posted images	none *tbk.**
random offset for geo-privacy as supported by citizen science projects	n/a	n/a	built-in	e. g., iNaturalist	n/a (coord. can be omitted)
cultural differences in use of geo privacy of citizen science projects	n/a	n/a	Fig. 19, 41–43	none *tbk.**	n/a
duplicated entries when combining independent datasets	bias	Fig. 73(a) vs. 73(b)		typical problem in data source merging	

* *tbk. = to the best of our knowledge*

sharing sites. On these social media sites, posters are not aware that we would use their data for analyzing habitats, and many of them may not even be aware that this is possible. Table 1 lists the respective past discussion, if we found any, in the last column. In particular, we demonstrated geo-coordinate manipulation in this data, likely with the intention of protecting the plant habitats from poaching. Also, we clearly demonstrated several aspects of popularity of certain regions, influenced by the population and popularity biases we discussed. Moreover, it is also interesting that we did explicitly fail to observe certain aspects such as coordinate errors through (well-intentioned) manual location entry without GPS data (e. g., coordinate flips, zeroed coordinates), which had otherwise been documented for social media data [37]. Table 1 visually showcases those biases and errors we newly demonstrate (green background), those that are due to known problems in data analysis but for which we demonstrate evidence visually in the context of citizen-contributed habitat data (light green), and those ones for which we demonstrate new aspects (yellow).

9 Conclusion

With our case study we thus, first, contribute to a better understanding of data biases and errors in citizen-contributed scientific datasets—yet not only with respect to traditional citizen science efforts but also and, in particular, with respect to "repurposed" data derived from social media posts. We showed that, in addition to existing explicit citizen science projects being able to fill gaps of biodiversity data [3], an additional "contribution model" exists in which data can be extracted from sources not targeted at citizen science. But in such cases the questions of data plausibility, data biases, and active data hiding have to be specifically investigated. In addition to providing new visual evidence for known biases and errors, we demonstrated biases and errors that had not yet been discussed. Moreover, we demonstrated and provided visual evidence for most of these biases and errors in habitat data derived from image-related data posted on social media. So, while admittedly laborious, our approach can allow scientists to gain insight to some dedicated sub-field of biodiversity if the need (or interest) exists and outweighs the needed labor. Moreover, our approach allows even *a single person* to collect data (because pictures exist for many types of subject matter—also beyond the field of biodiversity), as opposed to relying on *a whole community* to actively participate in a project such as iNaturalist.

Second, we contribute a new space-time visualization—the Motion Plausibility Profile—that allows us to show some of these errors in more detail. This visual representation qualitatively shows

certain data manipulations, for those dataset contributors with a high number of observations. Our representation is interesting from a visualization point of view because it shows space not in three or two dimensions but uses only one dimension (i. e., distance). Moreover, it represents two aspects of time: observations from several times in the day (shown implicitly as a representation of speed classes in a day's image sequence) and observation sessions from different days over the years. Our representation thus combines one-dimensional distance with speed (derived from distance and detailed time) and a coarse time representation—somewhat akin to train schedule maps (discussed, e. g., by Tufte [69])—, in contrast to other space-time cube representations. It also goes beyond "space flattening" [7] since we use the local coordinate system to encode both distance traveled in a day and precise time (the latter implicitly via color-coded speed).

Together with within- and between-poster validation, the motion plausibility profiles allow us to visually represent certain notions of trust into, in particular, the geographic locations of the observations. A neutral quantification of the errors and biases, however, is likely not possible and also probably not desirable because the visual representations depend on observation counts and viewer contributions, which themselves are biased as we have shown. So, in terms of whether a particular location report is plausible or not, our visual aids are instead more related to the notion of (qualitative) *implicit error* proposed by McCurdy et al. [50]. For example, while the insights on the plausibility of a particular contributor's data highlighted by our motion plausibility profiles can lead to disregarding the data if one looks to track down specific populations, they could still be valid for larger-scale distribution maps or for the yearly growth patterns of the plants due to the introduced error likely being small and geographic in nature and thus acceptable for some tasks and analyses. Moreover, evidence for data manipulation does not necessarily imply that the respective data is incorrect. Instead, the conclusions to be drawn from such evidence will always depend on the specific application case.

Finally, we demonstrated the role that visualization can play in this process of identifying and demonstrating errors and biases. We were able to find, observe, and showcase the many biases and errors as listed in Table 1, without much background in citizen science as we showed through the personal report of our scientific journey. Moreover, our specific observations and investigations of anecdotal evidence even led to a new generalizable visual representation that can be employed in future citizen science research and beyond. With it we were able to show and visualize intentional location data adjustment for geo-located social media posts, which, in our data, is often due to (well-intentioned) data adjustment. Citizen science projects like iNaturalist are more systematic about the data collection and can reduce location adjustment with geo-privacy mechanisms. They also reduce nomenclature errors through community corrections. While iNaturalist thus produces better data from more people interested in species habitats overall, the social media data repurposing approach still has merit: it "reaches" a different audience. Many plant amateurs are simply not aware of the specialized citizen science apps like iNaturalist (including some of us at the start of the project), or are interested to reach a different audience with their images (e. g., people on popular platforms).

Our specific motion plausibility profile representation design is also interesting in the context of space-time visualizations. It is an example where specific geographic coordinates are not needed for gaining insights into a person's activity—the one-dimensional traveled distance is sufficient. Moreover, we include two notions of passing time along a single axis: the time during a day in form of a sequence and speed, coupled with the covered distance, and a second notion of coarser time granularity to represent different days with pictures, which could be explored for extending existing notions of space-time representation generalizations [7].

Of course, many more ways exist to wrangle, compare, and analyze our data. In particular, it would be interesting to investigate the differences in distribution of selected (well-covered species) in our dataset and the iNaturalist data, potentially using the process of species distribution modeling mentioned in the introduction. Potentially this may lead to a better understanding of the habitat of certain species. We also specifically did not recommend any specific mechanism to compute a single measure of data plausibility because this plausibility varies depending on the application and data characteristics. So it would be interesting to investigate how such visual representations can be integrated into data tools used in different fields, in biodiversity research and beyond. It would specifically be interesting to see our motion plausibility profiles integrated into a citizen science tool such as iNaturalist, and investigate if it helps researchers to better understand the data or the project itself to understand the community's use of their geo-privacy feature. Finally, and relying on a hopefully growing Flickr/Panoramio dataset, it would be interesting to also attempt a quantitative analysis of some of the errors and biases.

ACKNOWLEDGMENTS

The authors thank Anastasia Bezerianos and Petra Isenberg for running a Data Fair as part of their 2019 InfoVis class, which started a student project that contributed to this work. Also thanks to Roderic Page for his analysis tool of Flickr's Encyclopedia of Life group [58]. We also thank Jennifer Preece, Ben Shneiderman, Pierre Dragicevic, and all anonymous reviewers for their valuable feedback on the paper as well as Christian Tominski for his help with space-time visualizations. Last but not least, thanks to the AVIZ team and to the audience of VieVisDays 2019 at TU Wien's computer graphics group for general feedback on the project.

IMAGE AND DOCUMENT LICENSES, DATA SHARING

REFERENCES

[1] W. Adlassnig, E. Mayer, M. Peroutka, W. Pois, and I. K. Lichtscheidl, "Two American *Sarracenia* species as neophyta in central Europe," *Phyton. Annales Rei Botanicae*, vol. 49, no. 2, pp. 279–292, Mar. 2010.

[2] J. Alstott, E. Bullmore, and D. Plenz, "powerlaw: A Python package for analysis of heavy-tailed distributions," *PLOS ONE*, vol. 9, no. 1, pp. e85 777:1–e85 777:11, Jan. 2014. doi: 10.1371/journal.pone.0085777

[3] T. Amano, J. D. L. Lamming, and W. J. Sutherland, "Spatial gaps in global biodiversity information and the role of citizen science," *BioScience*, vol. 66, no. 5, pp. 393–400, May 2016. doi: 10.1093/biosci/biw022

[4] G. Andrienko, N. Andrienko, P. Bak, S. Kisilevich, and D. Keim, "Analysis of community-contributed space- and time-referenced data (example of Flickr and Panoramio photos)," in *Proc. VAST (Posters)*. Los Alamitos: IEEE CS, 2009, pp. 213–214. doi: 10.1109/VAST.2009.5333472

[5] L. Anthony, "Introducing Fireant: A freeware, multiplatform social media data-analysis tool," *IEEE Trans Prof Commun*, vol. 61, no. 4, pp. 428–442, Dec. 2018. doi: 10.1109/TPC.2018.2870681

[6] N. D. Arnold, B. Jenny, and D. White, "Automation and evaluation of graduated dot maps," *Int J Geogr Inf Sci*, vol. 31, no. 12, pp. 2524–2542, Dec. 2017. doi: 10.1080/13658816.2017.1359747

[7] B. Bach, P. Dragicevic, D. Archambault, C. Hurter, and S. Carpendale, "A descriptive framework for temporal data visualizations based on generalized space-time cubes," *Comput Graph Forum*, vol. 36, no. 6, pp. 36–61, Sep. 2017. doi: 10.1111/cgf.12804

[8] J. Beck, M. Böller, A. Erhardt, and W. Schwanghart, "Spatial bias in the GBIF database and its effect on modeling species' geographic distributions," *Ecol Inf*, vol. 19, pp. 10–15, Jan. 2014. doi: 10.1016/j.ecoinf.2013.11.002

[9] R. Blanco, Z. Salazar, and T. Isenberg, "Exploring carnivorous plant habitats based on images from social media," in *IEEE VIS Posters*, 2019.

[10] G.-P. Bonneau, H.-C. Hege, C. Johnson, M. M. Oliveira, K. Potter, and P. Rheingans, "Overview and state-of-the-art of uncertainty visualization," in *Scientific Visualization: Uncertainty, Multifield, Biomedical, Scalable*, ser. Mathematics and Visualization, C. Hansen, M. Chen, C. Johnson, A. Kaufman, and H. Hagen, Eds. London: Springer, 2014, vol. 17, pp. 3–27. doi: 10.1007/978-1-4471-6497-5_1

[11] A. Bowser, A. Wiggins, L. Shanley, J. Preece, and S. Henderson, "Sharing data while protecting privacy in citizen science," *Interactions*, vol. 21, no. 1, pp. 70–73, Jan. 2014. doi: 10.1145/2540032

[12] J. Bright, S. De Sabbata, and S. Lee, "Geodemographic biases in crowdsourced knowledge websites: Do neighbours fill in the blanks?" *GeoJournal*, vol. 83, pp. 427–440, 2018. doi: 10.1007/s10708-017-9778-7

[13] C. T. Callaghan, J. J. L. Rowley, W. K. Cornwell, A. G. B. Poore, and R. E. Major, "Improving big citizen science data: Moving beyond haphazard sampling," *PLOS Biol*, vol. 17, no. 6, pp. e3 000 357:1–e3 000 357:11, Jun. 2019. doi: 10.1371/journal.pbio.3000357

[14] D. L. Campbell, A. E. Thessen, and L. Ries, "A novel curation system to facilitate data integration across regional citizen science survey programs," *PeerJ*, vol. 8, pp. e9219:1–e9219:25, Jul. 2020. doi: 10.7717/peerj.9219

[15] M. Chandler, L. See, C. D. Buesching, J. A. Cousins, C. Gillies, R. W. Kays, C. Newman, H. M. Pereira, and P. Tiago, "Involving citizen scientists in biodiversity observation," in *The GEO Handbook on Biodiversity Observation Networks*, M. Walters and R. J. Scholes, Eds. Cham, Switzerland: Springer International Publishing, 2017, ch. 9, pp. 211–237. doi: 10.1007/978-3-319-27288-7_9

[16] A. D. Chapman, "Principles of data quality," Global Biodiversity Information Facility, Copenhagen, Danmark, Tech. Rep., Jul. 2005. doi: 10.15468/DOCJRGG-A190

[17] S. Chen, L. Lin, and X. Yuan, "Social media visual analytics," *Comput Graph Forum*, vol. 36, no. 3, pp. 563–587, Jun. 2017. doi: 10.1111/cgf.13211

[18] S. Chen, X. Yuan, Z. Wang, C. Guo, J. Liang, Z. Wang, X. Zhang, and J. Zhang, "Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data," *IEEE Trans Vis Comput Graph*, vol. 22, no. 1, pp. 270–279, Jan. 2016. doi: 10.1109/TVCG.2015.2467619

[19] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proc. WWW*. New York: ACM, 2009, pp. 761–770. doi: 10.1145/1526709.1526812

[20] C. Darwin, *Insectivorous Plants*. London: John Murrey, 1875.

[21] M. Das, B. Hecht, and D. Gergle, "The gendered geography of contributions to OpenStreetMap: Complexities in self-focus bias," in *Proc. CHI*. New York: ACM, 2019, pp. 563:1–563:14. doi: 10.1145/3290605.3300793

[22] R. L. H. Dennis and C. D. Thomas, "Bias in butterfly distribution maps: The influence of hot spots and recorder's home range," *J Insect Conserv*, vol. 4, no. 2, pp. 73–77, Jun. 2000. doi: 10.1023/A:1009690919835

[23] J. L. Dickinson, B. Zuckerberg, and D. N. Bonter, "Citizen science as an ecological research tool: Challenges and benefits," *Annu Rev Ecol Evol Syst*, vol. 41, no. 1, pp. 149–172, Dec. 2010. doi: 10.1146/annurev-ecolsys-102209-144636

[24] C. Eaton, C. Plaisant, and T. Drizd, "The challenge of missing and uncertain data," in *Poster Compendium of IEEE Visualization*. Los Alamitos: IEEE CS, 2003, p. 100. doi: 10.1109/VIS.2003.10029

[25] C. Eaton, C. Plaisant, and T. Drizd, "Visualizing missing data: Graph interpretation user study," in *Proc. INTERACT*. Berlin, Heidelberg: Springer, 2005, pp. 861–872. doi: 10.1007/11555261_68

[26] S. Ferretti, S. Mirri, C. Prandi, and P. Salomoni, "Trustworthiness in crowd- sensed and sourced georeferenced data," in *Proc. PerCom WS*. IEEE, 2015, pp. 402–407. doi: 10.1109/PERCOMW.2015.7134071

[27] D. Fisher, "Hotmap: Looking at geographic attention," *IEEE Trans Vis Comput Graph*, vol. 13, no. 6, pp. 1184–1191, Nov. 2007. doi: 10.1109/TVCG.2007.70561

[28] Z. Gardner, P. Mooney, S. De Sabbata, and L. Dowthwaite, "Quantifying gendered participation in OpenStreetMap: Responding to theories of female (under) representation in crowdsourced mapping," *GeoJournal*, vol. 85, pp. 1603–1620, 2020. doi: 10.1007/s10708-019-10035-z

[29] T. Gschwandtner, W. Aigner, S. Miksch, J. Gärtner, S. Kriglstein, M. Pohl, and N. Suchy, "TimeCleanser: A visual analytics approach for data cleansing of time-oriented data," in *Proc. i-KNOW*. New York: ACM, 2014, pp. 18:1–18:8. doi: 10.1145/2637748.2638423

[30] T. Gschwandtner, J. Gärtner, W. Aigner, and S. Miksch, "A taxonomy of dirty time-oriented data," in *Multidisciplinary Research and Practice for Information Systems*. Berlin, Heidelberg: Springer, 2012, pp. 58–72. doi: 10.1007/978-3-642-32498-7_5

[31] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay, "In pursuit of error: A survey of uncertainty visualization evaluation," *IEEE Trans Vis Comput Graph*, vol. 25, no. 1, pp. 903–913, Jan. 2019. doi: 10.1109/TVCG.2018.2864889

[32] T. Isenberg, "Experiences with propagation of *Sarracenia flava* (*Sarraceniaceae*) through division with only one growing spot," *Carniv Plant Newsl*, vol. 33, no. 2, pp. 36–37, Jun. 2004.

[33] IUCN, "The IUCN Red List of Threatened Species," Online resource: http://www.iucnredlist.org/, 2019, version 2019-3, accessed Feb. 2020.

[34] L. A. Jackson and J.-L. Wang, "Cultural differences in social networking site use: A comparative study of China and the United States," *Comput Hum Behav*, vol. 29, no. 3, pp. 910–921, May 2013. doi: 10.1016/j.chb.2012.11.024

[35] D. E. Jennings and J. R. Rohr, "A review of the conservation threats to carnivorous plants," *Biol Conserv*, vol. 144, no. 5, pp. 1356–1363, May 2011. doi: 10.1016/j.biocon.2011.03.013

[36] A. Kamal, P. Dhakal, A. Y. Javaid, V. K. Devabhaktuni, D. Kaur, J. Zaientz, and R. Marinier, "Recent advances and challenges in uncertainty visualization: A survey," *J Vis*, vol. 24, no. 5, pp. 861–890, Oct. 2021. doi: 10.1007/s12650-021-00755-1

[37] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono, "Research directions in data wrangling: Visuatizations and transformations for usable and credible data," *Inf Vis*, vol. 10, no. 4, pp. 271–288, Oct. 2011. doi: 10.1177/1473871611415994

[38] S. Kim, R. Maciejewski, A. Malik, Y. Jang, D. S. Ebert, and T. Isenberg, "Bristle Maps: A multivariate abstraction technique for geovisualization," *IEEE Trans Vis Comput Graph*, vol. 19, no. 9, pp. 1438–1454, Sep. 2013. doi: 10.1109/TVCG.2013.66

[39] S. Kisilevich, M. Krstajic, D. Keim, N. Andrienko, and G. Andrienko, "Event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections," in *Proc. Information Visualisation*. Los Alamitos: IEEE CS, 2010, pp. 289–296. doi: 10.1109/IV.2010.94

[40] M. Kosmala, A. Wiggins, A. Swanson, and B. Simmons, "Assessing data quality in citizen science," *Front Ecol Environ*, vol. 14, no. 10, pp. 551–560, Dec. 2016. doi: 10.1002/fee.1436

[41] R. Krueger, G. Sun, F. Beck, R. Liang, and T. Ertl, "TravelDiff: Visual comparison analytics for massive movement patterns derived from Twitter," in *Proc. PacificVis*. Los Alamitos: IEEE CS, 2016, pp. 176–183. doi: 10.1109/PACIFICVIS.2016.7465266

[42] H. Laitinen, J. Lähteenmäki, and T. Nordström, "Database correlation method for GSM location," in *Proc. IEEE VTS*, vol. 4. Piscataway, NJ, USA: IEEE, 2001, pp. 2504–2508. doi: 10.1109/VETECS.2001.944052

[43] J. S. Lewis, J. L. Rachlow, E. O. Garton, and L. A. Vierling, "Effects of habitat on GPS collar performance: Using data screening to reduce location error," *J Appl Ecol*, vol. 44, no. 3, pp. 663–671, Mar. 2007. doi: 10.1111/j.1365-2664.2007.01286.x

[44] D. Ma, M. Sandberg, and B. Jiang, "Characterizing the heterogeneity of the OpenStreetMap data and community," *ISPRS Int J Geo-Inf*, vol. 4, no. 2, pp. 535–550, Apr. 2015. doi: 10.3390/ijgi4020535

[45] A. M. MacEachren, "Visualizing uncertain information," *Cartogr Perspect*, no. 13, pp. 10–19, Fall 1992. doi: 10.14714/CP13.1000

[46] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler, "Visualizing geospatial information uncertainty: What we know and what we need to know," *Cartogr Geogr Inf Sci*, vol. 32, no. 3, pp. 139–160, 2005. doi: 10.1559/1523040054738936

[47] L. Mair and A. Ruete, "Explaining spatial variation in the recording effort of citizen science data across multiple taxa," *PLOS ONE*, vol. 11, no. 1, pp. e0 147 796:1–e0 147 796:13, Jan. 2016. doi: 10.1371/journal.pone.0147796

[48] C. Maldonado, C. I. Molina, A. Zizka, C. Persson, C. M. Taylor, J. Albán, E. Chilquillo, N. Rønsted, and A. Antonelli, "Estimating species diversity and distribution in the era of Big Data: To what extent can we trust public databases?" *Glob Ecol Biogeogr*, vol. 24, no. 8, pp. 973–984, Aug. 2015. doi: 10.1111/geb.12326

[49] Global Biodiversity Information Facility, "Magnoliopsida," Dataset, 2017. doi: 10.15468/dl.wquvxb

[50] N. McCurdy, J. Gerdes, and M. Meyer, "A framework for externalizing implicit error using visualization," *IEEE Trans Vis Comput Graph*, vol. 25, no. 1, pp. 925–935, Jan. 2019. doi: 10.1109/TVCG.2018.2864913

[51] F. Morstatter and H. Liu, "Discovering, assessing, and mitigating data bias in social media," *Online Soc Networks Media*, vol. 1, pp. 1–13, Jun. 2017. doi: 10.1016/j.osnem.2017.01.001

[52] L. Muchnik, S. Pei, L. C. Parra, S. D. S. Reis, J. S. A. Jr., S. Havlin, and H. A. Makse, "Origins of power-law degree distribution in the heterogeneity of human activity in social networks," *Sci Rep*, vol. 3, pp. 1783:1–1783:13, May 2013. doi: 10.1038/srep01783

[53] R. F. C. Naczi, E. M. Soper, F. W. Case, Jr., and R. B. Case, "*Sarracenia rosea* (*Sarraceniaceae*), a new species of pitcher plant from the southeastern United States," *SIDA Contrib Bot*, vol. 18, no. 4, pp. 1183–1206, Dec. 1999.

[54] A. Olteanu, "Probing the limits of social data: Biases, methods, and domain knowledge," Ph.D. dissertation, École Polytechnique Fédérale de Lausane, Switzerland, 2016. doi: 10.5075/epfl-thesis-6892

[55] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, "Social data: Biases, methodological pitfalls, and ethical boundaries," *Front Big Data*, vol. 2, pp. 13:1–13:33, Jul. 2019. doi: 10.3389/fdata.2019.00013

[56] M. Oppermann and T. Munzner, "Data-first visualization design studies," in *Proc. BELIV*. Los Alamitos: IEEE Computer Society, 2020, pp. 74–80. doi: 10.1109/BELIV51497.2020.00016

[57] L. Padilla, M. Kay, and J. Hullman, "Uncertainty visualization," in *Handbook of Computational Statistics and Data Science*, W. W. Piegorsch, R. Levine, H. H. Zhang, and T. C. M. Lee, Eds. John Wiley & Sons, 2023, ch. 10, to appear. doi: 10.31234/osf.io/ebd6r

[58] R. Page, "flickreolmap," GitHub repository, https://github.com/rdmpage/flickreolmap, 2012.

[59] R. Page, "Where is the "crowd" in crowdsourcing? Mapping EOL Flickr photos," Blog post, https://iphylo.blogspot.com/2012/06/where-is-in-crowdsourcing-mapping-eol.html, Jun. 2012.

[60] C. Parisod, C. Trippi, and N. Galland, "Genetic variability and founder effect in the pitcher plant *Sarracenia purpurea* (*Sarraceniaceae*) in populations introduced into Switzerland: From inbreeding to invasion," *Ann Bot*, vol. 95, no. 2, pp. 277–286, Jan. 2005. doi: 10.1093/aob/mci023

[61] J. Pick, A. Sarkar, and J. Rosales, "Social media use in american counties: Geography and determinants," *ISPRS Int J Geo-Inf*, vol. 8, no. 9, pp. 424:1–424:25, 2019. doi: 10.3390/ijgi8090424

[62] K. Potter, P. Rosen, and C. R. Johnson, "From quantification to visualization: A taxonomy of uncertainty visualization approaches," in *Uncertainty Quantification in Scientific Computing*. Berlin, Heidelberg: Springer, 2012, pp. 226–249. doi: 10.1007/978-3-642-32677-6_15

[63] J. Preece, "Citizen science: New research challenges for human–computer interaction," *Int J Hum Comput Interact*, vol. 32, no. 8, pp. 585–612, Jun. 2016. doi: 10.1080/10447318.2016.1194153

[64] B. A. Renfro, M. Stein, N. Boeker, E. Reed, and E. Villalba, "An analysis of global positioning system (GPS) standard positioning service (SPS) performance for 2018," Space and Geophysics Laboratory, Applied Research Laboratories, University of Texas at Austin, USA, Tech. Rep. TR-SGL-19-02, Mar. 2019.

[65] M. P. Robertson, V. Visser, and C. Hui, "Biogeo: An R package for assessing and improving data quality of occurrence record datasets," *Ecography*, vol. 39, no. 4, pp. 394–401, Apr. 2016. doi: 10.1111/ecog.02118

[66] H. E. Roy, M. J. O. Pocock, C. D. Preston, D. B. Roy, J. Savage, J. C. Tweddle, and L. D. Robinson, "Understanding citizen science and environmental monitoring: Final report on behalf of UK Environmental Observation Framework," NERC Centre for Ecology & Hydrology and Natural History Museum, UK, Tech. Rep. N020679CR, Nov. 2012.

[67] F. Royer and M. Lutcavage, "Filtering and interpreting location errors in satellite telemetry of marine animals," *J Exp Mar Biol Ecol*, vol. 359, no. 1, pp. 1–10, Apr. 2008. doi: 10.1016/j.jembe.2008.01.026

[68] H. Sharp, J. Preece, and Y. Rogers, *Interaction Design: Beyond Human-Computer Interaction*, 5th ed. Indianapolis: Wiley & Sons, 2019.

[69] E. R. Tufte, *The Visual Display of Quantitative Information*, 2nd ed. Cheshire: Graphics Press, 2001.

[70] F. van Diggelen and P. Enge, "The world's first GPS MOOC and worldwide laboratory using smartphones," in *Proc. ION GNSS+*. Manassas, VA, USA: Institute of Navigation, 2015, pp. 361–369.

[71] J. Waller, "Data location quality at GBIF," *Biodivers Inf Sci Stand*, vol. 3, pp. e35 829:1–e35 829:3, Jun. 2019. doi: 10.3897/biss.3.35829

[72] Y. Wang, G. Baciu, and C. Li, "Cognitive exploration of regions through analyzing geo-tagged social media data," in *Proc. ICCI*CC*. Los Alamitos: IEEE CS, 2017, pp. 59–64. doi: 10.1109/ICCI-CC.2017.8109730

[73] A. Wiggins, G. Newman, R. D. Stevenson, and K. Crowston, "Mechanisms for data quality and validation in citizen science," in *Proc. eScienceW*. Los Alamitos: IEEE CS, 2011, pp. 14–19. doi: 10.1109/eScienceW.2011.27

[74] Y. Wu, N. Cao, D. Gotz, Y.-P. Tan, and D. A. Keim, "A survey on visual analytics of social media data," *IEEE Trans Multimedia*, vol. 18, no. 11, pp. 2135–2148, Nov. 2016. doi: 10.1109/TMM.2016.2614220

[75] N. Ytow, D. R. Morse, and D. M. Roberts, "Nomencurator: A nomenclatural history model to handle multiple taxonomic views," *Biol J Linn Soc*, vol. 73, no. 1, pp. 81–98, Jan. 2008. doi: 10.1111/j.1095-8312.2001.tb01348.x

[76] A. Zizka, D. Silvestro, T. Andermann, J. Azevedo, C. Duarte Ritter, D. Edler, H. Farooq, A. Herdean, M. Ariza, R. Scharn, S. Svantesson, N. Wengström, V. Zizka, and A. Antonelli, "CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases," *Methods Ecol Evol*, vol. 10, no. 5, pp. 744–751, 2019. doi: 10.1111/2041-210X.13152

Tobias Isenberg is a senior research scientist at Inria, France. Previously he held positions as post-doctoral fellow at the University of Calgary, Canada, and as assistant professor at the University of Groningen, the Netherlands. His research interests include scientific visualization, illustrative and non-photorealistic rendering, and interactive visualization techniques. He has been [32] and continues to be fascinated by the carnivorous plants of the world, and likes to visit their habitats.

Zujany Salazar is a PhD student at the Université Paris-Saclay, France. She received her M.Sc. in Computer Science for Communication Networks from Télécom SudParis in 2020. Her research covers the areas of simulation and emulation of network traffic patterns and cyberattacks, and risk assessment for 5G networks.

Rafael Blanco is a PhD student at the Polytechnic University of Catalonia, Spain. He received his M.Sc. in Computer Science for Communication Networks from Télécom SudParis in 2020. His main research interests include procedural modeling and authoring tools for crowd simulation and virtual environments.

Catherine Plaisant is a research scientist emerita at the University of Maryland Institute for Advanced Computer Studies and Associate Director of Research of the Human-Computer Interaction Lab. Catherine earned a Doctorat d'Ingénieur degree in France. In 1988 she joined the Human-Computer Interaction Laboratory where she has been working with multidisciplinary teams on designing and evaluating new interface technologies that are useful and usable.

Do you believe your (social media) data? A personal story on location data biases, errors, and plausibility as well as their visualization

Additional material

In this additional material we provide further examples. In Fig. 22–30 we show plots of some abstract data properties in the Flickr/ Panoramio data. Next, in Fig. 31–36 we provide further plots of the number of entries by contributors and by species for both datasets. Further, in Fig. 37–43 we show more plots about the global distribution of both datasets, while in Fig. 44–48 larger versions and additional plots for local distributions. Then in Fig. 49–64 we explicitly compare the two datasets for several world regions. Fig. 65 shows a graphical explanation of the construction of our motion plausibility profiles, followed by the motion plausibility profiles for all top 12 contributors in the Flickr/Panoramio dataset (Fig. 66–67) and all top 30 contributors of the iNaturalist dataset (Fig. 68–72). Finally, Fig. 73 compares contributions of the same person (who we dubbed "J. Doe" elsewhere in the paper) in either dataset.

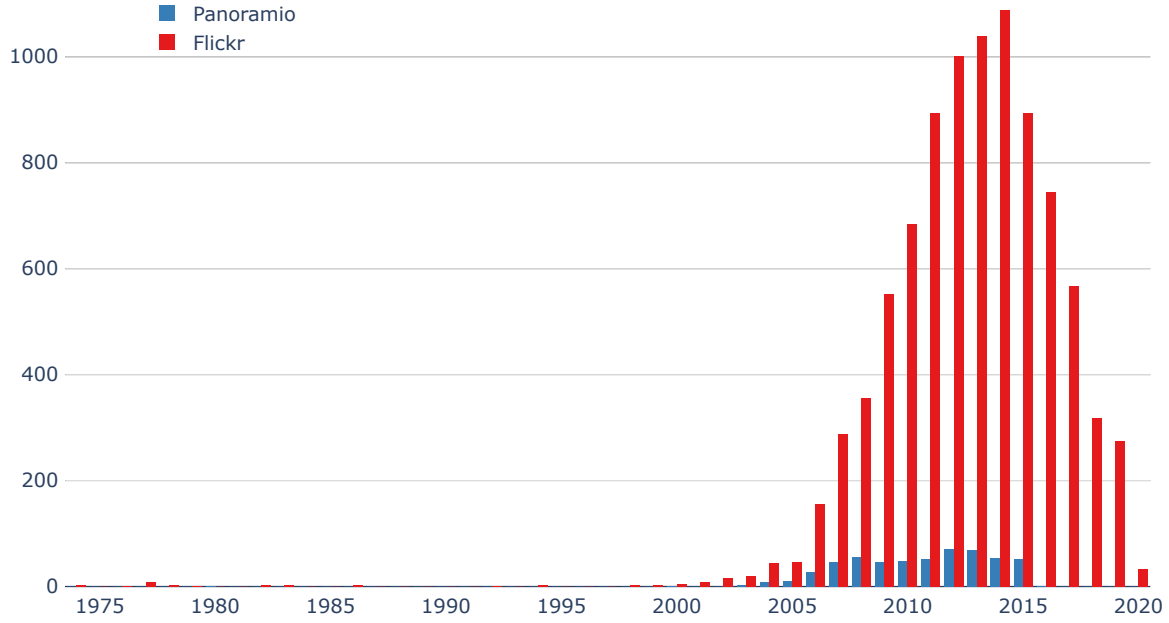


Fig. 22. Date histogram by year and service, complete Panoramio/Flickr dataset. Interestingly, the number of posts per year declines after 2014, possibly to due the fact that people post their pictures only after some time and not directly when taking them and/or due to Flickr's changed upload policies, which apparently have let quite a number people to delete their images from the service (see Fig. 27 for a graph that supports this hypothesis). Another reason may be the increased popularity of sites such as iNaturalist.

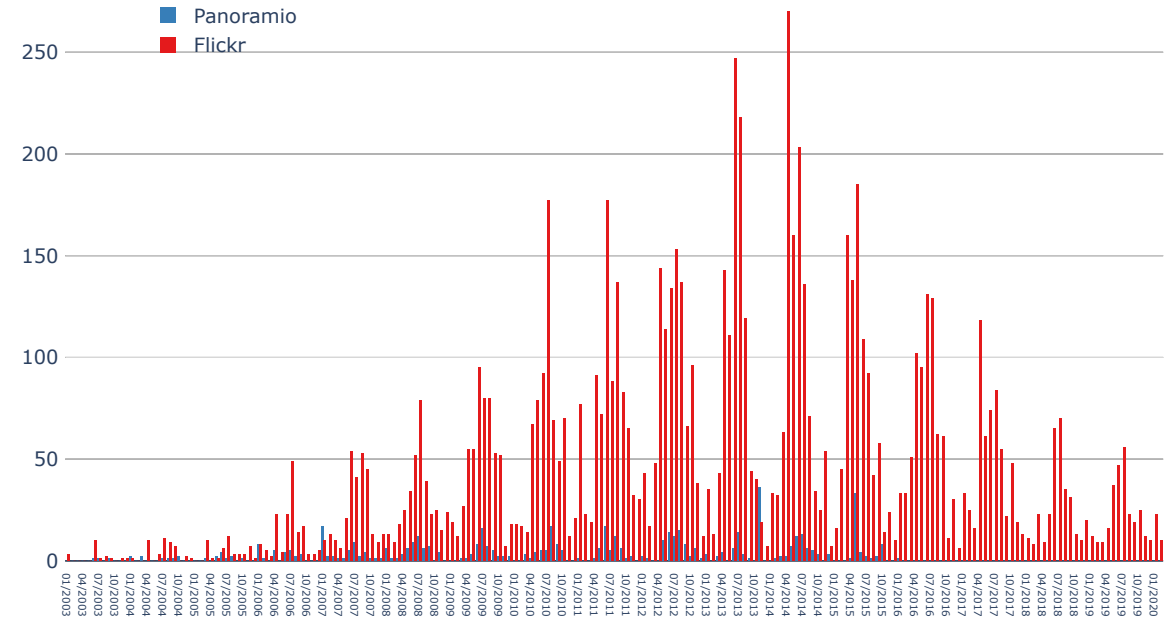


Fig. 23. Date histogram by month and service, complete Panoramio/Flickr dataset, ignoring pictures from before 2003.

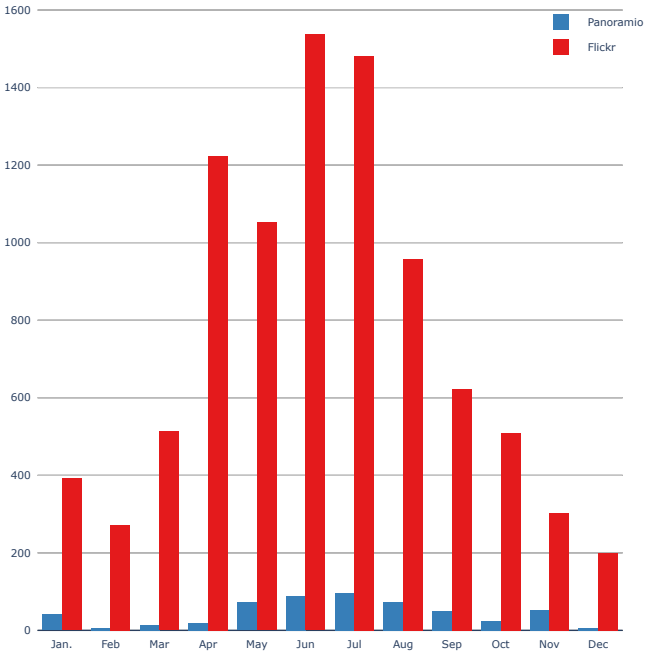


Fig. 24. Aggregated date histogram by month, complete Panoramio/ Flickr dataset.

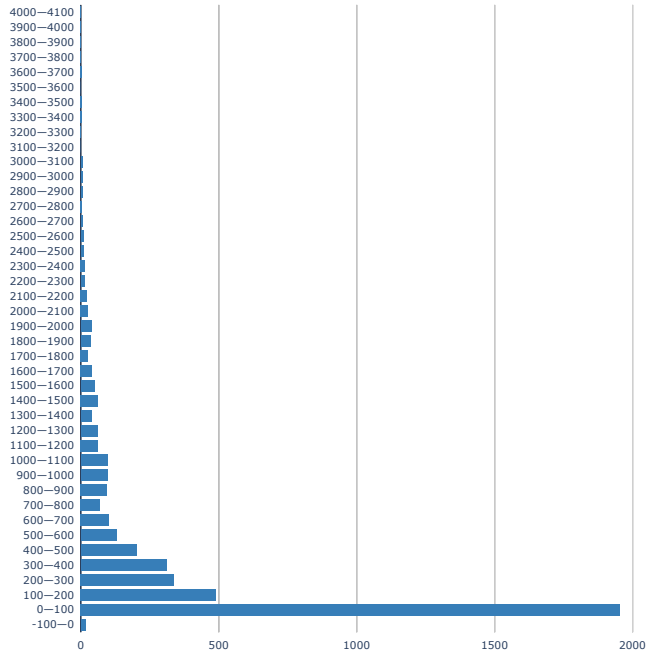


Fig. 25. Elevation histogram, complete Panoramio/Flickr dataset.

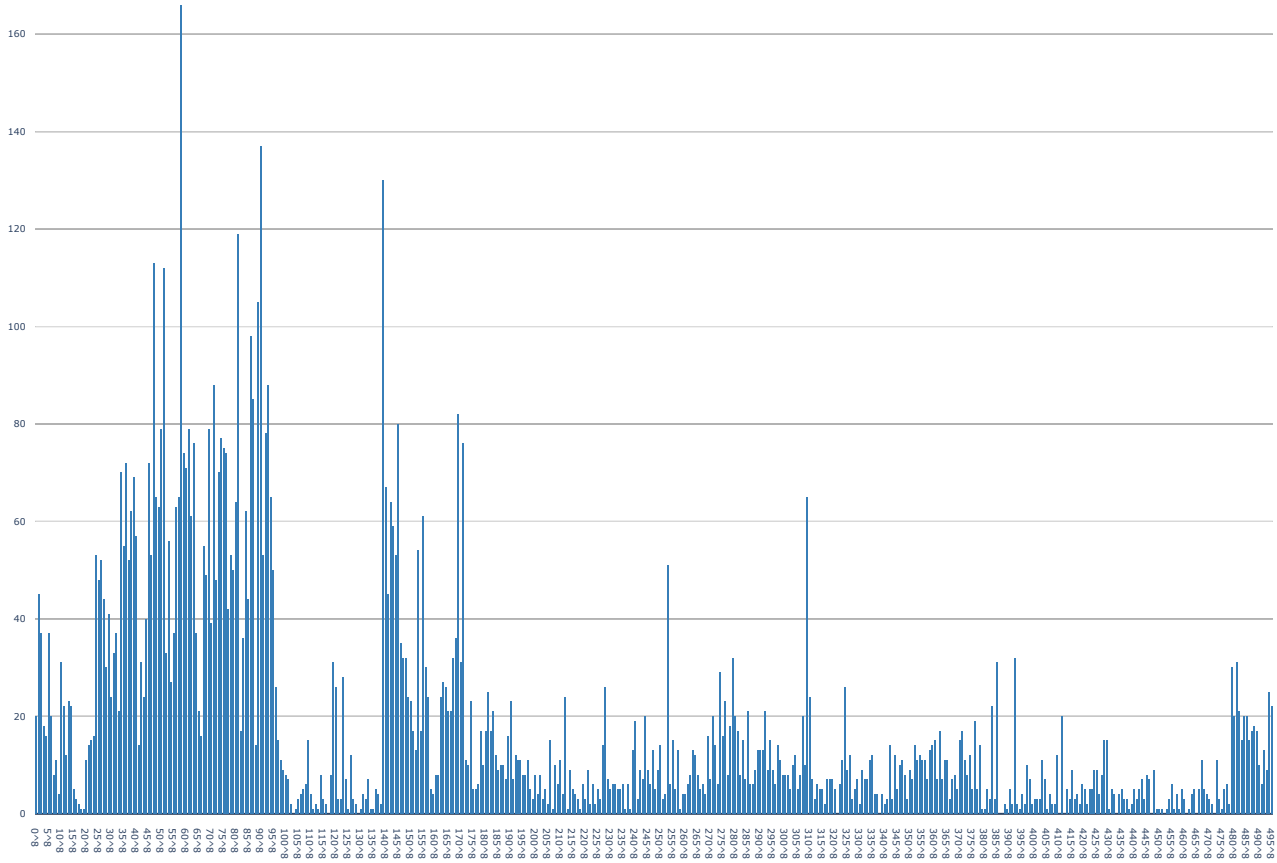


Fig. 26. Histogram of Flickr ID bins for the identified plausible habitat images, complete Flickr dataset. Each bin contains 10^8 Flickr IDs. This plot shows some interesting patterns: We are not clear about the reason for the drop at around $120 \cdot 10^8$ and why fewer entries exist in the more recent bins since all bins are of equal size. These effects may be caused by the employed search strategies for the images using our set of keywords, or potentially it is caused by when we conducted the searches for potential picture IDs on Flickr (which we then cached for later manual inspection).

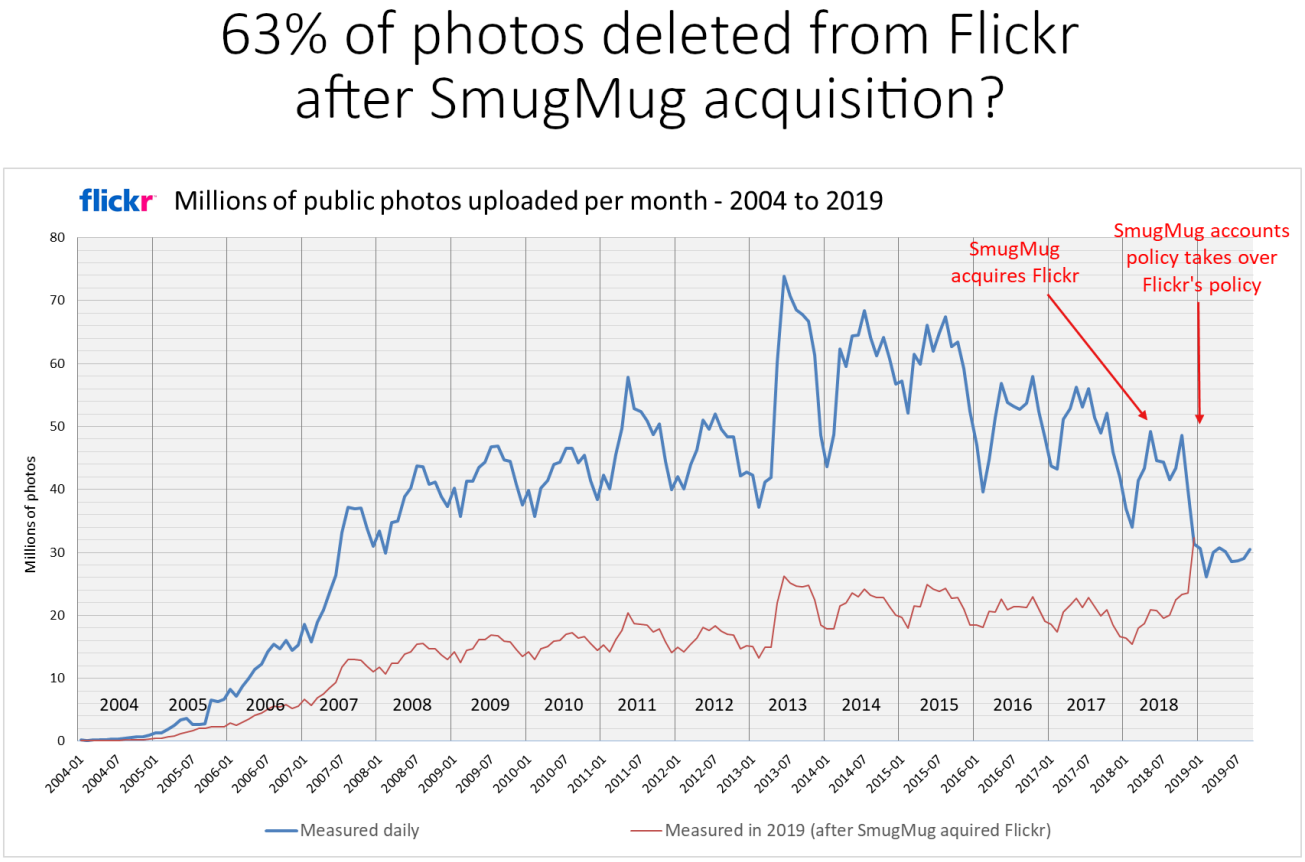


Fig. 27. Flickr's changed upload policies from January 2019 apparently have let quite a number people to delete their images from the service. This fact may have let us to "lose" some images in the process (see Fig. 22) since we do not continuously scrape the Flickr database but instead query it only in irregular intervals. Flickr image 6855169886 by Franck Michel (CC BY 2.0).

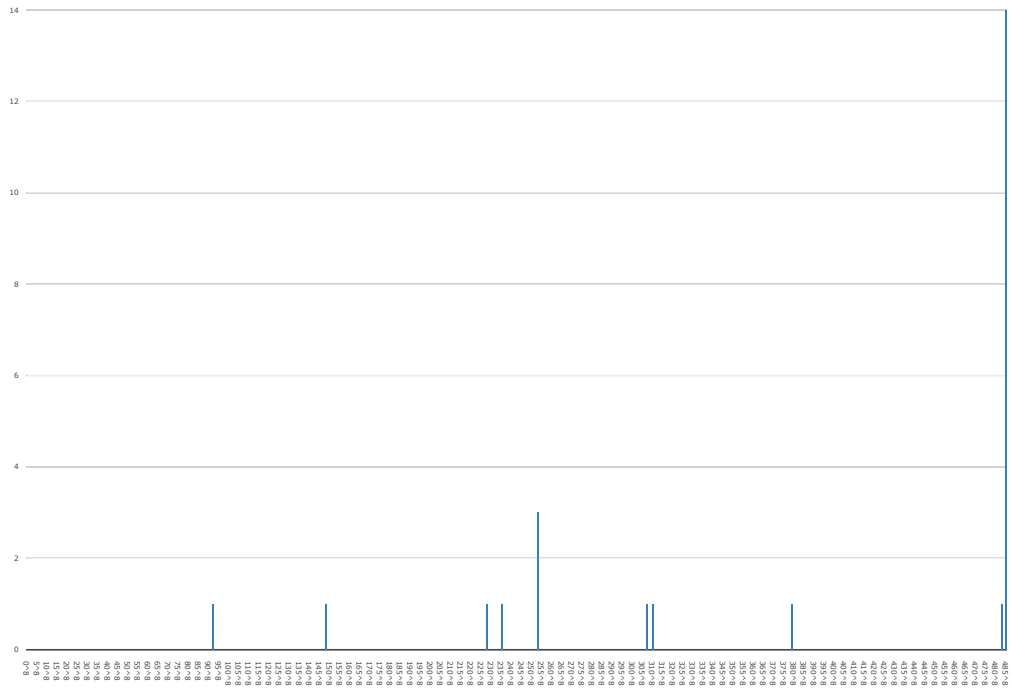
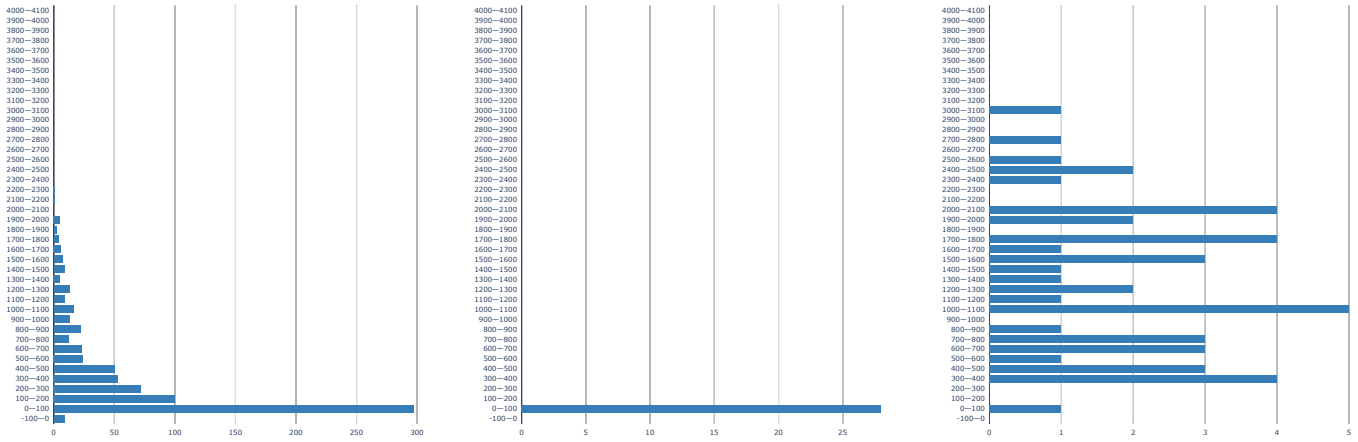
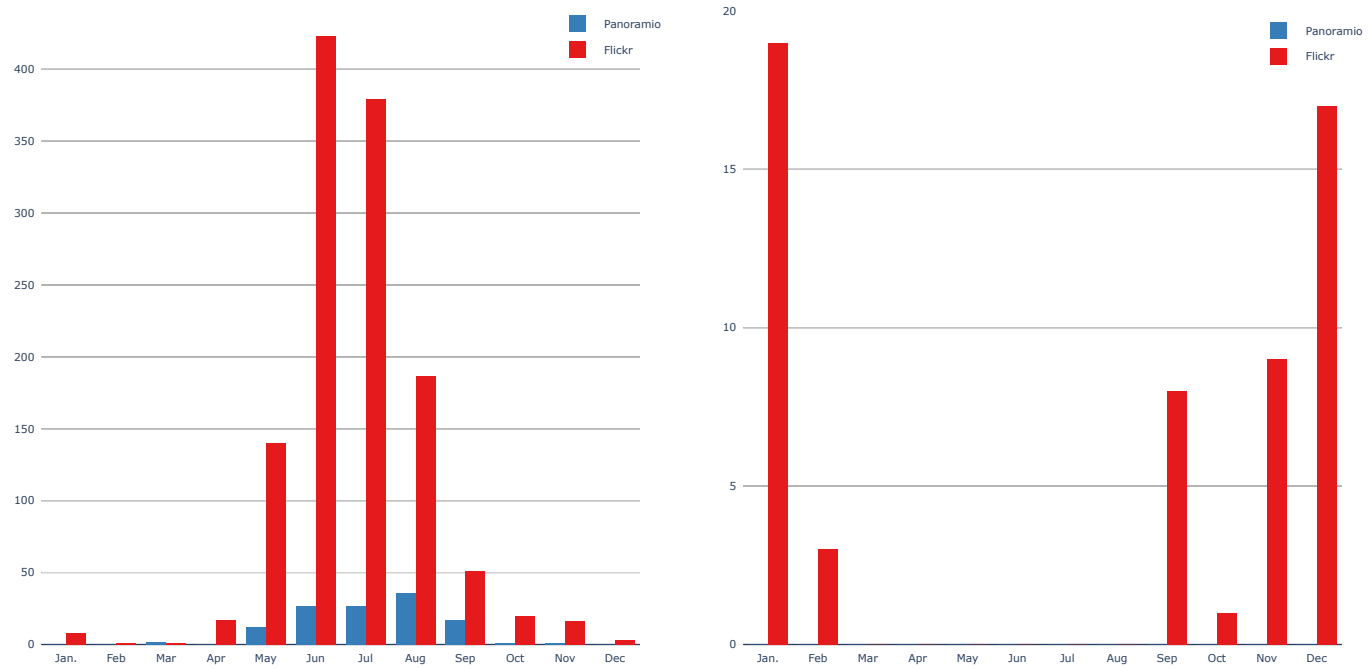


Fig. 28. Histogram of Flickr IDs of unclassified pictures returned by a search a few days after a previous search, both with the identical keyword set. The long peak on the right are newly uploaded images, but the small peaks represent older uploaded images that were not found by the first search.



(a) (b) (c)

Fig. 29. Example for analyzing the elevation distribution in the dataset for different species: Comparison of three elevation histograms with different characteristics for (a) *Drosera rotundifolia*, (b) *Dionaea muscipula*, and (c) *Pinguicula aplina*.



(a) (b)

Fig. 30. Example for analyzing the temporal distribution in a year in the dataset for different species: Comparison of the monthly sightings of two sundew species: (a) *Drosera rotundifolia* which occurs on the Northern Hemisphere and (b) *Drosera arcturi* from the Southern Hemisphere. The two plots clearly show the different growth periods in the Northern and Southern Hemispheres, respectively, but could also be influenced by when people travel to visit the respective habitats.

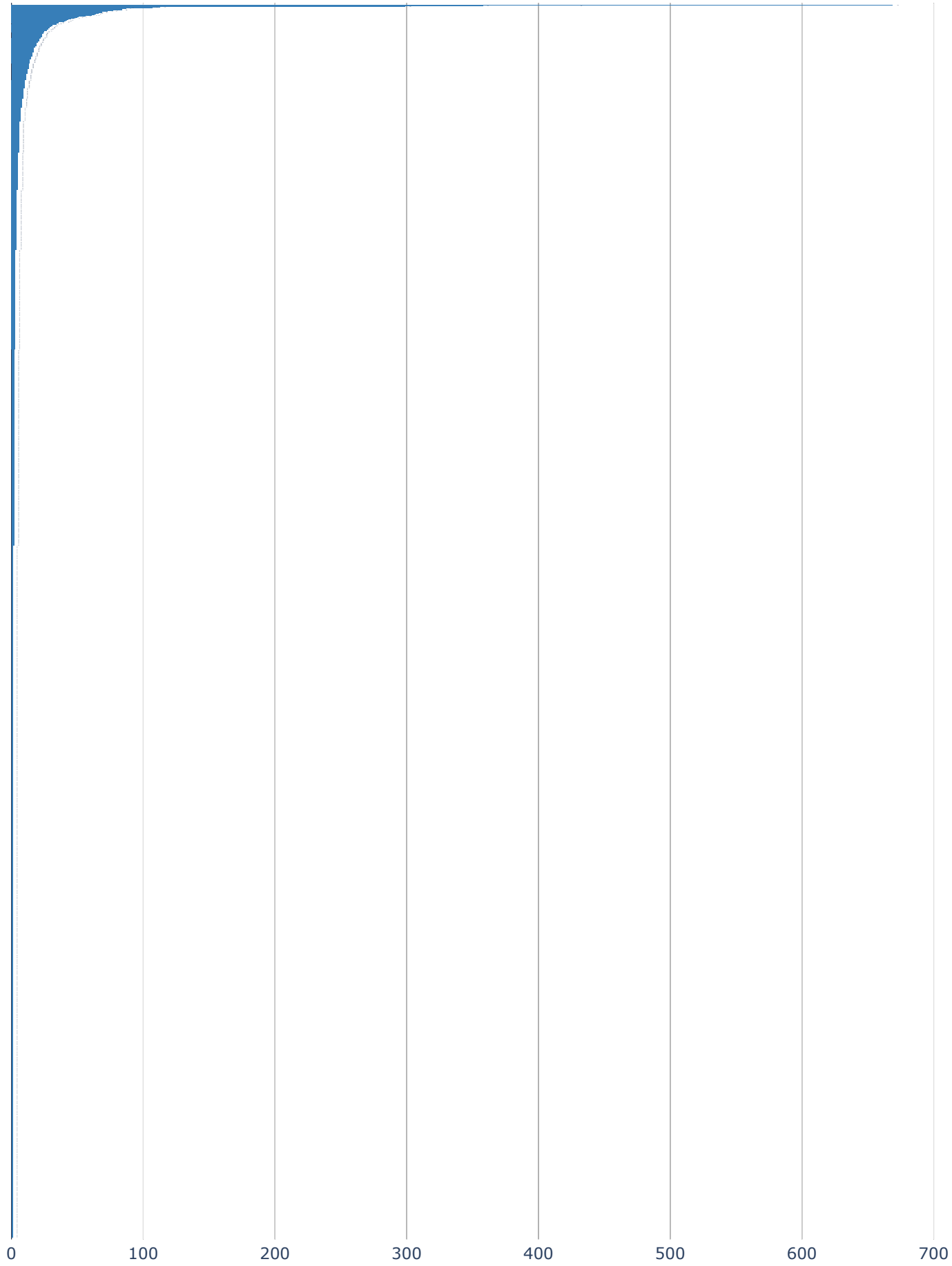


Fig. 31. Observation count by person in the complete database Panoramio/Flickr, sorted by rank.

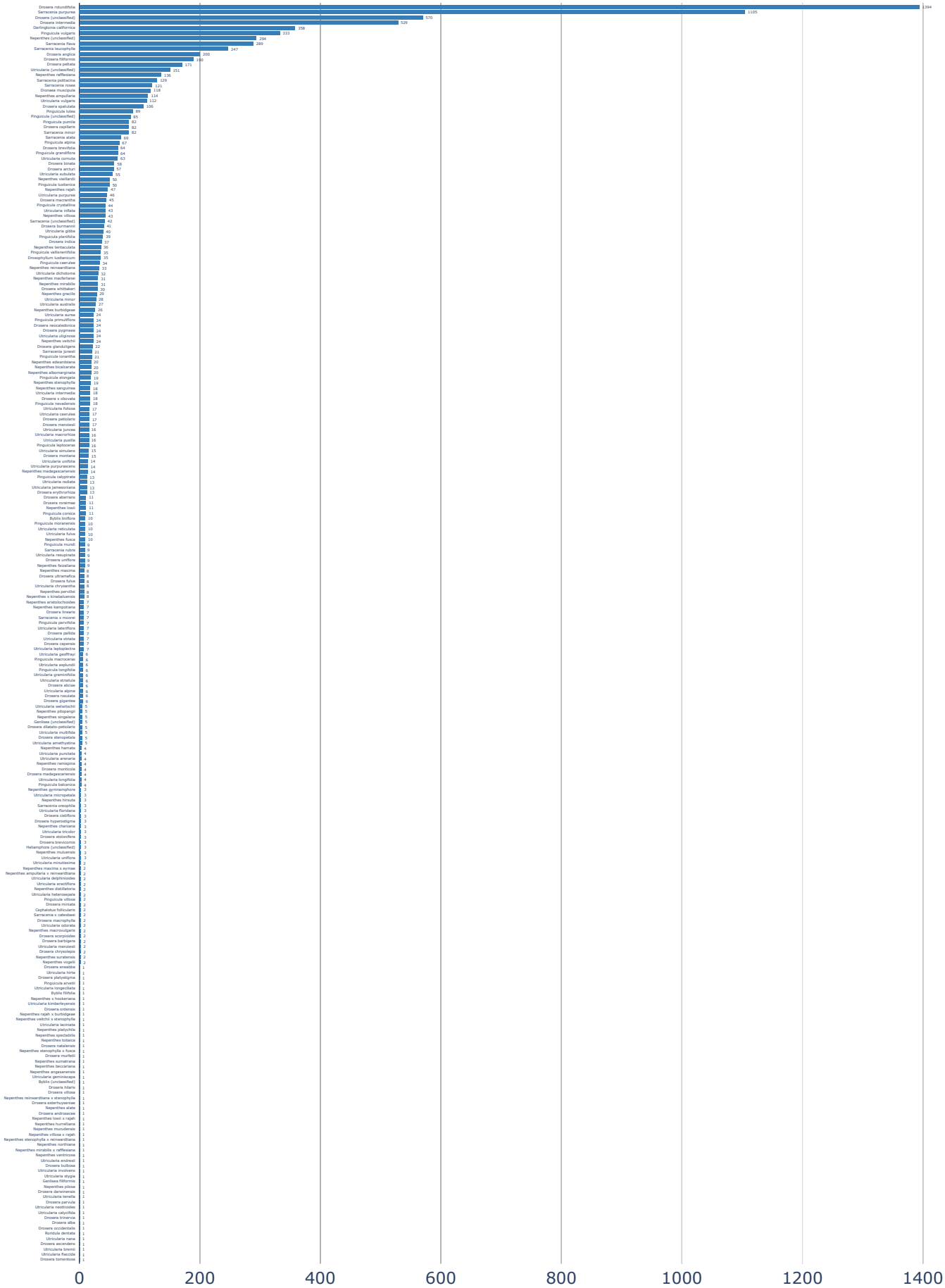


Fig. 32. Species count in the complete Panoramio/Flickr database of all species that we found, sorted by rank.

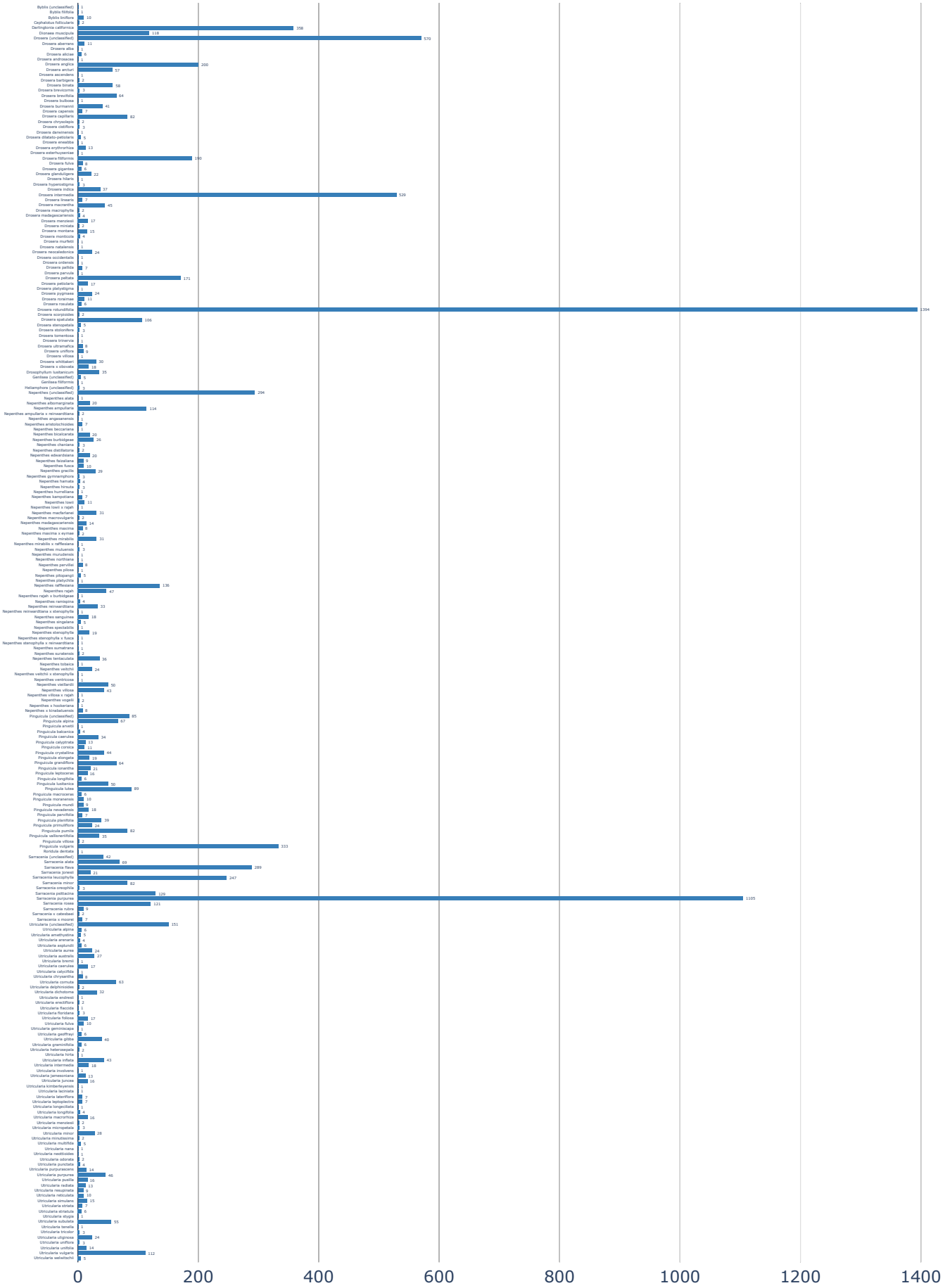


Fig. 33. Species count in the complete Panoramio/Flickr database of all species that we found, sorted by species name.

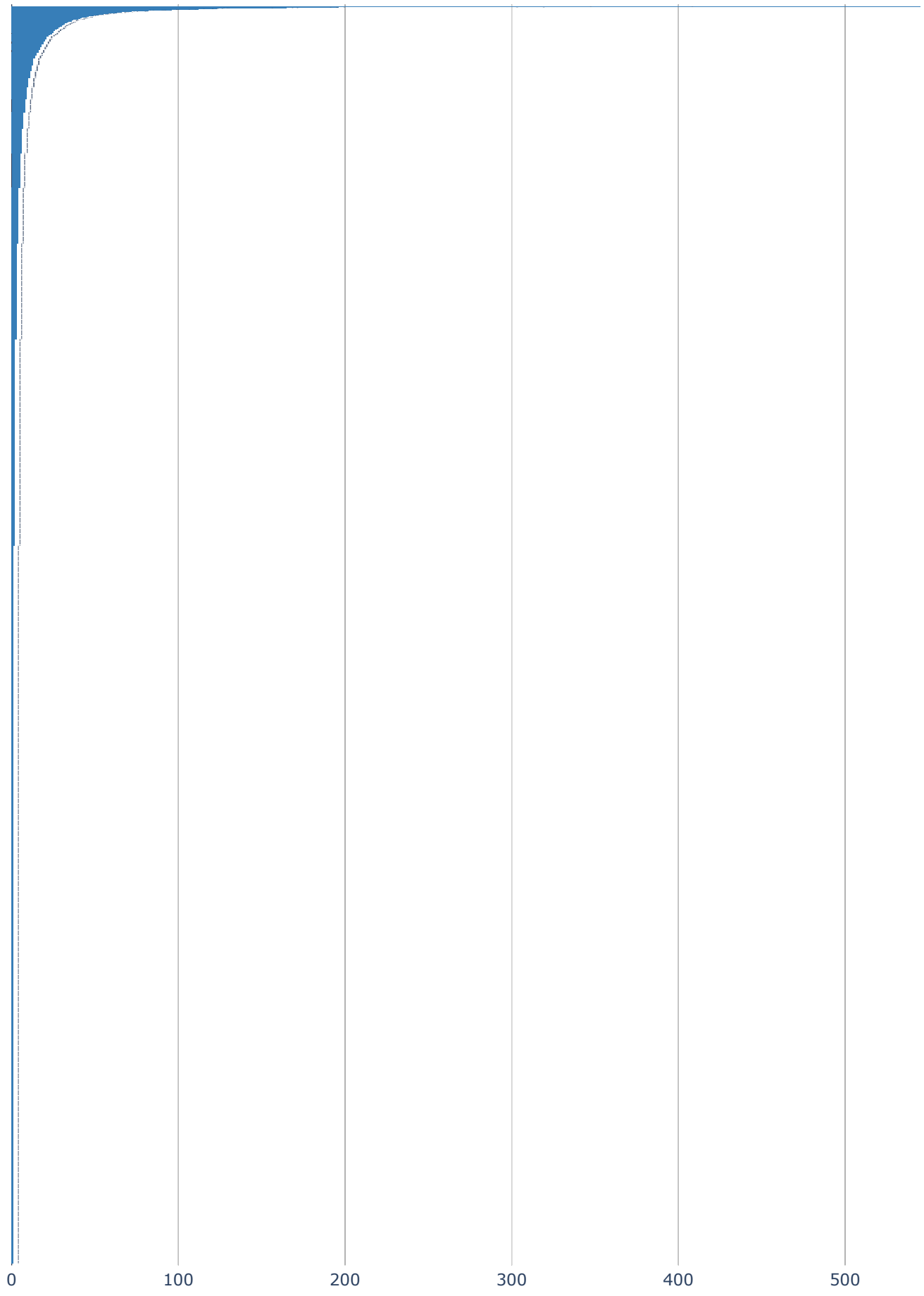


Fig. 34. Observation count by person in the complete database iNaturalist, sorted by rank.

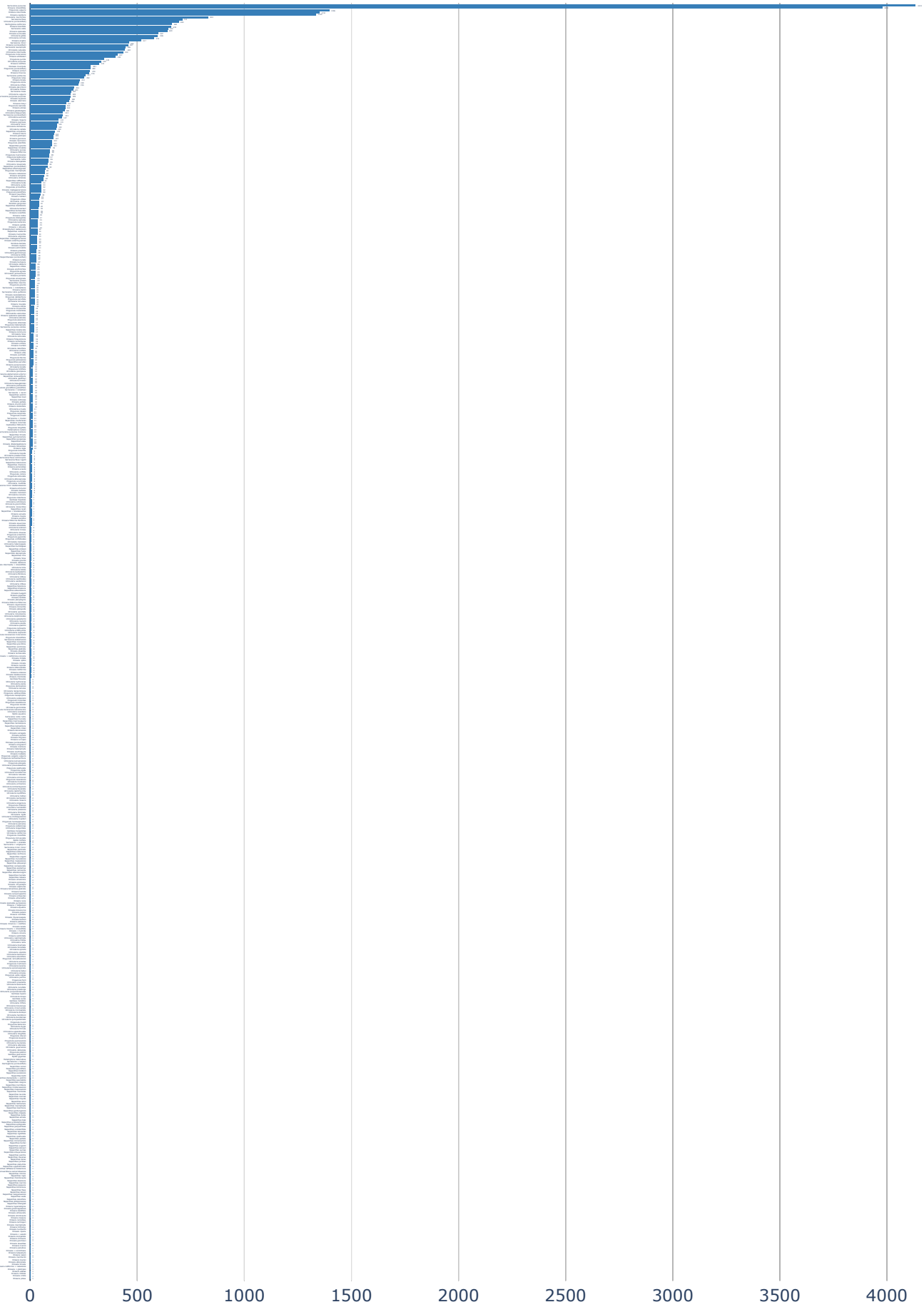


Fig. 35. Species count in the complete iNaturalist database of all species that we found, sorted by rank.

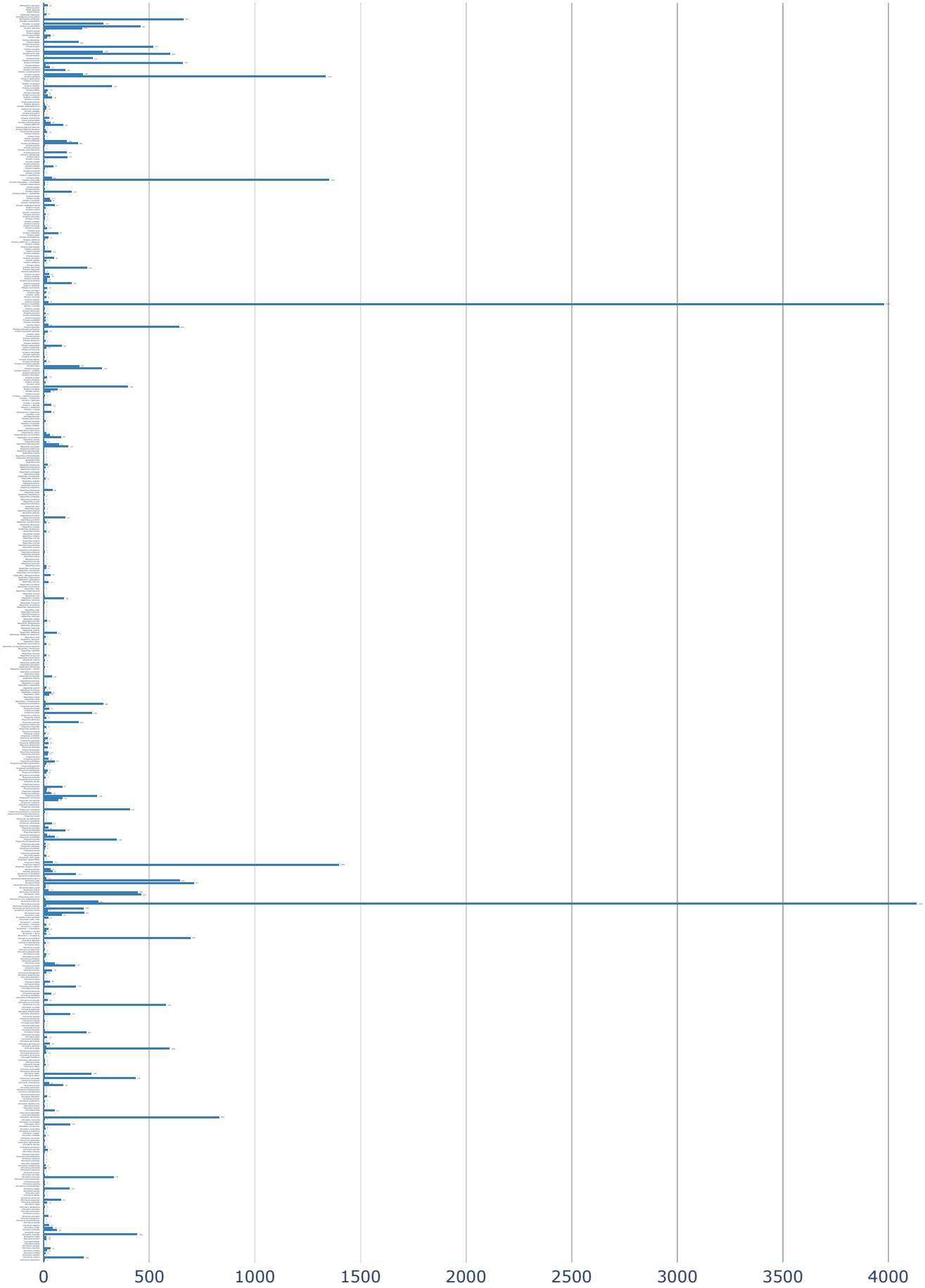


Fig. 36. Species count in the complete iNaturalist database of all species that we found, sorted by species name.

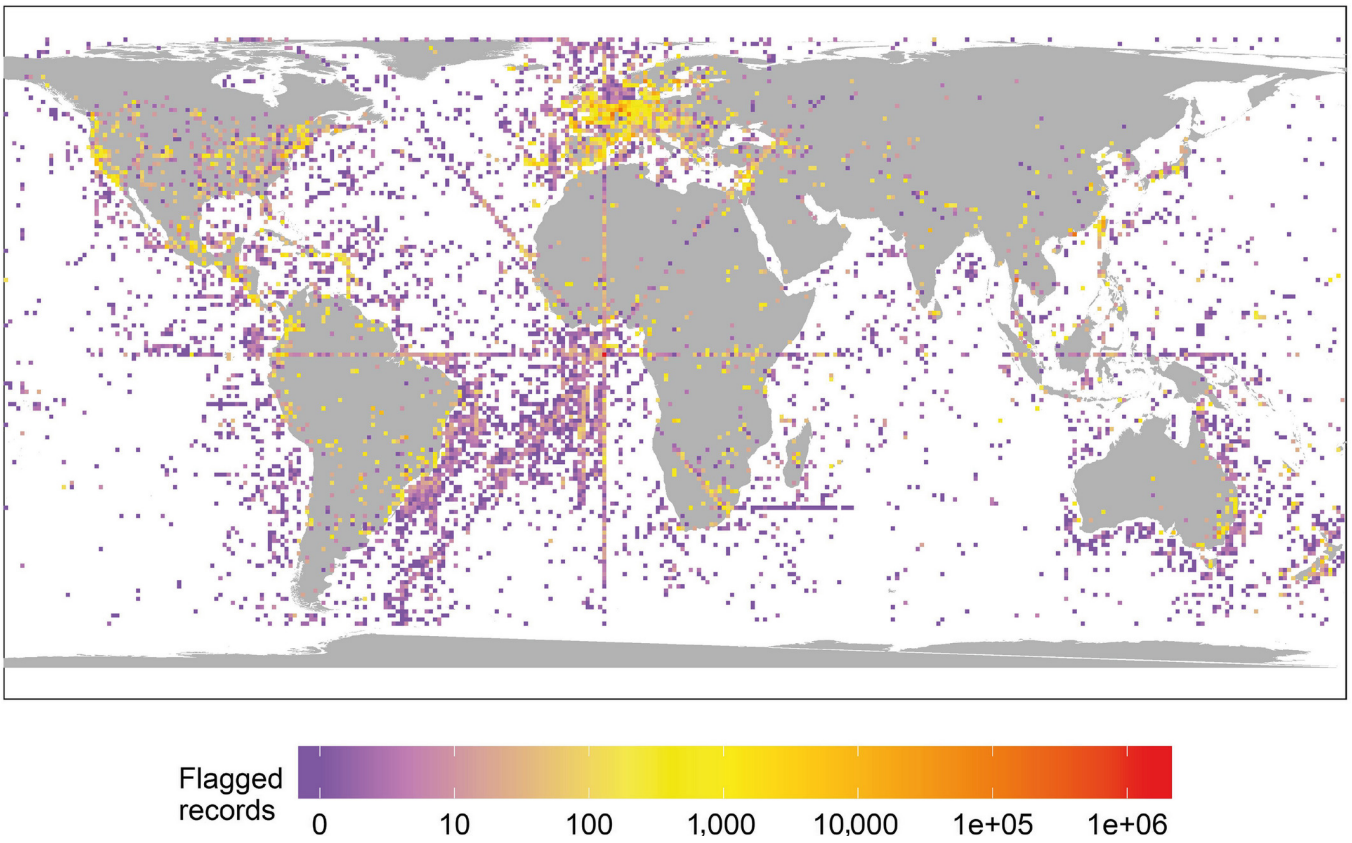


Fig. 37. Records flagged as likely incorrect by Zizka et al.'s CoordinateCleaner tool [76] applied to records for flowering plants from GBIF [49] (image from [76, Fig. 1(a)]; CC BY-NC 4.0). In our analysis we excluded many occurrences in botanical gardens similar to Zizka et al. but also automatically rejected locations in cities and towns that obvisously showed windowsills or gardens, once we saw a first evidence of such records. Moreover, we did not observe cases of zeroed coordinates as shown in this visualization (compare to Fig. 38 and Fig. 39). We also did not see many cases of locations in the oceans; the few that we saw we manually filtered out based on our manual data inspection process described in Sec. 3.

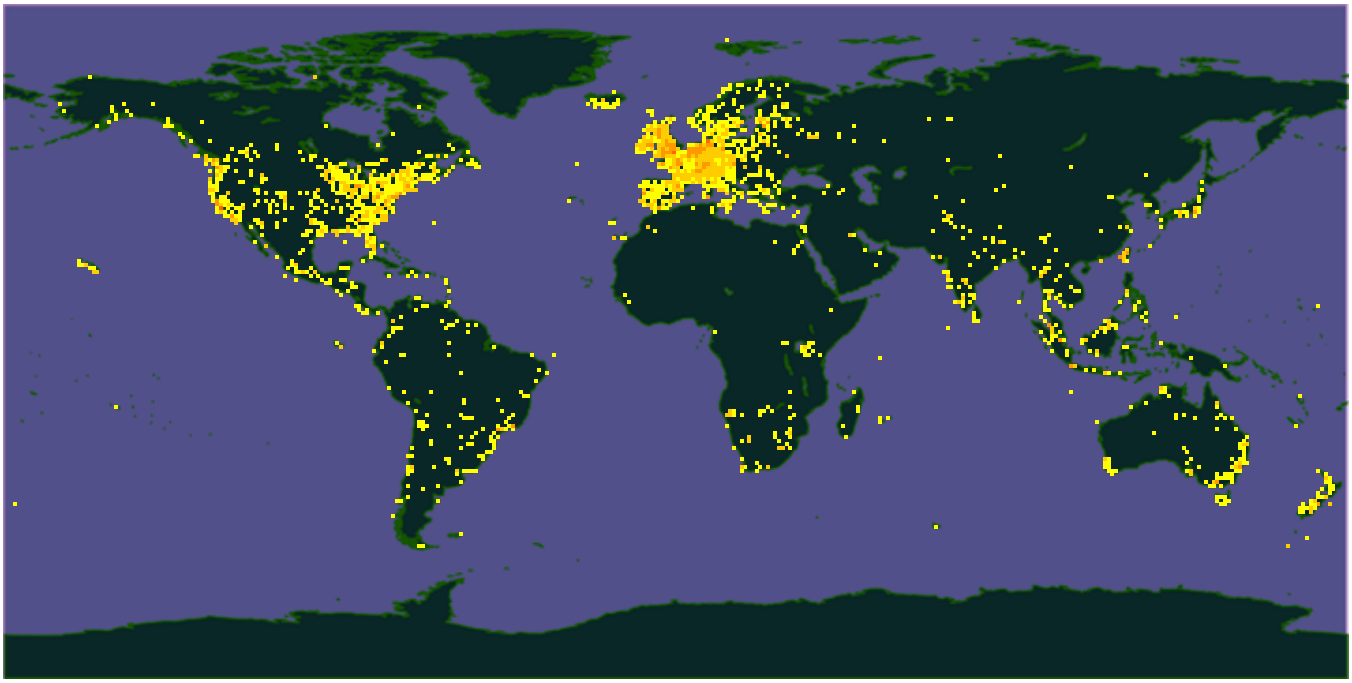
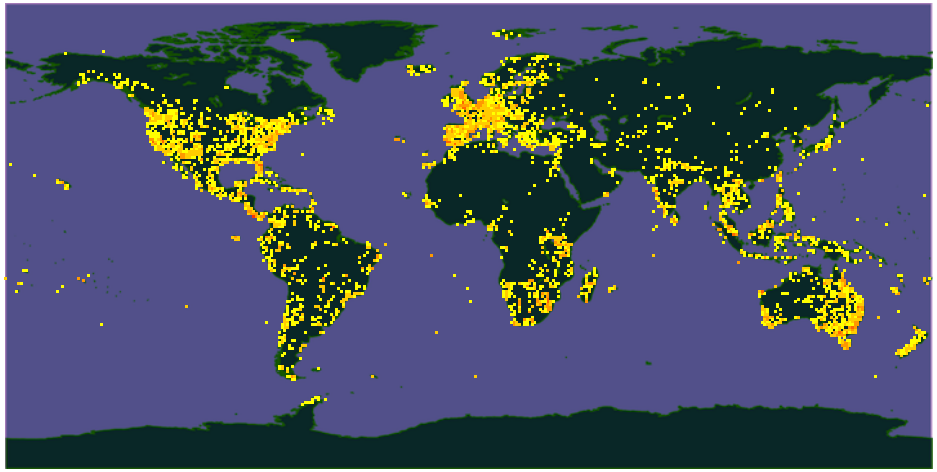
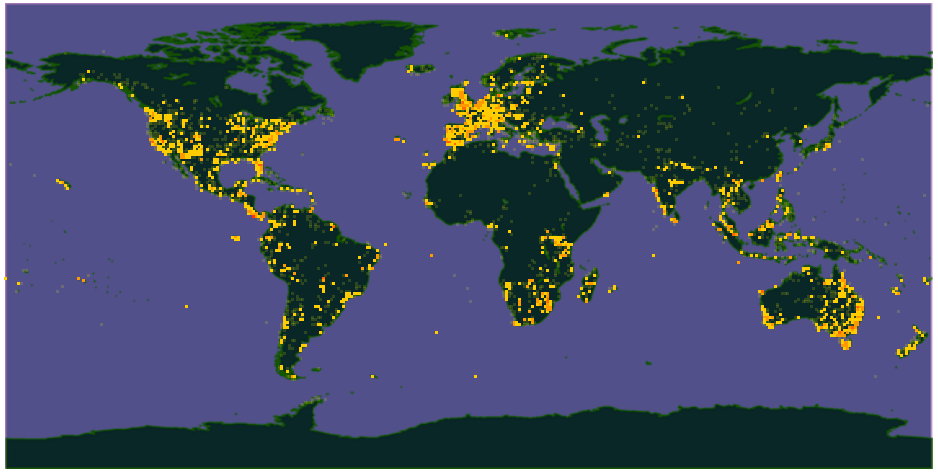


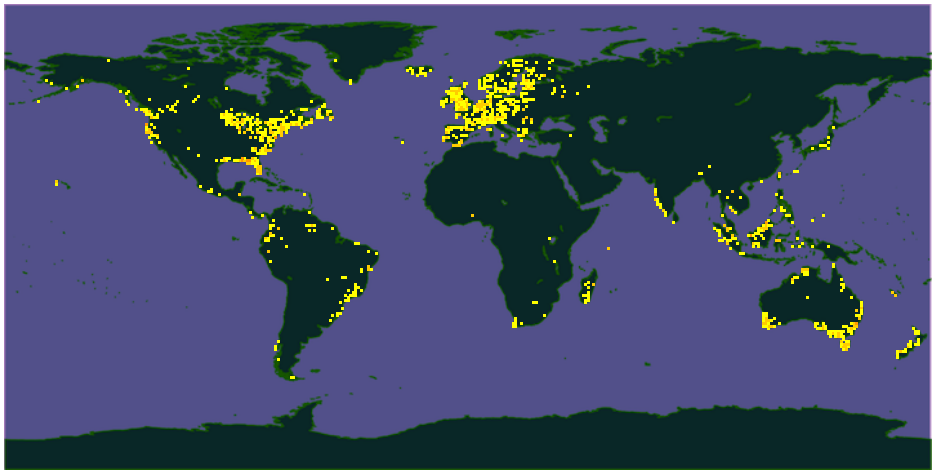
Fig. 38. Heatmap of the 38,169 images from Flickr returned by our search that we visually inspected but which we classified as not showing valid habitats. Note that this set also includes many locations with name/term/tag collisions with subject matter that has nothing to do with carnivorous plants (e. g., due to the many languages we used for the search). Image generated with Page's tool [58], colors indicate image count in a given region: yellow: 1–9; light orange: 10–99; medium orange: 100–999; dark orange: 1,000–9,999; red: $\geq$ 10,000. In regions where the background map is visible not a single image was recorded as ignored. Interestingly, as one can see in the heatmap we did not find patterns of zeroed coordinates, coordinates with identical longitude and latitude, or many locations in the oceans compared to the analysis by Zizka et al. [76] shown in Fig. 37.



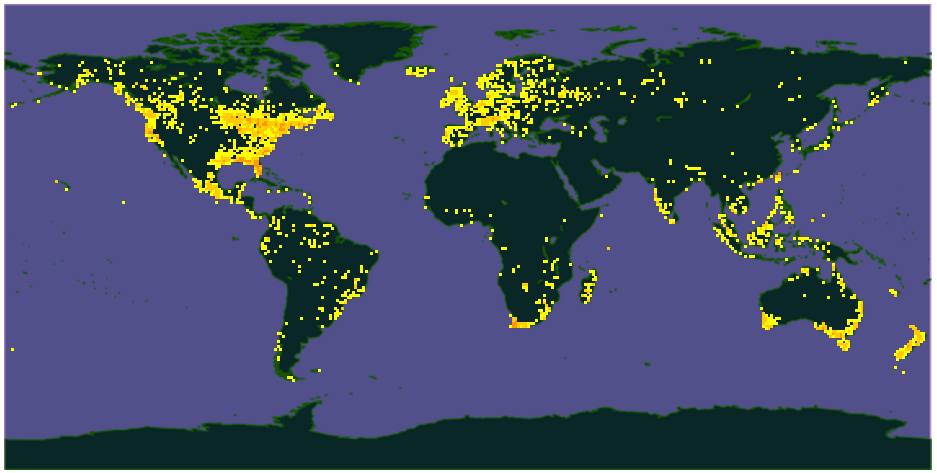
(a) Heatmap of all 172,083 geo-located EOL group posts, up to March 2020.



(b) Image from (a), with the smallest class (<10 images; formerly yellow) shown transparently to better emphasize the locations with multiple image posts.

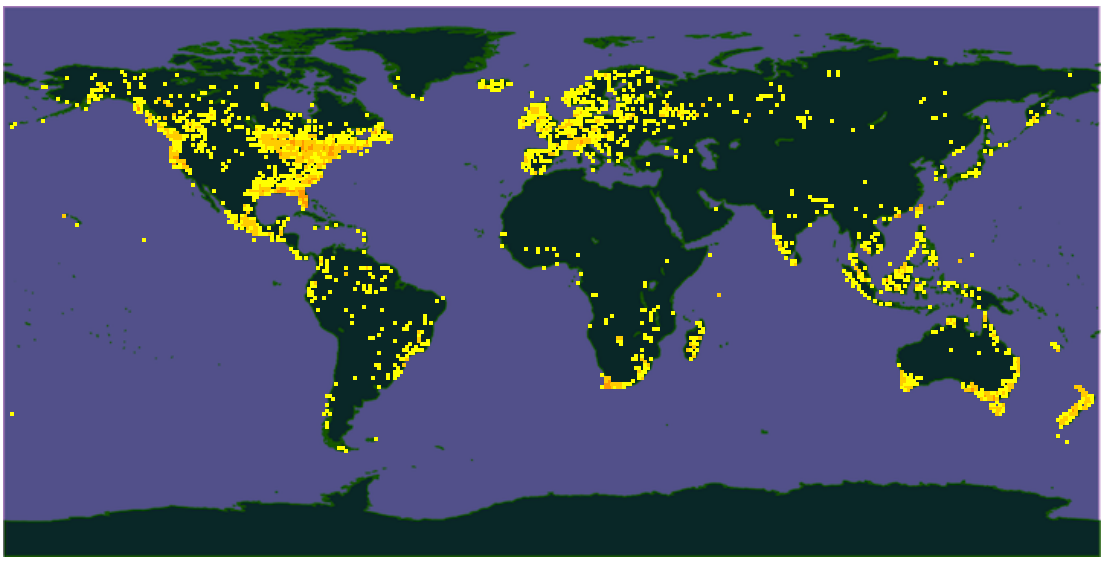


(c) Heatmap for the 9,720 entries of our own dataset (also up to March 2020).

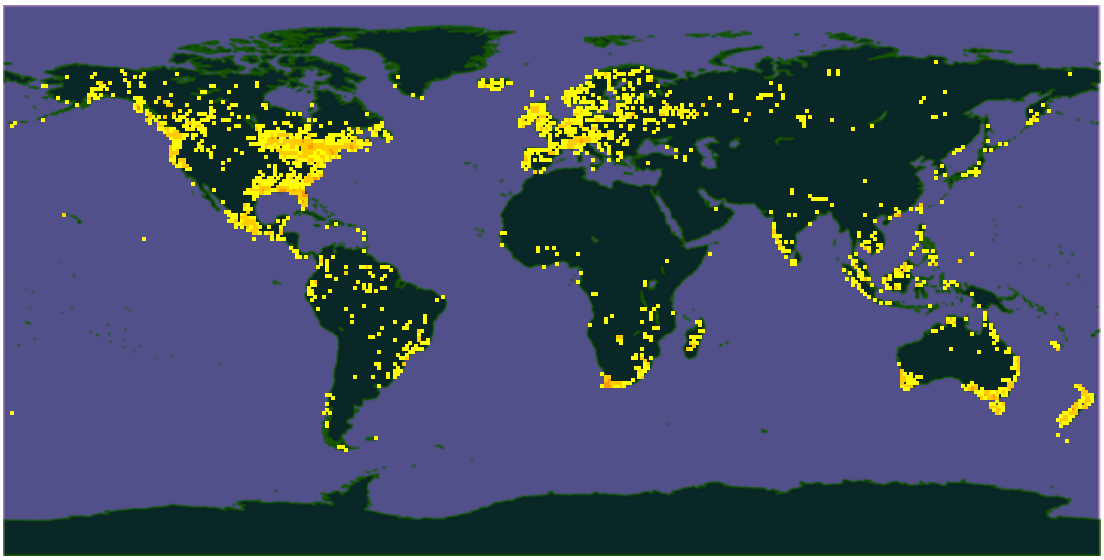


(d) Analysis of the geographic distribution of the iNaturalist carnivorous plant dataset (34,207 geo-located entries as of March 2020).

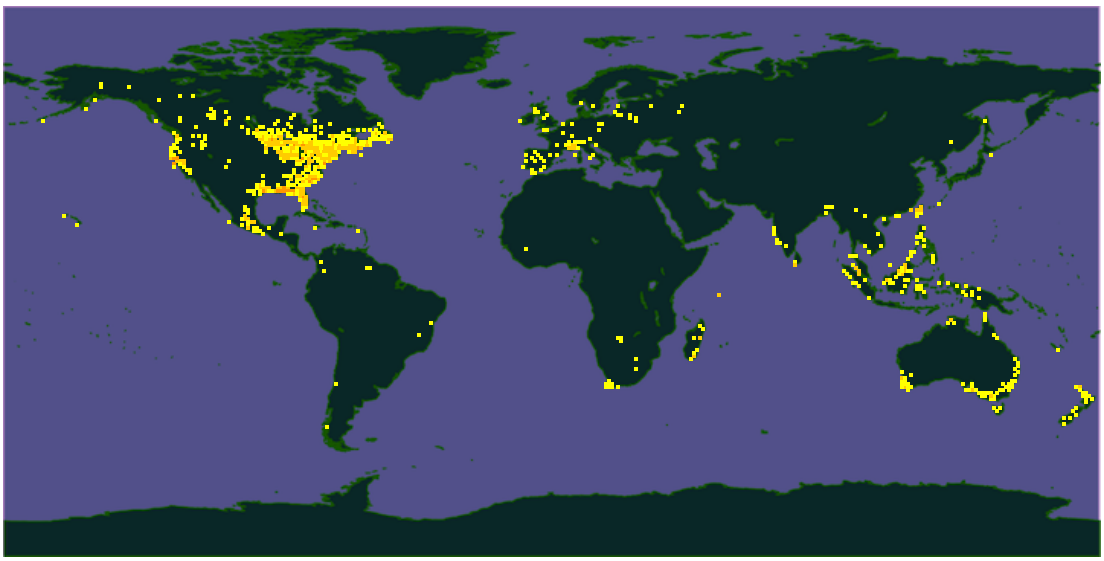
Fig. 39. Combination of larger versions of the images from Fig. 5 and Fig. 18: Geographic distribution analysis based on posts in the *Encyclopedia of Life* (EOL) Flickr group which collects images and videos of organisms: Due to the similar subject matter (for all species in general) we use the maps as an indication of Flickr's geographic bias for habitat pictures [59]. Images generated with Page's [58] tool, colors indicate image count in a given region: yellow: 1–9; light orange: 10–99; medium orange: 100–999; dark orange: 1,000–9,999; red: $\geq$ 10,000. In regions where the background map is visible not a single image was recorded.



(a) Geographic distribution of all observations in the iNaturalist carnivorous plant dataset (34,207 geo-located entries as of March 2020).

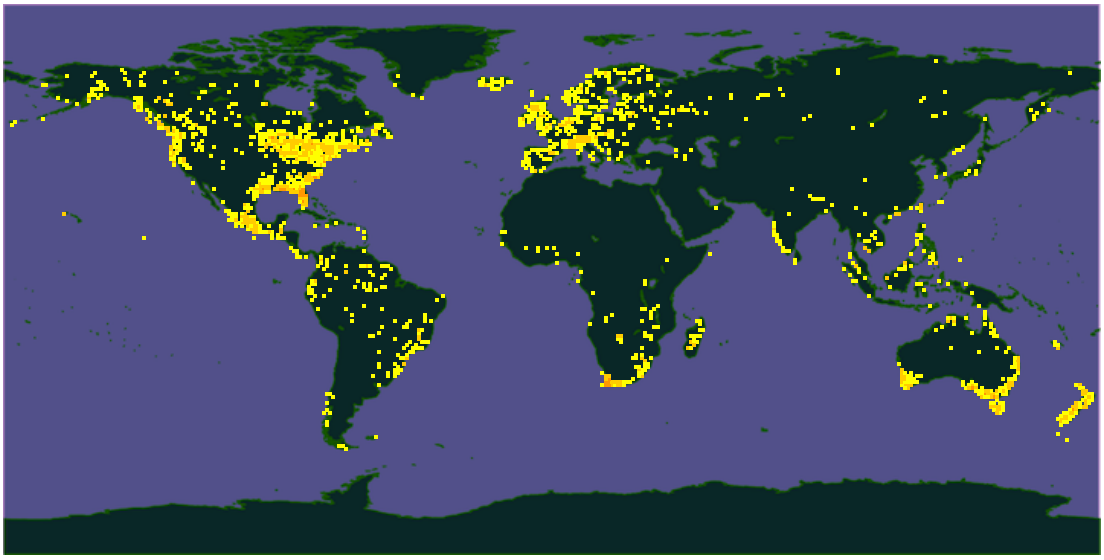


(b) Geographic distribution of the 22,513 precise observations in the iNaturalist carnivorous plant dataset (as of March 2020).

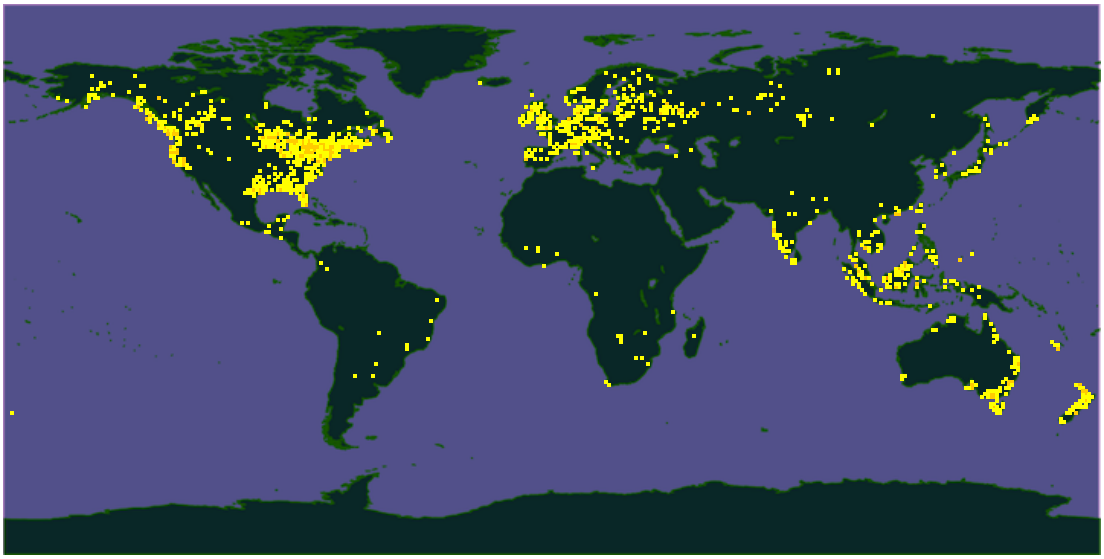


(c) Geographic distribution of the 11,694 obscured observations in the iNaturalist carnivorous plant dataset (as of March 2020).

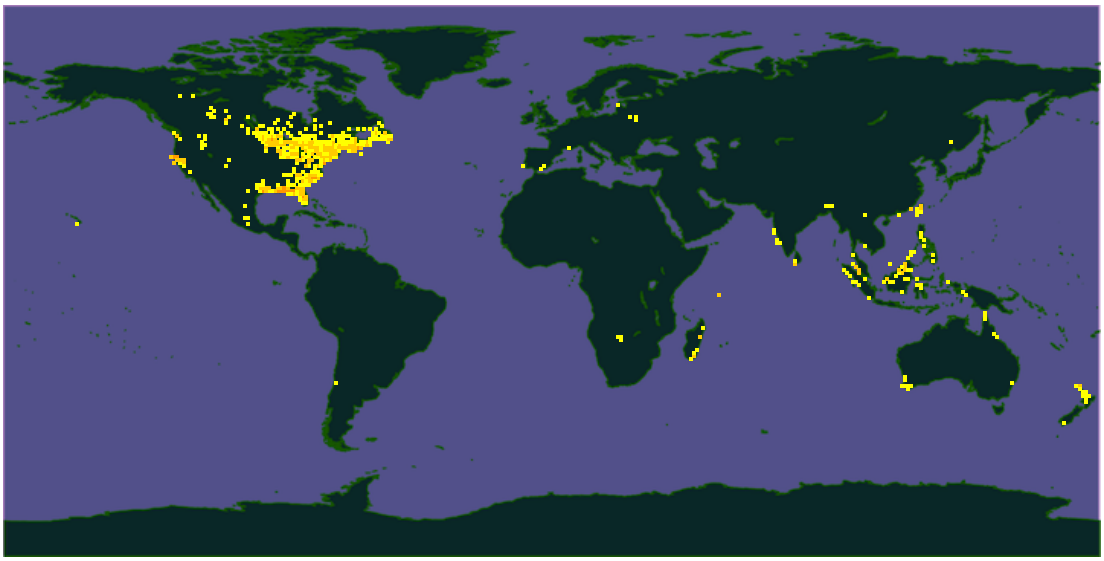
Fig. 40. Larger version of the images in Fig. 18 and Fig. 19 of the global distribution of entries in the iNaturalist carnivorous plant dataset.



(a) Entries for which `taxon_geoprivacy` is empty. Likely the same meaning as "open."

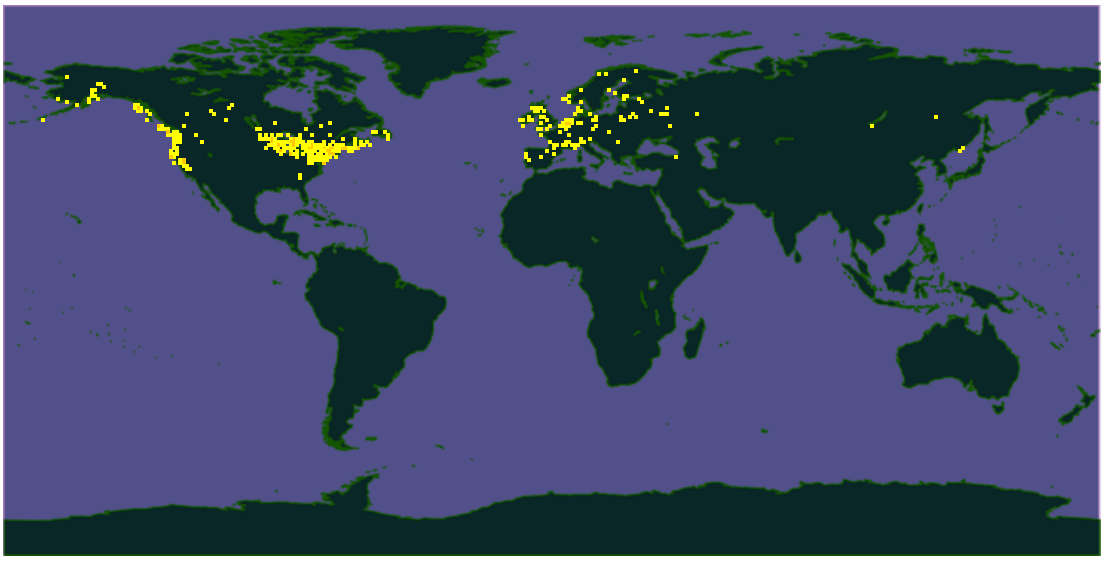


(b) Entries for which `taxon_geoprivacy` is "open."

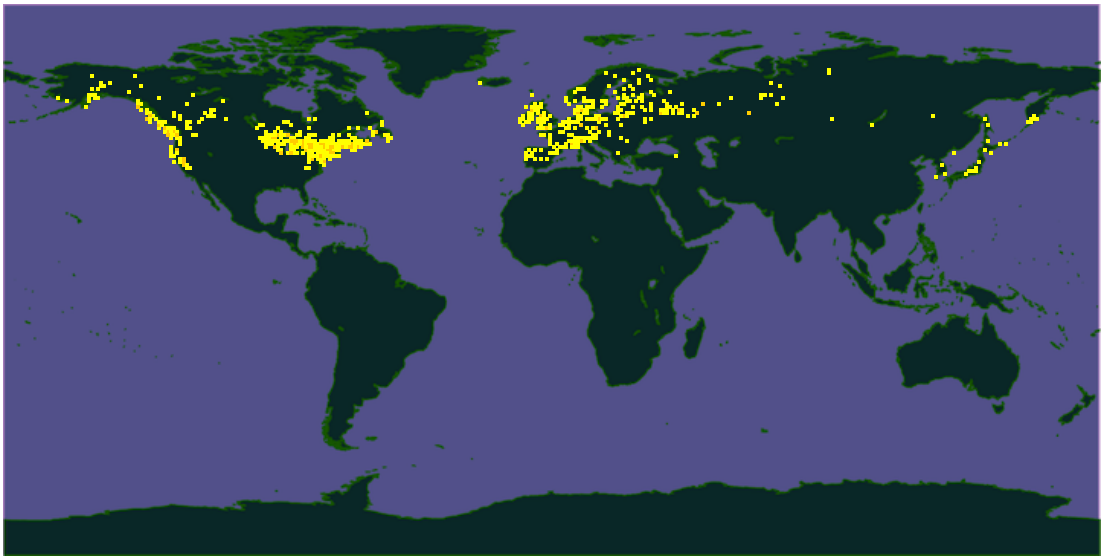


(c) Entries for which `taxon_geoprivacy` is "obscured."

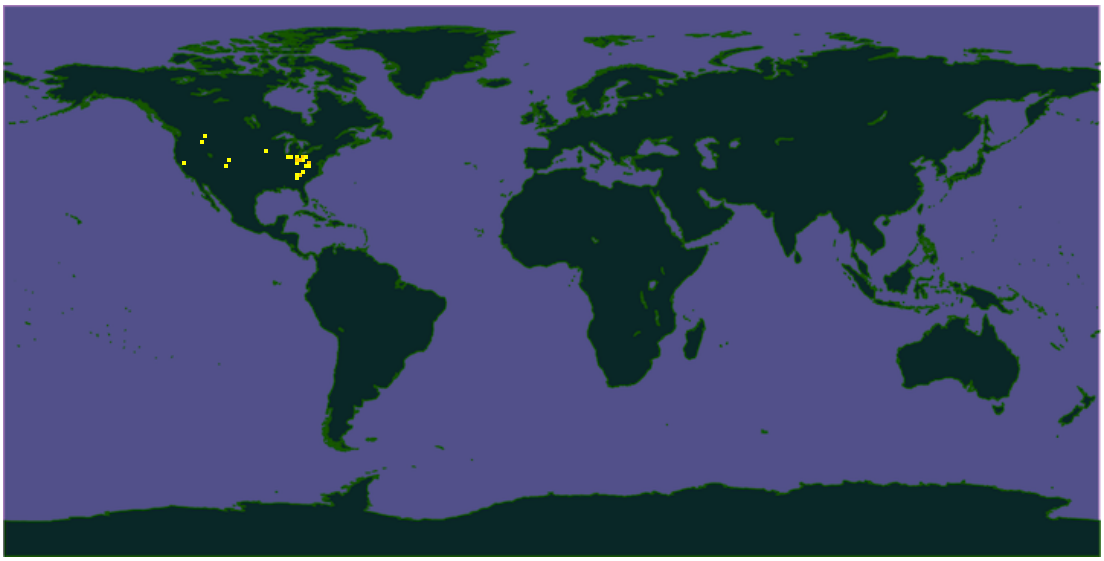
Fig. 41. Comparison of entries in the iNaturalist carnivorous plant dataset on the use of the `taxon_geoprivacy` tag. It seems that the `taxon_geoprivacy` tag is used much more frequently in North America.



(a) Entries for *Drosera rotundifolia* for which `taxon_geoprivacy` is empty. Likely the same meaning as "open."

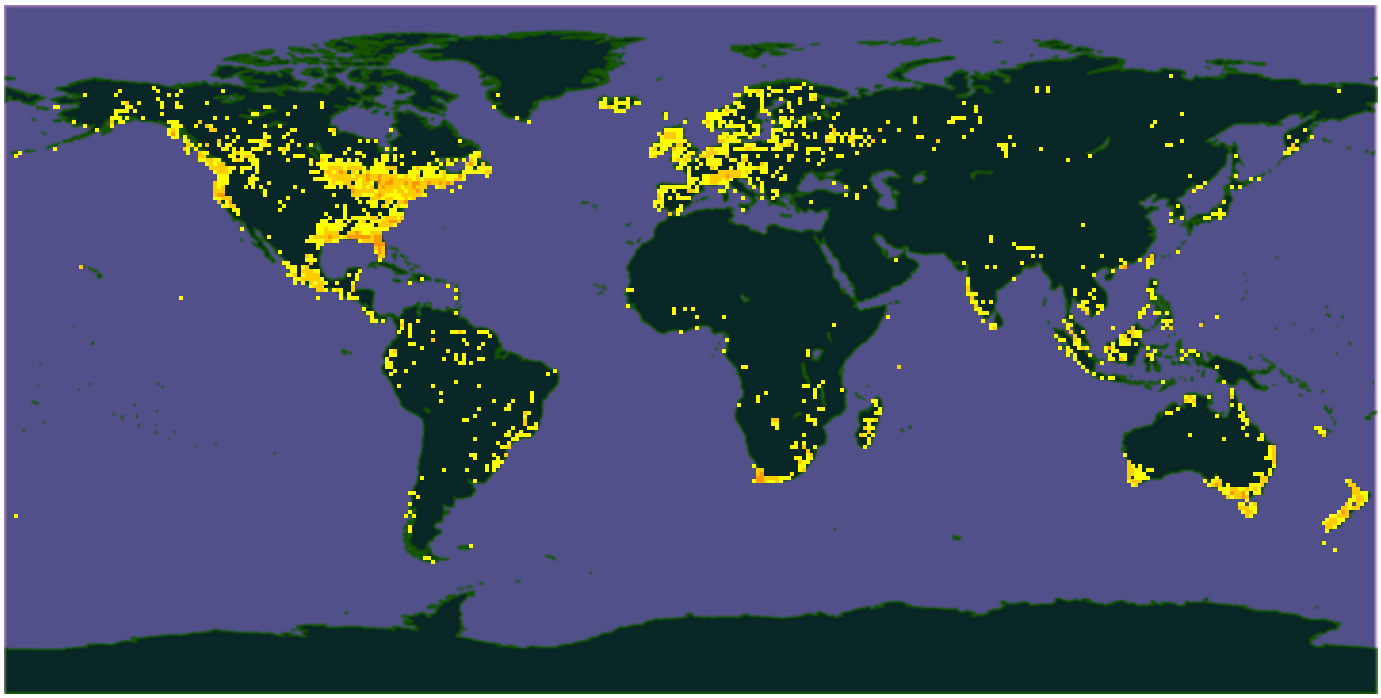


(b) Entries for *Drosera rotundifolia* for which `taxon_geoprivacy` is "open."

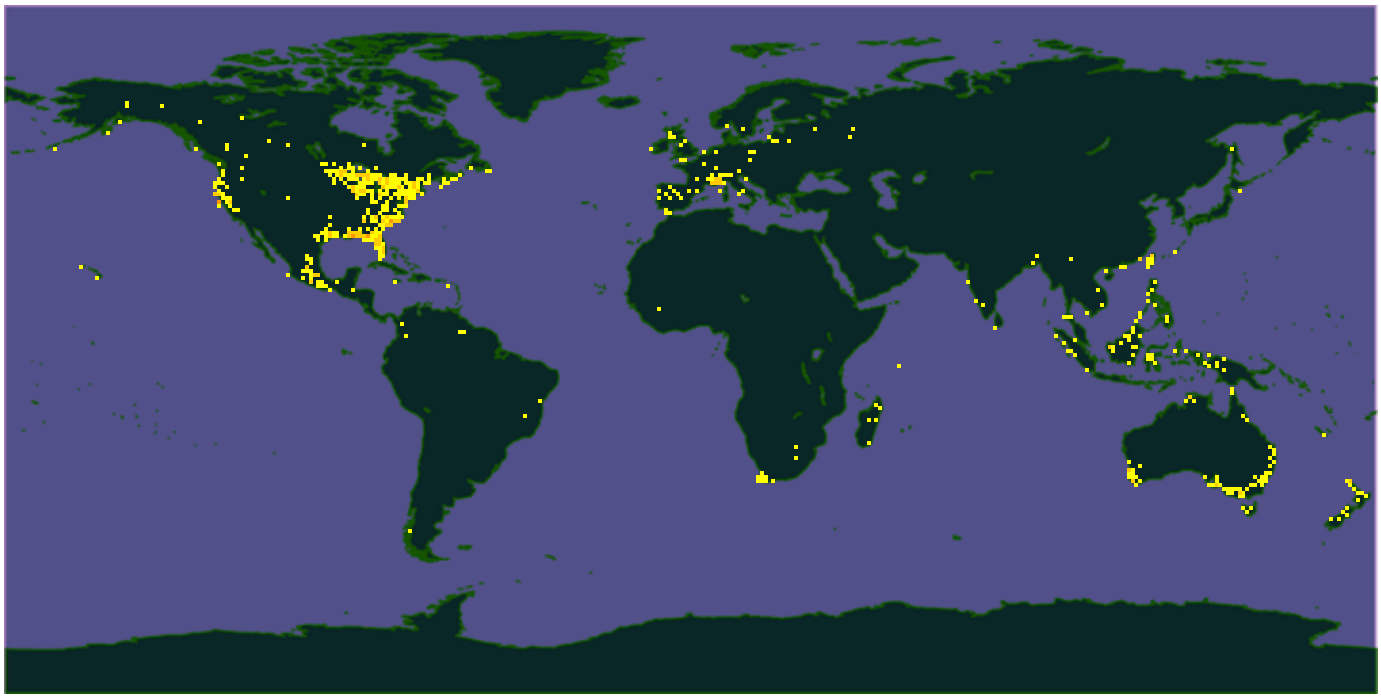


(c) Entries for *Drosera rotundifolia* for which `taxon_geoprivacy` is "obscured."

Fig. 42. Similar comparison of entries in the iNaturalist carnivorous plant dataset on the use of the `taxon_geoprivacy` tag (as in Fig. 41), but only for *Drosera rotundifolia* (the species with the second-most entries in the dataset that grows more or less in the entrire northern hemisphere, see Fig. 4). It seems that the `taxon_geoprivacy` tag is used in a location-specific way because there are literally no obscured entries in Europe and Asia.



(a) Entries for which `geoprivacy` is empty, which means "open."



(b) Entries for which `geoprivacy` is "obscured."

Fig. 43. Comparison of entries in the iNaturalist carnivorous plant dataset on the use of the `geoprivacy` tag which allows contributors to obscure the locations of their observations (shown as a random point within a 0.2 by 0.2 degree area that contains the true coordinates) or make them completely private (not shown at all). We do not show the latter here because the respective entries do not include latitude and longitude data (1.2% of the non-cultivated entries). It seems that the "obscured" option is used more by people in North America than in the rest of the world.

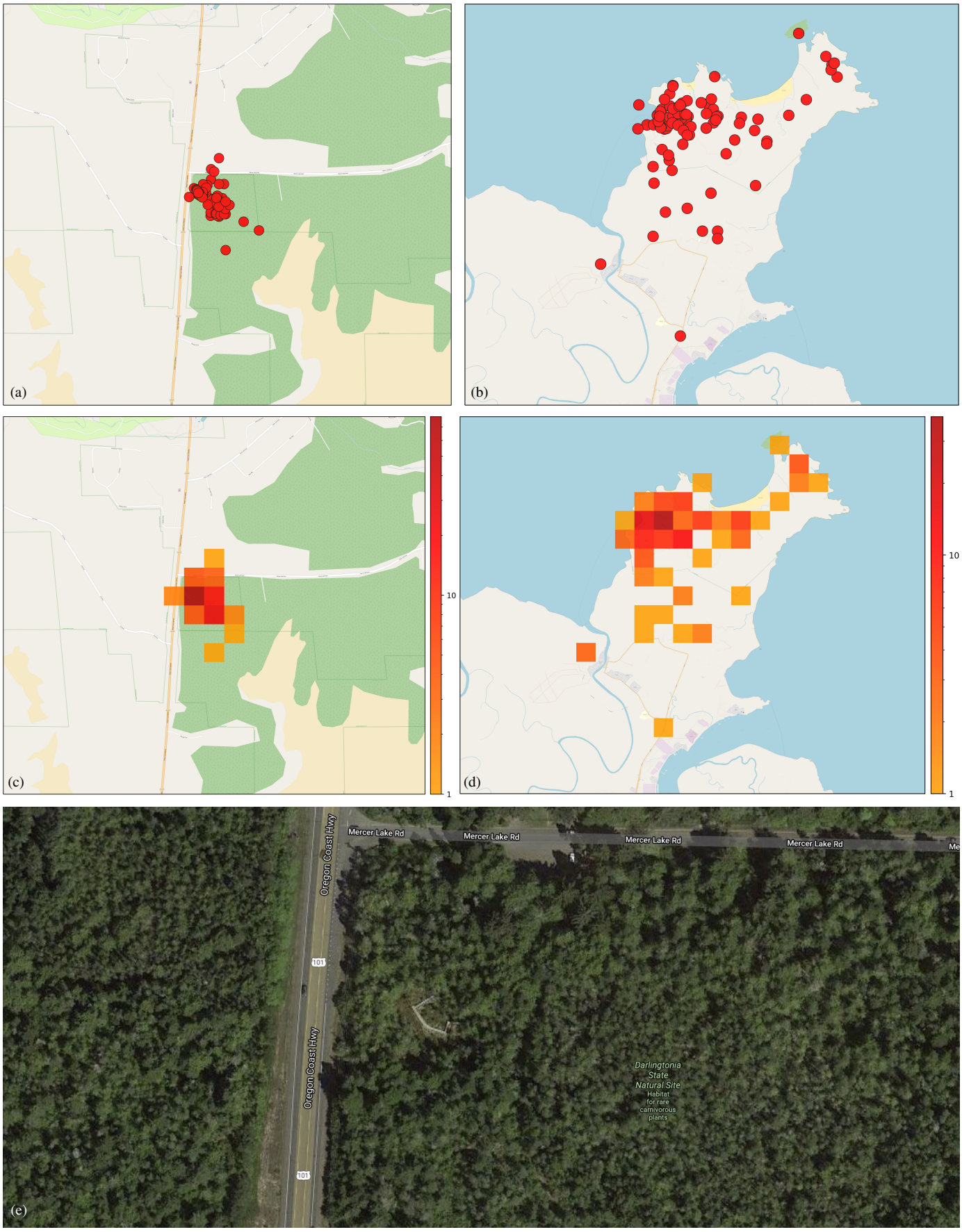


Fig. 44. Larger version of Fig. 8 (and comparison with a heatmap representation, all for the Panoramio/Flickr data): Local geographic bias due to the certain spots such as natural parks being more popular for visitors than others: (a),(c) Darlingtonia State Natural Site in the United States and (b),(d) Bako National Park in Malaysia. Interestingly, the case of Darlingtonia State Natural Site in (a),(c) also illustrates the GPS precision of our data: the plants are located in only a very small region such that the spread of the observations' geo-locations gives an indication of the expected precision of the GPS data, despite the presence of trees in the area as shown in (e).

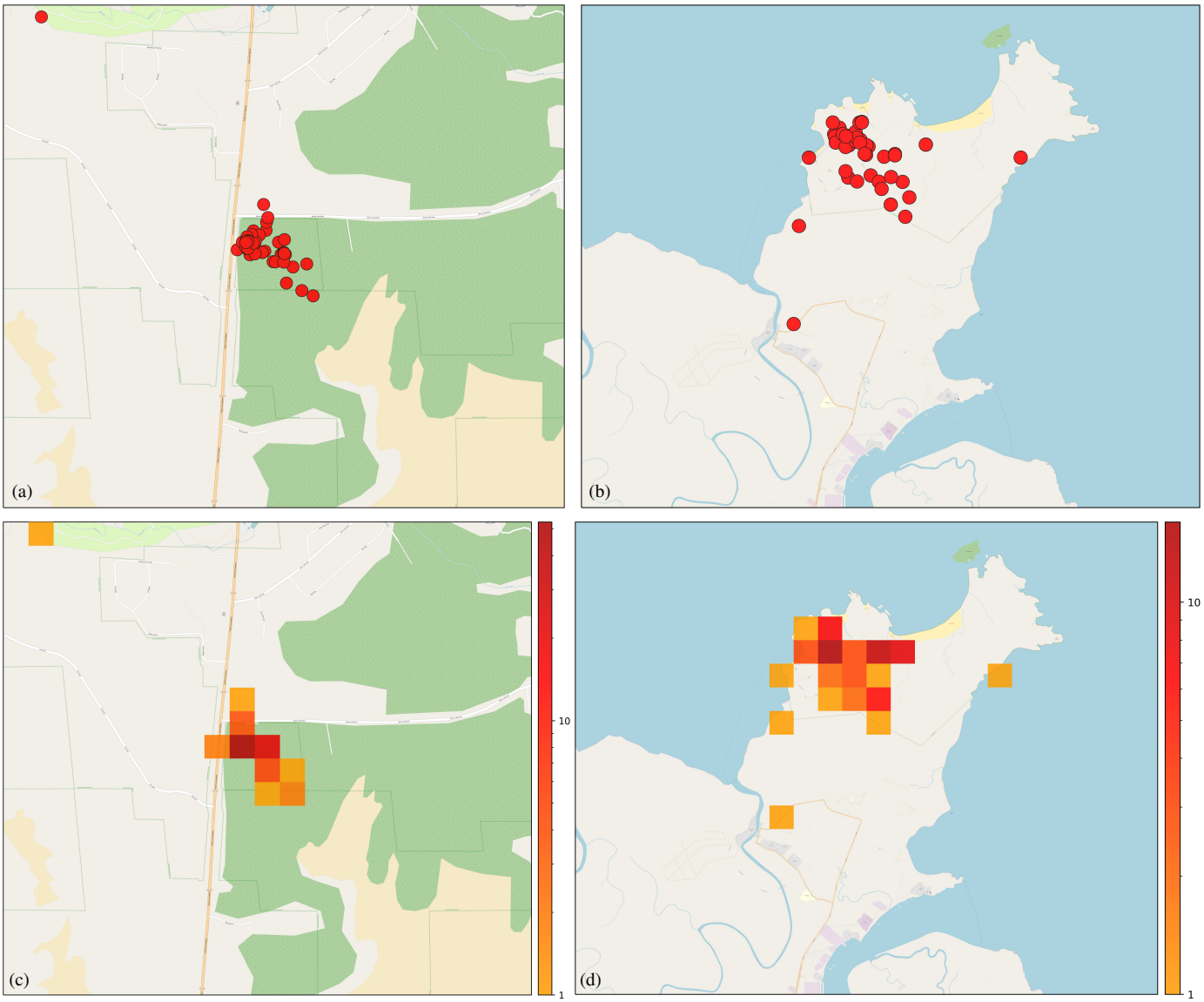
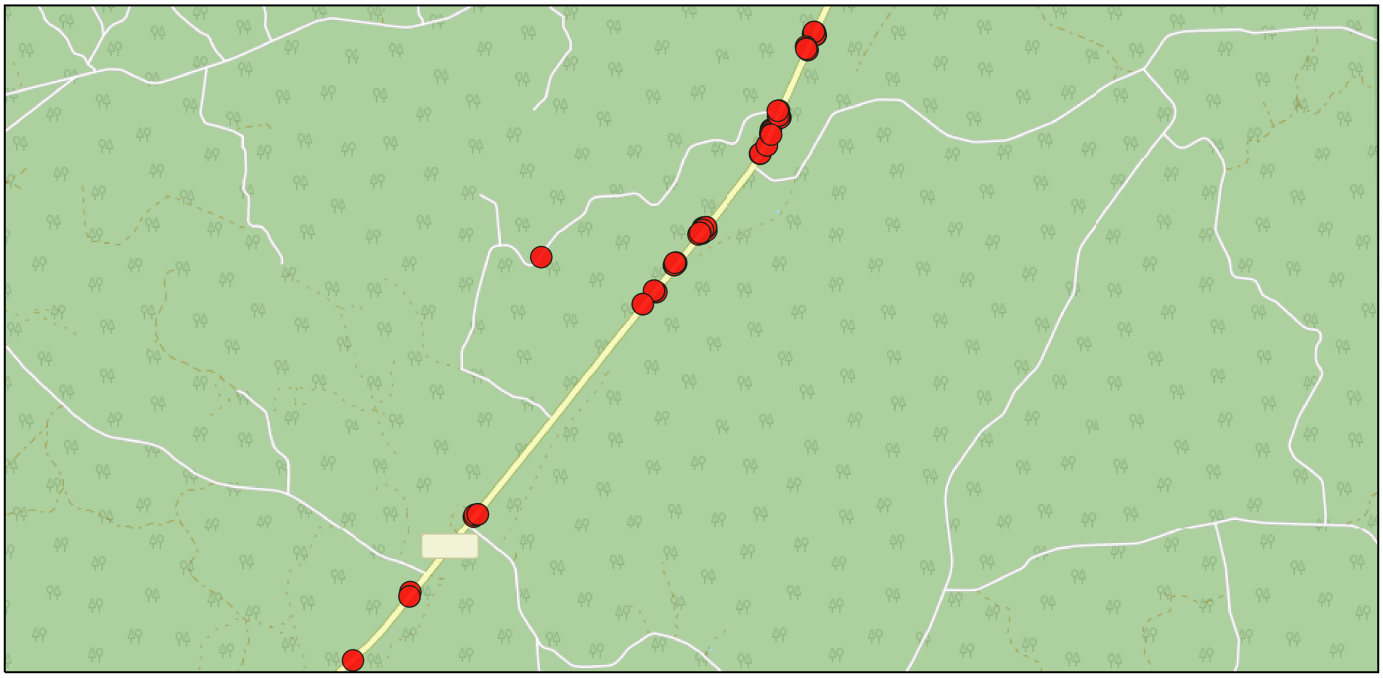
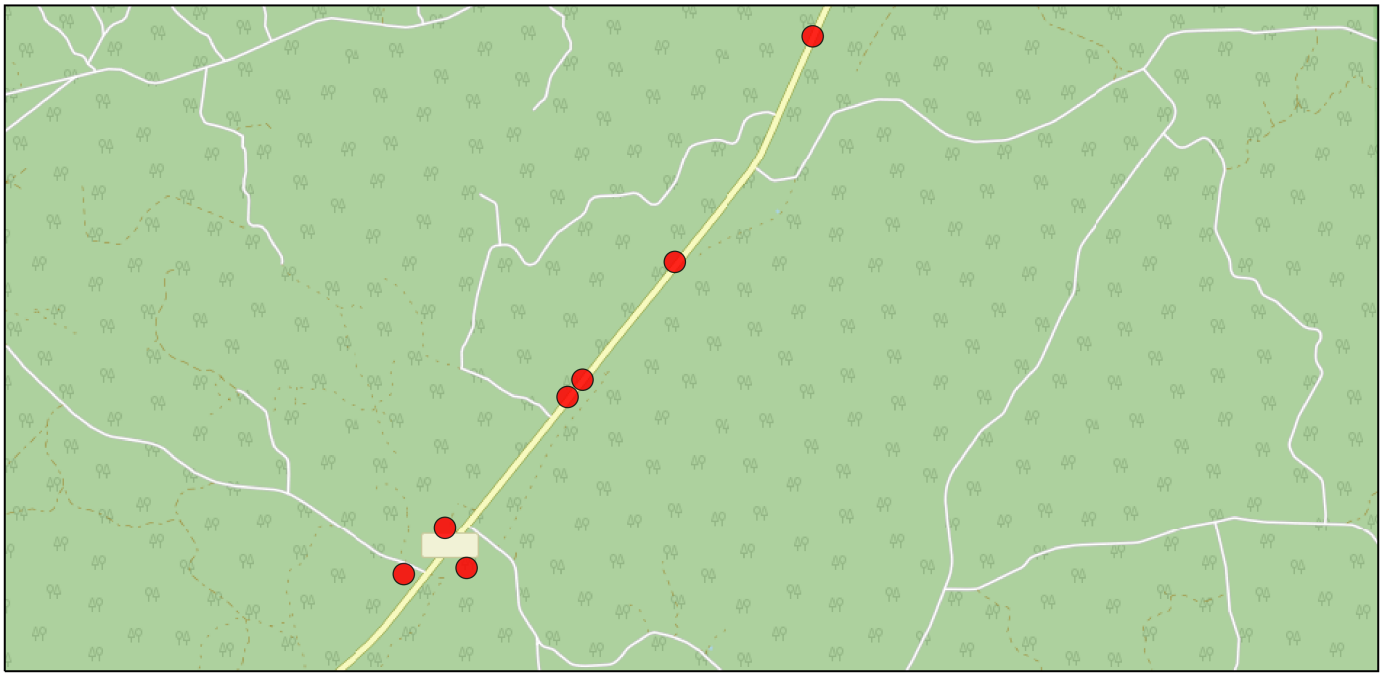


Fig. 45. Same visualization as in Fig. 44, but for the iNaturalist data (only observations without geo-privacy applied): Local geographic bias due to the certain spots such as natural parks being more popular for visitors than others: (a),(c) Darlingtonia State Natural Site in the United States and (b),(d) Bako National Park in Malaysia. Similar to Fig. 44, also in this figure (a),(c) illustrate the GPS precision of the iNaturalist data: the plants are located in only a very small region such that the spread of the observations' geo-locations gives an indication of the expected precision of the GPS data, despite the presence of trees in the area as shown in Fig. 44(e).



(a) Panoramio/Flickr data; larger version of Fig. 7.



(b) iNaturalist data (only observations without geo-privacy applied).

Fig. 46. Local access bias due to streets and paths. Notice that for the iNaturalist data many observations are not shown in (b) because they have a geo-privacy tag applied (because they would be randomly displaced and thus not show the local access bias of being close to roads and paths).

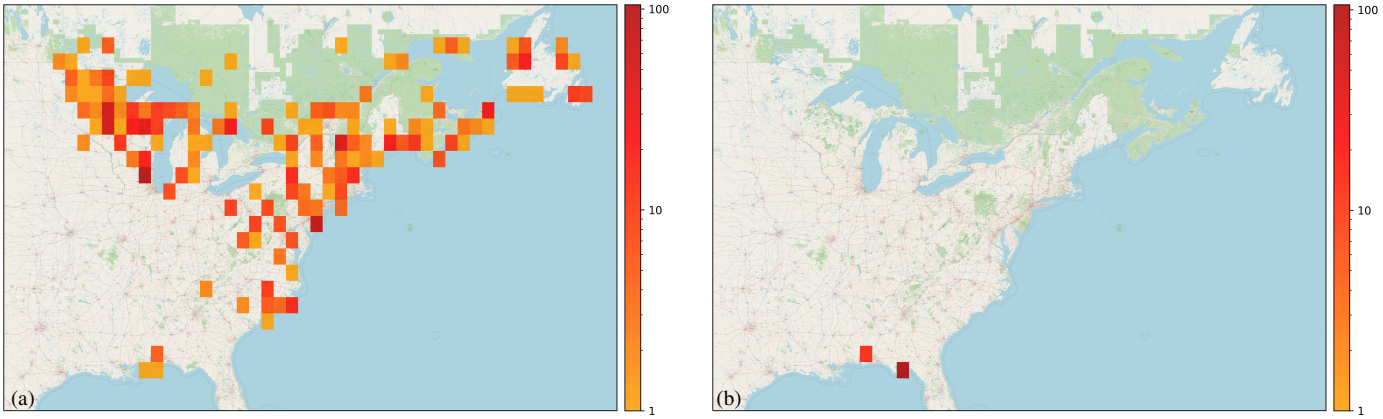


Fig. 47. Larger version of Fig. 11: Distribution of (a) *Sarracenia purpurea* and (b) *Sarracenia rosea*, according to the Panoramio/Flickr data. The cases of *S. purpurea* in the same area of *S. rosea* are likely, in fact, *S. rosea*, which was only elevated to species status in 1999.

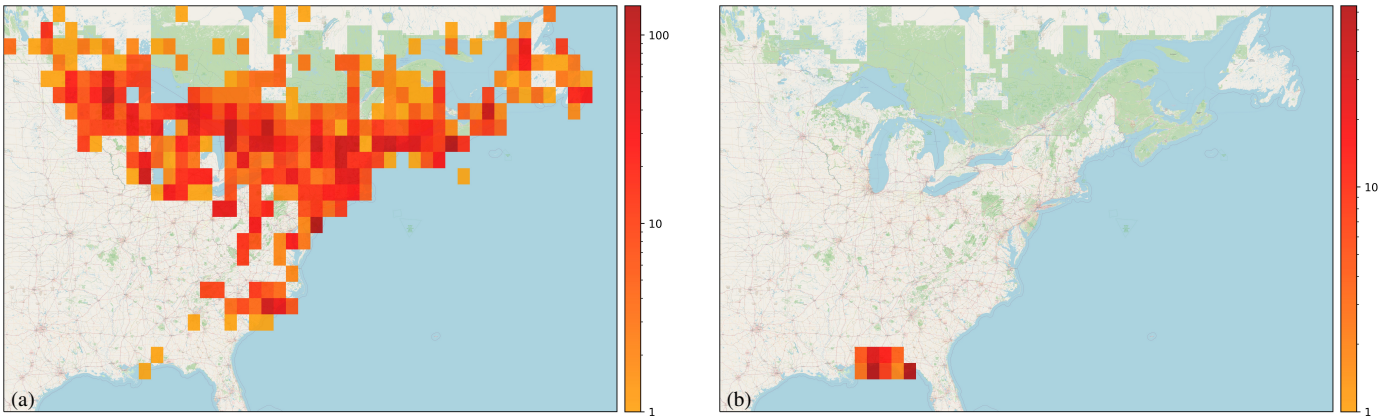


Fig. 48. Same visualization as in Fig. 47, but according to the iNaturalist data (including all observations, both with and without geo-privacy applied; in this large map section the random offset from geo-privacy is negligible): Distribution of (a) *Sarracenia purpurea* and (b) *Sarracenia rosea*. The cases of *S. purpurea* in the same area of *S. rosea* are likely, in fact, *S. rosea*, which was only elevated to species status in 1999. Figure (a) shows that the problem is not as bad as for the Panoramio/Flickr data (Fig. 48(a)), but it is still present.

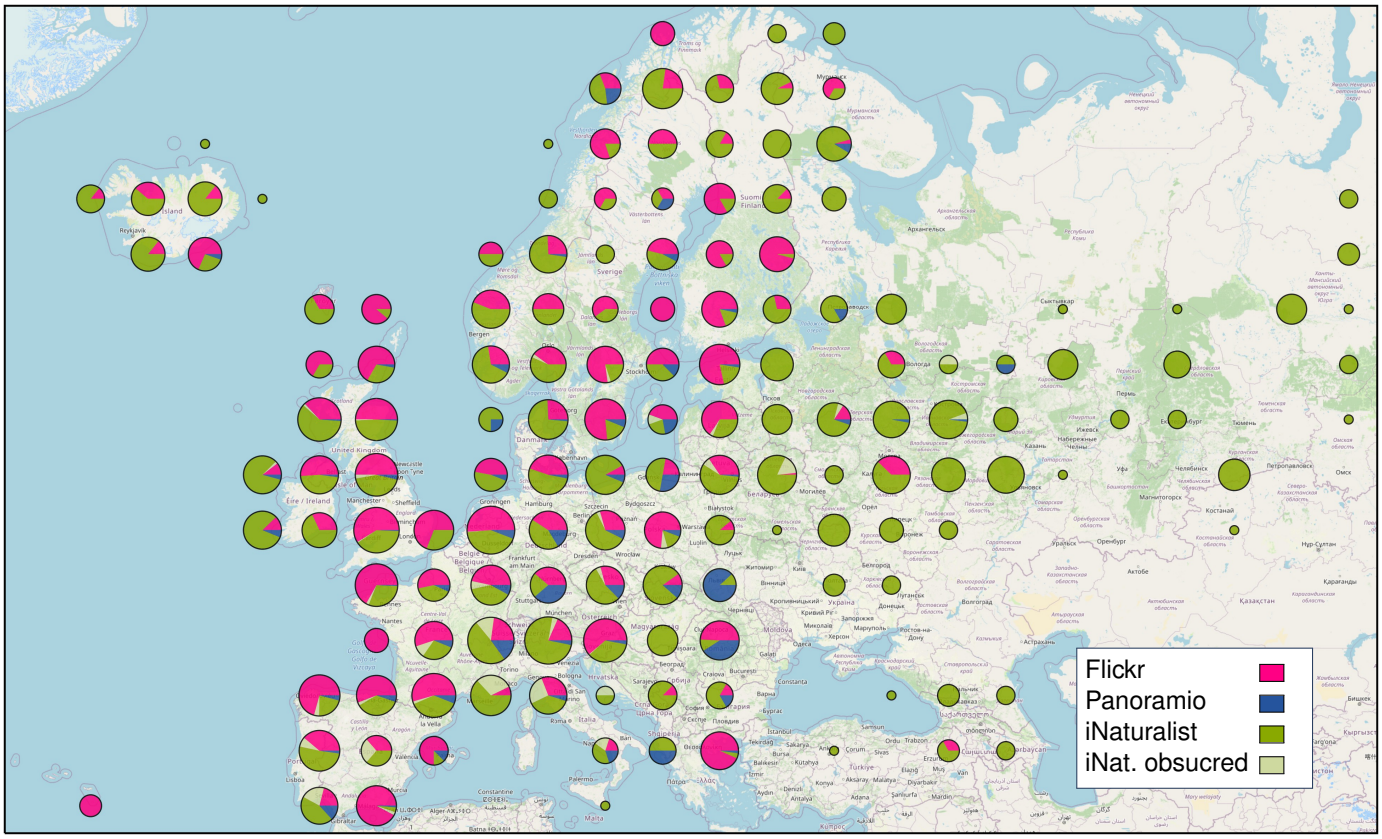


Fig. 49. Enlarged version of Fig. 21: Entries from both datasets shown via graduated [6] pie charts, scaled by the logarithm (base 1.2) of the entry count in the respective grid cell. Interestingly, the iNaturalist data also extends into central Russia, an area largely missing from the Panoramio/Flickr data. A heatmap version of this data is shown in Fig. 50–52.

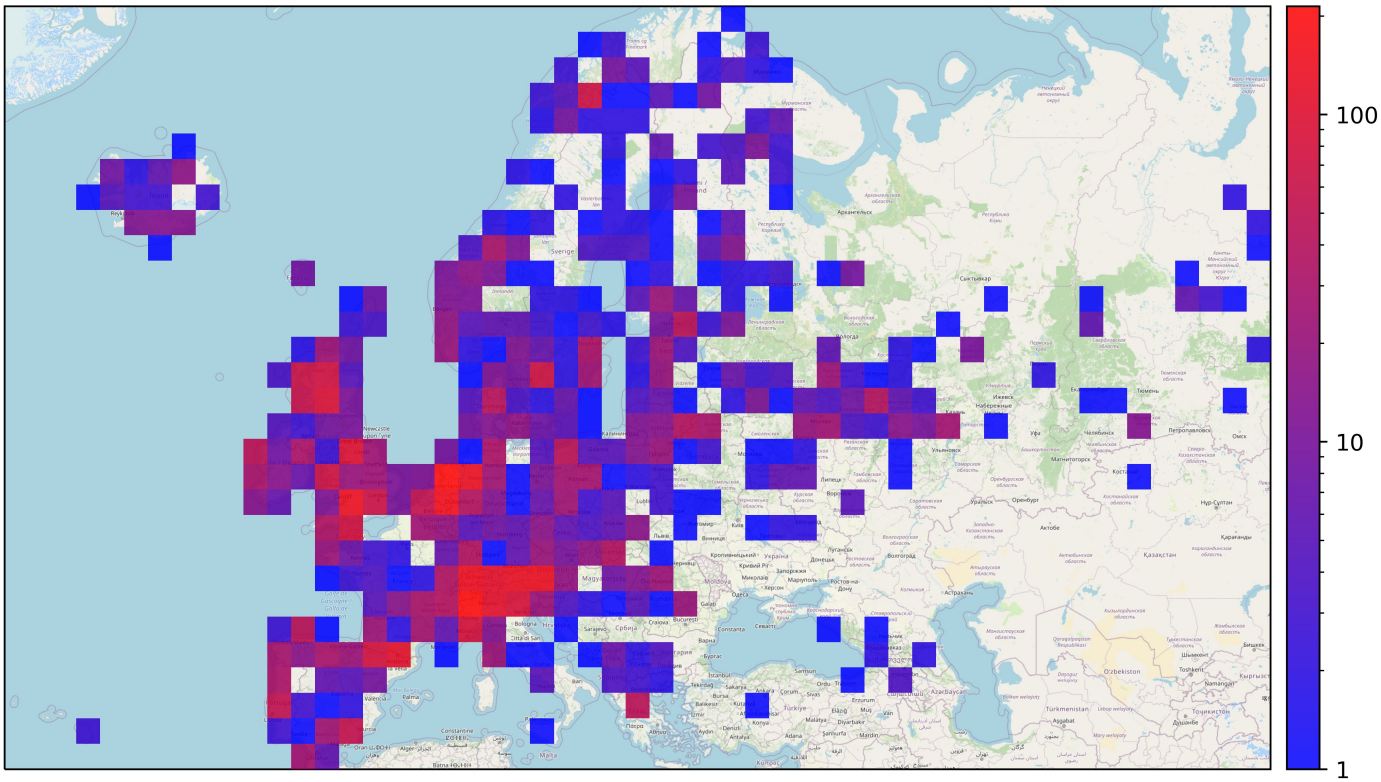


Fig. 50. Discrete heatmap of the number of entries of both datasets (same data as in Fig. 49). Separate plots in Fig. 51 and 52.

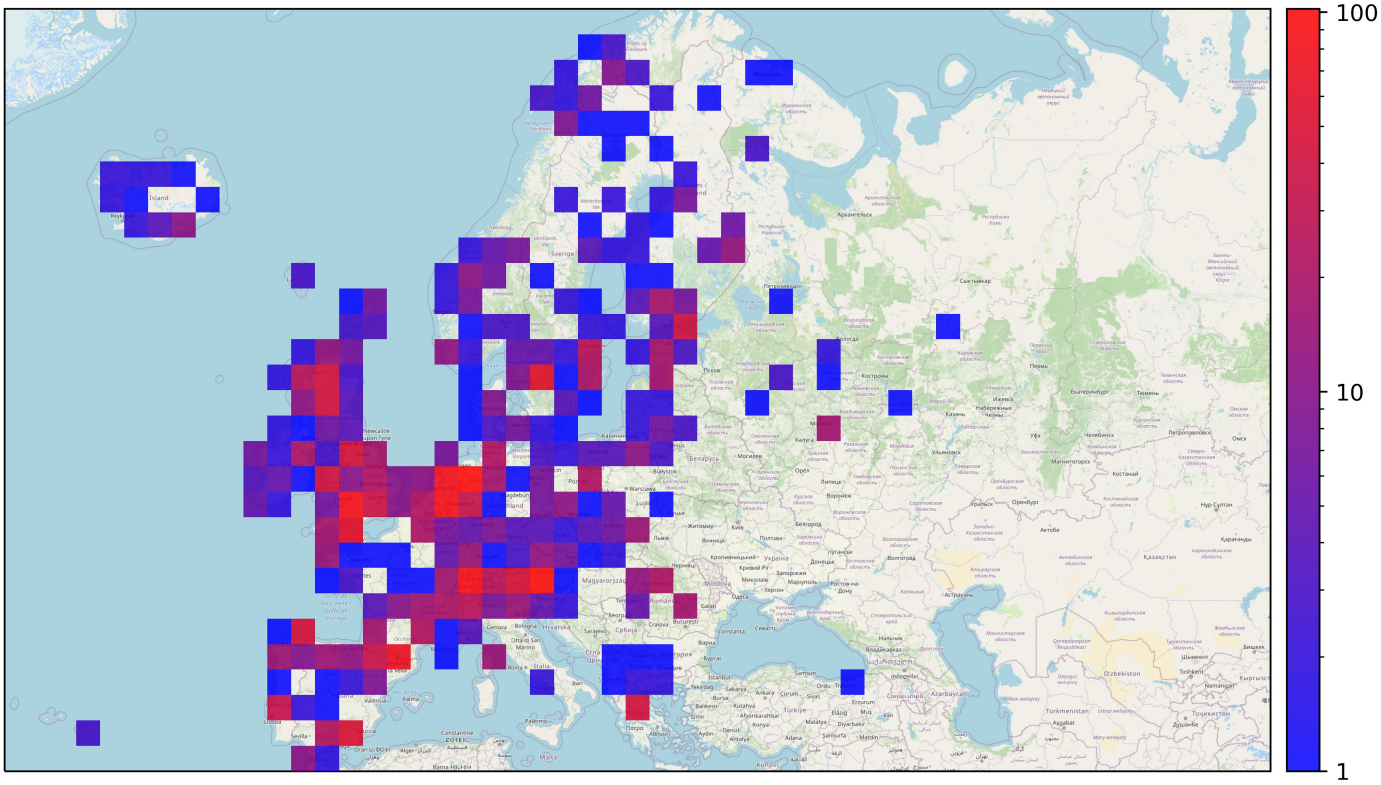


Fig. 51. Discrete heatmap of the number of entries of the Flickr/Panoramio data (part of the data from Fig. 49 and 50).

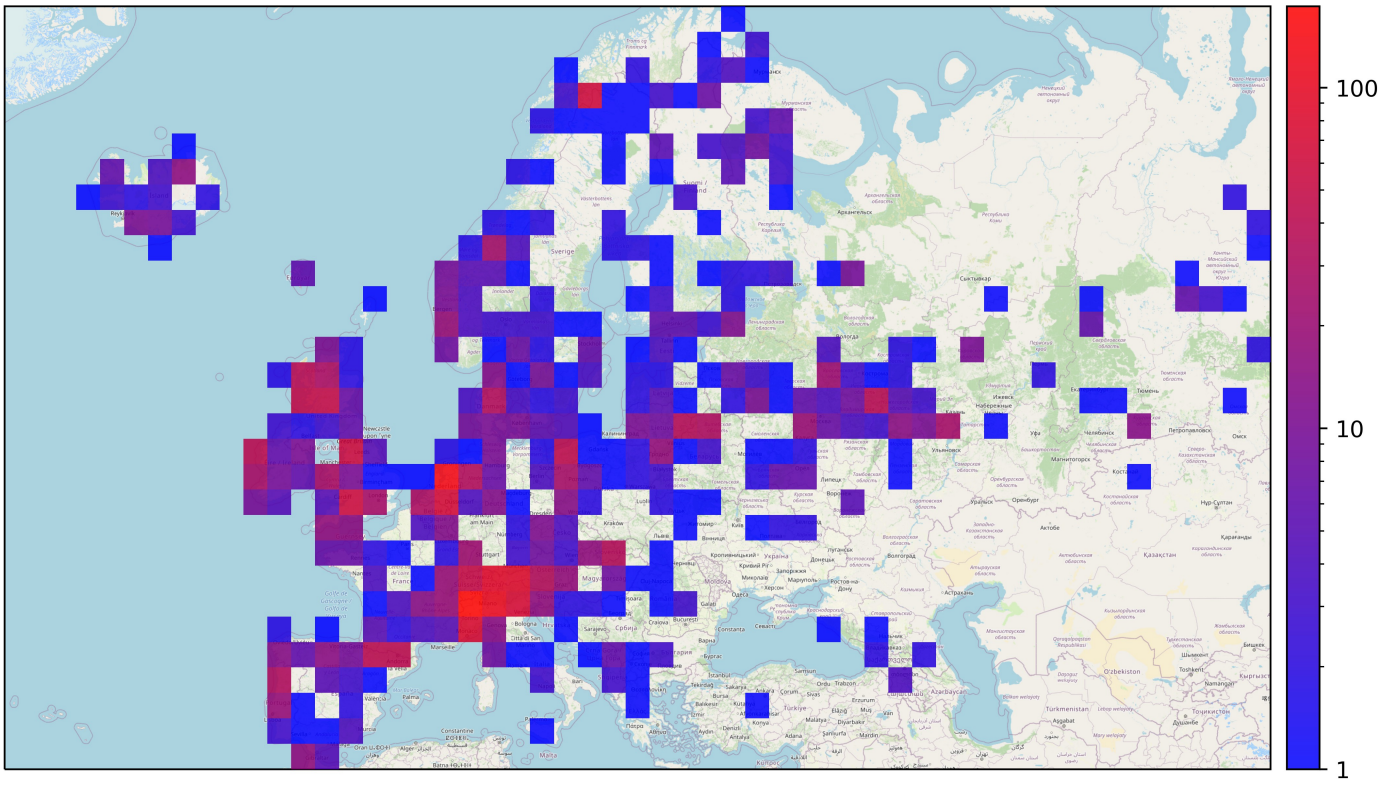


Fig. 52. Discrete heatmap of the number of entries of the iNaturalist data (part of the data from Fig. 49 and 50).

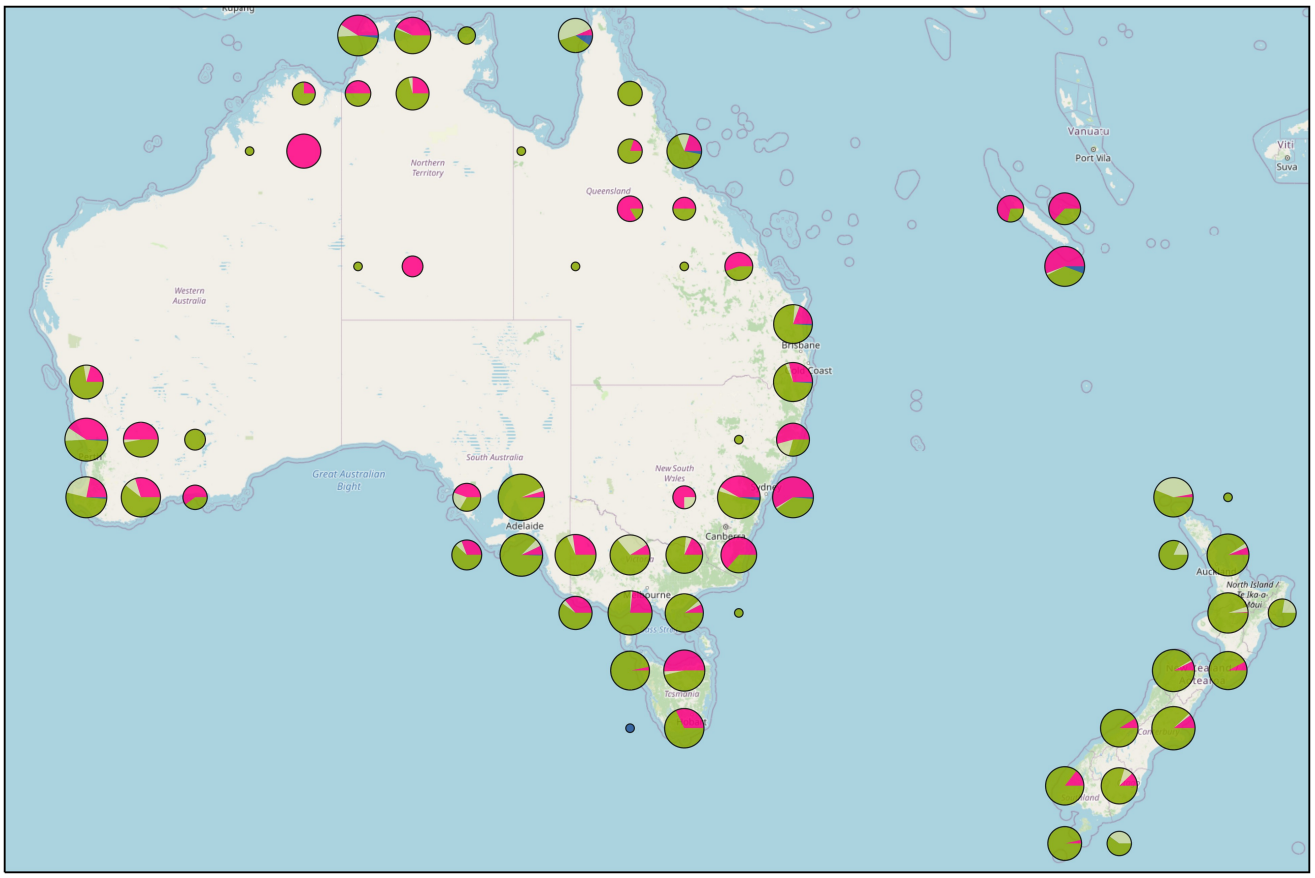


Fig. 53. Dataset contributions by the different services in our datasets: Entries from both datasets shown via graduated [6] pie charts, scaled by the logarithm (base 1.2) of the entry count in the respective grid cell. Legend as in Fig. 49. A heatmap version of this data is shown in Fig. 54–56.

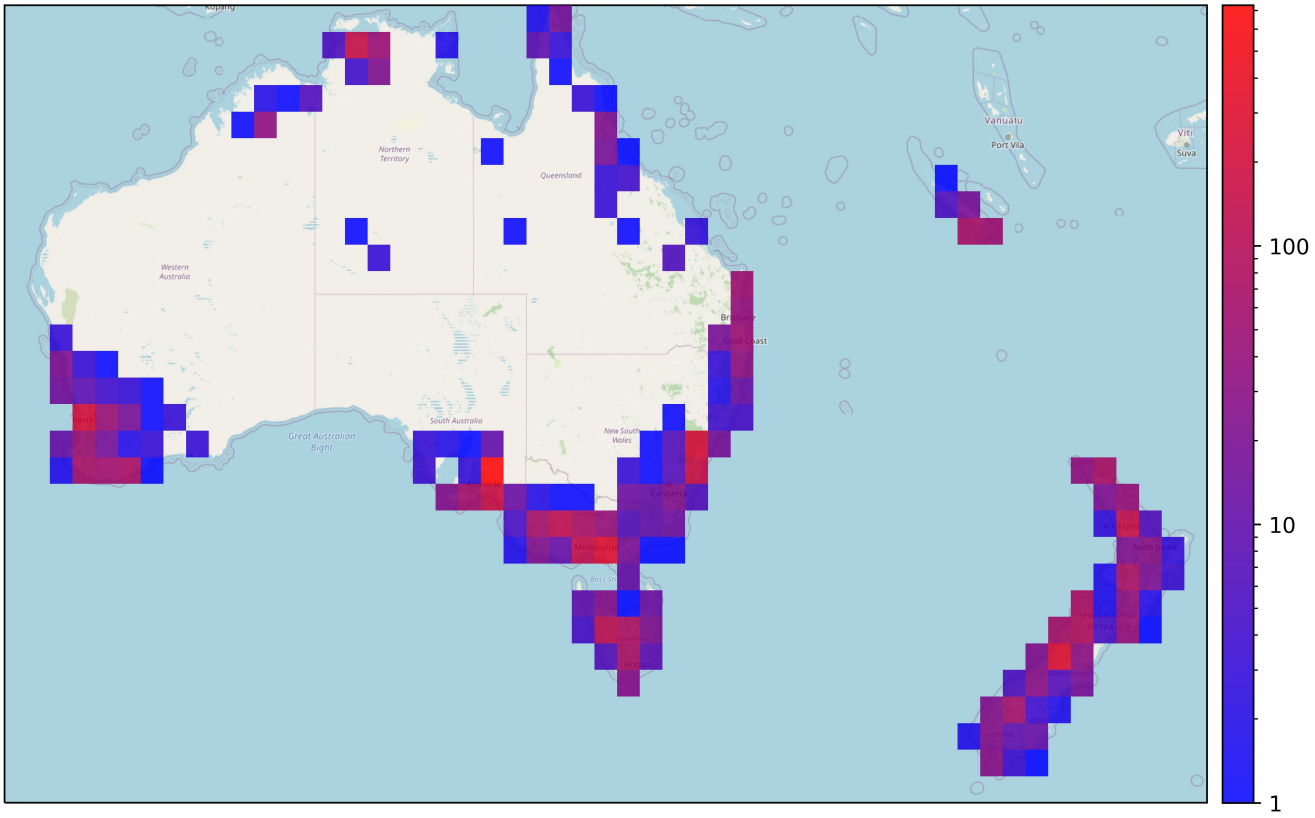


Fig. 54. Discrete heatmap of the number of entries of both datasets (same data as in Fig. 53). Separate plots in Fig. 55 and 56.

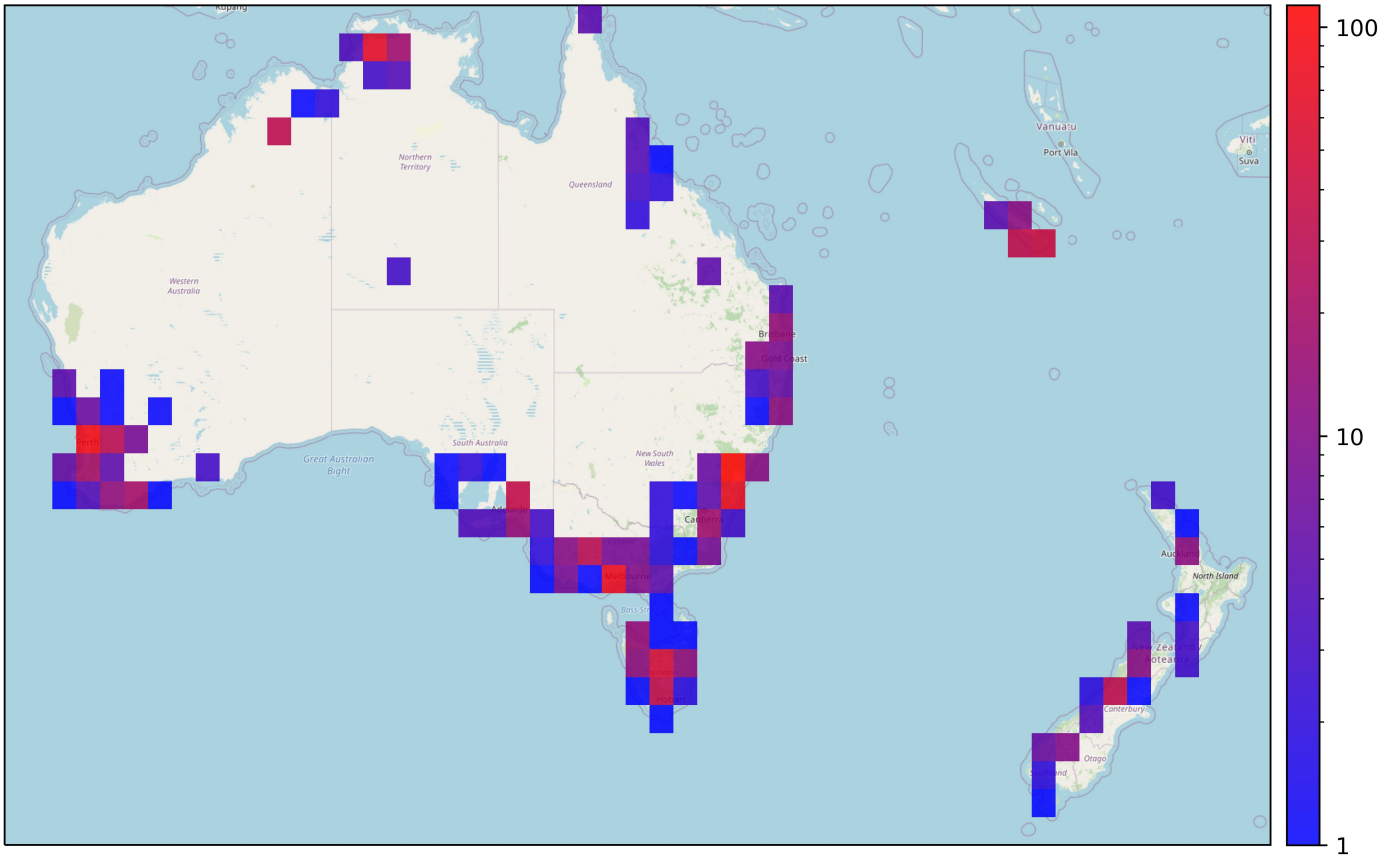


Fig. 55. Discrete heatmap of the number of entries of the Flickr/Panoramio data (part of the data from Fig. 53 and 54).

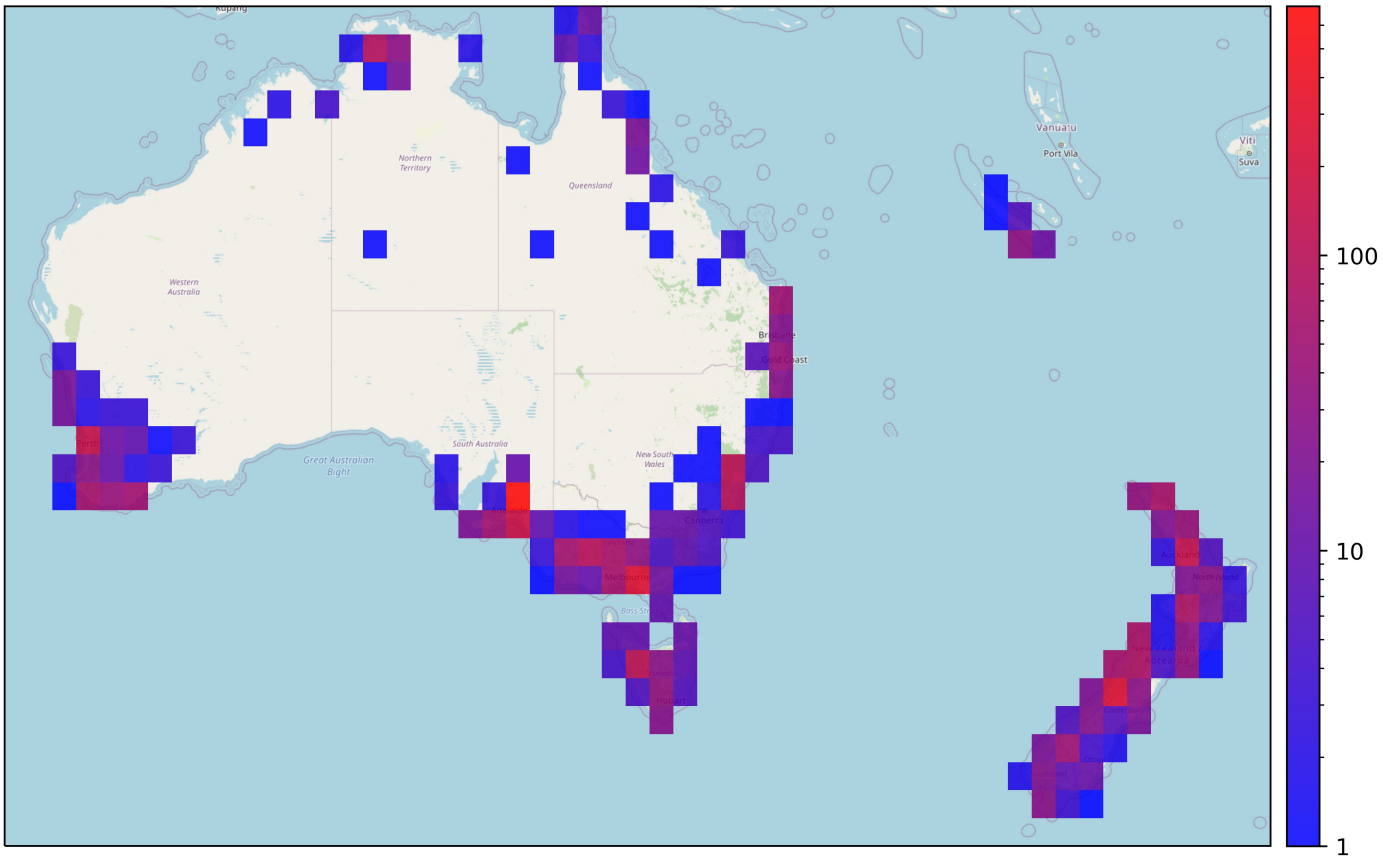


Fig. 56. Discrete heatmap of the number of entries of the iNaturalist data (part of the data from Fig. 53 and 54).

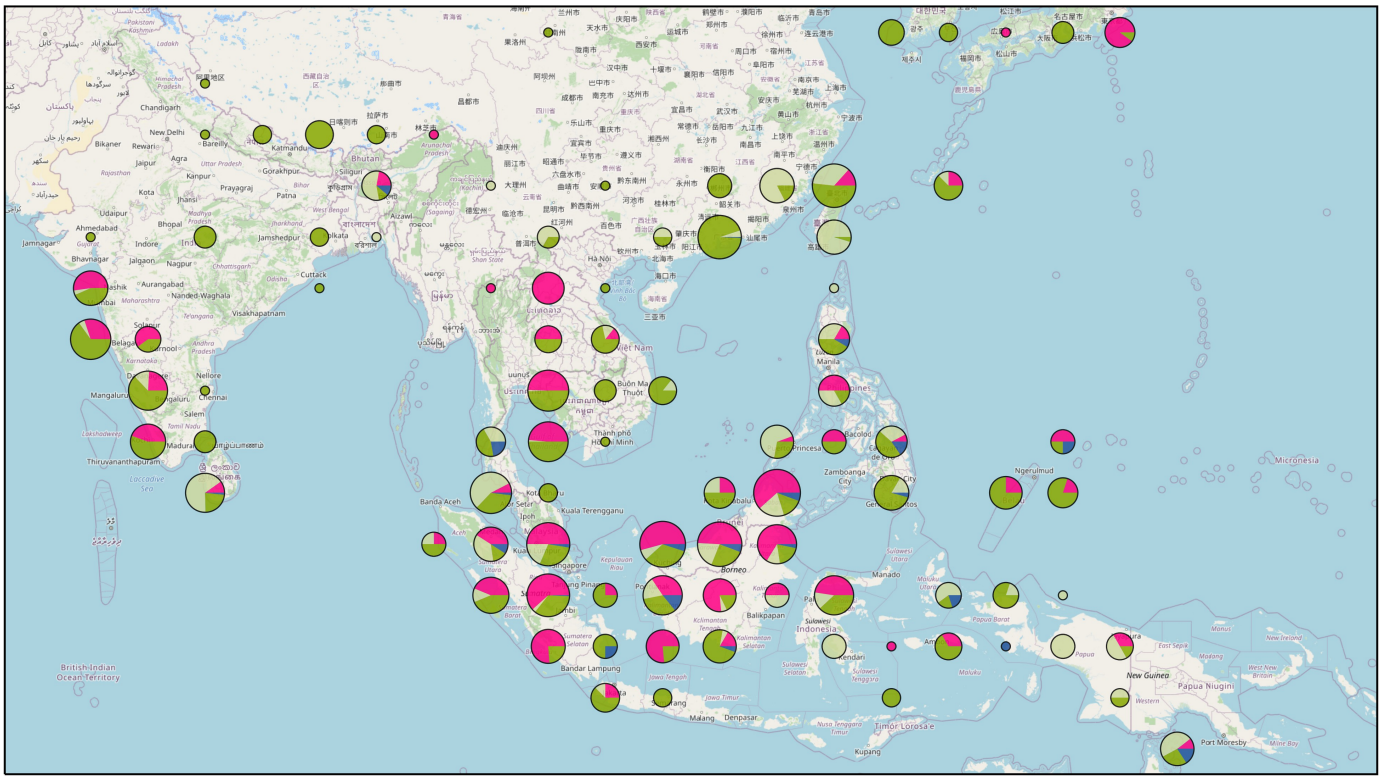


Fig. 57. Dataset contributions by the different services in our datasets: Entries from both datasets shown via graduated [6] pie charts, scaled by the logarithm (base 1.2) of the entry count in the respective grid cell. Legend as in Fig. 49. A heatmap version of this data is shown in Fig. 58–60.

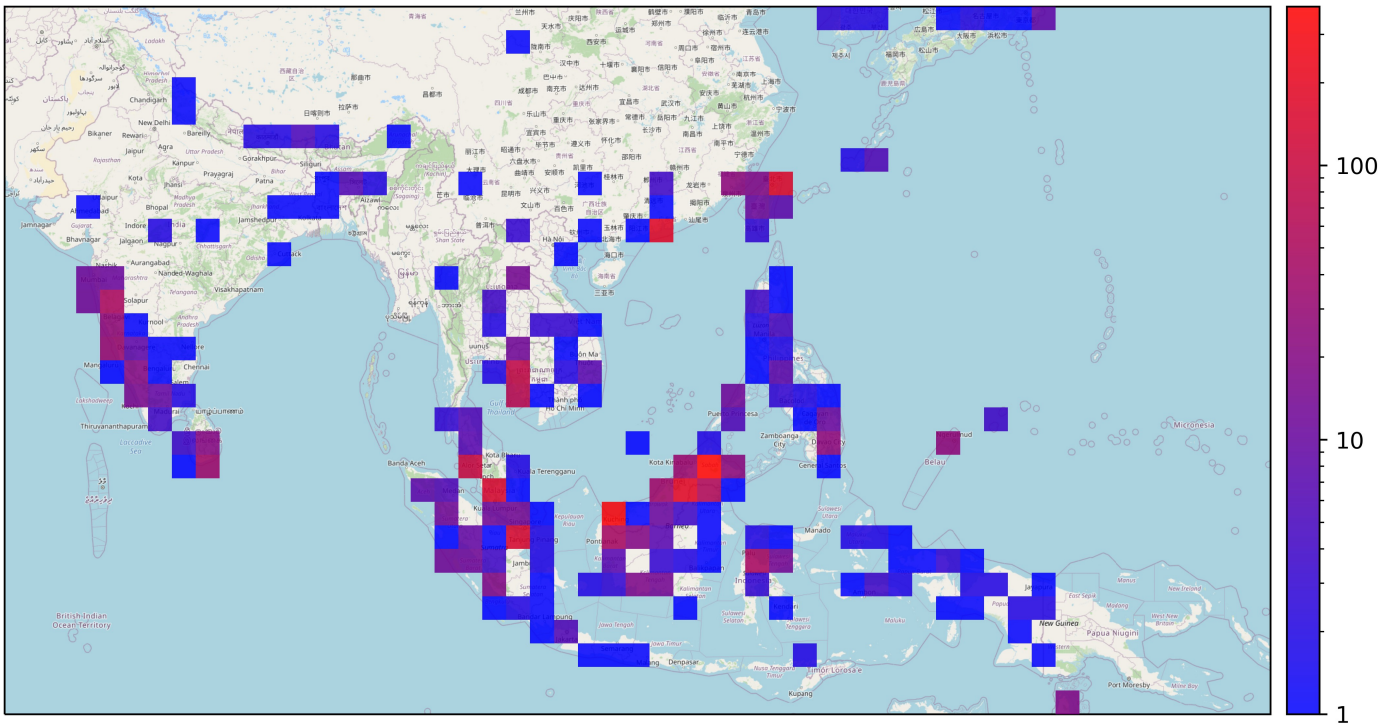


Fig. 58. Discrete heatmap of the number of entries of both datasets (same data as in Fig. 57). Separate plots in Fig. 59 and 60.

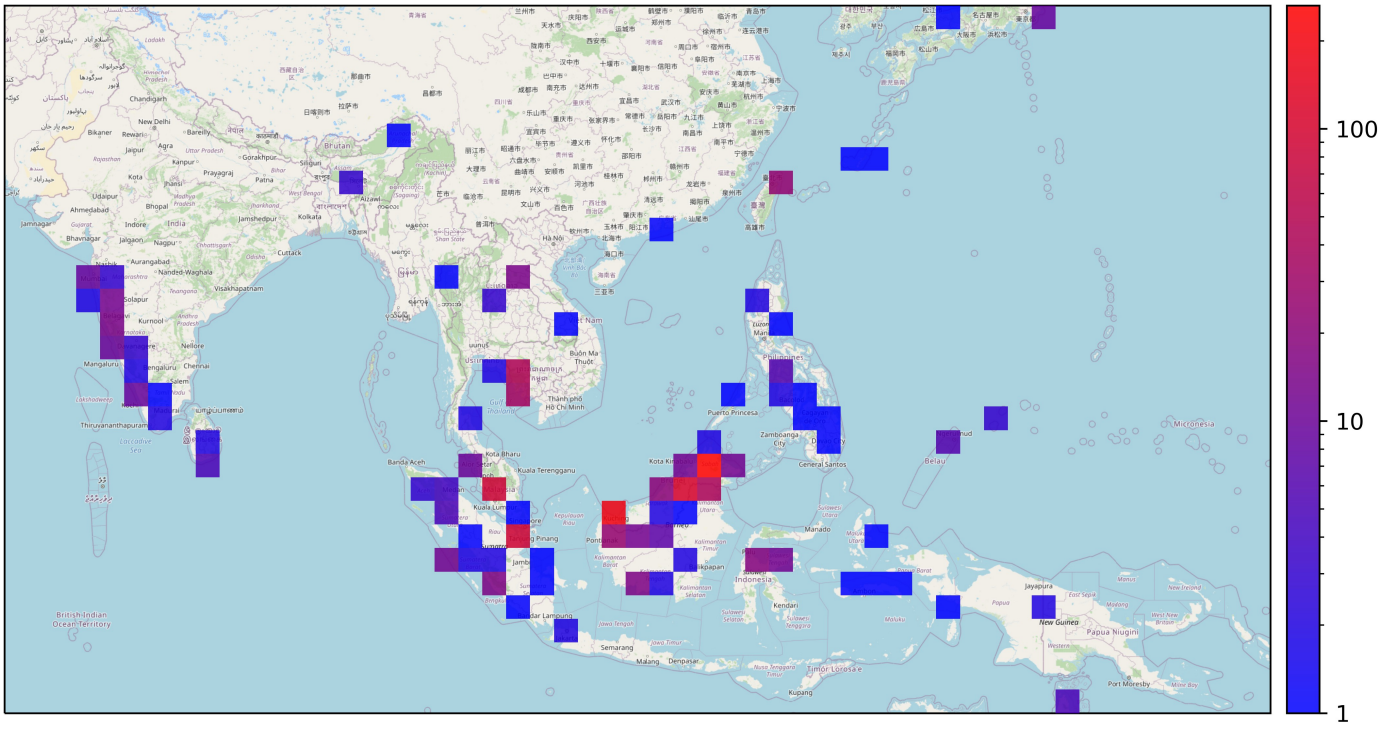


Fig. 59. Discrete heatmap of the number of entries of the Flickr/Panoramio data (part of the data from Fig. 57 and 58).

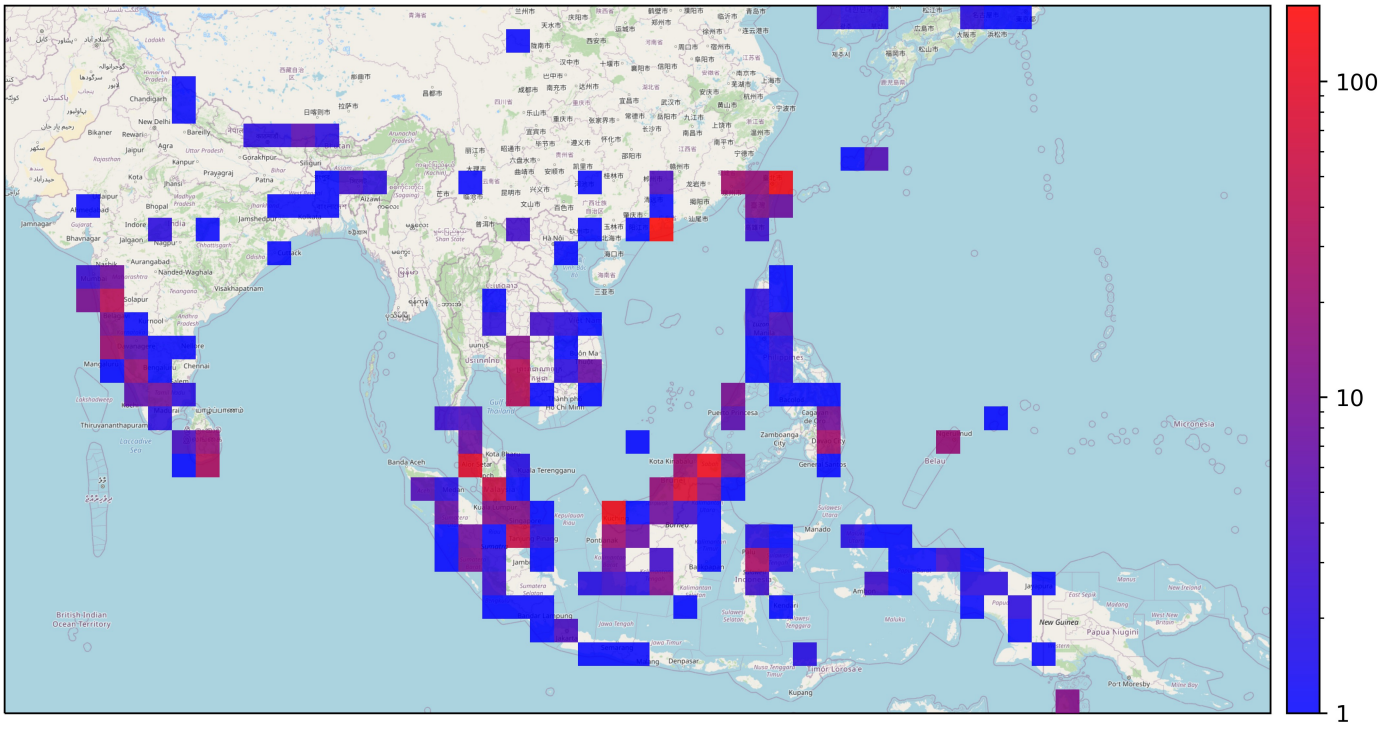


Fig. 60. Discrete heatmap of the number of entries of the iNaturalist data (part of the data from Fig. 57 and 58).

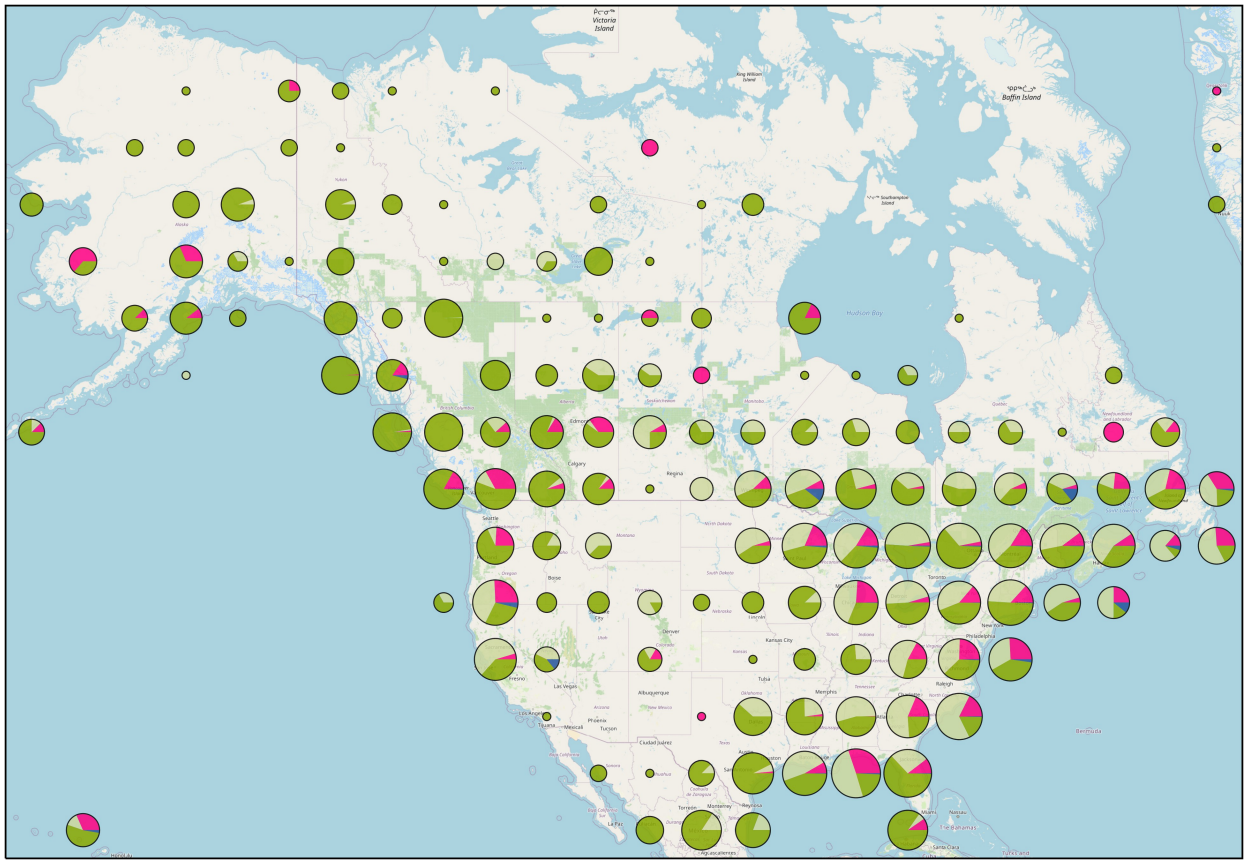


Fig. 61. Dataset contributions by the different services in our datasets: Entries from both datasets shown via graduated [6] pie charts, scaled by the logarithm (base 1.2) of the entry count in the respective grid cell. Legend as in Fig. 49. A heatmap version of this data is shown in Fig. 62–64.

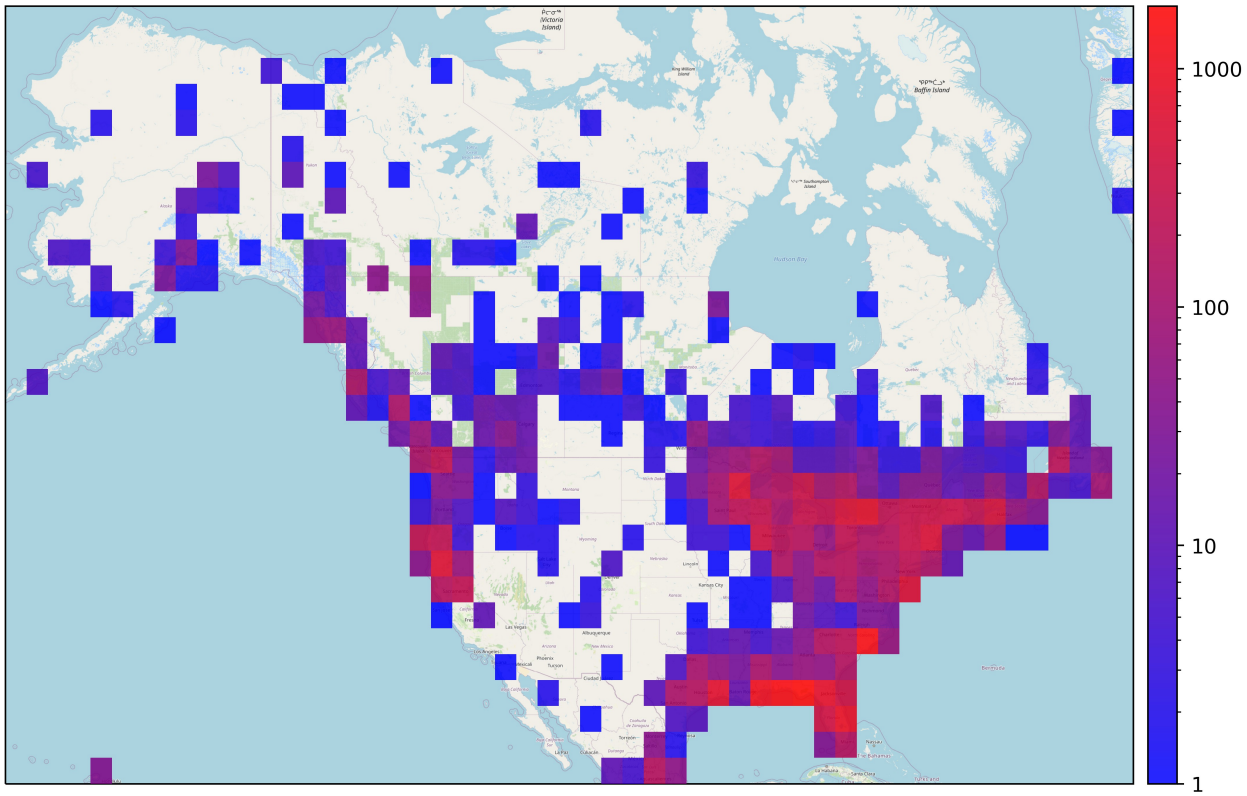


Fig. 62. Discrete heatmap of the number of entries of both datasets (same data as in Fig. 61). Separate plots in Fig. 63 and 64.

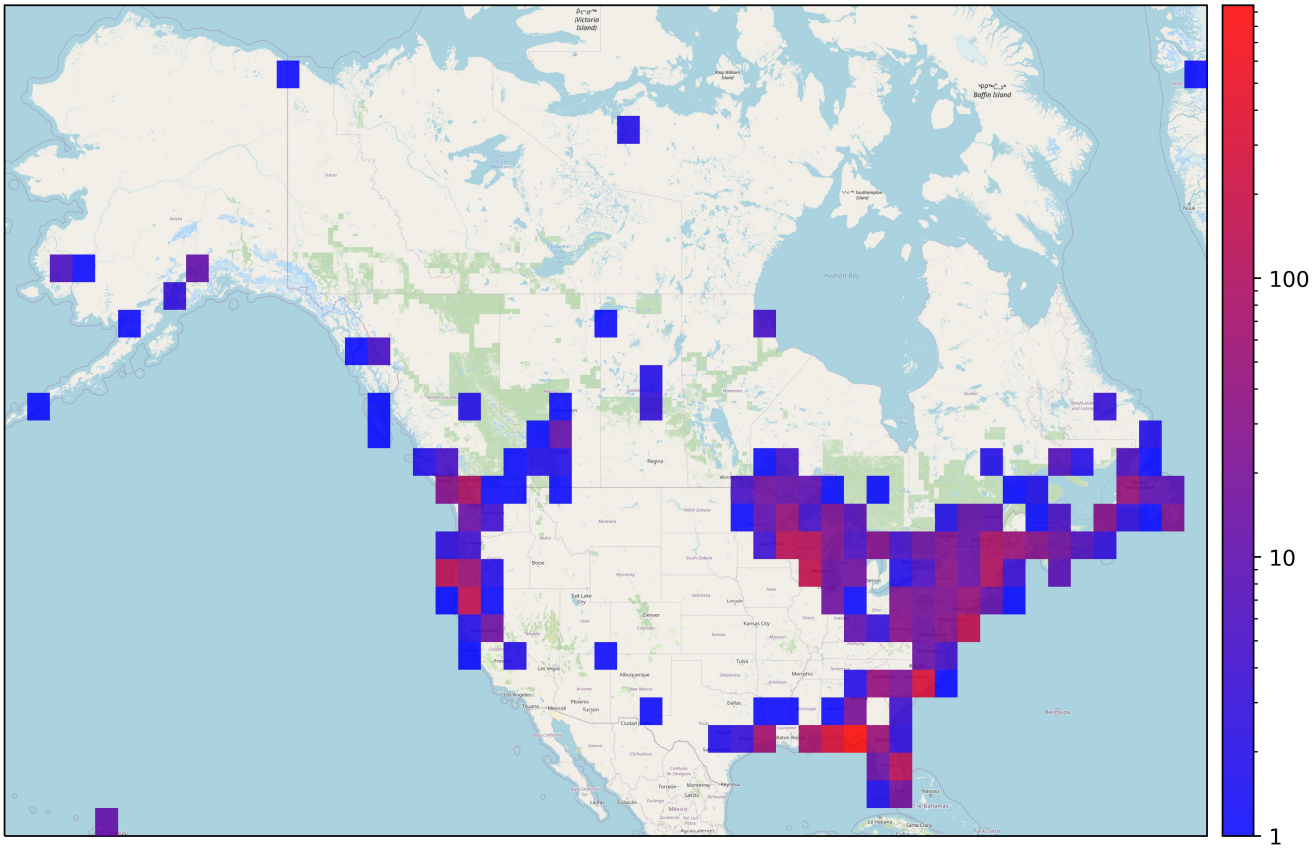


Fig. 63. Discrete heatmap of the number of entries of the Flickr/Panoramio data (part of the data from Fig. 61 and 62).

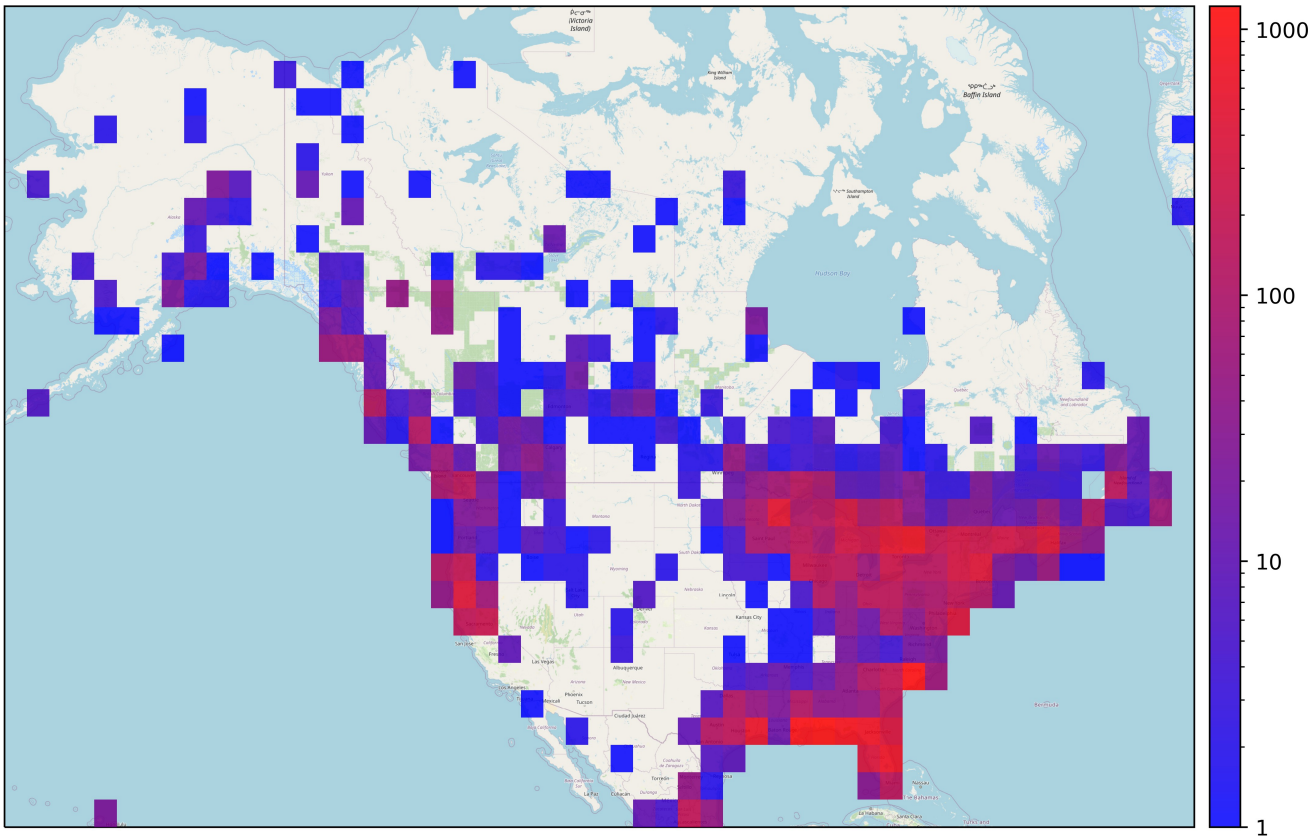
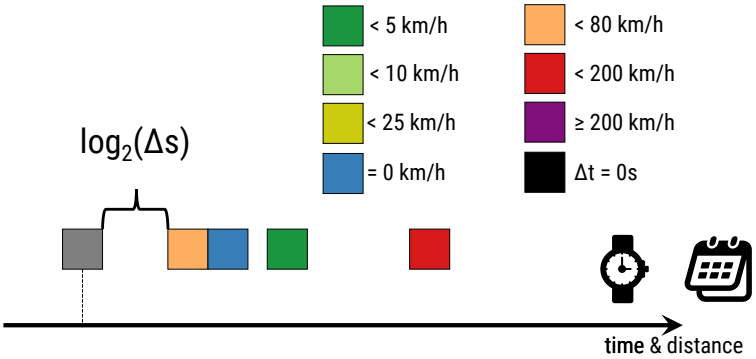
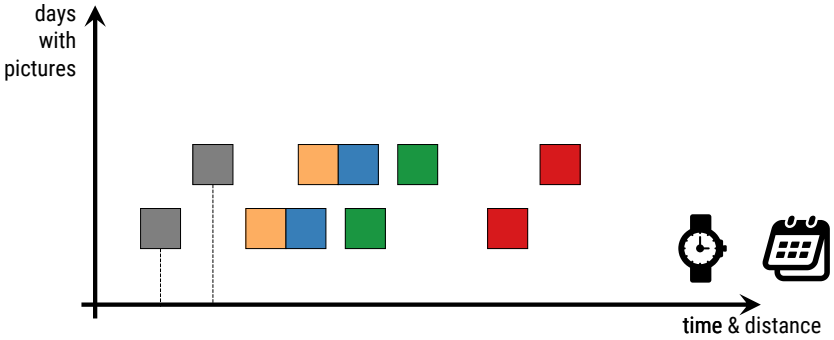


Fig. 64. Discrete heatmap of the number of entries of the iNaturalist data (part of the data from Fig. 61 and 62).

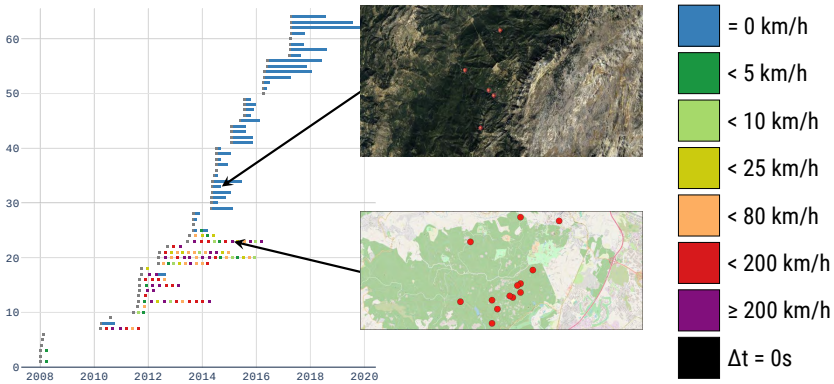
(a) Two successive markers in a line indicate two pictures, one taken after the other. The distance between them is the log-transformed geodesic distance between the places at which the pictures were taken, according to the meta data. While this log-transformation may be somewhat unusual, it ensures that we can show both small and large distances within a reasonably compact plot and are still able to make out differences between them. In a linear mapping already the visit of two habitat locations within the same day would have led to the large distance between the two sites fully dominating the small ones within each site. In an interactive software tool such mappings could be chosen interactively, also including linear mappings. Several markers in a row indicate several pictures taken on the same day, shown in sequence, with their respective distances. The first marker in a row, by its x-position, also marks the date of the series of pictures of that day.



(b) The color of each marker shows the apparent speed (using discretized speed ranges as explained in Sec. 6) that the photographer had to travel at least, if they traveled at constant speed and on the geodesic path. The real maximum speed is thus always higher because the photographer likely cannot take the direct way and also does not travel at constant speeds. The marked speeds are thus lower bounds for the average speed. We chose the specific colors based on a traffic light metaphor, green as believable, yellow/orange indicating caution, red meaning unbelievable, and purple as beyond unbelievable. The blue and black hues indicate conditions clearly different from the previous ramp (no motion or no time), so are outside of this metaphor.



(c) Several days worth of images of a given contributor are shown with several rows. We stack the rows directly on top of each other to save space, as continued periods without pictures can be read from the distances of the respective first markers of the rows on the x-axis.



(d) The motion plausibility profile of the example contributor discussed in Sec. 5.2 and 6 (i. e., Fig. 15), connected with the example observations from Fig. 13 and 14. The arrows point to the rows of the days of the respective sets of observation.

Fig. 65. Schematic depiction of how to read motion plausibility profiles and the link between Fig. 13, 14, and 15.

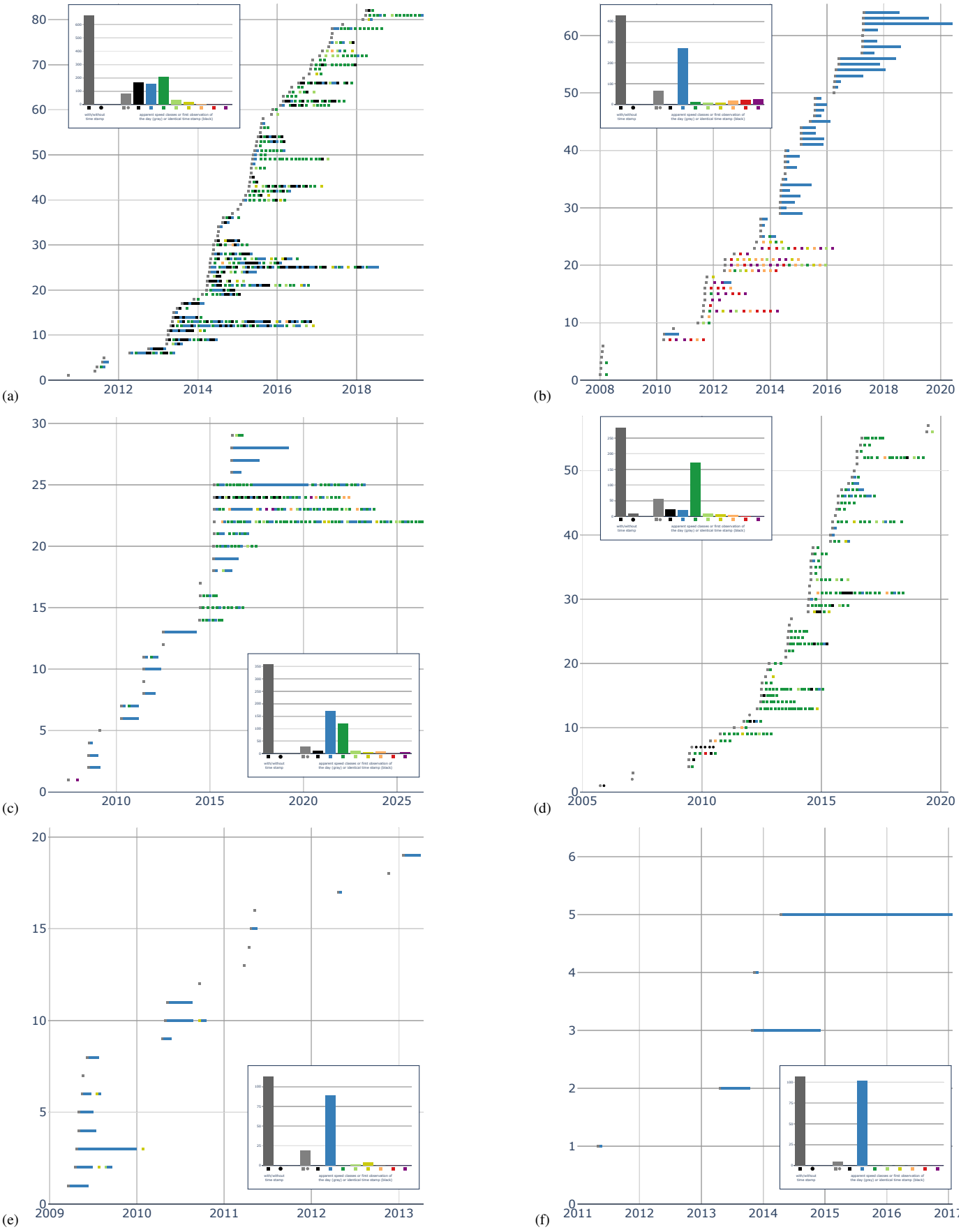


Fig. 66. The motion plausibility profiles of the top 1–6 contributors in the Panoramio/Flickr dataset, in decreasing order of number of observations. Same principle and color scale as in Fig. 15/16. The plot of (b) is a scaled version of Fig. 15. The plots of (a), (c), and (d) were already contained in Fig. 16 as Fig. 16(a)–(c), respectively.

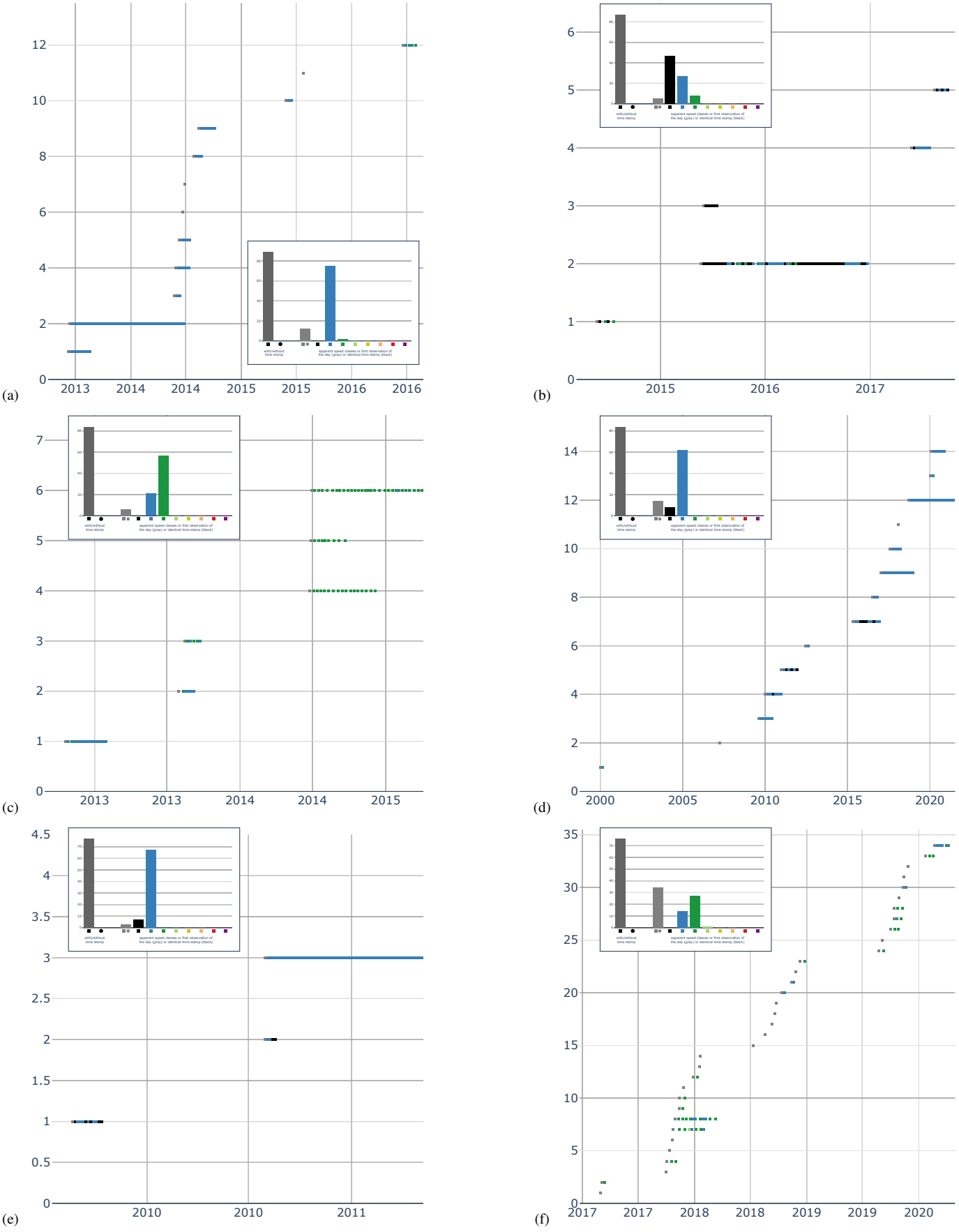


Fig. 67. The motion plausibility profiles of the top 7–12 contributors in the Panoramio/Flickr dataset, in decreasing order of number of observations. Same principle and color scale as in Fig. 15/16. The plots of (a), (b), and (c) were already contained in Fig. 16 as Fig. 16(d)–(f), respectively.

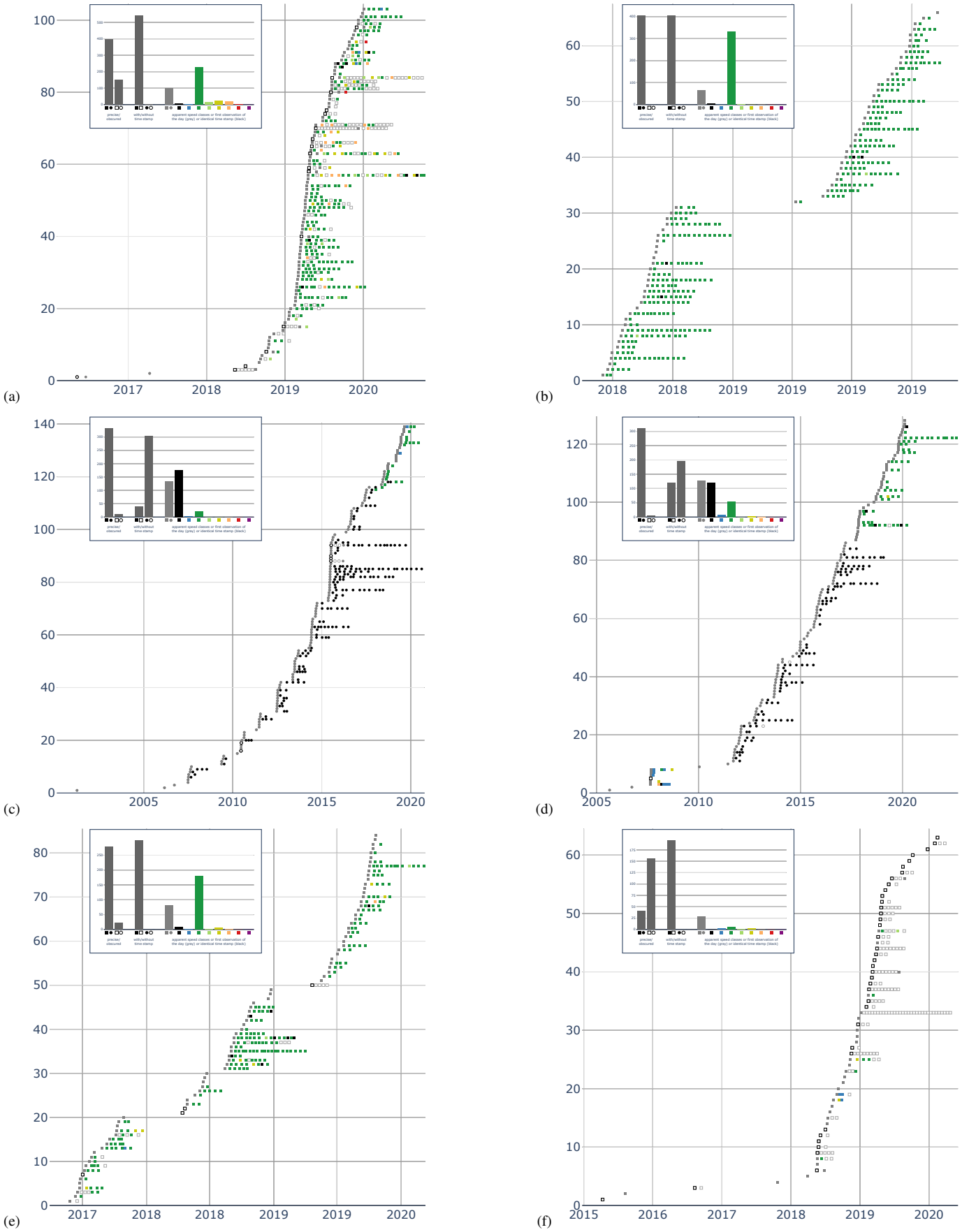


Fig. 68. The motion plausibility profiles of the top 1–6 contributors in the iNaturalist dataset, in decreasing order of number of observations. Same principle and color scale as in Fig. 20. Circles are observations without a valid time stamp, outlined shapes are obscured observations. The plots of (a), (b), (c), and (f) were already contained in Fig. 20 as Fig. 20(a)–(d), respectively.

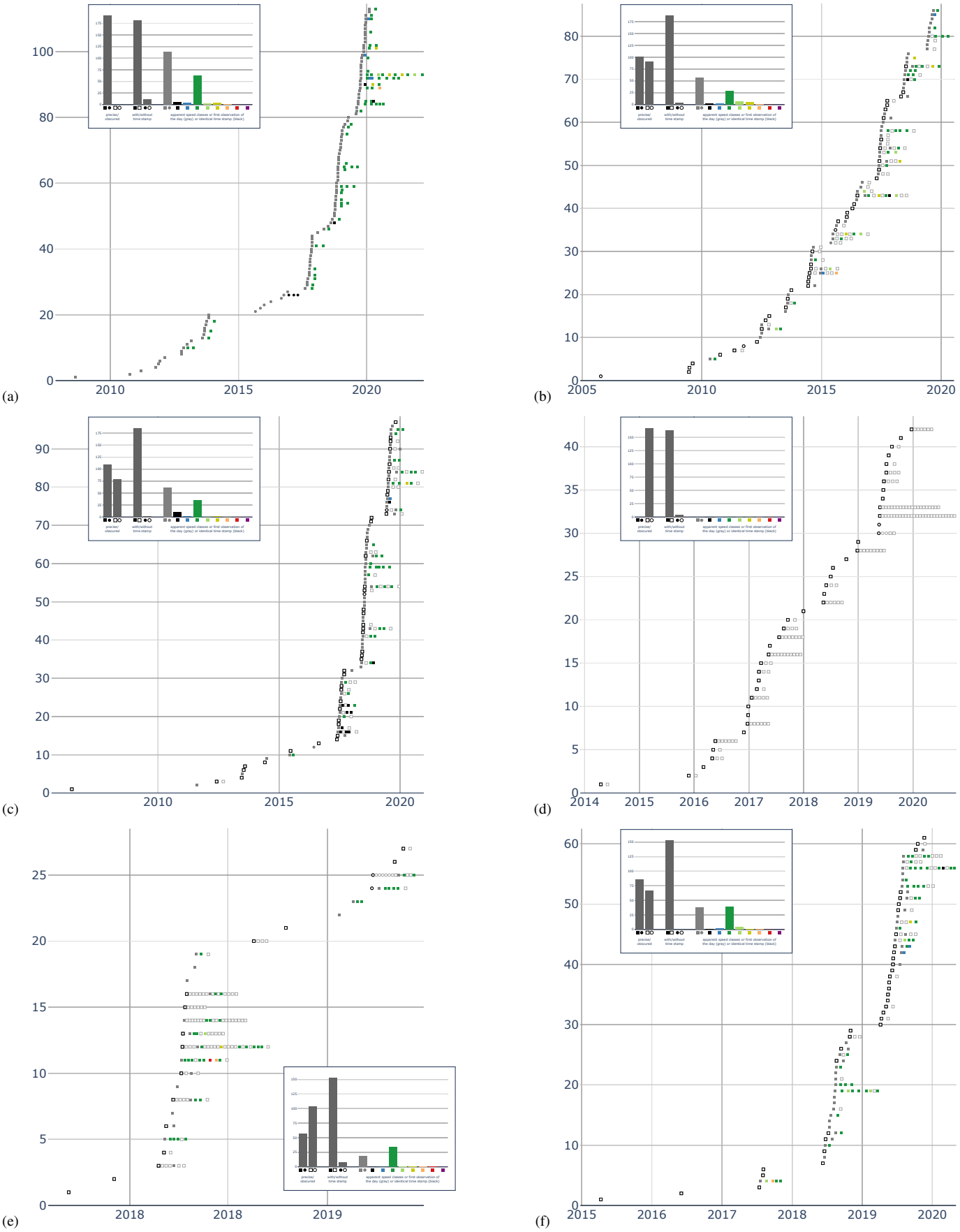


Fig. 69. The motion plausibility profiles of the top 7–12 contributors in the iNaturalist dataset, in decreasing order of number of observations. Same principle and color scale as in Fig. 20. Circles are observations without a valid time stamp, outlined shapes are obscured observations. The plots of (a) and (b) were already contained in Fig. 20 as Fig. 20(e) and (f), respectively.

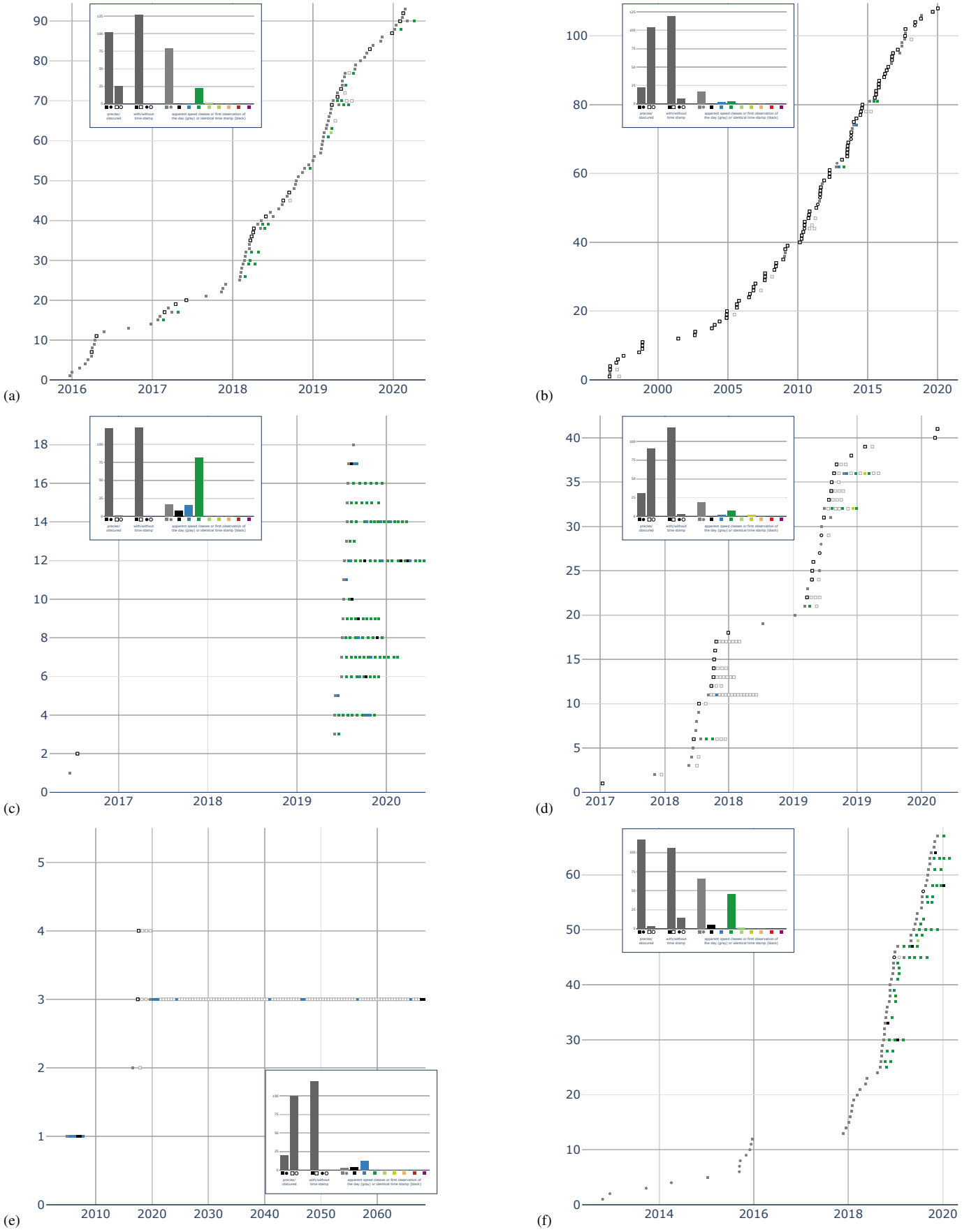


Fig. 70. The motion plausibility profiles of the top 13–18 contributors in the iNaturalist dataset, in decreasing order of number of observations. Same principle and color scale as in Fig. 20. Circles are observations without a valid time stamp, outlined shapes are obscured observations.

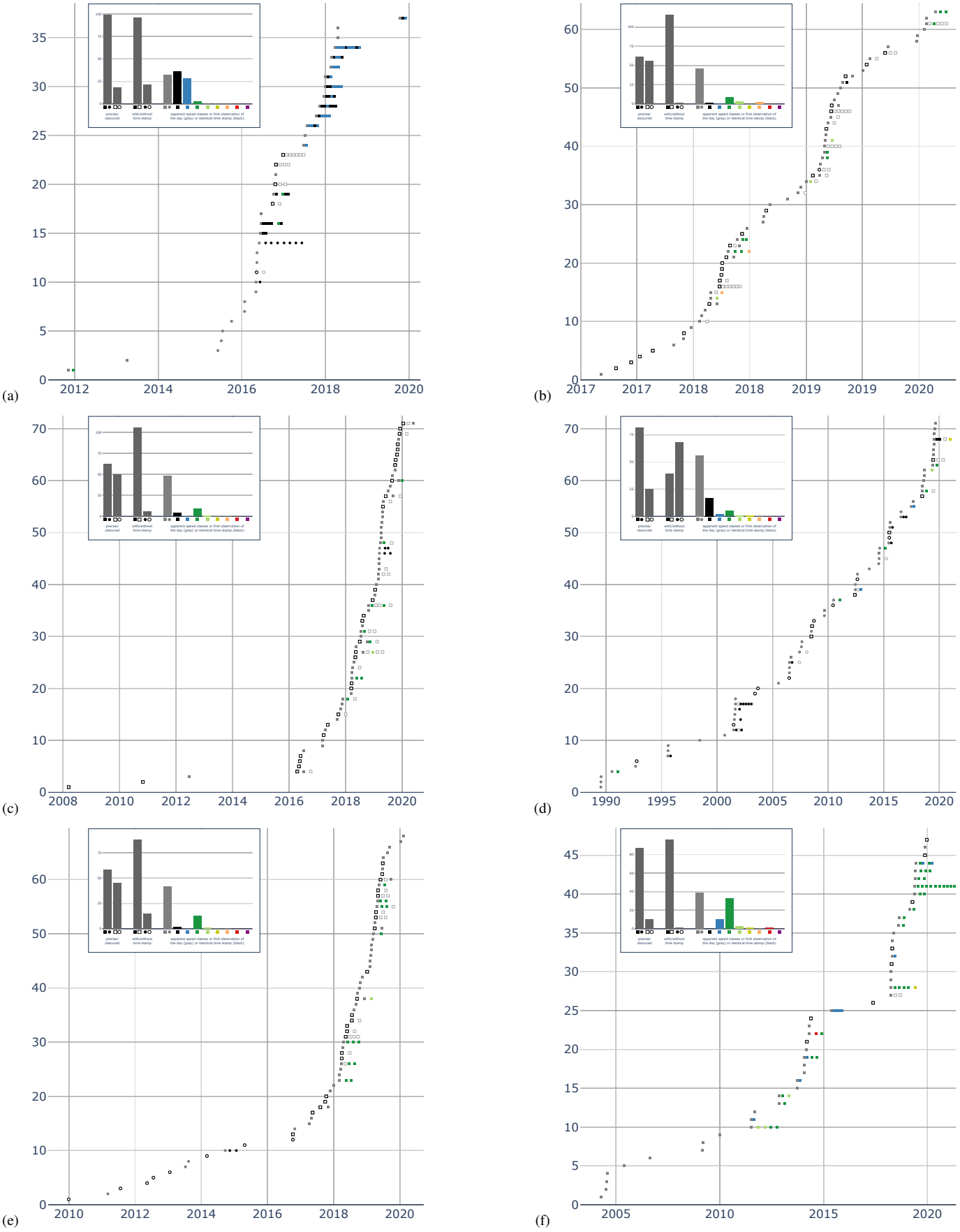


Fig. 71. The motion plausibility profiles of the top 19–24 contributors in the iNaturalist dataset, in decreasing order of number of observations. Same principle and color scale as in Fig. 20. Circles are observations without a valid time stamp, outlined shapes are obscured observations.

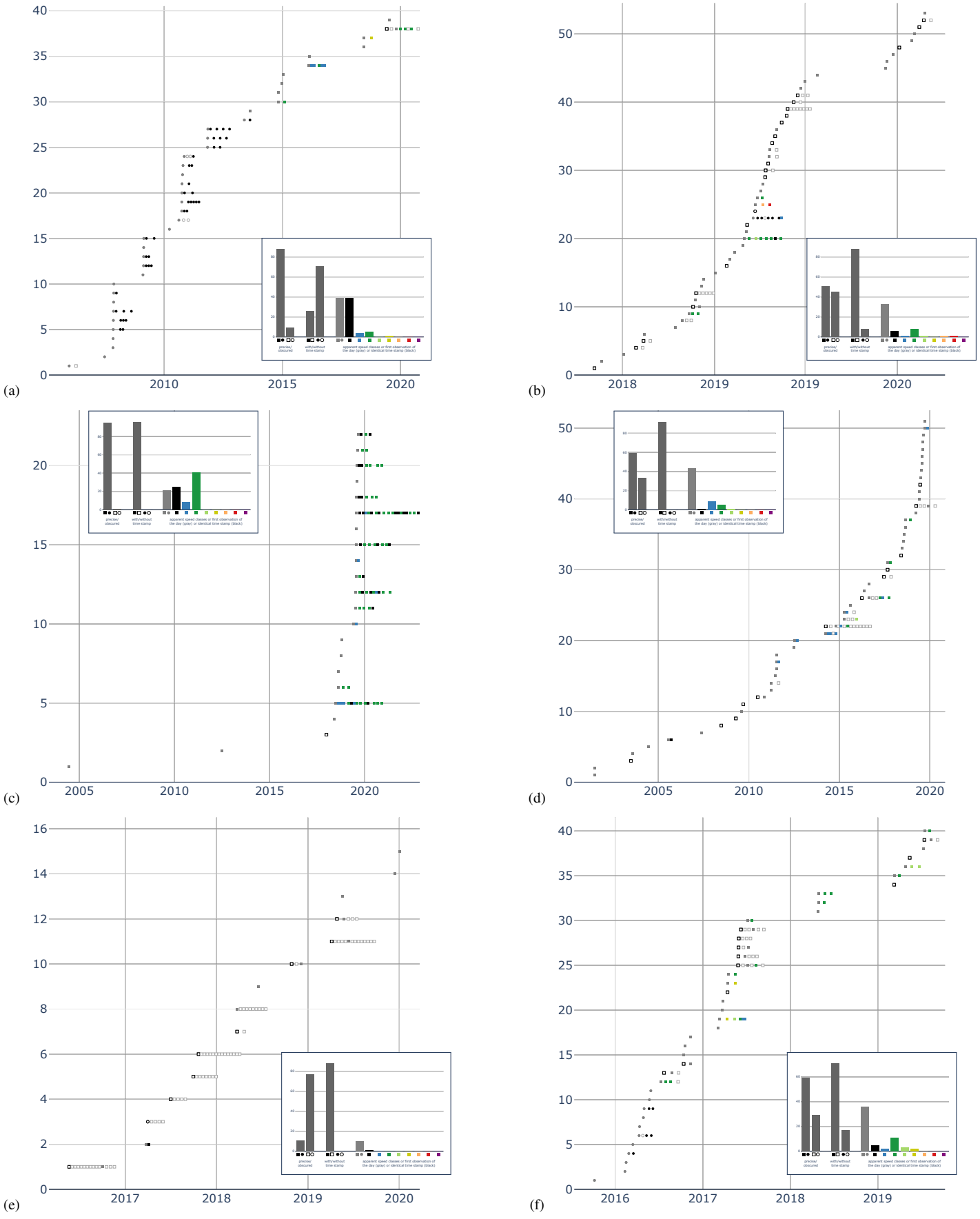


Fig. 72. The motion plausibility profiles of the top 25–30 contributors in the iNaturalist dataset, in decreasing order of number of observations. Same principle and color scale as in Fig. 20. Circles are observations without a valid time stamp, outlined shapes are obscured observations.

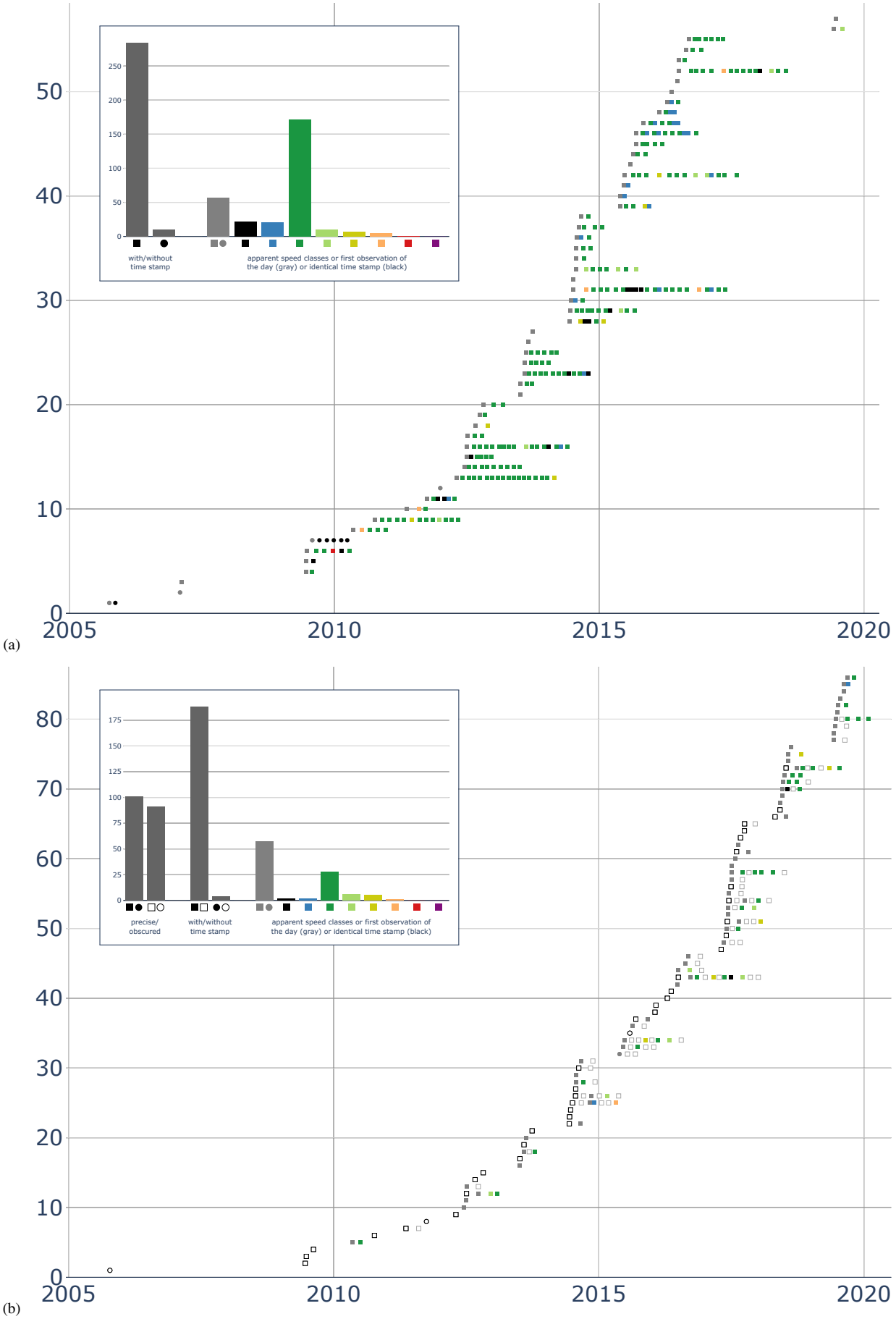


Fig. 73. Direct comparison of the motion plausibility profiles of the same person (“J. Doe”) based on (a) their Panoramio/Flickr data (same motion plausibility profile as in Fig. 16(c) and in Fig. 66(d)) and (b) their iNaturalist data (same motion plausibility profile as in Fig. 20(f) and in Fig. 69(b)).