Mapping Species Sightings via Webscraping

Alex Rufrano and Kate Blackwell

Project Objectives and Contributions

This project aimed to utilize data sourced from publicly available photos opportunistically collected by the public to assist with updating the known distribution of species. Species that were considered included species with higher numbers of sightings, like the American bumblebee (*Bombus pensylvanicus*), White-tailed deer (*Odocoileus virginianus*), Humpback whale (*Megaptera novaeangliae*), American Bison (*Bison bison*), Polar bear (*Ursus maritimus*), and Mallard duck (*Anas platyrhynchos*), and ones that were more rarely sighted, the Antarctic petrel (*Thalassoica antarctica*), or considered invasive in some locations, such as the Asian giant hornet (*Vespa mandarinia*). While a variety of sites are available for the public to share sightings on, ones such as eBird, iNaturalist, iGoTerra, Flickr, Project Noah, and iRecord, this project focused on making use of iNaturalist given its self-reported purpose to promote biodiversity research[1]. After creating a repository of documented sightings, data was sorted and cleaned before a map was generated of the reported sightings for each examined species. The final objective was to compare the publicly reported sightings to the reported distribution of each species from IUCN Red List.

Kate Blackwell contributed to the project by drafting the initial project proposal, which was later refined with Alex Rufrano by narrowing the focus to just the iNaturalist website and expanding the scope to include several species. Alex then worked on assembling the webscraper and sorting and cleaning of the pulled data as well as writing the code that generated the species occurrence map. Kate contributed during this stage with advice on what information to collect based on its use in future scientific analysis and worked with Alex on how best to present the information. From Alex, she learned how to make use of BeautifulSoup for data extraction from a website and then clean and process said data using Python along with utilizing GMAPs to generate a heatmap of the occurrence data. From Kate, he learned what information is of interest for scientific analysis and the need for transparency in data collection methods to ensure reproducibility by other analysts. Kate then modified the occurrence data and wrote a Python script to compare the iNaturalist occurrences of select species to said species' reported distribution on IUCN Red List, a website that acts as an inventory of the global conservation status for most species[2]. During this process, Alex followed along and learned more about the considerations of working with geometric data. Together, they finalized the project report and presentation.

Techniques and Tools

For the first part of the project, Python[3] in a Juypter Notebook[4] was used to navigate to and webscrape iNaturalist one single species at a time. Said species was selected using part of the URL identified as "taxon_id=56887", where 56887 is the species identification number (referring to the American bumblebee in this case). Changing the number in the relevant section in Python[3] changes the species for which data is being extracted. To extract data from the website, it was important to ensure a single page completely loaded by using a webdriver that opened a Google Chrome browser. Since iNaturalist was created via Javascript, the developers made it so that the viewer needs to scroll to the end of the page to load all of the data. Without navigating to the end of the page, only a portion of the data would be retrieved via BeautifulSoup. By making use of a script in the webdriver[5], the entirety of the webpage was loaded by activating the "END" button, which mimics the user function of pressing the "END" button on a webpage. Loading a page requires time[6], and if the page is still loading and the "END" button is automatically pressed, there is a risk of failing to scrap the data. Therefore, a timer set to approximately 5 seconds was added to the loop that presses the "END" button 10 times, and these parameters can be attuned to a user's internet connection speed and computer performance. Extracting the webpage's data via BeautifulSoup[7] required becoming familiar with HTML. The module's find_all method was used to extract all of the requested HTML data based on the specified search parameters. Data extracted for each sighting included the sighting's species name, sighting ID, listed location, and coordinate data. Code also had to be implemented to navigate across multiple pages of sightings as not all of the data was listed on one page. This was achieved using a for-loop that iterated from page 1 to page "n." "n" is the number of pages specified by the user and can be changed in "URL identified as (page=1)" (1 is the default and means only one page of sightings will be pulled).

The extracted sightings were then sorted and standardized using the Pandas[8] and NumPy[9] modules. Location information was consolidated and descriptions like "near" were removed. Removing such descriptions was not considered cause for concern as any sightings only containing location information without geographic coordinates were considered as approximate locations. Sightings without any location information were removed. The sightings data was then processed with an encoder. Some of the sightings had location information only consisting of an address without geographic coordinates or vice versa, but some had both. Location-only addresses were assigned geographic coordinates and location-only coordinates received addresses so that no data was missing one or the other when possible.

More specifically, this process was completed using GeoPy (the encoder) and Nominatim, (user creation for GeoPy). While GeoPy is free to use, there is a limit of up to 900 requests per day, and the user needs to wait at least a second between requests, which required adding a timer and exceptions to the Python[3] code since an HTTP exception will occur if the conditions are not met. As the location information is run through the Geocoder, it is stored in a dataframe. One important consideration was how the data was stored: if the Geocoder could not assign any location information, then nothing would be returned, leading to errors. This accumulation of errors was handled by adding an IF statement that when triggered, made it so that the returned location value from the Geocoder was "None." Location information obtained was returned as one value and needed to be separated into location address and geographic coordinates. Once separated, the two pieces of location information were added to the final dataframe containing the processed sightings. Separating the two pieces of location information was vital for later analysis.

Processed sightings were then imported into GMAPS to create a heatmap for easy viewing. Prior to import, any sightings with "None" in either piece of location information were removed. The generated map was then exported to the user's working folder by requesting an HTML file. To export the map via an HTML file, the IPWIDGETS module was used as its function "embed_minimal_html" permitted the embedding of two convenient parameters: the filename and generated heatmap. This technique also included exporting the final sightings data stored in a dataframe to an Excel document via the dataframe.to_excel function. The filename selected was the name of the species being examined and the time the data was extracted using the Datetime module in order to keep track of multiple extractions of the same species and not automatically replace prior extractions of said species.

The first part of this project was repeated for each species of interest. Each resulting Excel document was then converted into a .csv file after any sightings with missing location information were manually removed. Of the created .csv files, five species (American bison, Antarctic petrel, Humpback whale, Polar bear, and White-tailed deer) were analyzed for the second part of the project. The other species were not included because there were no listed distributions for them on the IUCN Red List site or the data was inaccessible (Mallard duck). Modules, some of which required required additional files, were installed and imported. The .csv file was then read into the Juypter Notebook[4] using the Pandas[9] module. The coordinate latitude and longitude for each sighting was isolated in a new dataframe that was then converted to a list. Using the "Geometry" function from the Shapely[10] module, coordinates were transformed into geometric points that were then compared to a polygon of the distribution data from the IUCN Red List that was made also using the "Geometry" function. The shapefile from the IUCN Red List site was read into using the shapefile[11] module. The number of points both in and out of the distribution polygon were totaled and the sightings data was assigned to the appropriate list based on its designation. The "In" and "Out" geometric points were then back converted to coordinates to make further investigation possible. For example, any sighting falling outside the IUCN Red List distribution could be further examined for possible species identification error on iNaturalist or as movement of the species to new, unknown locations outside its expected occurrence. From there, the polygons made of the IUCN Red List distribution and the iNaturalist sightings were viewed and their areas measured using the build-in Python[3] polygon function and Area[12] module.

Results and Conclusions

Originally, the project proposed to examine just the Antarctic petrel but across several websites (eBird, iNaturalist, iGoTerra, Flickr, Project Noah, and iRecord) used by the public to upload photographs of said species. This proposal originated from the fact that only a few studies actually describe the distribution of Antarctic petrels at sea[13], and the distribution of their breeding colonies on the continent is poorly known. Since solid data on which to base such evidence are often lacking from remote polar areas, particularly in Antarctica, data sourced from publicly available photos opportunistically collected by birdwatchers and tourists can further assist with updating the known distribution of the Antarctic petrel, specifically assisting work identifying breeding and feeding sites.

However, the initial scope of the project changed as it was realized that the task of webscraping each individual site would require significantly more time than was allotted since each website has its own quirks of access and thus each would require unique code to access and extract the required data. The choice was made to focus on iNaturalist given its extensive number of observed species with crowdsourced identification verification. The site was of particular interest due to its start as a Master's project at UC Berkley in 2008 before becoming a joint initiative overseen by CA Academy of Sciences and National Geographic in 2017[1].

Part 1 of the exercise proved to be an excellent challenge and lesson in data extraction and preparation (Table 1). Unlike the other species, the Antarctic petrel had few reported sightings, but a little less than half of them contained geographic coordinates in the initial report. This last detail was of interest as only the humpback whale and polar bear also had geographic coordinates reported, with the rest of the species' location information reported in address form. In future work, a refinement of the extraction process would be to, if possible, extract the EXIF data from posted photos as it may contain the geographic coordinates of where the image was taken. Geographic coordinates are of greater use in scientific research as they are most exact in terms of an animal's location. For this project, geographic coordinates were assigned based on the provided address information, but such data would be considered more unreliable than ones with geographic coordinates, even reported ones not extracted from the sighting photo itself.

Extracting the geographic coordinates from the photo itself may also increase the number of opportunistically-collected sightings, and thus the amount of data with which to work with, for marine species. Of the species examined, barring the Antarctic petrel and Asian giant hornet that started with low sightings, the polar bear and humpback whale dropped from over 900 sightings to around 300 and 100 final sightings, respectively, after processing of the data. The sightings excluded were primarily due to a lack of usable location information that could be assigned a geographic coordinate.

Also of interest, some of the images from the species were not the species itself but footprints or scat. This along with the fact that species identification is crowdsourced may warrant further consideration when deciding which sightings to use in further analysis. In this project, all extracted sightings with location information were utilized, but images not of the species itself may need to be excluded unless the researcher can confirm the scat or footprint belongs to the species under consideration. Furthermore, there is a possibility that species identification is incorrect, so there may be a need to examine each pulled sighting to confirm it is actually the target species.

Species	*Raw # of sightings*	*# of sightings originally with only geographic coord.*	*# of sightings originally with only address then converted to geographic coord.*	*# of sightings originally with only address failed to convert to geographic coord.*	*# of sightings used in heatmap*
American bison	960	0	653	307	653
American bumblebee	960	0	775	185	775
Antarctic petrel	27	8	15	4	23
Asian giant hornet	360	0	179	181	179
Humpback whale	960	11	114	19	125
Mallard duck	960	0	781	179	781
Polar bear	926	23	285	618	308
White-tailed deer	960	0	788	172	788

Table 1. Results of Part 1 of project: details # of data at various stages of processing and analysis.

The possibility of misidentified species became more apparent in the second part of the project. When comparing sightings from iNaturalist to the species' distribution from IUCN Red List for the American bison and white-tailed deer, there were hundreds of sightings located outside of the expected distribution (Table 2). Some of these outside sightings are within range of the IUCN Red List distribution, but a detailed examination of the outside sightings revealed reports of sightings in locations well outside of where the species would be expected. For example, the white-tailed deer, native to North America, was reported to have been seen in countries such as Finland and India. These sightings are most likely misidentifications although there is the possibility that the identification is a true one but the animal is located within a zoo or animal safari of some sort, possibly even being kept as a pet. There is also the possibility that species is invasive in locations well outside of its natural occurrence. Barring outright exclusion, individual verification of each of these sightings would be warranted before inclusion in a final dataset used for scientific analysis. This finding made clear how much verification and additional processing of crowdsourced data is required before its inclusion in any scientific database, but such work may be necessary for species like the Antarctic petrel where so little is known about it that each piece of information, each sighting is essential in increasing our knowledge of the species' distribution.

Species	*# of iNaturalist sightings inside IUCN Red List Area*	*# of iNaturalist sightings outside IUCN Red List Area*	*Total area of IUCN Red List polygon (sq km)*	*Total area of iNaturalist sightings polygon (sq km)*
American bison	5	648	196,913	22,138,114
Antarctic petrel	2	21	4,809,157	5,984,679
Humpback whale	42	83	46,274,955	9,962,567
Polar bear	177	131	19,620,988	7,672,211
White-tailed deer	398	390	18,843,732	33,577,871

Table 2. Results of Part 2 of project: comparison of the # of iNaturalist sightings located within and out of a species' distribution according to the IUCN Red List along with the area of both.

The comparison process was useful for terrestrial species with non-patchy distributions as reported on IUCN Red List. For marine species like the humpback whale, the projected distribution polygon included landmasses, which is not where a whale is likely to be found unless it is dead or beached. Further refinement of the distribution polygon for marine species is thus warranted, and it probably requires the creation of multiple polygons built from subsets of the ICUN Red List data. The same applies to terrestrial species with patchy distributions due to only being found in select locations either due to requiring specialized habitat (e.g. tree frogs located in specific patches of rainforest) or having become so endangered that populations of the species are only found in a few areas. For example, the American bison is only found in a few select locations due to its low abundance, and therefore its projected IUCN Red List distribution would be better as multiple small polygons rather than one large one. That said, it is of interest that so many of the iNaturalist sightings were well outside of the range reported by IUCN Red List. After refinement of the polygon and exclusion of misreported American bison, there may be evidence that the American bison has a broader range than what is reported by the IUCN Red List.

Literature Cited

[1]iNaturalist. Available from https://www.inaturalist.org. Accessed [05/01/2021].

[2]IUCN 2021. The IUCN Red List of Threatened Species. Version 2021-1. https://www.iucnredlist.org. Accessed [05/01/2021].

[3]G. van Rossum, Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995

[4]Kluyver, T., Ragan-Kelley, B., Fernando P&#x27;erez, Granger, B., Bussonnier, M., Frederic, J., … Willing, C. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87–90).

[5]The Selenium Browser Automation Project. (n.d.). Retrieved May 2, 2021, from https://www.selenium.dev/documentation/en/

[6]Time - Time access and conversions¶. (n.d.). Retrieved May 2, 2021, from https://docs.python.org/3/library/time.html

[7]Richardson, L. (2007). Beautiful soup documentation. *April*.

[8]McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

[9]Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

[10]Gillies, S., & others. (2007). *Shapely: manipulation and analysis of geometric objects*. Retrieved from https://github.com/Toblerity/Shapely

[13]Lawhead, J. (n.d.). Retrieved May 2, 2021, from https://pythonhosted.org/Python Shapefile Library/

[12]Area. (n.d.). Retrieved May 2, 2021, from https://pypi.org/project/area/

[13]Tancell, C., Sutherland, W. J., & Phillips, R. A. (2016). Marine spatial planning for the conservation of albatrosses and large petrels breeding at South Georgia. *Biological Conservation*, *198*, 165-176.

Supplementary Figures

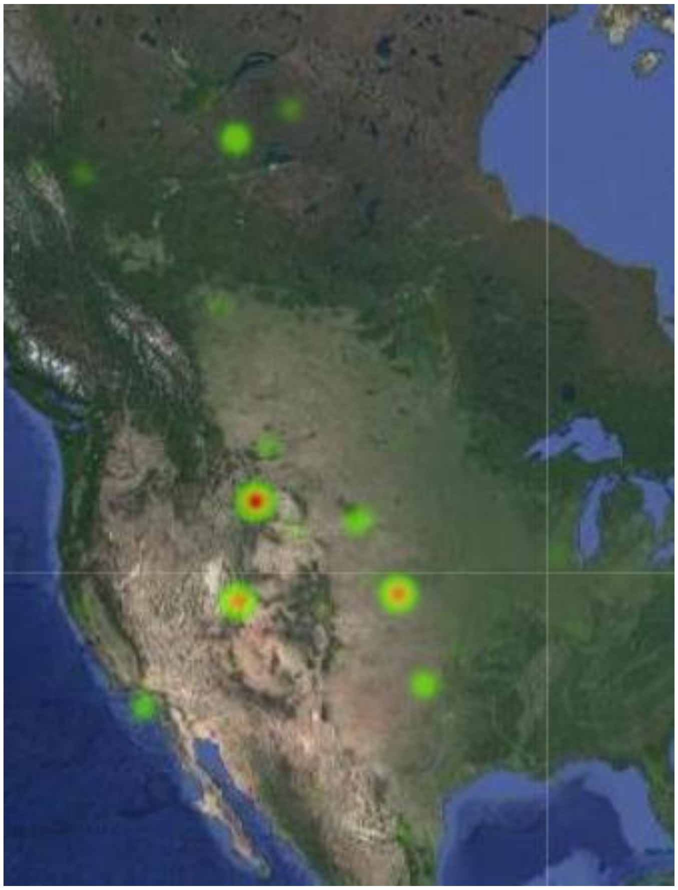

Figure 1. Heatmap of the American bison occurrence from low (green) to high (red).

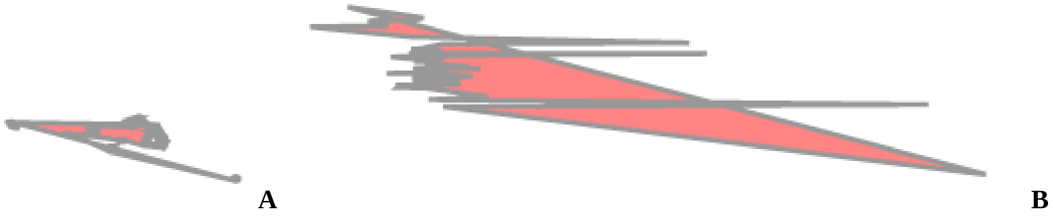

Figure 2. A) Polygon of the ICUN Red List American bison distribution. B) Polygon of the iNaturalist American bison distribution.

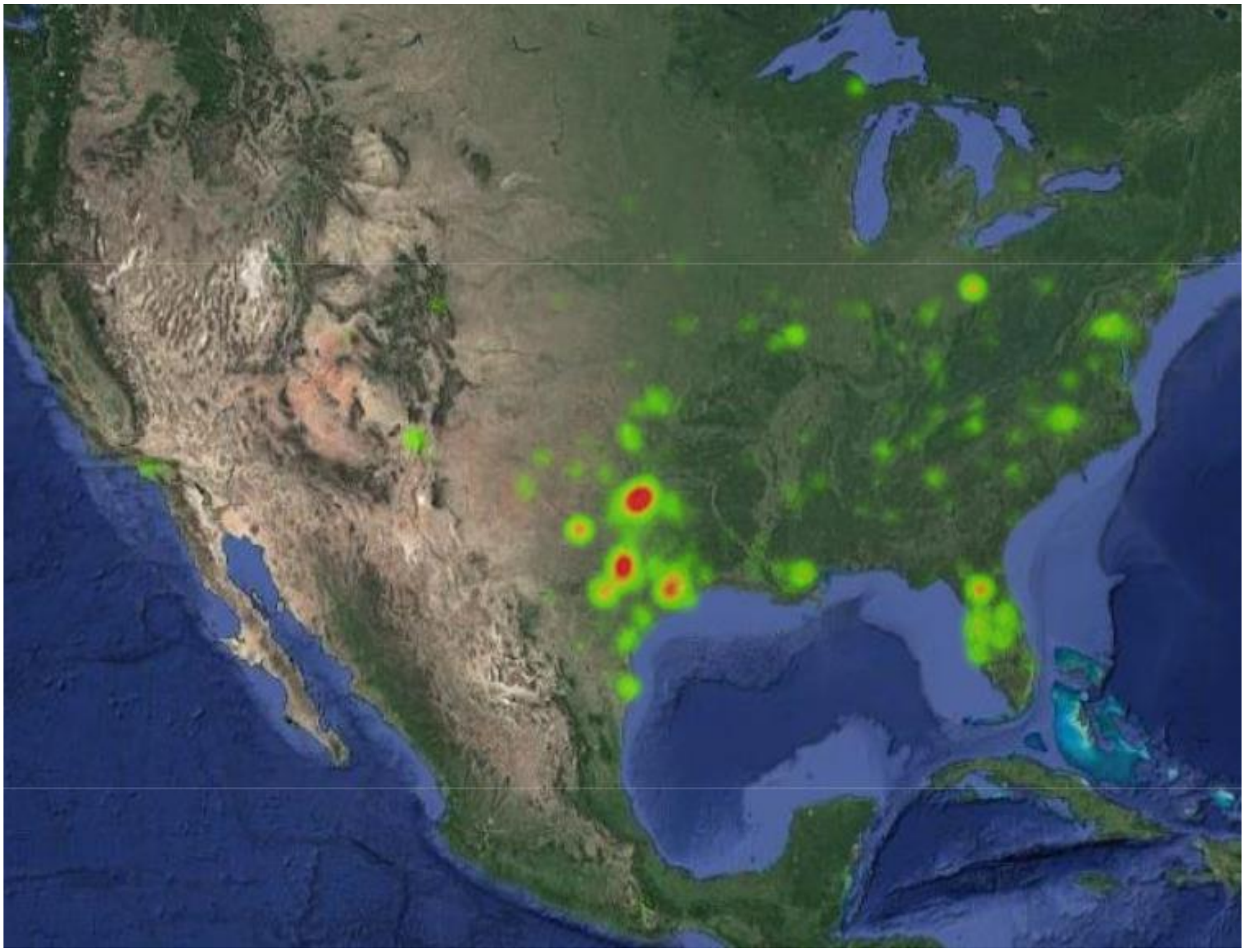

Figure 3. Heatmap of the American bumblebee occurrence from low (green) to high (red).

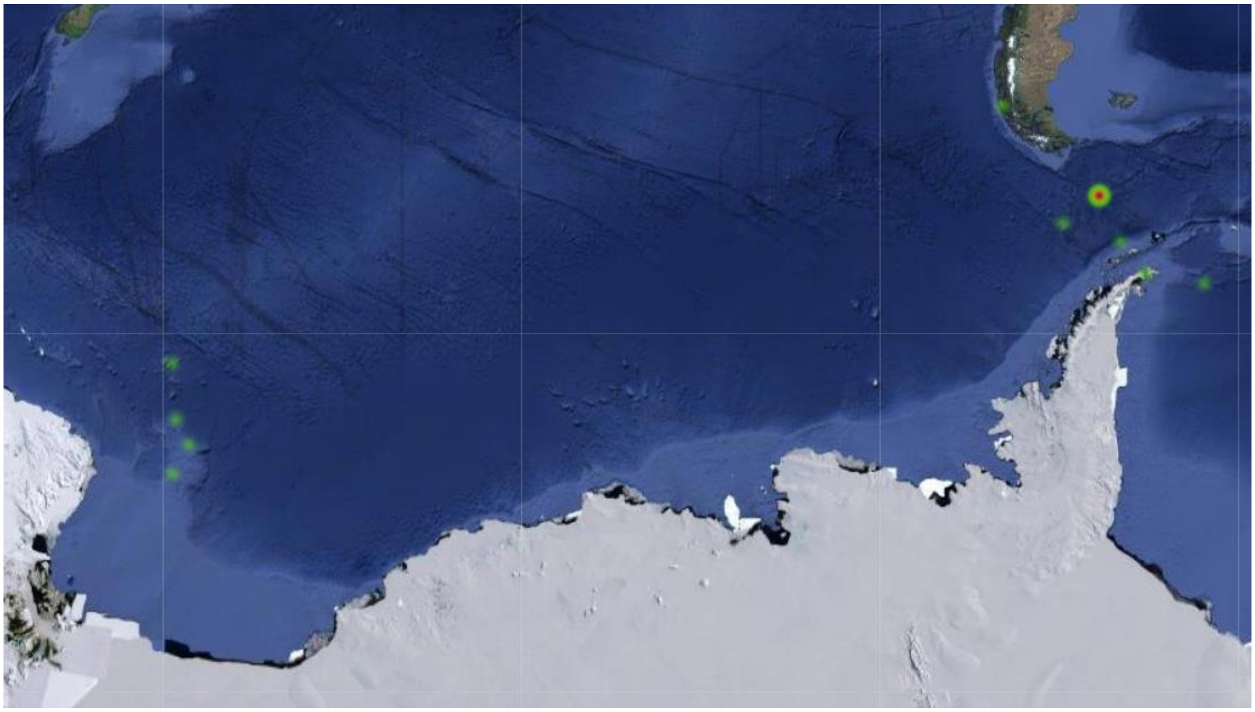

Figure 4. Heatmap of the Antarctic petrel occurrence from low (green) to high (red).

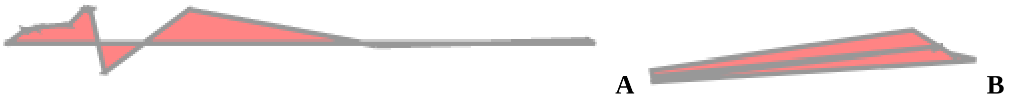

Figure 5. A) Polygon of the ICUN Red List Antarctic petrel distribution. B) Polygon of the iNaturalist Antarctic petrel distribution.

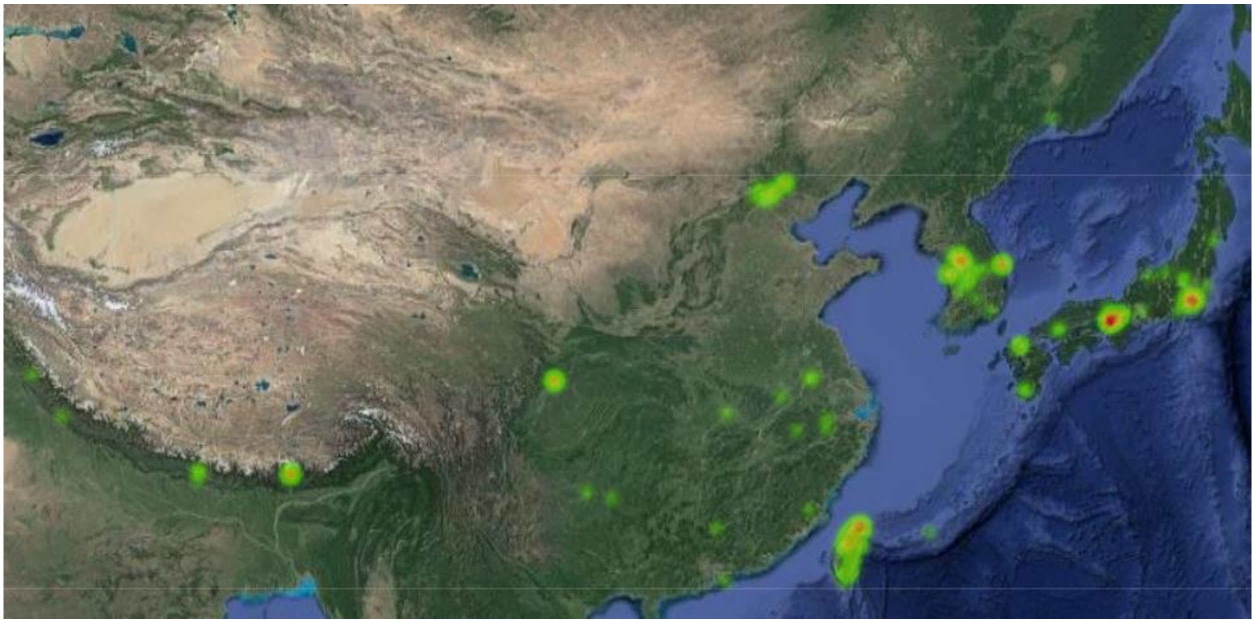

Figure 6. Heatmap of the Asian giant hornet occurrence from low (green) to high (red).

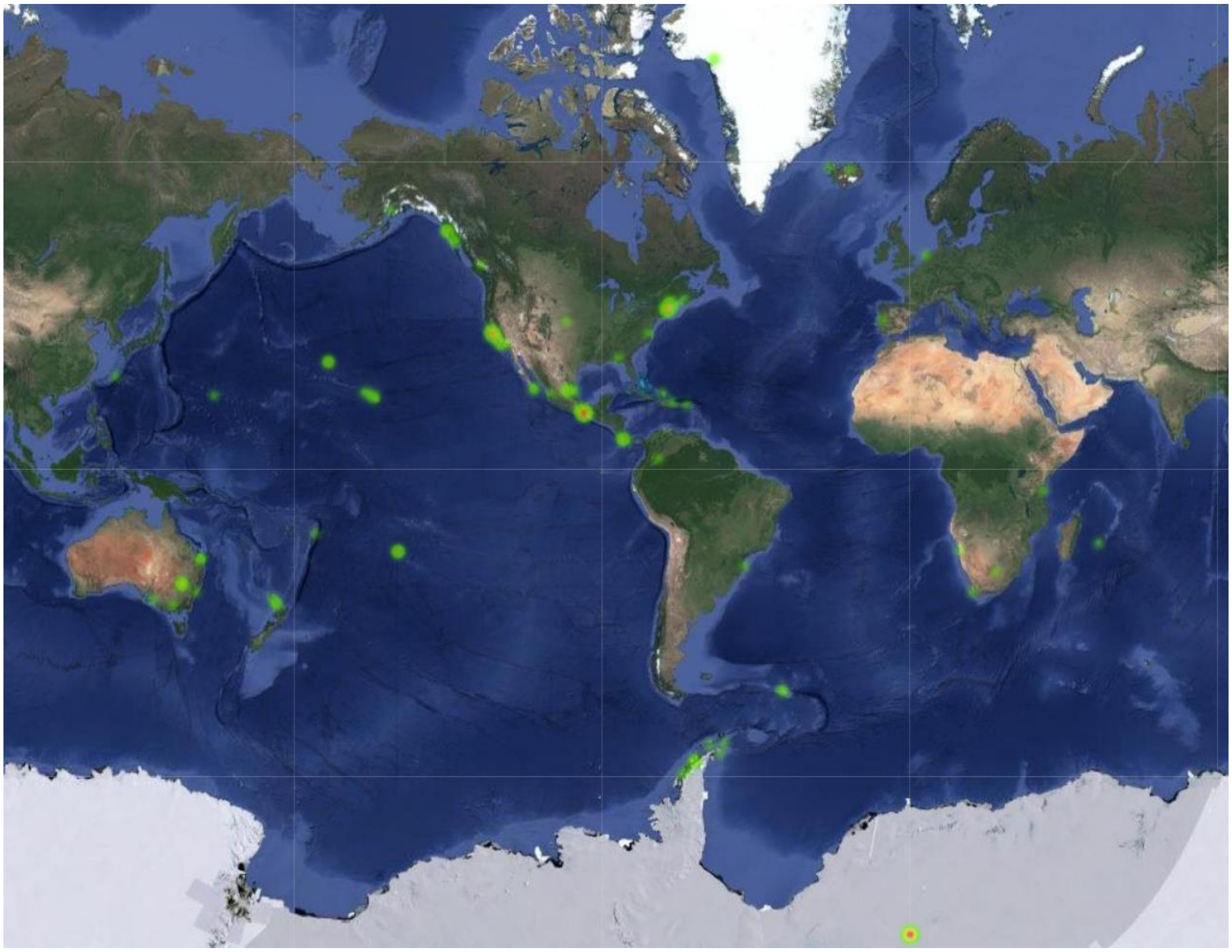

Figure 7. Heatmap of the humpback whale occurrence from low (green) to high (red).

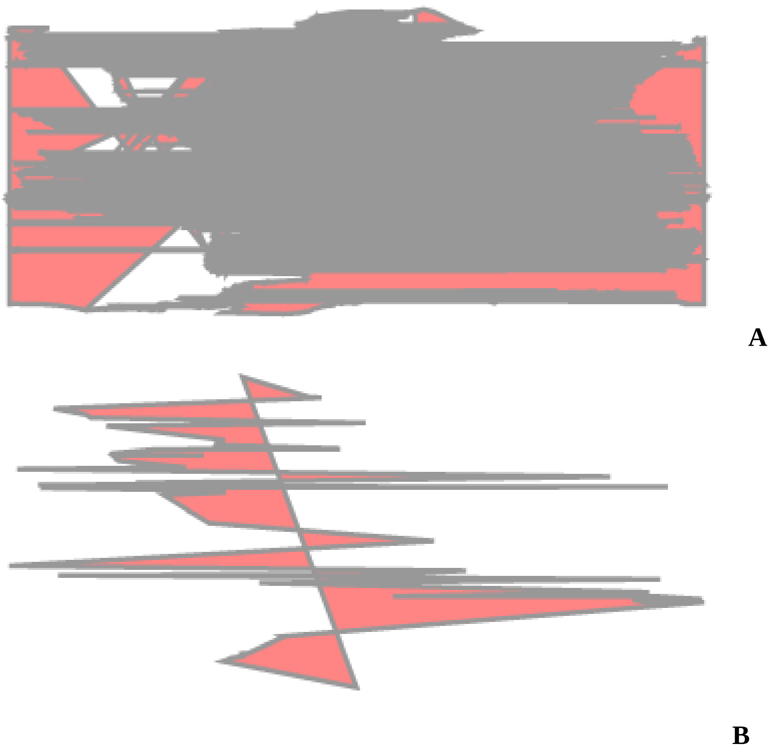

Figure 8. A) Polygon of the ICUN Red List humpback whale distribution. B) Polygon of the iNaturalist humpback whale distribution.

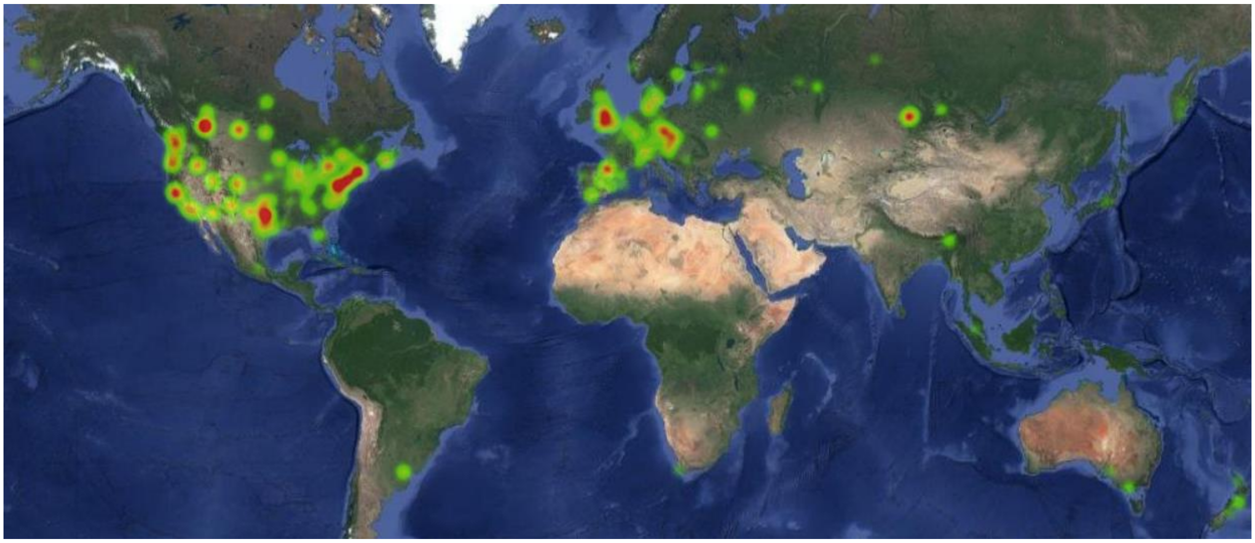

Figure 9. Heatmap of the mallard duck occurrence from low (green) to high (red).

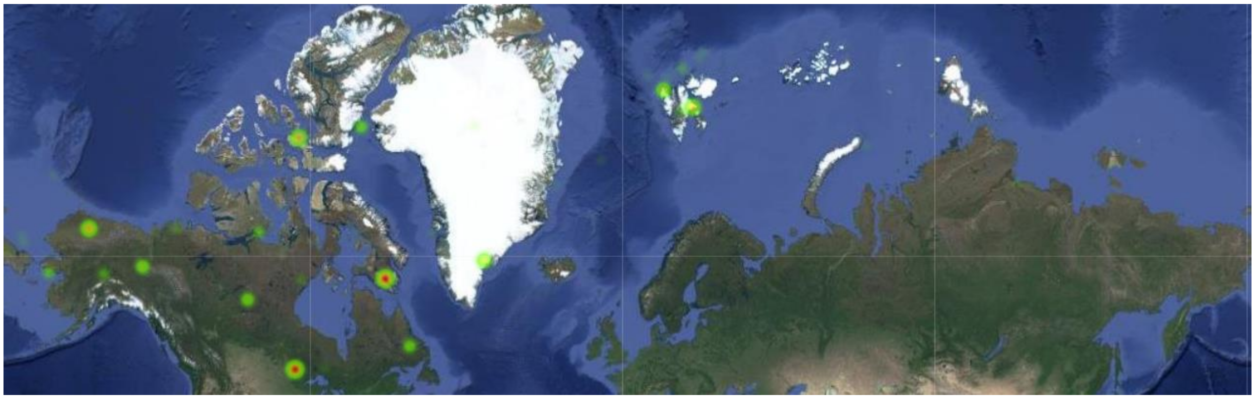

Figure 10. Heatmap of the polar bear occurrence from low (green) to high (red).

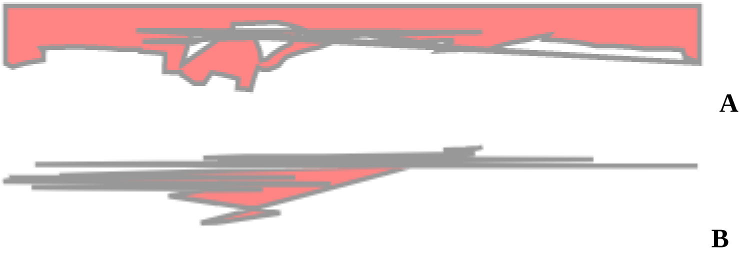

Figure 11. A) Polygon of the ICUN Red List polar bear distribution. B) Polygon of the iNaturalist polar bear distribution.

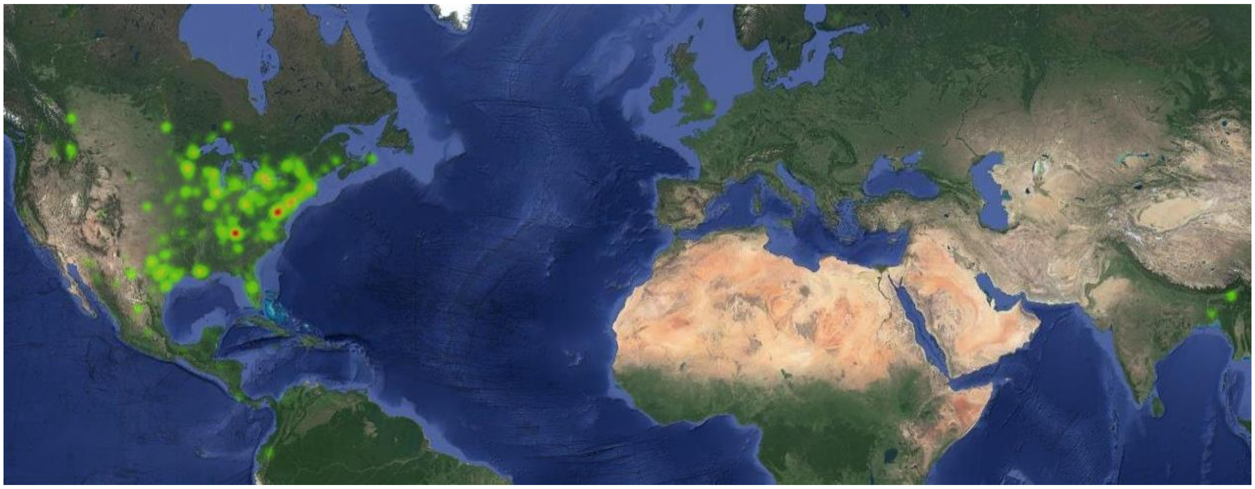

Figure 12. Heatmap of the white-tailed deer occurrence from low (green) to high (red).

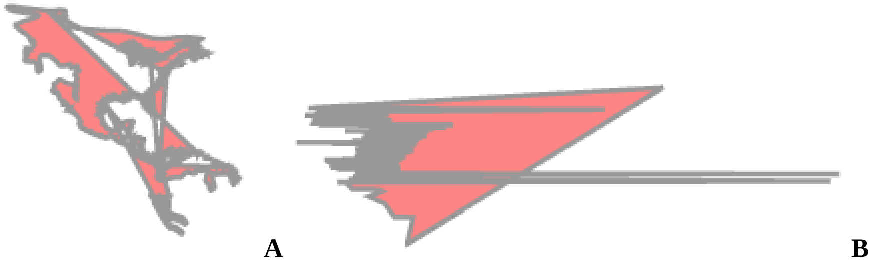

Figure 13. A) Polygon of the ICUN Red List white-tailed deer distribution. B) Polygon of the iNaturalist white-tailed deer distribution.