Using geo-tagged Flickr images to explore the correlation between land cover classes and the location of bird observations

M. Lotfian [1,2], J. Ingensand [1]

[1] University of Applied Sciences and Arts Western Switzerland, School of Business and Engineering Vaud, Institute INSIT, 1400, Yverdon-les-Bains, Switzerland - (maryam.lotfian, jens.ingensand)@heig-vd.ch
[2] Department of Civil and Environmental Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32,20133 Milano, Italy – maryam.lotfian@polimi.it

KEY WORDS: social networks, volunteered geographic information, Flickr, biodiversity, citizen science, geotagged images

ABSTRACT: Social media data are becoming potential sources of passive VGI (Volunteered Geographic Information) and citizen science, in particular with regard to location-based environmental monitoring. Flickr, as one of the largest photo-sharing platforms, has been used in various environmental analyses from natural disaster prediction to wildlife monitoring. In this article, we have used bird photos from Flickr to illustrate the spatial distribution of bird species in Switzerland, and most importantly to see the correlation between the location of bird species and land cover types. A chi-square test of independence has been applied to illustrate the association between bird species and land cover classes and results illustrated a statistically significant association between the two variables. Furthermore, species distributions in Flickr were compared to eBird data, and the results demonstrated that Flickr can be a possible complementary source to citizen science data.

1. INTRODUCTION

Volunteered Geographic Information (VGI), a term first introduced by Goodchild (Goodchild, 2007), is defined as the voluntarily creation or collection of geographic data by individuals. While there are various well-known VGI projects such as OpenStreetMap, there are other forms of VGI known as passive VGI (See et al., 2016) where the main objective of the contributors is not geospatial data collection. Social networks such as Facebook, Twitter, Flickr, etc., which have become popular information sharing sites in the past few years, are a major source of passive VGI (Campagna, 2016) as the majority of the shared information are geo-located thanks to GPS equipped smartphones and cameras. Flickr is one of the largest photo sharing platforms with more than 10 billion photographs. Most of the photos in Flickr are associated with textual data including title, description, and tags (which indicate what is present in the photo), and most of the photos are geo-located. As the majority of photos in Flickr include geolocation, they are being used for various analyses such as environmental and natural disaster monitoring (Sun et al., 2016), location-based behavioural analyses (Kisilevich et al., 2010), location prediction based on images (Weyand et al., 2016) to name a few. The use of social media data for biodiversity monitoring is not well supported by the experts in this field; however, previous studies suggest that Flickr images can be used as a complementary source to citizen science platforms of collecting biodiversity observations (ElQadi et al., 2017).
Although not all the VGI and citizen science data are being validated, data validation in such projects is one of the main concerns and many studies are focused on data quality assurance in such projects (Flanagin and Metzger, 2008; Fonte, Cidália Costa, 2017; Kosmala et al., 2016). In contrast, Flickr data does not go through any validation process, and thus the data requires several filtering steps before being used in scientific analysis. We present in this article the pre-processing steps needed to be taken on Flickr observations, prior to using the dataset for further analysis.

In this article, we aim to explore how Flickr bird images are distributed throughout Switzerland and to analyse the correlation between the distribution of species and land cover classes. In addition, we aim to investigate the level of validity at which Flickr images can be used for analyses regarding biodiversity observations, particularly with regard to the generation of species distribution models by comparing data from Flickr with more structured datasets collected from citizen science projects such as eBird (https://ebird.org/).

This article is structured as follows: In the next section, we will review some of the previous studies on the use of Flickr images for wildlife distribution. We then present our dataset and study area in section three. This is followed by the methodology applied to analyse the images in section four. In section five, we present the results and discussions, and finally the conclusions and future perspectives are presented.

2. RELATED WORK

Few studies have used geotagged Flickr images along with their textual data for analysing biodiversity distributions. In a study by Jeawak et.al. (Jeawak et al., 2017), the authors used georeferenced photos from Flickr to construct a predictive distribution map of the bird species “black woodpecker”. In this work besides the geolocation, they have used Flickr tags, to generate a model, which predicts the probability of a tag T to be associated with a specific location L. To do so, they counted the frequency of tags T reported within a distance D from location L. Depending on the distance of the tag to the location L, different weights are given to the tags using a model called Positive Pointwise Mutual Information (PPMI). This model compares the actual number of times tag T occurs within the distance D to location L, with the expected number of occurrences. The authors have used this model to predict species distribution and they have concluded that a combination of Flickr images with structured data (data collected through

traditional ways, considered as ground truth), showed a better result than using each of them separately. In another study by the same authors (Jeawak et al., 2018), they have used the same model (PPMI) for predicting birds species distribution, but this time rather than using the distance between tag T and location L, they have split their study area to grid cells, and analysed the probability of a tag T to be associated with cell C. They constructed models to verify the locations of species occurrence, once using only Flickr tags, which the target species names are explicitly mentioned, and another time using all Flickr tags, and they concluded that the model performed better when using all Flickr tags. Another study has used the geo-tagged images from Flickr in order to map the distribution of bees and flowering plants in Australia (ElQadi et al., 2017). In this work, they have investigated some elements causing unreliability in social media data and they have looked for ways to mitigate them. To verify the reliability of the images and tags, the authors used Google's reverse image-search tool to find similar images to what was obtained from Flickr along with their text-labels and thus filtered the unrelated Flickr images (for instance distinguishing the images of Honeybees species with the images of jars of honey). As a reference dataset, they have used data from a citizen science platform, which is validated by experts. They have overlapped the two distribution maps obtained from both datasets and similarly to the previous studies, the authors concluded that social media data can be a complementary source to the existing biodiversity data sources.

3. STUDY AREA AND DATASET

Two datasets were used in this study: the geo-tagged Flickr images, which were downloaded using the Flickr API (Application Programing Interface), and the CORINE land cover (https://land.copernicus.eu/pan-european/corine-land-cover) map of Switzerland for 2018. The initial dataset of Flickr images included only the north and central part of the canton of Vaud in Switzerland, but we later extended the dataset to include all of Switzerland. Figure 1 presents the study area as well as the data points that indicate the position of the collected Flickr images.

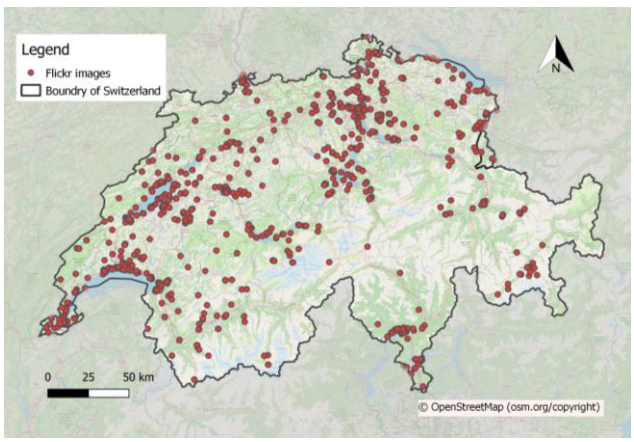

Figure 1: Location of Flickr images with bird tags in Switzerland

4. METHODOLOGY

4.1 Flickr data filtering

The first step was to download the images and to apply filters to them in order to obtain clean data for our analyses. As previously mentioned, we used the Flickr API and set the following requirements before beginning to download the images:

- The media was set to download only photos and not videos
- The starting date was set as the first of January 2018
- Only images with geolocation were downloaded
- Due to the limit of Flickr API in returning up to 500 images per API call, the bounding box was not set to include the whole of Switzerland. We divided Switzerland to ten regions with equal areas, and set the bounding box coordinates per region.
- Finally, in order to find only photos of birds, we needed to identify the correct tag. Flickr has two types of tags: user generated tags, which are added by Flickr contributors, and machine generated tags, which are added to images using Flickr's artificial intelligence. We set the machine generated tags as "any", and the user generated tags as "bird" in four languages of English, German, French, and Italian.

As a result, we obtained the images as well as their metadata, which includes but is not limited to geolocation, date, image URL, image ID, and a list of all tags for each image. Following the download of the images, we applied two major filters to the dataset: image filter and tag filter.

1) *Image filter*: Even though the search was performed using the "bird" tag, there are some other images, which have the same tag but are not birds (e.g. Figure 2). Thus, we used an API from a computer vision platform called Clarifai (https://www.clarifai.com/) to filter out the images that do not include a bird (e.g. statues or drawings of birds that had bird tag, or there were images where the presence of bird was not clear enough). The Clarifai platform, provides a set of pre-trained machine learning (ML) models, which can be used for various objectives, and we used their *general model* as it was the most suitable one for our use case. Using the Clarifai API, we could call the model, which obtains the images and predicts a set of tags of the elements present in the image along with their probabilities (Figure 3). We excluded an image if the probability of a bird to be present (prediction of the tag "bird") was less than 90%.

Figure 2: An example of an image with bird tag, which was filtered out using Clarifai (Flickr image source: (Jag9889, 2018))

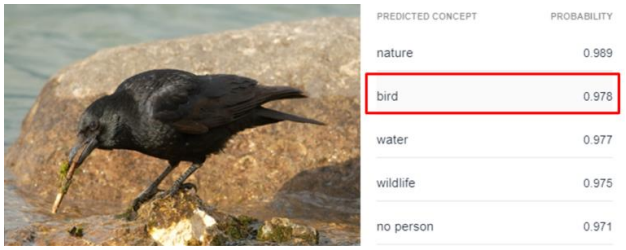

Figure 3: An example of Clarifai predicted tags and their probabilities for an image

2) *Tag Filter:* We had to filter the tags in addition to the images to only have the names of the species, since the tags contained not only the names of the species but also other keywords such as the camera type, the name of the place, general tags (for example, bird in different languages), and sometimes the photographer's name. In order to filter the tags, we used a dataset of bird species names in Switzerland, provided by The Swiss Ornithological Institute (https://www.vogelwarte.ch/). In this dataset there are names of bird species in all four official languages in Switzerland (German, French, Italian, and Rumantsch) plus English. Therefore, we implemented a script to build a matching string function to filter out the tags, which are in close match with the list of bird species names. We excluded the tags, which had a match less than 85% with the species names. Once the automatic tag filtering was completed, we performed a manual verification of the filtered tags to remove the possible duplicates and also to remove the tags which had a close match with bird names but were not a bird species (e.g. species name Verdone (European greenfinch) was matched with city name Yverdon).

Finally, we were able to obtain the unique number of Flickr bird species observed in our study area. However, it is important to note that certain observations were filtered out due to a mismatch in the species name, and in order to provide a list of all observations, a better approach is to train a convolutional neural network (CNN) model (or to use a pre-trained CNN) to extract the species name from the images and then to perform text matching. Our final dataset included the species names and species ID from Vogelwarte, the Flickr tag, image ID, and the geo-locations for the image.

4.2 Species distribution analysis

After obtaining the filtered dataset, in order to visualize the density of distribution of bird observations in our study area we used kernel density analysis (KDE). Moreover, to explore the distribution of the data within various land cover classes in our study area, an additional dataset was created including the CORINE land cover values for each observation point. Thus, the frequency of birds' observations within different land cover types was observed, and a chi-square test of independence was performed to explore the association between the bird species and land cover types. All the analyses from data pre-processing to statistical tests were performed in Python, and Figure 4 shows the workflow applied in this study.

Finally, to evaluate Flickr data using another dataset of bird observations, which is validated by experts (eBird in this case), the species distribution models (SDM) for a bird species called Common Kingfisher (https://en.wikipedia.org/wiki/Common_kingfisher), were generated for both Flickr and eBird data. The Common Kingfisher datasets for eBird and Flickr included 239 and 51 unique observation points respectively, and only the land cover map was used as the input environmental variable to generate the model. To generate the SDMs, we used the Maxent algorithm (Phillips and Dudík, 2008), and to compare the performance of the two models, the AUC (Area under the ROC Curve) metric was used (Bradley, 1997). Furthermore, the correlation between the two raster maps using their pair pixel values was computed to assess the similarity of the two species distribution maps.

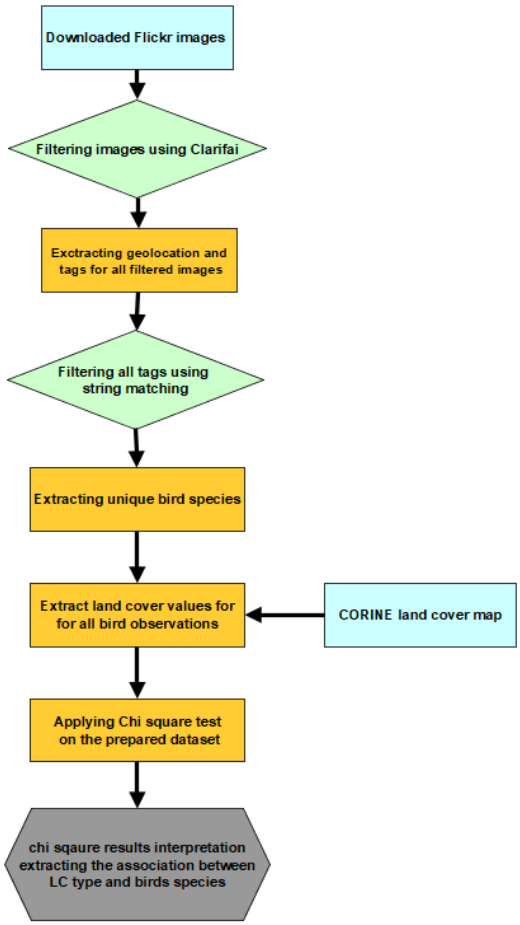

Figure 4: The methodology workflow

5. RESULTS AND DISCUSSIONS

The initial dataset, the images from the central and northern parts of the canton of Vaud, included 282 images which were reduced to 264 after filtering. However, after expanding the study area to whole Switzerland, we obtained 7719 images, which were then reduced to 4610 images after image and text filtering, and it included 2604 unique geolocations. The findings of this study are therefore adapted to the expanded dataset. As mentioned earlier, the majority of user-generated tags include the locations where images are taken, the model of the camera, or the general tags. Few tags, however, are including the names of species, and in most cases, the names are added with a shortened version of the species common names, or misspelled which is why the utilization of Flickr images to

conduct distribution analysis must be done with caution. Figure 5 illustrates the most frequently used tags in this study's downloaded images. The final dataset after tag filtering included 170 unique species with at least five observation points.

The KDE analysis performed for visualizing the density of the distribution is illustrated in Figure 6. From the density map, it can be noticed that the majority of images are concentrated near lakes or around the big cities. This is a common pattern found as well in data from citizen science projects such as eBird (Strimas-Mackey et al., 2020), and it is due to the tendency of the majority of contributors to gather observations in places close to where they live or in more accessible regions, which causes a spatial bias in such datasets (both in Flickr and citizen science data).

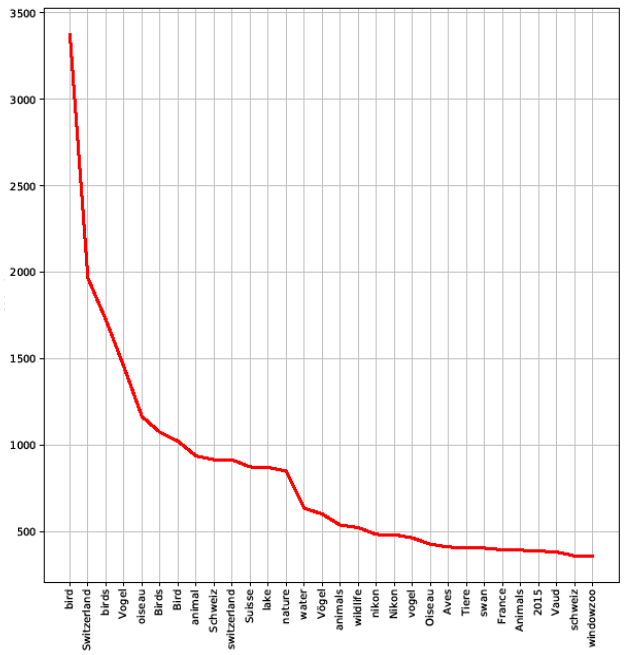

Figure 5: The most frequent tags from the downloaded Flickr images (x-axis: Flickr tags, y-axis: frequency of tags)

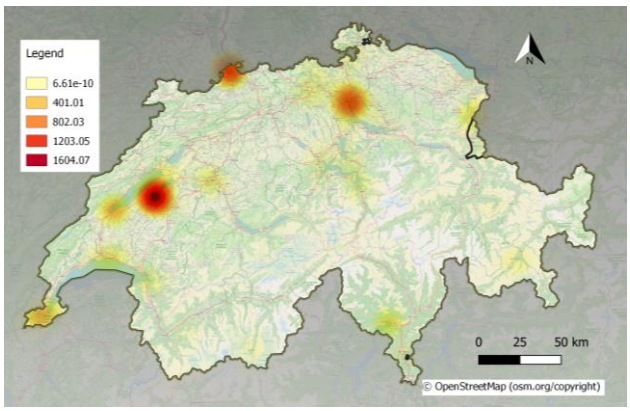

Figure 6: Density map of the Flickr images with bird tags

The frequency of observations within land cover classes is illustrated in Figure 7. The distribution indicates that the majority of observations are in areas with *discontinuous urban fabric* with 738 observations, followed by *water bodies*, *non-irrigated arable land*, and *inland marches* with 271, 175, and 122 observations respectively (The observations with several images from the same point with the same user ID but different image ID's were counted as one record in counting the frequency). This distribution, as expected can be due to collection of data in more accessible areas, or areas where contributors usually spend their leisure time. While a similar distribution pattern is observed for eBird observations in Switzerland, the disparity of observations in some land cover types is more visible in Flickr images than in eBird. This can be due to the difference between the objectives and motivations of social networks and citizen science participants, where in citizen science projects participants have other motivations rather than only data collection for leisure. That is consistent with our hypothesis in the introduction that Flickr data can be a complementary source to citizen science data, but should not be used as the sole source for scientific study of species distributions at this point. Another important point to consider is that certain species that live near human settlements are more acquainted with humans (Stephan et al., 2012), and this familiarity makes approaching and photographing them easier. Other species, on the other hand, can only be captured by experienced bird watchers, and therefore Flickr images could be biased in this case.

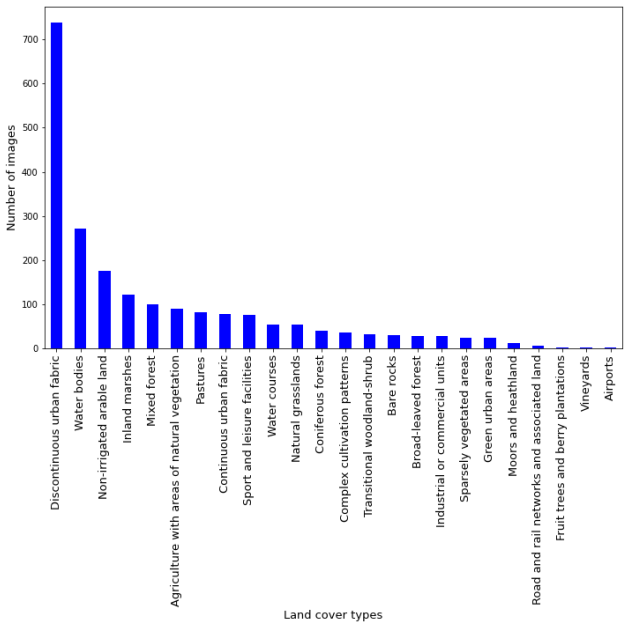

Figure 7: The distribution of Flickr images within various land cover classes from CORINE land cover map of Switzerland

Furthermore, the chi-square test was performed to measure the association between land cover types and bird species, and the Cramer's V metric was computed as a result of the test. Cramer's V is a metric to measure the strength of association between two variables. It ranges between 0 to 1, which values above 0.5 indicate strong association. Thus, the result of chi-square test illustrated a statistically significant association between the land cover types and birds species with Cramer's V = 0.5209 and p-value < 0.0001.

The SDM maps obtained for both datasets are illustrated in Figure 8. The model generated using eBird data performed better with AUC=0.86 compared to the one generated using Flickr data with AUC=0.7, which is reasonable given the number of records in Flickr which was nearly four times less than eBird. While the distribution patterns in both maps look

similar, the distribution from Flickr illustrates higher probability of occurrence in areas with discontinuous urban zones compared to eBird. Table 1 illustrates the statistics comparing the two raster maps, and it shows a very high correlation among the pixel values, supporting the similarity of the distribution between the two maps. From these analyses it can be discussed that Flickr data might be a potential source to address the issue of lack of occurrence species data particularly in SDM studies, given that necessary filtering steps are applied to the data. Moreover, informing the contributors about the value of their data in helping scientific projects can motivate them to contribute higher quality data (Lotfian et al., 2020). However, it is essential to note that a large number of species had few data points (less than 5), and thus we could not evaluate or make any comparisons of such data with eBird observations, and it remains a point for future investigations.

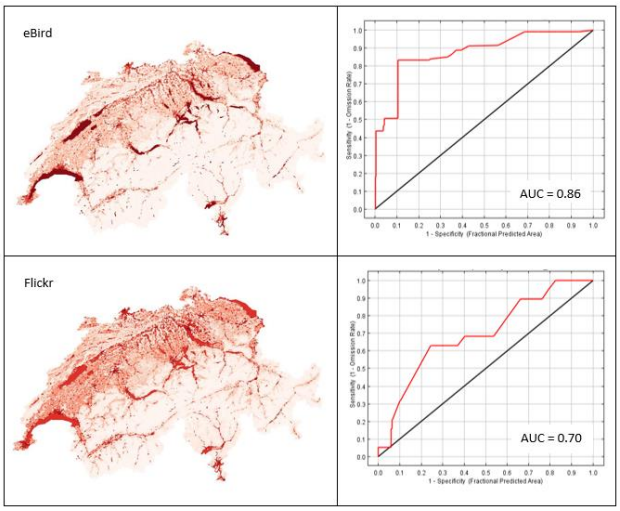

Figure 8: Species distribution maps and the models' performances generated using Maxent for Common Kingfisher using the datasets of eBird (top), and Flickr (down)

Table 1: Statistical comparison of the species distribution maps generated using eBird and Flickr datasets

Statistics of each species distribution model

Layer	MIN	MAX	MEAN	STD
SDM_eBird	0.0923	1.0000	0.2146	0.2091
SDM_Flickr	0.4281	1.0000	0.5331	0.1336

COVARIANCE MATRIX

Layer	SDM_eBird	SDM_Flickr
SDM_eBird	0.00847	0.01222
SDM_Flickr	0.01222	0.02076

CORRELATION MATRIX

Layer	SDM_eBird	SDM_Flickr
SDM_eBird	1.00000	0.92135
SDM_Flickr	0.92135	1.00000

6. CONCLUSIONS AND FUTURE PERSPECTIVES

Due to the developments of mobile technology in the last few years, the number of VGI projects is increasing considerably as many people are now able to collect/contribute geospatial information using their mobile phones. Social networks, as a source of passive VGI, are also growing and becoming the main information-sharing tools among people. As one of the largest websites for photo sharing, Flickr is attracting the attention of scientists as it offers geo-located photos along with textual information that can be used for many scientific analyses.

In this paper, we have used Flickr bird images for Switzerland to observe the distribution of bird species as well as to determine whether or not there is any association between the distribution of different birds species and land cover types. The results illustrated that, the data are more concentrated near lakes, and low-density urban areas. Moreover, a statistically significant association was observed between land cover types and bird species data from Flickr. In this article, we illustrated that the Flickr dataset can be useful in identifying spatial patterns of observations and behaviour of observers. However since the data are not expert-verified it cannot be used exclusively (in the absence of other structured datasets) to produce distribution models of species. The results showed that for common species, the SDM can give results close to citizen science data, however, as many species had very few observations, the evaluation of Flickr data for those species remain unclear and no comparison could be made. This remains a point for future investigation. Another interesting argument for future analyses is to look for alternative approaches for tag filtering and extracting useful information from Flickr tags, such as using CNN to predict species names from images and then to compare them with Flickr tags. Finally, we aim to replicate the analyses in other areas and to see if we can reach similar conclusions, and we plan to extend this study by looking deeper into some research questions, including but not limited to the following:

- To what extent would such species distribution studies based on Flickr images be interesting or useful for Flickr users?
- How would we profit more from using Flickr data in scientific biodiversity studies, given the disparity between data contributed to Flickr (with a focus on contributing more artistic images such as well-composed scenery) and data contributed to standard citizen science projects (with a focus on contributing species data rather than beautiful photos)?

REFERENCES

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7): 1145–1159. DOI: 10.1016/S0031-3203(96)00142-2.

Campagna M (2016) Social Media Geographic Information: Why social is special when it goes spatial? *European Handbook of Crowdsourced Geographic Information* (August): 45–54. DOI: 10.5334/bax.d.

ElQadi MM, Dorin A, Dyer A, et al. (2017) Mapping species distributions with social media geo-tagged images: Case studies of bees and flowering plants in Australia. *Ecological Informatics* 39: 23–31. DOI: 10.1016/j.ecoinf.2017.02.006.

Flanagin AJ and Metzger MJ (2008) The credibility of volunteered geographic information. *GeoJournal* 72(3–4): 137–148. DOI: 10.1007/s10708-008-9188-y.

Fonte, Cidália Costa et al. (2017) Assessing VGI Data Quality. *Mapping and the Citizen Sensor*: 137–163. DOI: https://doi.org/10.5334/bbf.g.

Goodchild MF (2007) Citizens as sensors: The world of volunteered geography. *GeoJournal* 69(4): 211–221. DOI: 10.1007/s10708-007-9111-y.

Jag9889 (2018) "Corbeau et Fromage." Available at: https://www.flickr.com/photos/jag9889/49916653626/.

Jeawak SS, Jones CB and Schockaert S (2017) Using flickr for characterizing the environment: An exploratory analysis. *Leibniz International Proceedings in Informatics, LIPIcs* 86(21): 1–13. DOI: 10.4230/LIPIcs.COSIT.2017.21.

Jeawak SS, Jones CB and Schockaert S (2018) Mapping wildlife species distribution with social media: Augmenting text classification with species names. *Leibniz International Proceedings in Informatics, LIPIcs* 114(45): 1–6. DOI: 10.4230/LIPIcs.GIScience.2018.34.

Kisilevich S, Krstajic M, Keim D, et al. (2010) Event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections. *Proceedings of the International Conference on Information Visualisation*. IEEE: 289–296. DOI: 10.1109/IV.2010.94.

Kosmala M, Wiggins A, Swanson A, et al. (2016) Assessing data quality in citizen science. *Frontiers in Ecology and the Environment* 14(10): 551–560. DOI: 10.1002/fee.1436.

Lotfian M, Ingensand J and Brovelli MA (2020) A Framework for Classifying Participant Motivation that Considers the Typology of Citizen Science Projects. *ISPRS International Journal of Geo-Information* 9(12): 704. DOI: 10.3390/ijgi9120704.

Phillips SJ and Dudík M (2008) Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography* 31(2): 161–175. DOI: 10.1111/j.0906-7590.2008.5203.x.

See L, Mooney P, Foody G, et al. (2016) Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information* 5(5). DOI: 10.3390/ijgi5050055.

Stephan C, Wilkinson A and Huber L (2012) Have we met before? Pigeons recognise familiar human faces. *Avian Biology Research* 5(2): 75–80. DOI: 10.3184/175815512X13350970204867.

Strimas-Mackey M, Hochachka WM, Ruiz-Gutierrez V, et al. (2020) *Best Practices for Using EBird Data v1.0*. Zenodo. DOI: 10.5281/zenodo.3620739.

Sun D, Li S, Zheng W, et al. (2016) Mapping floods due to Hurricane Sandy using NPP VIIRS and ATMS data and geotagged Flickr imagery. *International Journal of Digital Earth* 9(5). Taylor & Francis: 427–441. DOI: 10.1080/17538947.2015.1040474.

Weyand T, Kostrikov I and Philbin J (2016) Planet - photo geolocation with convolutional neural networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9912 LNCS: 37–55. DOI: 10.1007/978-3-319-46484-8_3.