Mapping Wildlife Species Distribution With Social Media: Augmenting Text Classification With Species Names

Shelan S. Jeawak[1]
Cardiff University, School of Computer Science and Informatics, Cardiff, UK
JeawakSS@cardiff.ac.uk

Christopher B. Jones
Cardiff University, School of Computer Science and Informatics, Cardiff, UK
JonesCB2@cardiff.ac.uk

Steven Schockaert[2]
Cardiff University, School of Computer Science and Informatics, Cardiff, UK
SchockaertS1@cardiff.ac.uk

Abstract

Social media has considerable potential as a source of passive citizen science observations of the natural environment, including wildlife monitoring. Here we compare and combine two main strategies for using social media postings to predict species distributions: (i) identifying postings that explicitly mention the target species name and (ii) using a text classifier that exploits all tags to construct a model of the locations where the species occurs. We find that the first strategy has high precision but suffers from low recall, with the second strategy achieving a better overall performance. We furthermore show that even better performance is achieved with a meta classifier that combines data on the presence or absence of species name tags with the predictions from the text classifier.

2012 ACM Subject Classification Computing methodologies → Machine learning, Information systems

Keywords and phrases Social media, Text mining, Volunteered Geographic Information, Ecology

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.34

Category Short Paper

1 Introduction

The value of social media to assist in mapping and predicting geospatial phenomena has been demonstrated in areas including the occurrence of disease, social unrest, natural disasters, levels of wellbeing and characteristics of the man-made and natural environment [7, 8]. In the fields of environmental monitoring and wildlife observation there is clearly strong potential for exploiting social media, reflected in the fact that searching for named species on photo-sharing websites such as Flickr often reveals thousands of results, many of which are associated with coordinates and almost all with time stamps. It can be envisaged that these observations could complement the many effective citizen science campaigns that record aspects of the natural environment and assist environmental scientists in understanding the

[1] Shelan S. Jeawak has been sponsored by HCED Iraq.
[2] Steven Schockaert has been supported by ERC Starting Grant 637277.

10th International Conference on Geographic Information Science (GIScience 2018).
Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 34; pp. 34:1–34:6
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

occurrence and behaviour of animals and plants [4]. Although many mentions of species names in social media might not correspond to records of actual occurrences, several studies have confirmed the validity of significant numbers of species observations in social media [1, 2]. While these studies highlight the potential value of such data, little progress has been made to date on developing reliable automated methods for exploiting all the textual content of social media postings for tasks such as mapping species distributions.

Here we present the results of experiments to predict species distribution based on geocoded social media postings from the Flickr website. As a baseline approach we study the performance of a method that predicts the occurrence of a species in a given region if there is at least one photograph on Flickr from that region which has been tagged with the name of the species (using either its common name or scientific name). This method is then compared with a standard machine learning based text classification approach, in which all Flickr tags are used, and in which a species may be predicted to occur in a region even if no photographs in that region have been tagged with its name. For the text classifier, we follow the method from [6]. In particular, we show that the best results are obtained by a meta-classifier, which combines the prediction of the text classifier with information about the occurrence of the species name in or near the given region. These results clearly show that better distribution models can be found by taking explicit account of the occurrence of the species name as a tag, in combination with exploiting all other tags.

2 Related Work

An overview of the potential for exploiting social media in conservation and biodiversity was provided by Di Mini et al [3], who conducted a study of the use of social media platforms for posting observations of nature. The most commonly used platforms were, in order of level of sharing of nature related content: Facebook, Instagram, Twitter, Youtube, Flickr and LinkedIn. The potential of Flickr for mapping wildlife observations was illustrated by Barve [1] who mapped geotagged postings that included the scientific or common names for the Monarch Butterfly and the Snowy Owl, although that study did not conduct any systematic evaluation of the quality of the retrieved data. Daume [2] performed a manual evaluation of a sample of Twitter postings that named three invasive species (using associated photos for validation). They identified factors correlated with valid observations, such as the presence of a linked photo and tags that describe the environment (e.g. ‘leaves’ and ‘tree’). The present work exploits such associated tags in predicting species distribution. An approach to validating individual observations in Flickr was described by ElQadi et al [5] who used Google’s reverse image-search service to find photos similar to those in Flickr postings. The tags of the Google photos were then compared with those in Flickr in an attempt to filter out non-wildlife images. In our work we learn an association between all Flickr tags and the presence of particular species at a location.

The methods presented here build on the work of [6] which exploited weighted values of all tags to train an SVM (support vector machine) classifier to predict the presence of various environmental phenomena including species. In looking at species distribution no distinction was made in [6] between whether the species name was present or not and the focus was on the additional value that Flickr tags provide relative to scientific data such as climate and landcover.

3 Methodology

The objective of this paper is to find a method that can use Flickr tags for predicting the occurrence of wildlife species. To this end, we split the target spatial area into grid cells

$C = \{c_1, ..., cx_m\}$ and associate each cell with all the georeferenced Flickr tags that occur within the cell. Following [6], we use Positive Pointwise Mutual Information (PPMI) to weight how strongly tag t is associated with cell c. In particular, PPMI compares the actual number of occurrences with the expected number of occurrences (given how many tags occur overall in c and how common the tag t is). Let *f(t,c)* be the number of times tag t (from the set of all tags T) occurs in the cell c. Then the weight *PPMI(t,c)* is given by $\max\left(0, \log\left(\frac{P(t,c)}{P(c)P(t)}\right)\right)$ where:

$$P(t,c) = \frac{f(t,c)}{N} \quad P(t) = \frac{\sum_{c' \in C} f(t,c')}{N} \quad P(c) = \frac{\sum_{t' \in T} f(t',c)}{N} \quad N = \sum_{t' \in T} \sum_{c' \in C} f(t',c')$$

Each cell c is now represented as a sparse vector V_p, encoding the PPMI weight of all the tags in c. We assume that a training set $K \subset C$ is available which contains cells with known ground truth species observations and a testing set $U \subset C \setminus K$ containing cells whose species presence our method will try to estimate.

Our method of estimating the presence of a particular species s in cell c involves learning two classifiers $SVM1$ and $SVM2$. The aim of the first classifier $SVM1$ is to make initial predictions for the cells in the testing set U using the feature vector representation V_p. To give a higher confidence to tags that correspond to the name of the species, we combined the output of $SVM1$ (i.e. classifier confidence score value) with information about the presence or absence of the *Common Name* or the *Scientific Name* of that species in the cell c or the neighboring cells. In particular, the cell c is now represented as a feature vector V_m which contains three features: the confidence value predicted by $SVM1$, the presence of the species actual name in c as a binary feature (being 1 if the c contains the actual name and 0 otherwise), and the percentage of neighbours that contain the species name (again as a common or scientific name) as tag. The second classifier $SVM2$ is learned using the feature vector V_m to give the final estimation.

4 Experimental Evaluation

4.1 Data Acquisition

In this work we use two datasets: the ground truth species distribution from the National Biodiversity Network Atlas (NBN Atlas)[3] and the geocoded social media postings from the photo sharing website Flickr[4]. The NBN is a collaborative project committed to making biodiversity information available via the NBN Atlas. This dataset covers the UK and Ireland. We used the Flickr API to collect approximately 12 million georeferenced Flickr photographs within the UK and Ireland in September 2015. However, our analysis in this paper will focus only on the tags associated with these photographs. The NBN Atlas dataset contains a total of 302 birds with at least 1000 observations, of which 200 have a name that occurs in at least 100 Flickr photographs. Among these, we have considered a random sample of 50 birds for our experiments. Note that even species with a large number of occurrences may possibly only occur in a few cells.

[3] NBN Atlas occurrence download at `http://nbnatlas.org`. Accessed 19 April 2018.
[4] `http://www.flickr.com`

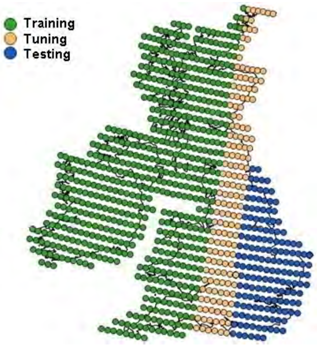

Figure 1 Training, Tuning, and Testing regions.

4.2 Experimental Settings and Baselines

In the experiments, we consider a binary classification problem for each of the selected birds. Specifically, the task we consider is to predict in which of the grid cells the bird occurs (i.e. for which grid cells the NBN Atlas data contains at least one observation). We test our method at three levels of granularity, considering grid cells of size 10, 20 and 30 kilometers. The set of cells C was split into two-thirds for training, one-sixth for testing, and one-sixth for tuning the SVM parameters. It is known that the quality of any supervised model is strongly affected by the way in which the data are divided. Therefore, we split the study area into geographically separated regions, as shown in Figure 1, to test the ability of our method to make predictions about geographic regions for which no observation records are given. This makes the task more challenging than choosing the cells randomly, due to possible differences between the training and testing regions. Finally, for formal evaluation we compared the results of three different methods: "Species Names" which predicts that the species occurs if its common or scientific name appears in at least one Flickr photo in the test cell, "All Flickr Tags" ($SVM1$) which uses the PPMI-based feature vector modelling all Flickr tags to train an SVM classifier using the cells in the training set and predict labels for the cells in the testing cells, and finally "Meta features"($SVM2$) which is our proposed method, as described in Section 3.

4.3 Results and Discussion

The results of predicting species distribution are reported in Table 1 in terms of the average accuracy, average precision, average recall, average F1 score, and average Area Under the ROC Curve (AUC) over the 50 birds. The results clearly show that "All Flickr Tags" significantly outperforms "Species Names". However, the proposed meta-classifier leads to the best results overall, especially in terms of F1 score.

While the "All Flickr Tags" approach works well overall, we found a few cases where using only the species names led to better performance. Perhaps unsurprisingly, this is mostly the case when the number of NBN records (i.e. True labels) in the training region is low, as there may not be enough training data to effectively learn an SVM classifier in such cases. To illustrate such issues, Table 2 shows the F1 scores of 5 individual species. As can be seen, for common species such as Mallard, Dunlin, and Green Sandpiper, the "All Flickr Tags" method performs rather well. In contrast, for some less common species (or species which only occur in particular geographic contexts), such as Atlantic Puffin and Nightingale, we found better results when using the "Species name" method. Interestingly, our proposed meta classifier, which takes account of both the species presence data and the

Table 1 Results for predicting the distribution of 50 species across the testing area.

Dataset	Cell Size	Accuracy	Precision	Recall	F1 Score	AUC
Species Names	10 km	0.520	0.876	0.109	0.183	0.550
All Flickr Tags	10 km	0.779	0.787	0.500	0.560	0.801
Meta features	10 km	0.825	0.820	0.603	0.637	0.850
Species Names	20 km	0.501	0.943	0.241	0.355	0.613
All Flickr Tags	20 km	0.784	0.852	0.639	0.705	0.893
Meta features	20 km	0.870	0.907	0.811	0.832	0.917
Species Names	30 km	0.567	0.970	0.384	0.515	0.684
All Flickr Tags	30 km	0.831	0.868	0.758	0.795	0.943
Meta features	30 km	0.919	0.943	0.896	0.905	0.952

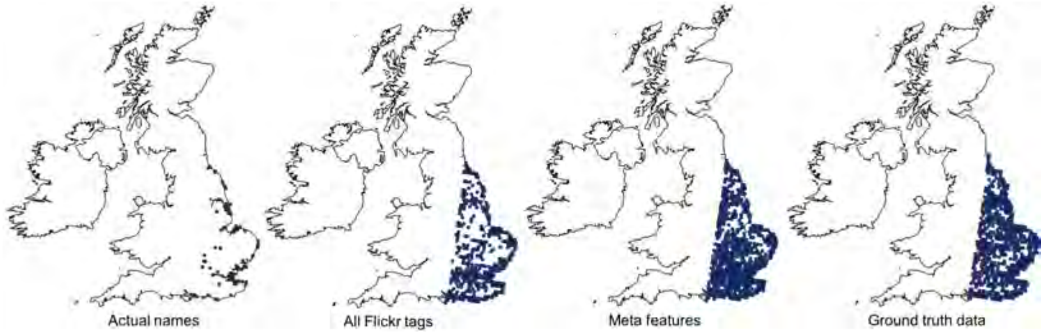

Figure 2 Prediction of the Dunlin distribution across the testing area with 10km grid cells.

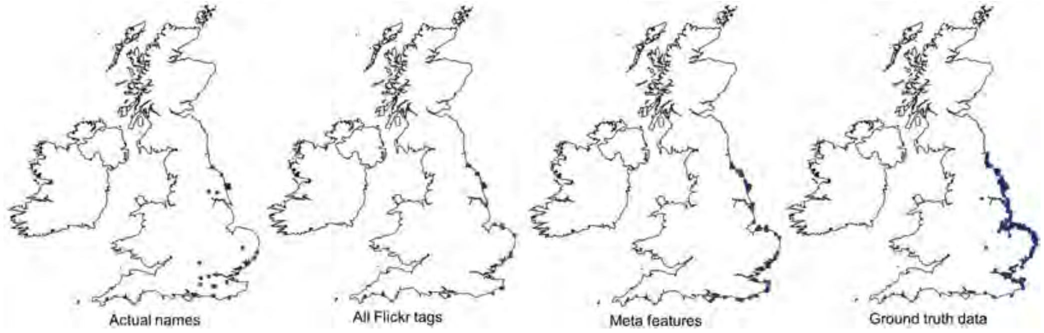

Figure 3 Prediction of the Atlantic Puffin distribution across the testing area with 10km grid cells.

all tags classification for nearby regions, outperforms both of the other methods for almost all the considered species.

Figures 2 and 3 visually illustrate the performance of our method. Note that these species (like most of the considered birds) occur in fewer than 50% of the cells, which is intuitively why the "All Flickr Tags" method is more cautious in predicting occurrence (i.e. in absence of any reason to predict occurrence, it is safer for a classifier to predict non-occurrence).

5 Conclusions and Future Work

In this paper we have presented a method for mapping the location of wildlife species occurrence using the evidence of tags from the photo sharing web site Flickr. We have shown

Table 2 F1 scores for predicting the distribution of individual species using different methods.

	No.NBN	No.Flickr	Cell	Species	All Flickr	Meta
	records	photos	size	Names	Tags	features
Mallard	1718823	11831	10 km	0.640	0.978	0.985
(Anas platyrhynchos)			20 km	0.899	0.974	0.986
			30 km	0.955	0.988	0.992
Dunlin	278872	796	10 km	0.196	0.630	0.744
(Calidris alpina)			20 km	0.346	0.920	0.969
			30 km	0.553	0.980	0.996
Green Sandpiper	103295	187	10 km	0.077	0.610	0.806
(Tringa ochropus)			20 km	0.195	0.849	0.955
			30 km	0.367	0.906	0.980
(Common) Nightingale	24437	383	10 km	0.128	0.0	0.401
(Luscinia megarhynchos)			20 km	0.326	0.0	0.705
			30 km	0.512	0.0	0.835
(Atlantic) Puffin	11551	2512	10 km	0.152	0.136	0.367
(Fratercula arctica)			20 km	0.173	0.359	0.518
			30 km	0.264	0.476	0.630

that while a method based simply on the presence or absence of the species name provides good precision, much better overall accuracy, with similar precision, can be achieved with a machine learning classifier that combines the presence-absence data with predictors based on all the textual tags of the photos.

One line of future work is to investigate the use of a text classifier to estimate confidence in observations of wildlife species in individual social media postings. This could be of particular value when considering postings that mention a species name but in a context that might be unrelated to its occurrence in nature.

References

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

2 Stefan Daume. Mining twitter to monitor invasive alien species?An analytical framework and sample information topologies. *Ecological Informatics*, 31:70–82, 2016.

3 Enrico Di Minin, Henrikki Tenkanen, and Tuuli Toivonen. Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, 3:63, 2015.

4 Janis L. Dickinson, Benjamin Zuckerberg, and David N. Bonter. Citizen science as an ecological research tool: Challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics*, 41:149–172, 2010.

5 Moataz Medhat ElQadi, Alan Dorin, Adrian Dyer, Martin Burd, Zoe Bukovac, and Mani Shrestha. Mapping species distributions with social media geo-tagged images: Case studies of bees and flowering plants in australia. *Ecological Informatics*, 39:23–31, 2017.

6 Shelan S. Jeawak, Christopher B. Jones, and Steven Schockaert. Using flickr for characterizing the environment: An exploratory analysis. In *13th International Conference on Spatial Information Theory, COSIT 2017, September 4-8, 2017, L'Aquila, Italy*, pages 21:1–21:13, 2017.

7 Philip Lei, Gustavo Marfia, Giovanni Pau, and Rita Tse. Can we monitor the natural environment analyzing online social network posts? a literature review. *Online Social Networks and Media*, 5:51–60, 2018.

8 Anthony Stefanidis, Andrew Crooks, and Jacek Radzikowski. Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 78(2):319–338, 2013.