# Predicting the environment from social media : a collective classification approach

Shelan S. Jeawak, Christopher B. Jones and Steven Schockaert

*Cardiff University, School of Computer Science and Informatics, Cardiff, UK*

ARTICLE INFO

*Keywords*:
Collective prediction
Text mining
Social media
Ecology
Geographic information systems

ABSTRACT

We propose a method which uses Flickr tags to predict a wide variety of environmental features, such as climate data, land cover categories, species occurrence, and human assessments of scenicness. The role of Flickr tags in our method is two-fold. First, we show that Flickr tags capture information which is highly complementary to what is found in traditional structured environmental datasets. By combining Flickr tags with traditional datasets, we can thus make more accurate predictions than is possible using either Flickr tags or traditional datasets alone. Second, we propose a collective prediction model which crucially relies on Flickr tags to define a neighbourhood structure. The use of a collective prediction formulation is motivated by the fact that most environmental features are strongly spatially autocorrelated. While this suggests that geographic distance should play a key role in determining neighbourhoods, we show that considerable gains can be made by additionally taking Flickr tags and traditional data into consideration.

## 1. Introduction

When users share their photos on websites such as Flickr[1], they often provide tags (i.e. short textual descriptions) to make these photos discoverable. In many cases[2], latitude and longitude coordinates are also provided, describing where the photographs were taken. Since the tags associated with such *georeferenced* photographs often correspond to descriptions of the corresponding locations, Flickr can effectively be seen as a source of environmental information. The usefulness of Flickr tags has already been demonstrated in several disciplines. For example, in geography, Flickr tags have been used to construct approximate boundaries for (vernacular) regions (Cunha and Martins, 2014; Grothe and Schaab, 2009) and for describing properties of places (Bahrehdar and Purves, 2018). In environmental science they have been used for early warning of flooding (Tkachenko et al., 2017). In linguistics, the tags of georeferenced Flickr photos have been found useful for generating vector space representations of perceptual terms (Bolognesi, 2016). In the domain of ecology, Flickr has been used to study species distribution (Barve, 2015; Jeawak et al., 2018, 2020).

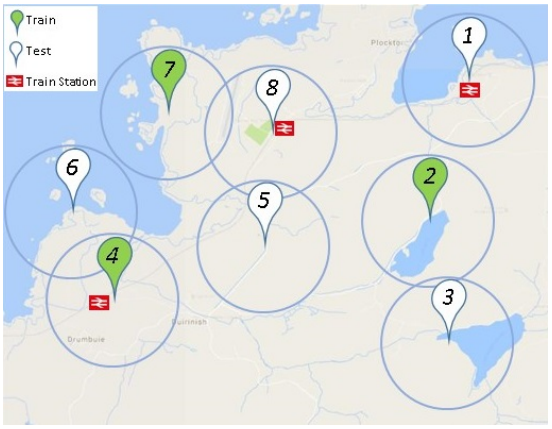
The aim of this paper is to study the usefulness of Flickr tags for predicting environmental features, such as climate features (e.g. average temperature, wind speed, precipitation, solar radiation, water vapor pressure), land cover categories, species occurrence, and scenicness. In seeking to exploit Flickr tags for mapping environmental features, it may be remarked that we work with all tags that have been used in the vicinity of each predicted location. The nature of these tags can be expected to vary widely, with many different phenomena being photographed and tagged with a wide variety of possible terminology. We cannot know whether a user is intending to describe accurately the subject of the photo as opposed for example to making some arbitrary observation or comment about it. In some sense the tags might therefore be regarded as potentially misleading. The methods that we employ here are however designed to identify those tags that are consistently positively correlated with our target environmental features, and hence serve as reliable indicators, while downgrading or ignoring the significance other tags.

Our contribution is two-fold. First, we analyze to what extent Flickr tags can provide information that is complementary to the datasets that are traditionally used for ecological analyses. To this end, we represent each location as the concatenation of a feature vector derived from Flickr tags and a feature vector that encodes available information from traditional structured datasets, and then train a Support Vector Machine (SVM) or Support Vector Regression (SVR) model to predict the features of interest. Note that while the feature vectors include large numbers of tags, the training process enables the classifier to determine which particular tags are specifically associated with the target class, thus distinguishing them from the more generic and hence less indicative tags. Second, we propose a collective prediction model, which takes advantage of the fact that most environmental features are strongly spatially autocorrelated (e.g. climate features typically do not vary much between places that are just a few kilometres apart). Inspired by Angelova and Weikum (2006) and Dori-Hacohen et al. (2016), a key feature of our approach is that the neighbourhood structure of the collective prediction model does not only depend on purely metric geographic distance but also on attribute similarity, which is estimated in our case from the Flickr tags associated with each location. This use of attribute similarity can be thought of as a form of categorical, i.e. attribute-based, geographic distance that takes into account non-spatial properties of location. Our use here of attribute

*Corresponding author
JeawakSS@cardiff.ac.uk (S.S. Jeawak)
ORCID(s): 0000-0002-4753-8997 (S.S. Jeawak)

[1] http://www.flickr.com

[2] We were able to crawl around 350M georeferenced Flickr photos in September 2015.




| | | Flickr tags | | | | | | Structured data | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Location ID | Ground truth label | mountain | sea | loch | station | railway | ..... | Lat | Lon | Elev | Temp | Precip | ..... |
| 1 | 3.4 | 0.178 | 1.154 | 0.101 | 3.060 | 2.626 | ..... | 1.65747 | -1.77379 | -0.8433 | -1.0558 | 2.8741 | ..... |
| 2 | 8.2 | 0.104 | 0.013 | 9.43 | 0 | 0 | ..... | 1.65147 | -1.77825 | -0.2046 | -0.9029 | 2.9580 | ..... |
| 3 | 6.5 | 0 | 0 | 3.258 | 0 | 0 | ..... | 1.64780 | -1.77317 | -0.1837 | -0.9029 | 2.9580 | ..... |
| 4 | 4.6 | 0.562 | 0.173 | 0.044 | 1.325 | 2.556 | ..... | 1.64928 | -1.79716 | -0.6888 | -0.7430 | 2.7344 | ..... |
| 5 | 7.2 | 0.342 | 0.553 | 0.624 | 0 | 0.253 | ..... | 1.65098 | -1.79174 | -0.5734 | -0.7430 | 2.7344 | ..... |
| 6 | 7.8 | 1.450 | 2.335 | 1.018 | 0.083 | 0 | ..... | 1.65282 | -1.80714 | -0.9263 | -0.7430 | 2.7344 | ..... |
| 7 | 8.3 | 2.133 | 4.032 | 0.046 | 0 | 0 | ..... | 1.65671 | -1.7987 | -0.9263 | -1.2286 | 2.9626 | ..... |
| 8 | 2.8 | 0.033 | 0 | 0 | 6.160 | 5.657 | ..... | 1.65626 | -1.78728 | -0.7761 | -1.0558 | 2.8741 | ..... |

**Figure 1:** Modeling locations based on Flickr tags, structured features, and neighbourhood structure.

similarity can be regarded as analogous to the way in which measures of spatial autocorrelation are used in geostatistics for purposes of interpolation (Webster and Oliver, 2007). Geographic attribute similarity complements other alternative measures of geographic distance such as network distance and cognitive distance (Montello, 1991; Sack, 1980). In this way, our model essentially uses Flickr tags to improve how known measurements, as well as predictions, of a given environmental feature are interpolated.

The problem we consider is to predict the value of a given feature (e.g. average temperature or land cover category) for a given set of locations, where we assume that for a subset of these locations (i.e. the training data), the correct value of the considered feature is available (e.g. temperature measurements). The method proceeds in two steps. First, in the bootstrap stage, an SVM model (for discrete features) or SVR model (for numerical features) is learned from the training data. To this end, each location is represented using a feature vector, which encodes how strongly that location is related to each Flickr tag, as well as the available structured information about the location. This is illustrated in the table in Figure 1 (where, in practice, the ground truth data is only available for items from the training data). This model is then used to predict the value of the considered feature for the locations which are not in the training data. In the second step a collective classification approach is adopted in which for each location a set of neighbours is selected, and a new classifier is trained, which aims to improve the predictions by taking into account the earlier predictions in addition to the true labels of the selected neighbours when they are available. This whole process is then iterated until the predictions converge.

The second step crucially relies on how the neighbours are selected. As a baseline, we could choose the neighbours of a given location as those locations which are geographically closest. For example, consider the locations shown on the map in Figure 1 for the task of predicting scenicness. To improve the prediction for location 8, based on geographic distance, we could select location 2, 5 and 7 as neighbours. However, location 1 and 4 are actually more relevant for the purposes of prediction, as they are both more similar to the target location in that, like location 8, they are close to railway train stations, which is an important indicator of low scenicness. To determine these more relevant locations, we first apply a term selection method to identify those Flickr tags that are most strongly related to the considered feature. For example, when predicting scenicness, relevant tags include 'mountain' (which is predictive of high scenicness) and 'station' (which is predictive of low scenicness). Then, from the geographically sufficiently close locations, as neighbours we select those locations whose associated tags (after term selection) are sufficiently similar.

The remainder of the paper is organized as follows. Section 2 gives an overview of related work. Then, in Section 3 we explain how locations are modeled using both Flickr tags and structured data, while Section 4 describes our collective prediction framework. Subsequently, in Section 5 we provide a detailed discussion about our experimental results. Finally, Section 6 summarizes our conclusions.

Some parts of this paper extend our work in Jeawak et al. (2017), where we demonstrated the complementarity of Flickr tags and structured information. However, the collective prediction model from Section 4 is completely new.

## 2. Related work

The related work falls broadly under three themes: citizen science, geo-spatial analysis of social media, and collective prediction.

### 2.1. Citizen science

Considerable progress has been made in recent years in citizen science projects in the environmental sciences that recruit participants to actively contribute to particular campaigns such as in land cover mapping (Fritz et al., 2012), hydrological surveys (Lowry and Fienen, 2013), ornithology and many forms of ecological study (Dickinson et al., 2010). In parallel with these initiatives there is growing interest in the potential of "passive" survey methods that exploit social media to provide additional useful data. For instance,

Wang et al. (2013) analysed the visual features of the photographs on Flickr (in an automated way) to observe natural world features such as snow cover and particular species of flowers. In (Zhang et al., 2012) photos from Flickr were used to estimate snow cover and vegetation cover, and to compare these estimations with fine-grained ground truth collected by earth-observing satellites and ground stations. Both the text associated with Flickr photographs and their visual features were used in Leung and Newsam (2012) to perform land-use classification. The approach was evaluated on two university campuses and three land-use classes were considered: Academic, Residential, and Sports. In Estima et al. (2014) and Estima and Painho (2014), they classified a sample of georeferenced Flickr photos according to CORINE land cover classes. They also evaluated the use of Flickr photos in supporting Land Use/Land Cover (LULC) classification for the city of Coimbra in Portugal and for comparison with Corine Land Cover (CLC) level 1 and level 2 classes. Note that their approach did not use machine learning and the results were evaluated manually by experts. Their results suggest that Flickr photos cannot be used as a single source to achieve this purpose but they could be helpful if combined with other sources of data.

The authors of (Stadler et al., 2011) explored the relationship between CORINE land cover classes and the valuation of natural scenery, namely scenicness, scenic beauty, landscape beauty, aesthetics, or cultural ecosystem services (CES), through user evaluated georeferenced photos from the ScenicOrNot[3] website. They employed the user's rating of a photo in a specific area as an evaluation of the land cover of that area. The results of this study showed that the highest rated areas belong to the forest and semi natural areas, and water bodies classes. In another work, Chesnokova et al. (2017) developed and evaluated a predictive model to predict the average scenicness of 5km grid cells. They used text describing the rated images in the ScenicOrNot website as input to train a Random Forest regression model. Measures of scenicness are important since they reflect human well-being and can be taken into consideration in land planning and decision-making processes. Nonetheless, people's perceptions of landscapes are subjective and cannot easily be quantified (Stadler et al., 2011). Some authors have assessed the beauty of the landscape through groups of evaluators using images, videos and/or questionnaires (Stadler et al., 2011; Pierskalla et al., 2016), while others used geographic information system (GIS) data such as elevation together with visual assessments and/or questionnaires to predict the scenicness (Bishop and Hulse, 1994; Schirpke et al., 2013). Another group of works, such as Casalegno et al. (2013), Gliozzo et al. (2016), and Tenerelli et al. (2016), quantify landscape aesthetics according to the number of photos taken near a given location (Casalegno et al., 2013) or the number of people who published photos (Gliozzo et al., 2016) in photo-sharing websites such as Flickr and Panoramio. Considering popularity on social media as a surrogate for the level of appreciation of a place might work with some types of landscapes, but the results might be liable to be biased towards more accessible places (one of our experiments reported in Section 5.1 provides evidence to that effect).

Another growing area of interest is in the use of social media data for ecological monitoring. For example, Barve (2015) examined Flickr biodiversity data quality by analysing its metadata and comparing it with ground-truth data, using Snowy owls and Monarch butterflies as a case study. They concluded that Flickr data has potential to add to knowledge of these species in terms of geographic, taxonomic, and temporal dimensions, which tends to be complementary to the information contained in other available sources. In another similar work, based on a manual analysis of Twitter posts, Daume (2016) confirm that social media mining for ecological analysis is as important as traditional monitoring and the features derived from Twitter could be integrated with and hence improve the value of existing sources of such information. In Richards and Friess (2015) the content of the Flickr photos was analysed manually to assess the quality of cultural ecosystem services and derive useful information to manage Singapore's mangroves.

## 2.2. Geo-spatial analysis of social media

Many recent studies have focused on analysing social media data, with the aim of extracting useful information in domains such as geography (e.g. Hollenstein and Purves (2010)). In particular, there is a large number of studies that derive such information from georeferenced Flickr photos. For example, Grothe and Schaab (2009) describe two methods for the automatic delineation of imprecise regions based on geotagged photos. The first one is a method based on kernel density estimation (KDE) and the second is based on one class support vector machines (SVMs). Similarly, Cunha and Martins (2014) present an approach for automatically defining the geographic boundaries of vague regions by using one class support vector machines (SVMs) and learning multiple kernels. To describe regions, they rely on a combination of the Flickr tags of the photos that were tagged with the region's name, and external features such as the land cover data, population count, elevation and the geographical coordinates (latitude and longitude) of Flickr photos that are tagged with the region's name. They showed that their method performs better than the simpler methods described by Grothe and Schaab (2009). The first step of our method is analogous to these approaches, in applying support vector machine learning methods to Flickr tags in combination with other geo-spatial data, but we are concerned with characterizing and predicting information about the environment.

The authors of (Serdyukov et al., 2009) presented and evaluated methods for automatically geo-referencing Flickr photos using the textual annotations of photos to predict the single most probable location where the image was taken. They showed that location-specific language models, based on sets of distinctive tags, can be estimated effectively by analysing the terms people use to describe images taken at particular locations. They demonstrated how to incorporate the GeoNames database and they defined extensions to

[3] http://scenic.mysociety.org/

improve their language models using cell based smoothing and tag based smoothing, and by leveraging spatial ambiguity. In Van Canneyt et al. (2014), a language modelling approach was used to discover and characterize places of interest (POIs). They experimented with both Flickr data and Twitter data, finding that Flickr data on its own is more useful than Twitter data for this task, while combining both sources led to the best results. Similar to this latter work, we explore the possibility that sets of tags cannot just distinguish one location from another, but can contribute to classifying aspects of the environment.

Our aim of predicting environmental features from various attributes might be regarded as analogous to the use of geostatistical methods such as Co-Kriging and Geographically Weighted Regression, but in the application of these methods the numbers of predictor, explanatory or other independent variables is very small compared to our task, often being less than 10 (e.g. Ristea et al. (2018); Liu et al. (2016)). This contrasts with our data in which, following weighting and filtering there may be hundreds of thousands of social media tags each of which is treated as a potentially useful predictor in addition to conventional environmental data attributes. It should also be noted that the benefits for environmental data interpolation of machine learning classifiers such as support vector machines and random forests in combination with an interpolation method such as inverse distance squared, or ordinary kriging, have been demonstrated relative to a variety of conventional interpolation methods (Li et al., 2011). The latter study used structured environmental data as the predictors.

## 2.3. Collective prediction

Many machine learning problems involve making predictions about networks of entities, where links in the network connect entities that are related in some way. The idea of collective prediction is to incorporate additional features in the learning process, summarizing information about the entities that are related to the considered one. A standard example is the problem of web page categorization (Chakrabarti et al., 1998; Angelova and Weikum, 2006; Dori-Hacohen et al., 2016): to determine the category of a website, in addition to the contents of the website itself, we can also take into account the categories of the websites it links to. Note that this creates a cyclic dependency between the predictions for the different entities in the network. To address this, a variety of collective prediction methods have been proposed. In this paper, we will use the Iterative Classification Algorithm (ICA) from Neville and Jensen (2000), which is conceptually simple but often highly effective. Other approaches are based on inference in joint probabilistic models using Gibbs sampling (Geman and Geman, 1984). However, Gibbs sampling tends to be slow (Sen et al., 2008), which is an important limitation in our setting, as we will consider hundreds of thousands of regions.

The authors of (Chakrabarti et al., 1998) experimentally demonstrated the effectiveness of taking into account link structure for web page categorization. More recent methods often take into account content similarity to improve the network structure, i.e. better results can often be obtained by only taking into account links from websites that are sufficiently similar. For example, Angelova and Weikum (2006) select a reliable set of neighbours for each test document by means of a similarity threshold. They only consider the links for which the similarity between the contents of the two documents (nodes) is sufficiently high. In Dori-Hacohen et al. (2016), a method is proposed which classifies Wikipedia pages as controversial or not, using a combination of intrinsic features (page meta-data) and predictions of controversy from related pages. They constructed a subnetwork by choosing for each page the $k$ most similar in-links (in terms of cosine similarity between the text of the pages) and the $k$ most similar out-links, where $k$ was chosen as either 10 or 300. They then use a stacked model on top of this constructed network. The stacked approach introduced in (Kou and Cohen, 2007) uses a non-relational base model to produce inferred class labels on related instances where the stacked relational model is trained on these predicted labels rather than the true labels. In Jiang et al. (2017), a collective prediction algorithm based on community structure (CPC) was proposed. Firstly, they obtained the community that each node belongs to by using a community detection algorithm. Then they used the node attribute features and community structure features as inputs to the local classification model in an iterative way. Their experimental results show that CPC performs better than both a standard prediction method which only utilizes the node attributes and an iterative classification algorithm which utilizes neighbour features in addition to the node attributes.

Although many studies have been conducted in collective classification, less effort has been focused on collective regression. Chopra (2008) proposed a relational factor graph framework for performing regression on relational data. The proposed models are learned with collective inferences which take a single instance of the entire collection of samples along with their relationship structure as input. The framework was applied to the problem of predicting house prices, taking into account spatiotemporal influences on the price of every house. Their experiments demonstrate that identifying and using the relational structure associated with this problem considerably improves performance. The authors of (Loglisci et al., 2016) presented an algorithm called CORENA (COllective REgression in Network dAta) which studies the transduction of collective regression in a sparsely labeled network. In particular, they iteratively augmented the descriptive and the target information of the labeled node set, the descriptive information of the unlabeled node set, as well as the link structure of the network, in order to collectively determine the numerical targets of the unlabeled part of the network. Thus, their proposed method can detect the autocorrelations of labels over a group of related instances and feed back the reliably predicted labels only. They (Loglisci et al., 2016) show that their proposed method is able to improve regression performance in the areas of social and spatial networks.

In this paper, we focus on both collective classification and regression problems by developing SVM/SVR models in an iterative way. We consider several nested sets of neighbours for each location based on their spatial and attribute similarity. Then, we aggregate the true and the predicted labels of these selected neighbours to generate the collective features.

## 3. Modelling locations

In this section we explain how locations are represented in our framework. Section 3.1 explains how feature vectors describing locations can be obtained from the tags associated with georeferenced Flickr photos. In Section 3.2 we then give an overview of the structured information sources that we will additionally consider.

### 3.1. Modelling locations using Flickr tags

Many of the tags associated with Flickr photos tell us something about the locations where these photos were taken. For example, tags might correspond to city and region toponyms (e.g. United Kingdom, England, London), landmarks (e.g. London Eye, Westminster Abbey, Hyde Park) or land cover types (e.g. forest, beach, airport). Using the Flickr API, we collected the metadata of all geo-referenced Flickr photos that were uploaded before the end of September 2015, leading to a total of over 70 million photos with coordinates in Europe (which is the region our experiments will focus on).

Let $L = \{l_1, ..., l_m\}$ be a set of locations, each characterized by latitude and longitude coordinates. Our aim is to associate with each of these locations a weighted bag of tags, intuitively encoding for each tag how strongly it is associated with photos near that location. To this end, we first use a BallTree[4] to retrieve the set $F_l$ of all Flickr photos whose distance to the considered location $l$ is at most $D$. Let us write $U_{t,c}$ for the set of users who have assigned tag $t$ to a photo with coordinates $c$. Then we define $n(t,l) = \sum_{d(c,l)\leq D} |U_{t,c}|$, with $d$ the Haversine distance. Intuitively, $n(t,l)$ is the number of times tag $t$ appears among the photos in $F_l$. However, to reduce the impact of bulk uploading, we count a tag occurrence only once for all photos by the same user at the same location.

One problem with using $n(t,l)$ to measure the importance of tag $t$ for location $l$ is that it gives equal weight to all photos, whereas intuitively we want photos which are closer to $l$ to influence our characterization of $l$ more than photos which are further away. To this end, following Van Canneyt et al. (2014), we use a Gaussian kernel to weight the tag occurrences:

$$w(t,l) = \sum_{d(c,l)\leq D} |U_{t,c}| \cdot \exp\Big(-\frac{d^2(l,c)}{2\sigma^2}\Big)$$

where $\sigma$ is a bandwidth parameter.

[4] http://scikit-learn.org/stable/modules/generated/sklearn.neighbours.BallTree.htm

The weight $w(t,l)$ still has the problem that common words (e.g. *iphone*) are given the same importance as more specific words. Intuitively, we want the weight of tag $t$ to reflect how strongly it is associated with location $l$. A standard way of measuring this in bag-of-words models is to use Positive Pointwise Mutual Information (PPMI), which is based on comparing the actual number of occurrences with the expected number of occurrences (given how many tags occur overall near $l$ and how common the tag $t$ is). Specifically, the weight of tag $t$ in our bag-of-words representation of $l$ is then given by:

$$PPMI(t,l) = \max\left(0, \log\left(\frac{P(t,l)}{P(t)P(l)}\right)\right)$$

where:

$$P(t,l) = \frac{w(t,l)}{N} \qquad P(t) = \frac{\sum_{l'\in L} w(t,l')}{N}$$
$$P(l) = \frac{\sum_{t'\in T} w(t',l)}{N} \qquad N = \sum_{t'\in T}\sum_{l'\in L} w(t',l')$$

with $T$ the set of all tags that appear in the collection. Finally, each location $l$ is represented as a sparse vector, encoding the weights $PPMI(t,l)$ for all the tags in $T$.

### 3.2. Modelling locations using structured data

There is a wide variety of structured data that can be used to describe places. The most obvious type of structured data are the coordinates of the photo itself. Clearly, latitude and longitude degrees can be helpful for predicting a range of environmental phenomena (e.g. southern areas of Europe tend to be warmer than northern areas). In addition to geographic coordinates, we will consider the following sources of scientific data:

- CORINE Land Cover 2006[5] is a European dataset which describes land cover with a 100 meter spatial resolution. CORINE uses three levels of description: a top level with 5 classes, an intermediate level with 15 classes and a detailed level with 44 classes.
- SoilGrids[6] is a global raster dataset, which classifies locations into 116 types of soil, using a 250 meter spatial resolution.
- The Digital Elevation Model over Europe (EU-DEM)[7] is a Europe-wide digital surface model, encoding elevation with a spatial resolution of about 30 meter.
- European Population Map 2006[8] is a digital raster grid that reports the number of residents (night-time population) with a 100 meter spatial resolution.

[5] http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-2
[6] https://www.soilgrids.org
[7] http://www.eea.europa.eu/data-and-maps/data/eu-dem
[8] http://data.europa.eu/89h/jrc-luisa-europopmap06

- WorldClim[9] is a global raster dataset, containing average monthly recordings of the following climate features, over the period 1970-2000, using a 1 km spatial resolution: temperature, precipitation, solar radiation, wind speed and water vapor pressure. In this work, we convert the monthly averages reported in the dataset to a single overall average.

To encode locations, we consider a feature vector that contains one binary feature for each CORINE land cover class (being 1 if the location belongs to that class and 0 otherwise), one binary feature for each SoilGrids class, and 9 real-valued features (encoding latitude, longitude, elevation, population, and the 5 climate features). The real-valued features have been normalised using the standard z-score.

In most of our experiments, we will use Flickr tags in combination with structured information. In such cases, we simply concatenate the PPMI-based feature vector from Section 3.1 with the feature vector modeling the structured information.

## 4. Collective prediction model

Many real world problems can be described as graphs, where the nodes correspond to objects about which we want to predict something, and edges denote relationships between these objects. In collective prediction frameworks, the class label (in classification problems) or feature value (in regression problems) of a given object can be used to improve the predictions about related objects.

In particular, the goal of collective prediction is to jointly determine the labels of all nodes in the graph, taking into account their interrelationships. To apply the collective prediction framework to our setting, we consider each of the locations $l_i \in L$ as a node. Two nodes are connected by an edge if they represent sufficiently similar locations. The underlying notion of similarity will be partially based on geographic closeness, but will also take the Flick tags and structured data that are associated with these locations into account. We assume that a partition $L = T1 \cup T2 \cup T3 \cup T4$ of the locations is given, where $T1 \cup T2 \cup T3$ will be used as training data and $T4$ will be used as testing data. The locations in $T1$ will be used for training a bootstrap classifier, while those in $T2$ will be used for learning how to improve predictions based on related locations. The locations in $T3$, finally will be used for tuning the classifiers.

The overall method involves the following steps, which are illustrated in Figure 2.

**Bootstrap:** In this step, we use the training data in $T1 \cup T2$ to learn an SVM or SVR model, using the feature vector representation for each location as explained in Section 3. When applying our overall model, this classifier ($P1$ in Figure 2) will be used to make an initial prediction for the unlabeled locations (i.e. for the locations from $T4$). This prediction will later be used to generate the collective features. We also learn a second classifier ($P2$ in Figure 2), which is trained in the same way as $P1$ but only using the locations from $T1$ as training data. This variant is needed to allow us to train an iterative collective classifier, which will intuitively be learned by comparing the true labels of $T2$ with the predictions that are made by classifier $P2$.

**Identifying distinctive Tags:** A key property of our method is that it uses Flickr tags to find relevant neighbours, i.e. to find nearby locations that are sufficiently related to the considered target location. Clearly, the required notion of relatedness depends on what we are trying to predict. For example, when predicting scenicness as in the locations shown on the map in Figure 1, we may want to select location 1 and 4 as the most relevant neighbours to location 8 because all three of them are close to train stations. To estimate relatedness, we therefore first determine which tags are most relevant for the considered prediction problem, using a term selection method based on Kullback-Leibler (KL) divergence. Let us first consider a classification problem with classes $C_1, ..., C_n$. Given that we are interested in predicting properties of locations, each class $C_i$ here corresponds to a subset of locations from $L$ that share a particular property (such as, for example, having a type of land cover).

In particular, we select the 1000 tags that score highest on the following score:

$$KL(t) = \sum_{i=1}^{n} P(C_i|t) \log \frac{P(C_i|t)}{Q(C_i)} \quad (1)$$

where $P(C_i|t)$ is the probability that the location a photo with tag $t$ belongs to $C_i$, whereas $Q(C_i)$ is the probability that an arbitrary tag occurrence is assigned to a photo with location $l$. We estimate $Q(C_i)$ as follows:

$$Q(C_i) = \frac{1}{N} \sum_{l \in C_i} \sum_{t' \in T} w(t', l)$$

$$N = \sum_{j=1}^{n} \sum_{l \in C_j} \sum_{t' \in T} w(t', l)$$

Since $P(C_i|t)$ often has to be estimated from a small number of tag occurrences, it is estimated using Bayesian smoothing:

$$P(C_i|t) = \frac{\left(\sum_{l \in C_i} w(t, l)\right) + \beta \cdot Q(C_i)}{N + \beta}$$

where $\beta$ is a parameter controlling the amount of smoothing, which will be tuned in the experiments. Intuitively, we can think of $\beta$ as a number of samples from the background distribution $Q$ that are added to our data about tag $t$. Larger values of $\beta$ will have a penalizing effect on rare terms.

For regression problems, we discretize the feature values and then proceed in the same way. In particular, we discretize the feature values into three classes $C_1$, $C_2$ and $C_3$ based on feature dependent thresholds. For example, to identify a set of tags that are related to scenicness, we classify tags into $C_1$ if they occur in locations whose scenicness rate is at least 7 , $C_2$ for the tags that occur in locations whose scenicness rate is between 3 and 7, and $C_3$ for the tags that

[9] http://worldclim.org

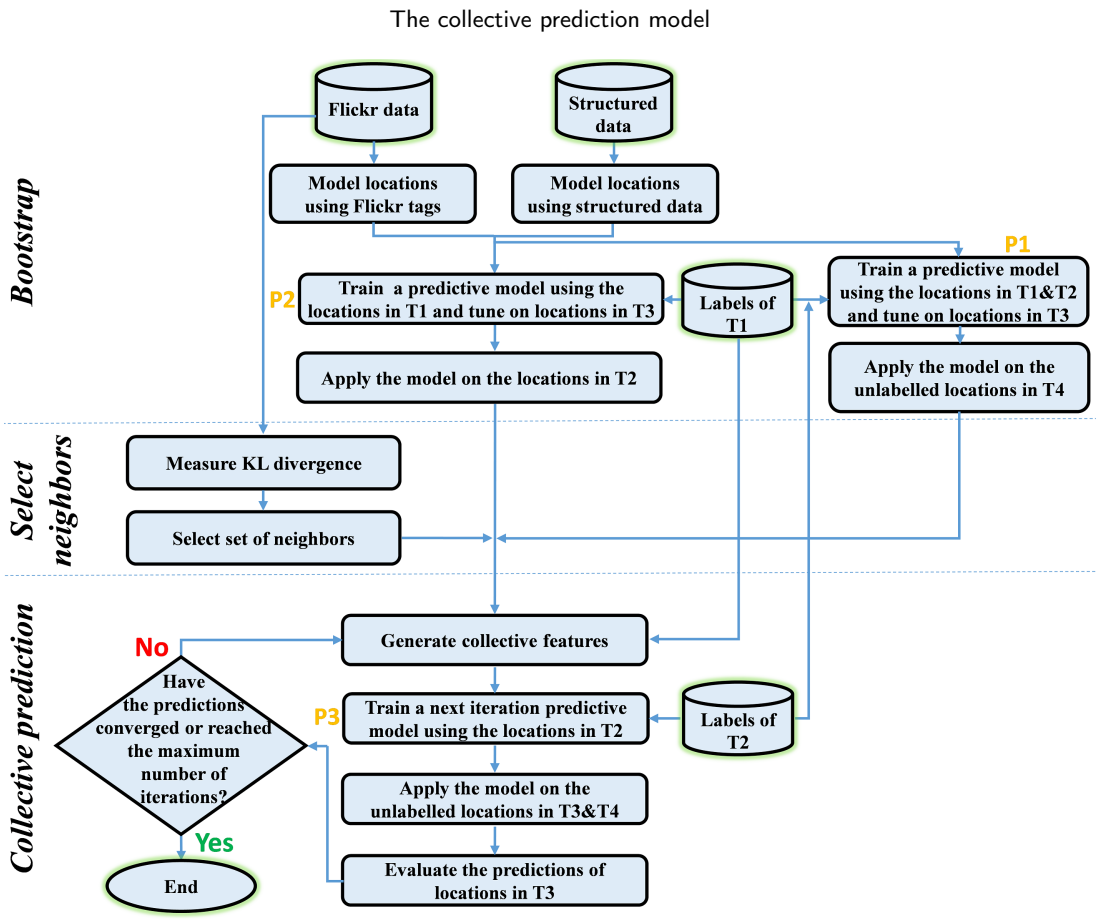



**Figure 2:** The collective prediction model.

occur in locations whose scenicness rate is at most 3. Then, because the most informative tags are likely to be found in the extreme cases, we only consider tags that are distinctive for classes $C_1$ and $C_3$ which are determinant as for classification problems.

**Selecting neighbours:** The effectiveness of collective prediction relies on the assumption that neighbouring nodes have similar labels. Since environmental features tend to be spatially autocorrelated, in our setting it is natural to choose nearby locations as neighbours. However, while only taking into account geographic closeness already leads to a strong baseline, as we will see in the experiments, further improvements are possible by additionally taking into account the structured environmental data and Flickr tags. The underlying motivation is that such tags can reveal whether nearby locations are actually similar. Consider, for example, a train station which is located very close to a beach. Despite their close locations, these places belong to different land cover classes, and may have a considerably different scenicness degree. Specifically, to select the neighbours of a given location $l$, we first determine the set of nearby locations (i.e. those whose location is within a given radius $r$) and then pick the $k$ most similar ones among these nearby locations. For this last step, locations are represented as PPMI-weighted feature vectors from Flickr data, as in Section 3.1 but only considering the 1000 tags that were selected based on (1), concatenated with the structured feature vectors from Section 3.2. These feature vectors are then compared using the cosine similarity.

**Iterative inference:** In order to improve the predictions for a given target location, we train a classifier whose input is derived from the earlier predictions of that location and its neighbours (see below). Note that all locations from $L$ are considered as possible neighbours, including the locations from the training data $T1$ and the tuning data $T3$. For neighbours that come from $T1$ and $T3$, we use the corresponding ground truth instead of a predicted value. In this sense, we could intuitively think of our proposed method as a refinement of the $K$-nearest neighbours method. Note that while we are using the actual ground truth for neighbours from $T1$, we cannot do the same for neighbours from $T2$ during the training phase, since that would lead the iterative SVM/SVR model ($P3$ in Figure 2) to simply pick $p_L$ as the only relevant feature, given that this value would correspond to the ground truth for all training items.

In standard collective prediction only a single set of neighbours is considered, but in this paper we instead consider several nested sets of neighbours for each target location. To determine the neighbours of a target location, we have to choose a radius $r$ and the desired number of neighbours

$k$. Rather than fixing a single value for these parameters, we consider a sequence of radii $r_1, ..., r_n$ and a corresponding sequence of numbers $k_1, ..., k_n$. Let $N_i$ be the $k_i$ most similar locations within the radius $r_i$ (i.e the set of neighbours corresponding to the choice $(r_i, k_i)$). With each set $N_i$ we associate a corresponding prediction $x_i$, which is the average prediction for the locations in $N_i$ in the case of regression problems, and the average of the confidence scores associated with each class in the case of classification problems. We can give higher weight for those neighbours that have ground truth (i.e. locations from $T1$ and $T3$) when computing $x_i$. Let $ground(l)$ be the ground truth value of location $l$, $N_i^G$ be the set of neighbouring locations for which the ground truth is known, while $pred(l)$ be the prediction value or confidence score of the unlabeled neighbouring location $l$. We estimate $x_i$ as follows:

$$x_i = \frac{\sum_{l \in N_i^G} \lambda \cdot ground(l) + \sum_{l \in N_i \setminus N_i^G} pred(l)}{\lambda \cdot |N_i^G| + |N_i \setminus N_i^G|}$$

where the weight $\lambda$ is used to control how much we want to boost the evidence coming from neighbours with a known ground truth.

For this iterative classification step ($P3$ in Figure 2), the location $l$ is represented as the $n$-dimensional feature vector $(p_l, x_1, ..., x_n)$, where $p_l$ is the earlier prediction for the location $l$ itself. From these feature vectors, we learn an SVM or SVR model, using the locations from $T2$ as training data, to find an improved prediction for the unlabeled locations (i.e. for the locations from $T4$). This step is then repeated, using the new predictions as input, until convergence or reach the maximum number of iterations. We evaluate the convergence here according to the locations in $T3$ set. This is illustrated in Figure 2, which provides an overview of the whole process.

## 5. Experiments

In the following experiments, we evaluate how well we can predict a number of environmental features using Flickr tags and the considered structured environmental data. For the variables in Section 3.1, we have set the maximum Haversine distance D (cluster radius) to 1 kilometre and the bandwidth $\sigma$ to D/3. The choice of $D$ represents a trade-off, where larger values can potentially lead to better results but also lead to a higher computational cost. The choice of $\sigma = D/3$ was found to be reasonable in a small set of initial experiments. For the variables in Section 4, we examined various smoothing values to select the distinctive tags in KL divergence ($\beta = 10, 100, 1000$) and chose the best value for each experiment separately based on held-out tuning data ($T3$). The feature dependent thresholds used to discretize the regression problem data into $C_1$ and $C_3$ classes when computing KL divergence are listed in Table 1, these values having been chosen as reasonable values from initial experiments. To generate the collective feature vector, we combine the earlier prediction $p_l$ with seven collective features where $r_1$-$r_7$ are chosen as $1, 2, 5, 10, 20, 50$ and $100$ kilometres for each location. We test with different numbers of similar neighbours, choosing $k_i$ as $r_i$+1, $r_i$+10 or $r_i$+100, again based on the held-out tuning data ($T3$). Figure 3 shows examples of the collective feature vectors of different locations with their ground truth labels. We set the ground truth labels weight $\lambda$ to 5. Finally, we set the maximum number of iterations to 10.

**Table 1**
High and low boundaries for discretizing the regression problems data into classes.

| | $C_1$ | $C_3$ |
|---|---|---|
| Scenicness | $\geq 7$ | $\leq 3$ |
| Temperature ($^\circ C$) | $\geq 15$ | $\leq 5$ |
| Precipitation (mm) | $\geq 100$ | $\leq 50$ |
| Solar Rad (kJ $m^{-2}day^{-1}$) | $\geq 17000$ | $\leq 10000$ |
| Wind Speed (m $s^{-1}$) | $\geq 5$ | $\leq 3$ |
| Water Vapor Press (kPa) | $\geq 1$ | $\leq 0.7$ |

**Collective features vector**

| Ground truth label | Earlier Prediction | Average values Within 1 km | Average values Within 2 km | Average values Within 5 km | Average values Within 10 km | Average values Within 20 km | Average values Within 50 km | Average values Within 100 km |
|---|---|---|---|---|---|---|---|---|
| 7.4 | 6.1 | 8.306 | 7.924 | 7.385 | 7.085 | 7.212 | 6.844 | 6.279 |
| 2.6 | 4.8 | 3.255 | 3.664 | 4.258 | 4.226 | 3.449 | 3.789 | 4.326 |
| 9.25 | 7.66 | 8.641 | 9.039 | 8.281 | 7.420 | 8.224 | 8.426 | 7.047 |

**Figure 3:** Modeling locations based on collective features.

To make predictions, we use Support Vector Machines (SVMs) for classification problems and Support Vector Regression (SVR) for regression problems. In both cases, we used the SVM$^{light}$ implementation[10] Joachims (1998). For each experiment, the set of locations $L$ was randomly split into training ($T1$ and $T2$), tuning ($T3$), and testing ($T4$) sets because the effectiveness of collective prediction may depend quite drastically on the amount of training/testing data that is available. In particular, we have considered three different training/test splits: 5/85, 20/70 and 80/10 while the remaining 10% of the data have used for tuning. Each training set has been split into two equal size subset $T1$ and $T2$. We compared the results for seven different variants and baseline methods:

- "Structured" uses the feature vector modeling the structured scientific information from Section 3.2 only to train SVM/SVR model using locations in $T1$ and $T2$, and predict label or feature value for locations in $T4$.
- "Flickr" uses the PPMI-based feature vector modeling Flickr tags from Section 3.1 only to train SVM/SVR model using locations in $T1$ and $T2$, and predict label or feature value for locations in $T4$.
- "Structured + Flickr" uses the combination of both Structured data and Flickr data by concatenating the

[10] http://www.cs.cornell.edu/people/tj/svm_light/

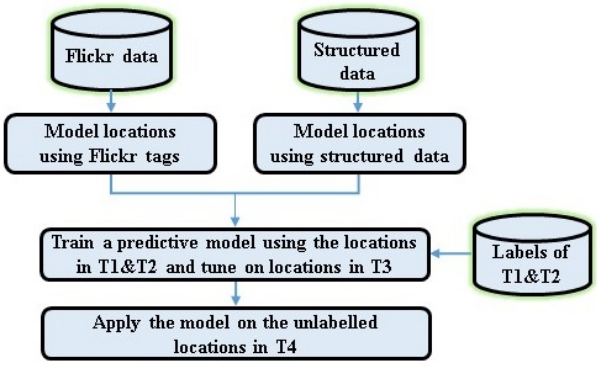



**Figure 4:** Structured+Flickr prediction model.

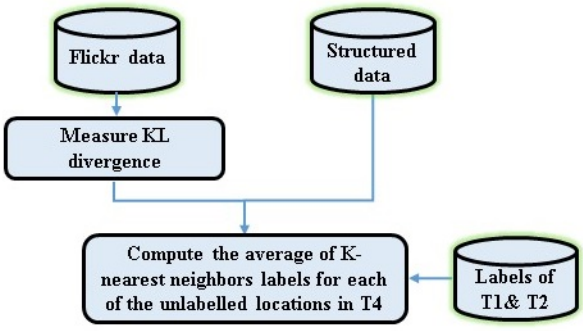



**Figure 5:** K nearest neighbours prediction model.

two corresponding feature vectors. This process is illustrated in Figure 4.

- "KNN-All" computes the average result (i.e. prediction values for regression problems and confidence scores for classification problems) over the K geographically nearest neighbours where these neighbours are selected according to the latitude and longitude coordinates only. We consider the neighbours from the training data $T1$ and $T2$ sets and tune the value of K using the tuning data $T3$.
- "KNN-K" computes the average result of the K most similar neighbours. Similarity is defined here as for our collective prediction method, i.e. based on a feature vector that contains the PPMI values of the 1000 top selected Flickr tags together with the structured data. Again, we consider the neighbours from the training data $T1$ and $T2$ sets and tune the value of K using the tuning data $T3$. This process is illustrated in Figure 5.
- "Collective-All" uses the collective features derived from all neighbours. It is very similar to the method described in Figure 2 except that the neighbours are selected according to their geographical distance (latitude and longitude coordinates) only.
- "Collective-K" is our proposed method, as described in Section 4.

### 5.1. Predicting the scenicness of a place

In this first experiment, we consider the problem of predicting people's opinions of landscape beauty, using the UGC dataset from the ScenicOrNot website[11] as ground truth. This website allows people to evaluate places in Britain by rating photos collected from Geograph[12]. The dataset ratings for 217,000 photos (at distinct locations), each of which has been rated by at least three people on a scale from 1 (not scenic) to 10 (very scenic). For 25,395 of the photos in this dataset, our Flickr collection did not contain any georeferenced photos within a 1 km radius. Therefore, we only report results for the remaining 191,605 photos (i.e. 88.3% of the full dataset). The number of Flickr photos within a 1km radius of these locations varies between 1 and 397982.

For this experiment, $L$ thus contains the locations of these 191,605 photos. To compute KL divergence, we discretize the locations that have been rated between 7 and 10 as high scenicness class while those that have been rated between 1 and 3 as low scenicness class. The results in Figure 6a show the mean absolute error between the predicted and actual scenicness scores, as well as the Spearman $\rho$ correlation between the rankings induced by both sets of scores for the seven considered methods. The mean and standard deviation of the data is shown in Table 2. Although the differences between the results are small, we find that using Flickr outperforms using structured data, and that combining both leads to better results than using them separately. We also find that all these setups (Structured, Flickr, and Structured+Flickr) perform better than the K Nearest neighbours (KNN) method even when we select the most related neighbours in KNN-K. The collective prediction method leads to the best results overall especially when selecting the K most similar neighbours (*Collective-K*). Looking at the top tags, in terms of KL divergence, we find terms relating to natural landscape which represent high scenicness such as *highlands*, *mountains*, and *beach* and names of artificial and urban phenomena which are representative of low scenicness such as *station*, *bus*, and *supermarket*. This reinforces the finding from Stadler et al. (2011) that land cover categories are strongly correlated with scenicness scores. We also tested whether the number of photos (or users) could be used to predict scenicness, as was suggested in Casalegno et al. (2013); Tenerelli et al. (2016); Gliozzo et al. (2016) for particular restricted settings. However, we actually found a negative correlation of around -0.12 (resp. -0.1) between scenicness and the number of photos (resp. users who have posted photos) near a given location.

[11] http://scenic.mysociety.org/
[12] http://www.geograph.org.uk/

**Table 2**
Mean and Standard deviation of regression problems data.

| | Mean | STDEV |
|---|---|---|
| Scenicness | 4.372 | 1.6 |
| Temperature (°C) | 9.268 | 3.490 |
| Precipitation (mm) | 66.625 | 24.827 |
| Solar Rad (kJ $m^{-2}day^{-1}$) | 11478 | 2388 |
| Wind Speed (m $s^{-1}$) | 3.605 | 1.126 |
| Water Vapor Press (kPa) | 0.958 | 0.186 |

We also analyzed whether the performance could be improved by using a different type of model to make the initial prediction. In particular, we first used the method from Jeawak et al. (2019), which learns a low-dimensional vector space embedding that jointly captures the information from the Flickr tags and the structured scientific data. The result of applying this method is that each location is represented as a dense low-dimensional vector, which has the advantage that a much wider range of regression models can be used. In Table 3, we compare our standard model (Collective-K) with the performance we get with these dense vectors, for two different cases. First, we again use Support Vector Regression (SVR). Second, we use a Multi-layer Perceptron (MLP). Note that an MLP model cannot be used directly on the bag-of-words representations, due to the excessive memory requirements this would entail. The results are presented in Table 3. They show that changes in the bootstrap classifier only have a minimal impact on the overall performance, with our standard approach performing slightly better in the 80% setting, and slightly worse in the other settings. Given the small differences, for the remaining experiments we will only focus on the bag-of-words based representation.

## 5.2. Predicting species distribution

The next experiment we considered was to predict the distribution of species across Europe, using as ground truth the dataset of the European network of nature protected sites Natura 2000[13]. This dataset contains information about 35,600 species from 7 classes: Amphibians, Birds, Fish, Invertebrates, Mammals, Plants and Reptilia. In particular, it specifies which species occur at 26,425 different sites across Europe. For this experiment, $L$ is defined as the set of these sites.

For species that only occur at a few of the sites in $L$, it is clearly not possible to estimate a reliable distribution model. Therefore, we focused our evaluation on 100 species which occur at more than 500 sites. For each of these species, we consider a binary classification problem, i.e. predicting at which of the sites the species occurs. Note that as in all analyses we use all Flickr tags, some of which might include the species name. The results of predicting species distribution are reported in Figure 7a in terms of the average precision, average recall and macro average F1 score over the 100 species. Note that we do not consider accuracy as it is not informative here, given the high class imbalance (i.e. a baseline classifier predicting that a species occurs nowhere would already have a very high accuracy). The results are clearly showing that combining structured data with Flickr data leads to substantially better results than both variants of structured data alone, Flickr data alone and K Nearest neighbours. However, the collective predictions (*Collective-K*) lead to the best results overall especially in term of F1 score. Note that we used the same set of structured and Flickr features in *KNN-K* and *Collective-K*. We compute KL divergence for each species separately to identify the most relevant Flickr tags. In this case, to use the KL-divergence feature selection method, we treat the locations where the species is present as one class and all the other locations where the species is not present at the second class. Table 4 contains examples of the top tags of some species as selected by the KL-divergence feature selection method. Interestingly, most of these tags are place names and land cover categories and this applies to many of the 100 species.

[13] http://ec.europa.eu/environment/nature/natura2000/index_en.htm

## 5.3. Predicting CORINE land cover classes

In this section we consider the task of predicting CORINE land cover classes. For this experiment, we have used the same set $L$ of locations as for species distribution. Since the task is about predicting CORINE land cover classes, for the results reported in this section we do not consider any CORINE features in the representations of the locations as the CORINE data serve as ground truth. We experimented with predicting CORINE land cover classification at level 1 (5 sub classes), level 2 (15 sub classes) and level 3 (44 sub classes), each time treating the task as a binary classification problem. The results of predicting CORINE land cover classification at levels 1, 2 and 3 are presented in Figure 7b, Figure 7c, and Figure 7d respectively in terms of the average precision, average recall and macro average F1 score. Again, the results show that combining structured data and Flickr data clearly leads to better performance than using them separately, and that the collective prediction method (*Collective-K*) leads to the best results overall. We compute KL divergence for each land cover classes separately where we treat the locations belonging to the target land cover type as one class and all the other locations as the second class.

To illustrate how Flickr tags are used to select the neighbours of CORINE land cover classes, Table 5 shows examples of the top 5 tags of some CORINE level 1 classes which are clearly informative and semantically related to those classes. For some classes, especially for CORINE level 3, we found that the collective prediction converged already after the first iteration. This seems related to the small number of locations belonging to these classes where it is not possible to find the optimal neighbours if only few locations belong to that class.

## 5.4. Predicting climate data

In the last experiment, we assess the usefulness of Flickr tags in the task of predicting climate data. We again use the same set of sites $L$ as in the species distribution experiment. In this case, we omit all the climate related features from

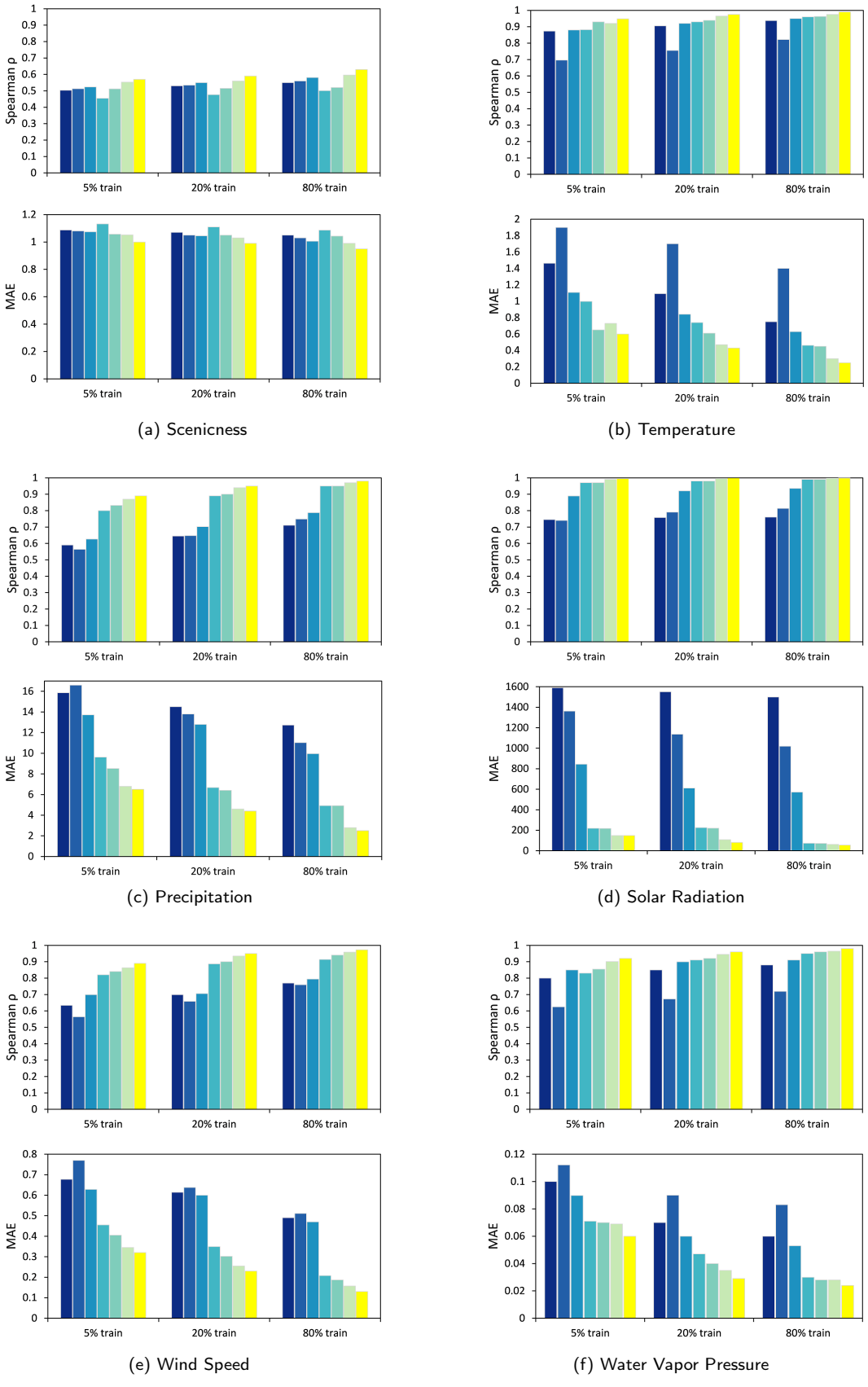



**Figure 6:** Results of regression problems.

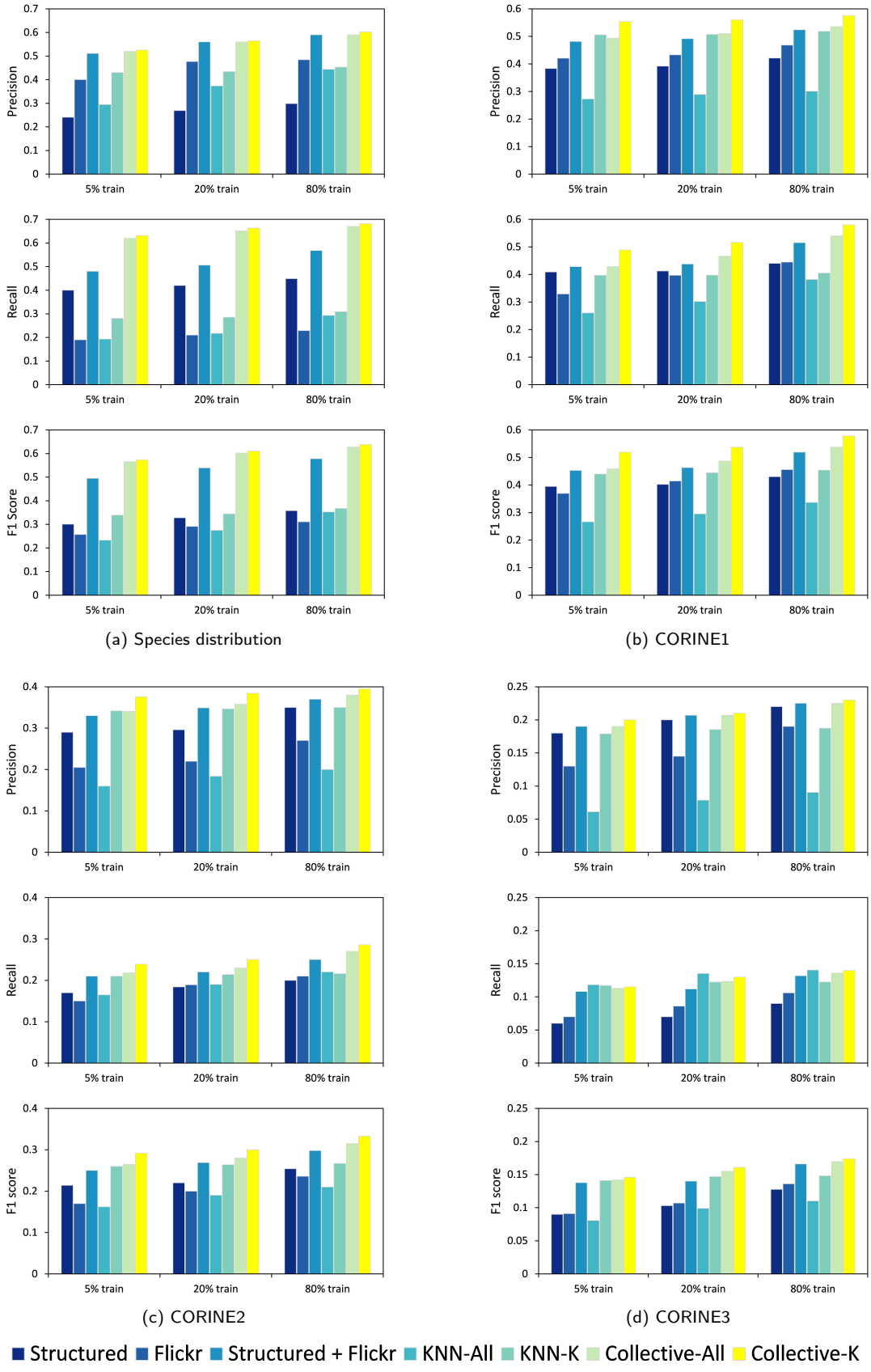



**Figure 7:** Results of classification problems.

**Table 3**
Results for predicting scenicness.

| | 5% | | 20 % | | 80 % | |
|---|---|---|---|---|---|---|
| | $\rho$ | MAE | $\rho$ | MAE | $\rho$ | MAE |
| Collective-K | 0.566 | 1.041 | 0.583 | 1.011 | 0.627 | 0.972 |
| Collective-K-embeddings-SVR | 0.571 | 1.038 | 0.591 | 0.998 | 0.615 | 0.984 |
| Collective-K-embeddings-MLP | 0.588 | 1.022 | 0.600 | 0.991 | 0.623 | 0.969 |

**Table 4**
Top 5 Flickr tags of Aquila chrysaetos, Dryocopus martius, and Lacerta bilineata species in terms of KL divergence.

| Aquila chrysaetos | Dryocopus martius | Lacerta bilineata |
|---|---|---|
| montagna | nationalpark | italy |
| spain | forest | tuscany |
| huesca | harz | umbria |
| aragon | mountains | lombardia |
| mountain | hautesavoie | lucertola |

**Table 5**
Top 5 Flickr tags for some CORINE level 1 classes in terms of KL divergence.

| Forest & semi nat. areas | Wetlands | Water bodies |
|---|---|---|
| forest | bog | sea |
| woods | moor | beach |
| mountains | marsh | coast |
| trees | swamp | lake |
| wald | saline | pier |

the feature vector representations as they are used as ground truth. We consider five different regression problems: predicting average annual temperature, average annual precipitation, average annual solar radiation, average annual wind speed, and average annual water vapor pressure. The results of these experiments are reported in Figure 6b, Figure 6c, Figure 6d, Figure 6e and Figure 6f respectively. The mean and standard deviation of each of those features are shown in Table 2. As regression problems, we evaluate the results in terms of Spearman $\rho$ and mean absolute error (MAE). Overall, combining both structured and Flickr data outperforms using them separately, showing that the information we obtain from Flickr is complementary to what is available as structured data and using collective prediction leads to an impressive improvement over the basic prediction methods, especially with the collective-K variant leading to the best results overall. Looking at the top selected tags in terms of KL divergence, we find names of countries, regions, or weather phenomena, which are indicative of either high or low values of the corresponding feature as shown in Table 6.

### 5.5. Qualitative analysis

We present two examples to illustrate how Flickr tags can help to determine the neighbourhood structure. First, looking at Figure 8, we can clearly see a scenic coastal area rated with 7.2 by the users in the SenicOrNot dataset. We note that considering the neighbours according to geographic proximity leads to a predicted value of 5.4. This is close to the average value, hence the model has failed to identify the location as being scenic in this case. However, using Flickr tags to filter these neighbours helps to make a more accurate prediction, with the value 7.8. The most characteristic tags used to select these neighbours are *sand*, *beach*, and *coast*, showing that the method has correctly identified that what matters most in determining the scenicness of the target location is that it is at the coast. Another interesting example is shown in Figure 9; looking at the image, we can see a housing development, which has been rated as not scenic with a value of 1.2. Considering the neighbours according to their geographic distance leads to a poor prediction, with the value of 5.7. However, using Flickr tags (and scientific features) to choose the neighbours leads to a much better prediction of 2.3. The most characteristic tags in this case were *road*, *houses*, and *buildings*.

## 6. Conclusions

In this paper, we have proposed a method which uses Flickr tags in tasks that rely on characterizing the environment. To this end, we have considered four different evaluation tasks. The first experiment aimed to predict the scenicness of a place, as assessed subjectively by humans on the ScenicOrNot website. In the second experiment, we focused on modelling the distribution of species across Europe, using observations from the Natura 2000 dataset as ground truth. The third experiment consisted in predicting CORINE land cover categories. Finally, we looked at predicting five climate related properties. The role of Flickr tags in our method is two-fold. First, we showed that Flickr tags can be used to supplement structured scientific data. We found that the combined model substantially and consistently outperformed the model that relied on either structured data sources or Flickr tags alone. This strongly suggests that Flickr can indeed be valuable, as a supplement to more traditional datasets in environmental analyses. Although all experiments demonstrated the benefits of using Flickr for selected environmental features, it may be speculated that in practice its use may be most beneficial in future for tasks such as species distribution modelling and scenicness prediction where, unlike temperature for example, there are no existing methods for direct instrumental recording of the phenomena. Second, we proposed a collective prediction model which again relied on both Flickr tags and structured data to define a neighbourhood structure. The use of a collective prediction formulation was motivated by the fact that most environmental features are strongly spatially autocorrelated. While this suggests that geographic distance should

**Table 6**
Top 5 Flickr tags for different climate related features in terms of KL divergence.

| Temperature | Precipitation | Solar Radiation | Wind Speed | Water Vapor Pressure |
|---|---|---|---|---|
| sweden | scotland | finland | island | sea |
| finland | ireland | sweden | sea | sardegna |
| snow | canaryislands | spain | denmark | mallorca |
| spain | nubes | italy | highlands | portugal |
| italy | clouds | france | beach | spain |

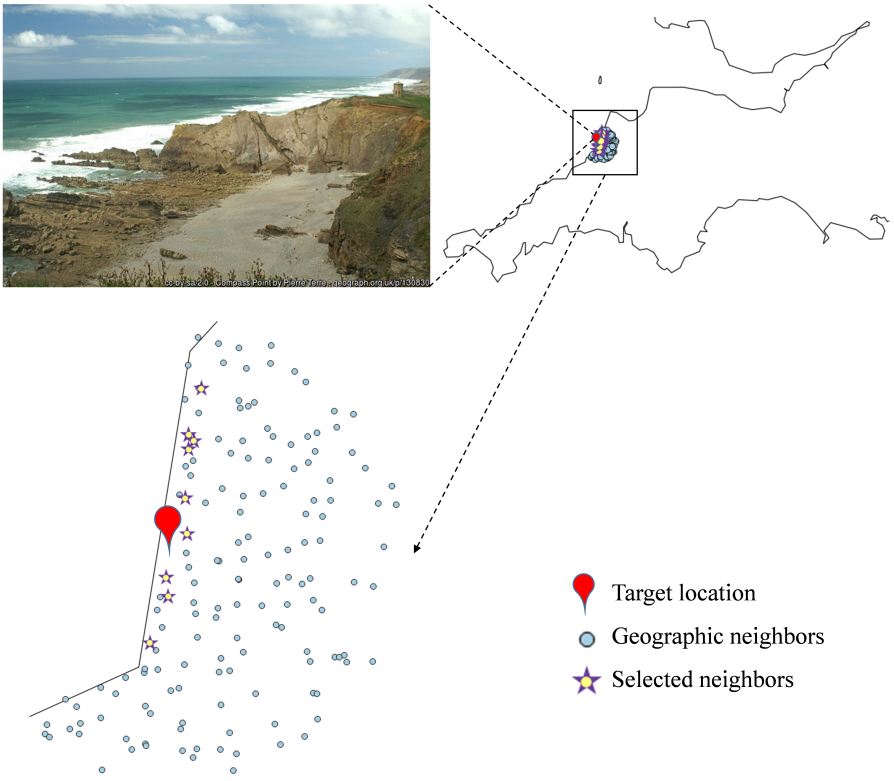



**Figure 8:** Neighbourhood structure for location coordinates (50.827,-4.559), photo link: https://www.geograph.org.uk/photo/130830

play a key role in determining neighbourhoods, we showed that considerable gains can be made by additionally taking Flickr tags and traditional data into consideration.

## Declarations of interest

None.

## Acknowledgments

Shelan Jeawak has been sponsored by HCED Iraq. Steven Schockaert has been supported by ERC Starting Grant 637277.

## References

Angelova, R., Weikum, G., 2006. Graph-based text classification: learn from your neighbors. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 485–492.

Bahrehdar, A. R., Purves, R. S., Jul. 2018. Description and characterization of place properties using topic modeling on georeferenced tags. Geo-spatial Information Science 21 (3), 173–184.
URL https://doi.org/10.1080/10095020.2018.1493238

Barve, V. V., 2015. Discovering and developing primary biodiversity data from social networking sites. Ph.D. thesis, University of Kansas.

Bishop, I. D., Hulse, D. W., 1994. Prediction of scenic beauty using mapped data and geographic information systems. Landscape and urban planning 30 (1-2), 59–70.

Bolognesi, M., 2016. Flickr distributional tagspace: evaluating the semantic spaces emerging from flickr tags distributions. Big Data in Cognitive Science.


**Figure 9:** Neighbourhood structure for location coordinates (56.138,-3.831), photo link: http://www.geograph.org.uk/photo/201043

Casalegno, S., Inger, R., DeSilvey, C., Gaston, K. J., 2013. Spatial covariance between aesthetic value & other ecosystem services. PloS one 8 (6), e68437.

Chakrabarti, S., Dom, B., Indyk, P., 1998. Enhanced hypertext categorization using hyperlinks. In: ACM SIGMOD Record. Vol. 27. ACM, pp. 307–318.

Chesnokova, O., Nowak, M., Purves, R. S., 2017. A crowdsourced model of landscape preference. In: LIPIcs-Leibniz International Proceedings in Informatics. Vol. 86. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Chopra, S. P., 2008. Factor graphs for relational regression. Ph.D. thesis, New York University.

Cunha, E., Martins, B., 2014. Using one-class classifiers and multiple kernel learning for defining imprecise geographic regions. International Journal of Geographical Information Science 28 (11), 2220–2241.

Daume, S., 2016. Mining twitter to monitor invasive alien species—an analytical framework and sample information topologies. Ecological Informatics 31, 70–82.

Dickinson, J. L., Zuckerberg, B., Bonter, D. N., Jan-12-2010 2010. Citizen science as an ecological research tool: Challenges and benefits. Annual Review of Ecology, Evolution, and Systematics 41, 149 – 172.

Dori-Hacohen, S., Jensen, D., Allan, J., 2016. Controversy detection in wikipedia using collective classification. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM, pp. 797–800.

Estima, J., Fonte, C. C., Painho, M., 2014. Comparative study of land use/cover classification using flickr photos, satellite imagery and corine land cover database. In: Proceedings of the 17th AGILE International Conference on Geographic Information Science, Castellon, Spain. pp. 1–6.

Estima, J., Painho, M., 2014. Photo based volunteered geographic information initiatives: A comparative study of their suitability for helping quality control of corine land cover. International Journal of Agricultural and Environmental Information Systems (IJAEIS) 5 (3), 73–89.

Fritz, S., McCallum, I., Schill, C., Perger, C., See, L., Schepaschenko, D., van der Velde, M., Kraxner, F., Obersteiner", M., 2012. Geo-wiki: An online platform for improving global land cover. Environmental Modelling & Software 31, 110 – 123.

Geman, S., Geman, D., 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence (6), 721–741.

Gliozzo, G., Pettorelli, N., Haklay, M. M., 2016. Using crowdsourced imagery to detect cultural ecosystem services: a case study in south wales, uk. Ecology and Society 21 (3).

Grothe, C., Schaab, J., 2009. Automated footprint generation from geotags with kernel density estimation and support vector machines. Spatial Cognition & Computation 9 (3), 195–211.

Hollenstein, L., Purves, R., 2010. Exploring place through user-generated content: Using flickr tags to describe city cores. Journal of Spatial Information Science 1.

Jeawak, S., Jones, C., Schockaert, S., 2018. Mapping wildlife species distribution with social media: Augmenting text classification with species names.

Jeawak, S. S., Jones, C. B., Schockaert, S., 2017. Using flickr for characterizing the environment: an exploratory analysis. In: 13th International Conference on Spatial Information Theory, COSIT 2017. Vol. 86 of Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, pp. 21:1–21:13.

Jeawak, S. S., Jones, C. B., Schockaert, S., 2019. Embedding geographic locations for modelling the natural environment using flickr tags and structured data. In: European Conference on Information Retrieval. Springer, pp. 51–66.

Jeawak, S. S., Jones, C. B., Schockaert, S., 2020. Predicting environmental features by learning spatiotemporal embeddings from social media. Ecological Informatics 55, 101031.

Jiang, Y., Li, T., Zhang, Y., Yan, Y., 2017. Collective prediction based on community structure. Physica A: Statistical Mechanics and its Applications 465, 587–598.

Joachims, T., 1998. Making large-scale svm learning practical. Tech. rep., SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund.

Kou, Z., Cohen, W. W., 2007. Stacked graphical models for efficient inference in markov random fields. In: Proceedings of the 2007 SIAM International Conference on Data Mining. SIAM, pp. 533–538.

Leung, D., Newsam, S., 2012. Exploring geotagged images for land-use classification. In: Proceedings of the ACM multimedia 2012 workshop on Geotagging and its applications in multimedia. pp. 3–8.

Li, J., Heap, A. D., Potter, A., Daniell, J. J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environmental Modelling & Software 26 (12), 1647–1659.

Liu, W., Du, P., Zhao, Z., Zhang, L., 2016. An adaptive weighting algorithm for interpolating the soil potassium content. Scientific reports 6, 23889.

Loglisci, C., Appice, A., Malerba, D., 2016. Collective regression for handling autocorrelation of network data in a transductive setting. Journal of Intelligent Information Systems 46 (3), 447–472.

Lowry, C. S., Fienen, M. N., 2013. Crowdhydrology: Crowdsourcing hydrologic data and engaging citizen scientists. Ground Water 51 (1), 151–156.

Montello, D. R., Jun. 1991. The measurement of cognitive distance: Methods and construct validity. Journal of Environmental Psychology 11 (2), 101–122.
URL https://linkinghub.elsevier.com/retrieve/pii/S0272494405800714

Neville, J., Jensen, D., 2000. Iterative classification in relational data. In: Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data. pp. 13–20.

Pierskalla, C. D., Deng, J., Siniscalchi, J. M., 2016. Examining the product and process of scenic beauty evaluations using moment-to-moment data and GIS: The case of Savannah, GA. Urban Forestry & Urban Greening 19, 212–222.

Richards, D. R., Friess, D. A., 2015. A rapid indicator of cultural ecosystem service usage at a fine spatial scale: content analysis of social media photographs. Ecological Indicators 53, 187–195.

Ristea, A., Kounadi, O., Leitner, M., 2018. Geosocial media data as predictors in a gwr application to forecast crime hotspots (short paper). In: 10th International Conference on Geographic Information Science (GIScience 2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Sack, R., Sep. 1980. Conceptions of Geographic Space. Progress in Human Geography 4 (3), 313–345.
URL https://doi.org/10.1177/030913258000400301

Schirpke, U., Tasser, E., Tappeiner, U., 2013. Predicting scenic beauty of mountain regions. Landscape and Urban Planning 111, 1–12.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., Eliassi-Rad, T., 2008. Collective classification in network data. AI magazine 29 (3), 93.

Serdyukov, P., Murdock, V., Van Zwol, R., 2009. Placing flickr photos on a map. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 484–491.

Stadler, B., Purves, R., Tomko, M., 2011. Exploring the relationship between land cover and subjective evaluation of scenic beauty through user generated content. In: Proceedings of the 25th International Cartographic Conference.

Tenerelli, P., Demšar, U., Luque, S., 2016. Crowdsourcing indicators for cultural ecosystem services: a geographically weighted approach for mountain landscapes. Ecological Indicators 64, 237–248.

Tkachenko, N., Jarvis, S., Procter, R., Feb. 2017. Predicting floods with Flickr tags. PLOS ONE 12 (2), e0172870. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0172870

Van Canneyt, S., Schockaert, S., Dhoedt, B., 2014. Discovering and characterizing places of interest using flickr and twitter. Hospitality, Travel, and Tourism: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications, 393.

Wang, J., Korayem, M., Crandall, D., 2013. Observing the natural world with flickr. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 452–459.

Webster, R., Oliver, M. A., Oct. 2007. Geostatistics for Environmental Scientists. John Wiley & Sons, google-Books-ID: WBwSyvIvNY8C.

Zhang, H., Korayem, M., Crandall, D. J., LeBuhn, G., 2012. Mining photo-sharing websites to study ecological phenomena. In: Proceedings of the 21st international conference on World Wide Web. pp. 749–758.