Exploiting Flickr Meta-Data for Predicting Environmental Features

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Shelan S. Jeawak

2019

Cardiff University

School of Computer Science & Informatics

STATEMENT 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD.

Signed ....

Date ....

STATEMENT 2

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is it being submitted concurrently for any other degree or award (outside of any formal collaboration agreement between the University and a partner organisation).

Signed ....

Date ....

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available in the University's Open Access repository (or, where approved, to be available in the University library and for inter-library loan), and for the title and summary to be made available to outside organisations, subject to the expiry of a University-approved bar on access if applicable.

Signed ....

Date ....

DECLARATION

This thesis is the result of my own independent work, except where otherwise stated, and the views expressed are my own. Other sources are acknowledged by explicit references. The thesis has not been edited by a third party beyond what is permitted by Cardiff University's Use of Third Party Editors by Research Degree Students Procedure.

Signed ....

Date ....

WORD COUNT............................

(Excluding Abstract, acknowledgements, declarations, contents pages, appendices, tables, diagrams and figures, references, bibliography, footnotes and endnotes)

To the soul of my mother the origin of my success.

To my beloved father; I hope I made you proud.

Abstract

The photo-sharing website Flickr has become used as an informal information source in disciplines such as geography and ecology. Many recent studies have highlighted the fact that Flickr tags capture valuable ecological information, which can complement more traditional sources. A shortcoming of most of these existing methods is that they rely on manual interpretation of Flickr content, with little automated exploitation of the associated tags. Therefore, they fail to exploit the full potential of the data. Automatically extracting and analysing information from unstructured and noisy data remains a hard task. This research aims to investigate the use of Flickr meta-data for predicting a wide variety of environmental phenomena. In particular, we consider the problem of predicting scenicness, species distribution, land cover, and climate-related features. To this end, we developed several novel machine learning methods that can efficiently utilise Flickr tags as a supplementary source to the structured information that is available from traditional scientific resources.

The first proposed method aims at modelling locations, and hence inferring environmental phenomena, using georeferenced Flickr tags. Our focus was on comparing the predictive power of Flickr tags with that of structured environmental data. This method represents each location as a concatenation of two feature vectors: a bag-of-words representation derived from Flickr and a feature vector encoding the numerical and categorical features obtained from the structured dataset. We found that Flickr was generally competitive with the structured environmental data for prediction, being sometimes better and sometimes worse. However, combining Flickr tags with existing ecological data sources consistently improved the results, which suggests that Flickr can indeed be regarded as complementary to traditional sources. The second method that we propose is based on a collective prediction model, which crucially relies on Flickr tags to define the 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.

The thesis considers two further novel methods which are based on a low dimensional vector space representation. The first model, called EGEL (Embedding Geographic Locations), learns vector space embeddings of geographic locations by integrating the textual information derived from Flickr with the numerical and categorical information derived from environmental datasets. We experimentally show that this method improves on bag-of-words representation approaches, especially in cases where structured data are available. This model has been extended by considering a spatiotemporal representation of regions. In particular, we propose a spatiotemporal embeddings model, called SPATE (Spatiotemporal Embeddings), which learns a vector space embedding for each geographic region and each month of the year. This allows the model to capture environmental phenomena that may depend on monthly or seasonal variation. Apart from extending our primary model, SPATE also includes a new smoothing method to deal with the sparsity of Flickr tags over the considered spatiotemporal setup.

The experimental results demonstrated in this thesis confirm our hypothesis that there is valuable information contained in Flickr tags which can be used to predict environmental features.

Acknowledgements

Undertaking this PhD has been truly one of the most challenging and rewarding experiences in my life. Despite the ups and downs, the disappointments and triumphs, the happy and sad moments, my journey has finally got to its end. I am certain that I would not have been able to complete this journey without the help and support of ALLAH and many people.

I am heartily thankful to my main supervisor Professor Steven Schockaert, for his uncountable encouragement, guidance and support from the beginning of my PhD. His professional excellence enabled me to develop my research skills and tackle the challenges of this research. Thanks for his great enthusiasm and patience with me. I would also like to thank my second supervisor Professor Christopher Jones, for his knowledgeable advice, help and guidance. I am much blessed at being under the supervision of such great people; they have ideally assisted me in all of my research needs.

I would also like to thank all the members of the School of Computer Science and Informatics at Cardiff University for their kind assistance. I must acknowledge the technical support of the ARCCA team, especially Mr Thomas Green. Thanks to all my friends and colleagues in Cardiff who have been positive and supportive during the PhD journey.

I want to extend my thanks to my sponsor in Iraq, HCED Iraq, for the financial support of my PhD study in Cardiff. I am also thankful to my friends and colleagues in the Department of Computer Science at Al - Nahrain University for their encouragement and friendship.

I am deeply grateful to my family, who always encourage and support me. To my parents, who provided endless support, encouragements and love, to make me who I am today. Words cannot express how I am grateful to them. I wish my mother was still alive and share this achievement with me. To my sisters Suzan and Rajuan, and my brothers Dr Mohammed and Muayed; you have always been my biggest cheerleaders. To my nieces Noor and Maryam and my nephews Ahmed and Ibrahim for their love. My last thankful words are kept to my husband Ahmed and my lovely son Mohammed for their unconditional love, support, and care. It is thanks to them I was able to come to the finish line of my PhD.
Thank You!

Contents

Abstract **vii**

Acknowledgements **ix**

Contents **xi**

List of Publications **xvii**

List of Figures **xix**

List of Tables **xxiii**

List of Acronyms **xxv**

1 Introduction **1**

1.1 Background and Motivation 1

1.2 Hypothesis and Research Questions 3

1.3 Contributions 4

1.4 Thesis Structure 6

1.5 Summary 8

2 Background and Related Work 9

2.1 Introduction 9

2.2 Social Media 10

2.2.1 Social Media Mining 12

2.3 Text Representation 17

2.3.1 Term Weighting 18

2.3.2 Term Selection 20

2.4 Low Dimensional Vector Space Representation 23

2.4.1 Vector Space Embeddings 23

2.4.2 Embedding Spatial or Spatiotemporal Information 25

2.5 Machine Learning 27

2.5.1 Supervised Machine Learning 28

2.5.2 Collective Prediction 31

2.5.3 Evaluation Measures 33

2.6 Summary 36

3 Data Acquisition and Preprocessing 37

3.1 Introduction 37

3.2 Flickr Data 38

3.2.1 Flickr Data Collection 38

3.2.2 Flickr Data Preprocessing 40

3.2.3 Exploratory Data Analysis 41

3.3 Structured Scientific Data 46

3.4 Ground Truth Data 48

3.4.1 The ScenicOrNot project 48

3.4.2 Natura2000 project 48

3.4.3 NBN Atlas 49

3.5 Summary 49

4 Modelling Locations using Bag-Of-Words Representation 51

4.1 Introduction 51

4.2 Methodology 52

4.2.1 Modelling Locations Using Flickr Tags 52

4.2.2 Modelling Locations Using Structured Data 54

4.3 Experiments 55

4.3.1 Predicting Scenicness 55

4.3.2 Predicting Species Distribution 57

4.3.3 Predicting CORINE Land Cover Classes 59

4.3.4 Predicting Climate Data 60

4.4 Evaluating the Role of Species Name Tags 62

4.4.1 Methodology and Data 64

4.4.2 Evaluation 65

4.4.3 Results and Discussion 66

4.5 Summary 69

5 Collective Prediction Model 73

5.1 Introduction 73

5.2 Collective Prediction 74

5.3 Experimental Evaluation 81

5.3.1 Experimental Settings 81

5.3.2 Variants and Baseline Methods 81

5.3.3 Experimental Results 84

5.4 Summary 93

6 Modelling Locations using Vector Space Embedding 95

6.1 Introduction 95

6.2 Embedding Geographic Location 96

6.2.1 Tag Based Location Embedding 97

6.2.2 Structured Environmental Data 99

6.3 Experimental Evaluation 101

6.3.1 Experimental Settings 101

6.3.2 Variants and Baseline Methods 102

6.3.3 Experimental Results 103

6.4 Summary 108

7 Spatiotemporal Embeddings Model 111

7.1 Introduction 111

7.2 Related Work: Spatiotemporal Modelling 112

7.3 Methodology 113
7.3.1 Spatiotemporal Modelling Using Flickr tags 113
7.3.2 Spatiotemporal Modelling Using Structured Environmental Data 119
7.3.3 Spatiotemporal Embeddings 121
7.4 Experimental Evaluation 123
7.4.1 Selecting the Optimal Bandwidth for Each Tag 124
7.4.2 Experimental Settings 126
7.4.3 Variants and Baseline Methods 127
7.4.4 Experimental Results 128
7.4.5 Location Similarity 134
7.5 Summary 141

8 Conclusions and Future Work 143
8.1 Introduction 143
8.2 Thesis Summary and Contributions 143
8.3 Research Questions 146
8.4 Future Work 148
8.5 Summary 149

Bibliography 151

List of Publications

The work introduced in this thesis is based on the following publications:

1. 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, volume 86, pages 21:1–21:13 [54].

2. Shelan S Jeawak, Christopher B Jones, and Steven Schockaert. Mapping wildlife species distribution with social media: Augmenting text classification with species names. In 10th International Conference on Geographic Information Science, GIScience 2018, August 28-31, 2018, Melbourne, Australia, volume 114, pages 34:1–34:6, [52].

3. Shelan S Jeawak, Christopher B Jones, and Steven Schockaert. Embedding Geographic Locations for Modelling the Natural Environment Using Flickr Tags and Structured Data. In 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings Part I, volume 11437, pages 51–66, [55].

4. Shelan S Jeawak, Christopher B Jones, and Steven Schockaert. Collective Prediction of Environmental Features using Flickr Tags, under review.

5. Shelan S Jeawak, Christopher B Jones, and Steven Schockaert. Predicting environmental features by learning spatiotemporal embeddings from social media, Accepted for publication in Ecological Informatics, Elsevier.

List of Figures

2.1 The geometric intuition of linear SVMs. 30

2.2 Artificial Neural Networks architecture. 31

3.1 Flickr photographs 39

3.2 The number of georeferenced photographs in each month among the collected Flickr data 42

3.3 Scientific species names frequency among Flickr data. 43

3.4 Number of photographs in each of the CORINE land cover class. . . . 45

3.5 CORINE Land Cover dataset. 46

4.1 Modelling locations using a Bag-Of-Words model. 52

4.2 Prediction of the black woodpecker distribution across Europe. 58

4.3 Prediction of subclasses of the CORINE class “Forest & semi natural areas” 61

4.4 Prediction of solar radiation. 63

4.5 The training process. 65

4.6 Training, Tuning, and Testing regions. 66

4.7 Prediction of the Dunlin distribution across the testing area with 10km grid cells 69

4.8 Prediction of the Atlantic Puffin distribution across the testing area with 10km grid cells 69

5.1 Modeling locations based on Flickr tags, structured features, and neighborhood structure. 75

5.2 The collective prediction model. 76

5.3 Structured+Flickr prediction model (baseline method). 83

5.4 K nearest neighbors prediction model (baseline method). 83

5.5 Modeling locations based on collective features. 83

5.6 Results of predicting species distribution. 85

5.7 Results of predicting CORINE land cover at level 1. 87

5.8 Results of predicting CORINE land cover at level 2. 88

5.9 Results of predicting CORINE land cover at level 3. 89

5.10 Results of predicting scenicness. 90

5.11 Results of predicting average annual temperature. 90

5.12 Results of predicting average annual precipitation. 91

5.13 Results of predicting average annual solar radiation. 91

5.14 Results of predicting average annual wind speed. 92

5.15 Results of predicting average annual water vapour pressure. 92

6.1 Modelling locations in vector space embedding 97

6.2 The geometric intuition of tags based embedding. 99

6.3 The geometric intuition of numerical features based embedding. . . . 100

6.4 The geometric intuition of categorical features based embedding. . . . 101

6.5 Comparison between the performance of the GloVe and bag-of-words models for predicting scenicness, as a function of the number of tag occurrences at the considered locations 105

7.1 Spatiotemporal grid cells. 114

7.2 The representation of the months as circular data. 117

7.3 The spatiotemporal embeddings (SPATE) model. 124

7.4 The average MISE of all the considered tags when using the rule of thumb (ROT), the adaptive kernel bandwidth (Adaptive), and the leave-one-out kernel estimator (LOO) 125

7.5 Prediction of the seasonal distribution of Swift across the UK and Ireland using 1% of the data for training/tuning 131

7.6 F1 score of predicting the monthly distribution of Swift. 132

7.7 The prediction confidence score for location coordinate (54.81503616, -2.086120293) over the testing months. Jan, Oct, Nov and Dec are not shown in the figure because the corresponding cells are in the training set for that location. Note that Swift has positive ground truth observations in that location in April, May, June, July and August, and negative ground truth observations in February, March and September 132

7.8 Prediction of the seasonal precipitation across the UK and Ireland using 1% of the data for training/tuning. 136

7.9 The monthly prediction results of precipitation. 137

7.10 The monthly average value of predicting the amount of precipitation for location coordinate (latitude= 55.26469626 and longitude= -4.784080876) over the testing months. April, July, October, November, and December are not shown in the figure because the corresponding cells are in the training set for that location 138

7.11 Location's similarity maps in July. 139

7.12 Location's similarity maps in January 140

List of Tables

2.1 Confusion matrix for binary classification. 34

3.1 Top 10 tags with the highest KL divergence, in relation to the months of the year and the occurrence of species names 44

3.2 Top 10 tags related to forest in different term selection methods. . . . 45

4.1 Results for predicting scenicness. 56

4.2 Results for predicting species distribution. 58

4.3 Results for predicting CORINE land cover classes, at levels 1, 2 and 3. 60

4.4 Top 5 Flickr tags for CORINE level 1 classes in the SVM models. . . 60

4.5 Results for predicting average climate data. 62

4.6 Mean and Standard deviation of climate data. 62

4.7 Results for predicting the distribution of 50 species across the testing area 67

4.8 F1 scores for predicting the distribution of individual species using different methods 68

5.1 High and low boundaries for discretising the features from the regression problems . . . 84

5.2 Top 5 Flickr tags of Aquila chrysaetos, Dryocopus martius, and Lacerta bilineata species in terms of KL divergence . . . 86

5.3 Top 5 Flickr tags for some CORINE level 1 classes in terms of KL divergence . . . 86

5.4 Top 5 Flickr tags for different climate related features in terms of KL divergence . . . 93

6.1 Results for predicting species distribution. . . . 104

6.2 Results for predicting soil type. . . . 105

6.3 Results for predicting CORINE land cover classes, at levels 1, 2 and 3. 106

6.4 Results for predicting average climate data. . . . 107

6.5 Results for predicting scenicness. . . . 108

7.1 Top 10 Flickr tags in terms of KL divergence. . . . 127

7.2 Results for predicting the monthly distribution of 50 species across the UK and Ireland . . . 130

7.3 Results for predicting the monthly average climate features. . . . 135

7.4 Mean and Standard deviation of the monthly average climate data. . . 137

List of Acronyms

GPS Global Positioning System

API Application Programming Interface

VGI Volunteered Geographic Information

UGC User Generated Content

BOW Bag-Of-Words

CBOW Continuous Bag-of-Words

GloVe Global Vectors for Word Representation

NLP Natural Language Processing

POI Points-of-Interest

ML Machine Learning

TP True Positive

TN True Negative

FP False Positive

FN False Negative

SVM Support Vector Machines

SVR Support Vector Regression

EGEL Embedding Geographic Locations

SPATE Spatiotemporal Embeddings

Chapter 1

Introduction

1.1 Background and Motivation

Social media platforms such as Flickr[1], Twitter[2] and Facebook[3] have become popular vehicles for sharing and finding information. This led to the creation of a large amount of a new form of data, which is known as social media data. These data are mostly user-generated, informal, and unstructured, and is sometimes associated with information about time and location. For example, the photo-sharing platform Flickr hosts more than 10 billion photographs[4], most of which are associated with short textual descriptions in the form of tags to describe what is depicted in the photograph. Moreover, the time at which these photographs were taken is available as a meta-data. The Global Positioning System (GPS) support in current electronic devices such as smartphones means that latitude and longitude coordinates can be easily recorded as meta-data. For a large number of photographs on Flickr, these coordinates have been made publicly available[5]. Together with their textual descriptions, such photographs can thus be regarded as Volunteered Geographic Information (VGI [41]). VGI is a special case of the larger phenomenon known as User Generated Content (UGC) [41]. It allows people to voluntarily create, collect, and disseminate geographic information, which

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

[2] http://www.twitter.com

[3] http://www.facebook.com

[4] http://expandedramblings.com/index.php/flickr-stats

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

has played an active role for some applications such as urban planning and mapping [111]. The coordinates and textual meta-data associated with Flickr photographs have already proven valuable in many disciplines. For example, in geography, Flickr tags have been used to construct approximate boundaries for (vernacular) regions [19, 42] or to discover events that take place in a given city [91]. In linguistics, the tags of georeferenced Flickr photographs have been found useful for generating vector space representations of perceptual terms [7] and the correlations between the occurrence of Flickr tags and geographical location have been used to analyse colloquial language [29]. In the domain of ecology, Flickr has been used to study species distribution [5]. The use of Flickr for modelling urban environments has already received considerable attention. For instance, various approaches have been proposed for modelling urban regions [19], and for identifying points-of-interest [117] and itineraries [21, 90]. However, the usefulness of Flickr for characterising the natural environment, which is the focus of this thesis, is less well-understood.

Although there are many organisations that serve environmental data, the information they provide is far from complete [5]. The idea of using Flickr as a supplementary source of environmental information is appealing for several reasons. For example, due to the fact that photographs are often uploaded directly after they have been taken, Flickr can provide us with more up-to-date information than traditional citizen science datasets. This can be important, for instance, for monitoring the spread of invasive species and migration patterns of pollinators. Moreover, the information that is captured by Flickr tags is broader than what is normally recorded, and includes, for example, the subjective assessments about the scenicness of a landscape. In fact, Flickr has already proven valuable as a resource for ecological analysis. However, most of the recent studies rely on manual evaluations of image content with little automated exploitations of the associated tags [92, 30]. Manually analysing Flickr is clearly limited and time-consuming. Moreover, both the structure and the volume of the data present practical challenges [20], compared to formal or semi-formal citizen science monitoring data [107]. Nonetheless, these studies prove that Flickr contains valuable information which could be used to support the available sources [20, 5]. All this highlights the need for automated methods for extracting environmental information from Flickr.

The aim of the research presented in this thesis is to automate methods that can utilise Flickr meta-data as a supplementary source of environmental information. The idea has come from the fact that Flickr data are free to use, more up-to-date than traditional resources, and has already proven valuable as a resource for ecological information [5, 20]. In particular, this thesis studies the usefulness of Flickr tags for predicting a wide range of environmental phenomena, such as land cover categories, species occurrence, scenicness of place, and climate features which include average temperature, wind speed, precipitation, solar radiation and water vapour pressure. The most important challenge to handle in social media mining is that its content is often sparse and noisy. To mitigate this problem, our analysis in this research focuses on features that we can ascribe to locations, for example, "there is a coniferous forest at this location" rather than to individual photographs, for example, "this is a photo of a 7-spot ladybird". Flickr tags can be an extremely rich source of information. However, extracting useful knowledge from it can be difficult due to its unstructured nature. To this end, we need methods that efficiently extract and represent such potentially useful information hidden in Flickr. Moreover, to employ Flickr tags as a supplementary source to the structured scientific datasets, we need to develop methods that can efficiently combine those two diverse data sources. To achieve this, several novel text mining and machine learning algorithms have been proposed and developed in the present thesis.

1.2 Hypothesis and Research Questions

Our main hypothesis in this thesis is:

Social media can be used as a valuable source of ecological information. In particular, we can use the meta-data associated with the photographs on the photo-sharing platform Flickr as a complementary source to the publicly available scientific datasets in *order to predict spatially and temporally grounded information about the natural environment. This meta-data allows us to improve the prediction of features such as the scenicness of a place, species distribution, land cover categories, and several climate related features.*

In order to verify this hypothesis, the following set of research questions were addressed:
Research Question 1: *Is it possible to extract large amounts of high-quality environmental information from Flickr, and if so, how complementary is this information to publicly available scientific datasets?*
Research Question 2: *How can we deal with the sparsity of Flickr tags for location (and possibly time-dependent) representation?*
Research Question 3: *How can we best integrate these representations with the available structured environmental data to improve the predictive power?*

1.3 Contributions

The primary contribution of the present PhD research is the development of methods for utilising Flickr meta-data as a complementary source of environmental information. The contributions made through this research are:

1. We introduce a new method for modelling locations using georeferenced Flickr tags. The method is based on a spatially smoothed version of pointwise mutual information. The main aim of this work is to obtain a clearer picture about the kinds of environmental features that can be modelled using Flickr tags. To this end, we consider the problem of predicting scenicness, species distribution, land cover, and climate features. We focus on comparing the predictive power of Flickr tags with that of structured data from more traditional sources. We find that Flickr tags perform sometimes better and sometimes worse than the

considered structured data. Nonetheless, combining Flickr tags with structured data consistently improves the results. This suggests that Flickr can be used as a complementary source to traditional sources. This work was published in [54].

2. We compare and combine two main strategies for using Flickr to predict the spatial distribution of species. The first strategy is based on identifying postings that explicitly mention the target species name, while the second is based on exploiting all tags to construct a model of the locations where the species occurs. We find that the second strategy works well overall. However, in a few cases, the strategy that uses the species names only leads to better 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 using all Flickr tags. This work was published in [52].

3. We propose a novel collective prediction framework that relies on both Flickr tags and structured data to make initial predictions that are updated iteratively using a combination of neighbouring predictions and ground truth data. The motivation behind the use of collective prediction is, as in conventional spatial interpolation, that most environmental features are spatially autocorrelated. A key feature of the approach is that, in the collective prediction model, estimation of a location from its neighbouring data depends not only on geographic distance but also on attribute similarity, which is estimated in our case from the Flickr tags and structured data associated with each location. This work is still under review.

4. We develop a novel method for learning low-dimensional vector space embeddings of geographic locations, called EGEL, by combining textual information in the form of Flickr tags with the numerical and categorical information contained in structured scientific datasets. Our experimental evaluations show that using such low-dimensional vector space representations allows us to integrate the textual, numerical and categorical features in a more effective way than is possible with bag-of-words representations. This work was published in [55].

5. We extend the EGEL model to encode spatiotemporal information. To this end, we propose a novel spatiotemporal embeddings model, named SPATE, which is able to integrate Flickr tags and structured scientific information from more traditional environmental data sources. The novelty of this work is two-fold. First, we propose a new method based on spatiotemporal kernel density estimation to handle the sparsity of the tag distribution over space and time. Then, we efficiently integrate the spatially and temporally smoothed Flickr tags with the structured scientific data into low-dimensional vector space representations. The proposed model can be used for modelling and predicting a wide variety of ecological features such as species distribution, as well as related phenomena such as climate features. We experimentally show that our model is able to substantially outperform baselines that rely only on Flickr or only on traditional sources. This work is accepted for publication in [53].

1.4 Thesis Structure

The remaining chapters are organised as follows:

- Chapter 2 - Background and Related Work - provides an overview of social media in general and social media mining in particular. The chapter reviews the related work in this area and also defines the fundamental concepts of the relevant methods for text representation, vector space embedding, and machine learning.
- Chapter 3 - Data Acquisition and Preprocessing - introduces the datasets used in this work, covering both Flickr data and structured scientific data, and describes the methodology that was used for collecting each of those datasets. A primary analysis of the collected Flickr data is conducted to evaluate its usefulness toward the considered task. It also presents a set of ground truth datasets related to the environment and biodiversity that was used for experimentally evaluating the methods developed in this thesis.

- Chapter 4 - Modelling Locations using Bag-Of-Words Representation - describes the proposed methodology for modelling geographic locations using the structured data and using Flickr tags based on a bag-of-words (BOW) representation model. The chapter introduces a set of experiments based on supervised machine learning to evaluate, investigate, and compare the prediction power of using Flickr tags only, structured data only, and their combination. It also focuses on evaluating the role of tags that correspond to the name of the species for estimating species occurrence.

- Chapter 5 - Collective Prediction Model - presents the proposed collective prediction framework. It investigates the usefulness of Flickr tags to make the initial prediction and defines the neighbourhood structure of a given environmental feature. Detailed experiments are carried out to test the quality of the proposed model.

- Chapter 6 - Modelling Locations using Vector Space Embedding - describes the proposed *Embedding GEographic Locations* (EGEL) model that integrates the georeferenced Flickr tags and structured scientific data into a low-dimensional vector space embedding. It experimentally shows that EGEL model can integrate Flickr tags with structured information in a more effective way than the BOW model from Chapter 4.

- Chapter 7 - Spatiotemporal Embeddings Model - introduces the proposed *SPAtioTemporal Embeddings* (SPATE) model that handles the problem of Flickr data sparsity and is aimed at learning a low-dimensional vector space embedding of spatiotemporal regions based on the textual, numerical, categorical, spatial, and temporal information. The chapter qualitatively and quantitatively evaluates how well the SPATE model can predict the monthly and seasonal variation of a number of environmental phenomena.

- Chapter 8 - Conclusion and Future Work- concludes the thesis by summarising our contributions, findings, as well as highlighting proposals for future work.

1.5 Summary

In this chapter, we introduced the background and our motivation to work on the considered topic. We also discussed the hypothesis and the main research questions, and we gave an overview of the thesis contributions and structure. Before moving to the main technical contributions of this thesis, the next chapter will first provide more detailed background information and put the thesis in the context of existing work.

Chapter 2

Background and Related Work

2.1 Introduction

This thesis aims to explore the benefits of using social media, specifically the metadata associated with the photo-sharing platform Flickr, as a supplementary source of ecological information. Mining social media platforms has become a very active research area in many domains, as it coincides with the rapid growth and the availability of textual user-generated content on the web. This chapter will present the required background knowledge about the field of social media mining, which encompasses text mining and machine learning.

In particular, this chapter serves four main purposes that are directly related to the research presented in the thesis. First, Section 2.2 presents a general overview of social media and the process of mining social media platforms. To address the question of whether social media can be considered as a valuable source of information, two closely related research areas are discussed, which are: geospatial analysis of social media and citizen science. Then Section 2.3 attempts to define the common techniques that are used to generate bag-of-words representations of text documents. Specifically, we discuss methods that are used for term weighting and term selection. Subsequently, Section 2.4 describes the state-of-the-art methods used for learning low-dimensional vector space representations, including word embeddings and spatial or spatiotemporal information embeddings. Section 2.5 presents and compares some of the widely used machine learning approaches for the task of supervised learning. It also presents a more detailed review of the collective prediction model as well as the standard evaluation methods. Finally, Section 2.6 summarises the main topics discussed in this chapter. Note that there is an additional short related work section in Chapter 7, specifically Section 7.2, on spatiotemporal modelling.

2.2 Social Media

The rapid emergence and dissemination of Web 2.0 functionalities during the first decade of the 21st century has led to a leap in the social component of Web use. In contrast with the first generation of Web 1.0 where people were mostly limited to view the content of websites only, Web 2.0 applications allow users to communicate and share information through social media platforms. Social media can be defined as a class of web-based applications and information sources. They are typically characterised by collaborative content creation driven by explicit or implicit social networks that represent virtual communities of shared interest. The concept of social media has a broader meaning than the interaction platforms where a large amount of user-generated data are now available. Kaplan and Haenlein [59] define social media as *"a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content"*. These data are of great importance in many domains when mined and used for such purposes as analysis, modelling and prediction.

There are many types of social media that vary with regard to the level of personalisation and the richness of the media [59]. For example, the microblogging network is highly personalised as the authors provide content and information about themselves, whereas collaborative projects such as Wikipedia have a low degree of personalisation. It is difficult to propose a precise classification for social media types; however, the most popular classes include:

- Social Network Sites (SNSs) are a typical type of social media. The purpose of SNSs is to provide users with a platform to connect with others, such that they can share thoughts, knowledge, photographs and videos, and participate in discussions. Generally, a social networking service builds on and reflects the real-life social networks among people through online platforms. Facebook and Linkedin are well-known examples of SNSs worldwide. Some of these platforms provide the ability to share the user's location information. It is also possible to attach the location and the timestamp with the posts.

- Location-Based Social Networks (LBSNs) are social networks that use GPS coordinates such that users can share their location information. LBSNs are bridging the gap between the physical world and online social networking services. Examples of LBSN sites are Foursquare and Whrrl. LBSNs do not only add a location to an existing social network. They are also determined by the new social structure made up of individuals connected by their geographical locations as well as their location tagged content, such as text, photographs, and video. Furthermore, the physical location consists of the instantaneous position of a user at a given timestamp and the location history of the user over a specific time period [133].

- Blogging Networks, also called weblogs, blogs or online diaries, are informational or discussion platforms that are often informal and loosely connected, with a high level of personalisation diary-style posts. Examples are Wordpress and Blogspot.

- Microblogging Networks are blogging platforms where the amount of information that can be shared per user is very short. The most common examples of microblogs are Twitter and Tumblr. For instance, Twitter limits users' posts, which are called "tweets", to 280 characters. These tweets can come in the form of a variety of content formats, including text, images, video, audio, and hyperlinks as well as the location tag.

- Wikis are platforms that allow users to collaborate in creating and editing the content of a website which can include text, photographs and external links. The best-known example is Wikipedia, which is a free multilingual online encyclopaedia written by users.

- Photo-Sharing Sites are platforms that give people a place to share, store and find media online, with an emphasis on photographs. While the posts on the majority of social media platforms of the aforementioned types start with a text that may be supplemented with photographs or videos, posts on photo-sharing sites start with a photograph (or video) as the main posts. This post can then be supplemented with textual data in the form of a title, description and a set of tags that express what is in the photographs. Highly popular photo-sharing sites are Flickr, Instagram, and Snapchat.

The photo-sharing platform Flickr has been chosen as the site to be exploited in the present research. Both professional photographers and amateurs widely use Flickr as a platform for sharing their photographs. There are more than 90 million monthly active users on Flickr[1], and more than 10 billion photographs have been uploaded so far on Flickr [2], many of which are publicly available. The nature of the data that is available from Flickr will be described in more detail in Chapter 3.

2.2.1 Social Media Mining

The process of extracting useful information from large-scale user-generated data on social media sites is usually known as social media mining. This term is an analogy to the process of mining to extract rare minerals. Resources mining requires professional specialists and advanced technologies to sift through a vast amount of raw ore. Similarly, social media mining requires data analysts and automated software to sift

[1] `https://blog.statusbrew.com/social-media-statistics-2019/`

[2] `http://expandedramblings.com/index.php/flickr-stats`

through massive amounts of raw social media data. It usually uses a range of data mining and machine learning techniques for analysing, representing, and extracting trends and patterns.

The availability of GPS systems in current electronic devices such as smartphones enables GPS coordinates (latitude and longitude) to be recorded as meta-data for social media posts. Thus, this data can be regarded as Volunteered Geographic Information (VGI) [41]. Since time is also recorded meta-data, social media can also be utilised as a real-time data source. The problem of mining social media data has received significant research attention in recent years and has led to valuable contributions in critical fields such as monitoring public health [17, 61], detecting natural disasters such as earthquakes [116, 28, 34] and typhoons [96], or predicting criminal activities [122].

With reference to these successful applications, ecological observations shared via social media could contribute to public participation in scientific research, something that is often named “citizen science”. In this sense, this thesis deals with social media data, specifically Flickr meta-data, as a passive form of citizen science. To this end, we need a conceptual framework within which this data can be explored, assessed, and used as an additional source of data. Below we will discuss some previous works that use social media data in general, and Flickr data in particular for applications, including geospatial analysis and citizen science, which are closely related to this research.

Geospatial Analysis

Geospatial analysis is the process of gathering, displaying, and manipulating GPS data, satellite imagery or historical data in a way that can be applied to geographic models. Many recent studies have focused on analysing georeferenced social media data, with the aim of extracting useful geographic information. In particular, there is a large number of studies that derive such information from georeferenced Flickr photographs. For example, [42] described two methods for the automatic delineation of imprecise regions based on geotagged photographs. 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, [19] presents 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 photographs that were tagged with the region's name, and external features such as land cover data, population count, elevation and the geographical coordinates (latitude and longitude) of Flickr photographs that are tagged with the region's name. They showed that their method performs better than the simpler methods described by [42]. Our work in this thesis is analogous to these approaches, in applying support vector machine learning methods to Flickr tags in combination with other geospatial data, but we are concerned with characterising and predicting information about the environment.

The authors of [105] presented and evaluated methods for automatically georeferencing Flickr photographs using the textual annotations of photographs 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 furthermore demonstrated how to incorporate the GeoNames database and they defined extensions to improve their language models. In [117], a language modelling approach was used to discover and characterise 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.

Citizen Science

Citizen science, also known as community science, crowdsourced science, or volunteer monitoring [27], refers to scientific research conducted by members of the general public, typically as part of a collaborative project with professional scientists. Considerable progress has been made in recent years in citizen science projects in the environmental sciences, where participants are recruited to actively contribute to particular campaigns such as in land cover mapping [37], hydrological surveys [75], ornithology and many forms of ecological study [25]. In parallel with these initiatives, there is a growing interest in the potential of “passive” survey methods that exploit social media to provide additional useful data. For instance, [120] 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 [131] photographs 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 [69] 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 [31] and [32], they classified a sample of georeferenced Flickr photographs according to CORINE land cover classes. They also evaluated the use of Flickr photographs 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 (see Chapter 3 for more details on the CORINE dataset). Note that their approach did not use machine learning and the results were evaluated manually by experts. Their results suggest that Flickr photographs 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 [110] 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 photographs 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 [14], they developed and evaluated a model to predict the average scenicness of 5km × 5km grid cells. They used text describing the rated images in the ScenicOrNot website as input to train a 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 [110]. Some authors have assessed the beauty of the landscape through groups of evaluators using images, videos and/or questionnaires [110, 88], while others used geographic information system (GIS) data such as elevation together with visual assessments and/or questionnaires to predict the scenicness [6, 100]. Another group of works, such as [12], [40], and [114], quantify landscape aesthetics according to the number of photographs taken near a given location [12] or the number of people who published photographs [40] in photo-sharing sites 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 biased towards more accessible places (one of our experiments reported in Section 4.3.1 provides evidence to that effect).

Another growing area of interest is in the use of social media data for ecological monitoring. An overview of the potential for exploiting social media in conservation and biodiversity was provided by [24], 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. In [5], they examined Flickr biodiversity data quality by analysing its metadata and comparing it with ground-truth data, using Snowy owls

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

and Monarch butterflies as a case study. They concluded that Flickr data has the potential to add to the 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, [20] performed a manual evaluation of a sample of Twitter postings that named three invasive species (using associated images for validation). They identified factors correlated with valid observations, such as the presence of a linked photograph and tags that describe the environment (e.g. 'leaves' and 'tree'). They 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. An approach to validating individual observations in Flickr was described by [30] who used Google's reverse image-search service to find photographs similar to those in Flickr postings. The tags of the Google photographs were then compared with those in Flickr in an attempt to filter out non-wildlife images. In [92] the content of the Flickr photographs was analysed manually to assess the quality of cultural ecosystem services and derive useful information to manage Singapore's mangroves. The research presented in this thesis is different from these works, where we do not focus on the content of a particular photograph (e.g. which species it may show). Instead, we focus on exploiting and utilising the tags associated with Flickr photographs for predicting a wide range of ecological phenomena such as species distribution, scenicness of a place, soil type, land cover type, and climate features.

2.3 Text Representation

Text representation is one of the most fundamental problems in text mining. It aims to numerically represent unstructured text documents to make them mathematically computable. In the past decades, various strategies for text representation have been proposed for different application problems such as text classification, clustering, and information retrieval. The problem of text representation is to represent each text document as a vector, such that the distance between document vectors is representative of their intuitive degree of dissimilarity.

A popular and simple method for representing text is called the bag-of-words (BOW) model, which is commonly used in natural language processing, machine learning, and information retrieval. This model represents a document by encoding the number of times each word appears in it, thus disregarding any information related to word order. Formally, we will treat bag-of-words representations as vectors, where each coordinate captures the weight of a given word. Therefore, next in Sections 2.3.1 and 2.3.2 we will discuss some of the widely used methods for term weighting and term selection in the context of BOW representations.

2.3.1 Term Weighting

Term weighting is the process of assigning numerical values to terms which represent their importance in a document [98]. It is a crucial component of any machine learning system, which has shown great potential for improving the effectiveness of the system [97]. For a given text document d, where we write $w(t, d)$ to encode the weight of term t in d. The following statistical methods are the most common examples of term weighting:

- Term Presence (also known as Boolean, One-hot or Binary vector) is a binary weight, taking the value of 1 or 0 based on the term's presence or absence in the text. If text document d does not contain term t then weight $w(t, d) = 0$, otherwise the term is assigned value 1.
- Term Count, the weight of a term is the count of its occurrences in the text. If a term t occurs five times in the text document d then weight $w(t, d) = 5$.
- Term Frequency (TF), similar to the 'Term Count' method, but taking into account the length of the document. Because every document is different in length,

it is possible that a term would appear more times in long documents than in shorter ones. Thus, the term frequency is often divided by the document length (i.e. the total number of terms present in the document) as a way of normalisation. The weight $w(t,d)$ is computed as:

$$\textit{TF}(t,d) = \frac{c(t,d)}{\sum_{t' \in d} c(t',d)} \tag{2.1}$$

where $c(t,d)$ is the number of times term t appears in a document d and $c(t',d)$ is the total number of terms in d.

- Term Frequency-Inverse Document Frequency (TF-IDF) is a more sophisticated measure which reflects how important a term t is to document d given how often t occurs in d and how frequently it appears in the entire document collection. The TF-IDF weight is computed by:

$$\textit{TF-IDF}(t,d) = \textit{TF}(t,d) \cdot \textit{IDF}(t,d) = \frac{c(t,d)}{\sum_{t' \in d} c(t',d)} \cdot \log \frac{n}{n_t} \tag{2.2}$$

  with n the number of the documents in the document collection and n_t the number of documents containing the term t.

- Pointwise Mutual Information (PMI) is a measure of association between the term and the text. It is similar to TF-IDF, essentially comparing the actual number of occurrences with the expected number of occurrences given how many terms occur in document d and how common the term t is:

$$\textit{PMI}(t,d) = \log \frac{P(t,d)}{P(t)P(d)} \tag{2.3}$$

  where:

$$\begin{aligned} P(t,d) &= \frac{c(t,d)}{m} \\ P(t) &= \frac{\sum_{d' \in D} c(t,d')}{m} \\ P(d) &= \frac{\sum_{t' \in T} c(t',d)}{m} \\ m &= \sum_{t' \in T} \sum_{d' \in d} c(t',d') \end{aligned}$$

with T the set of all terms that appear in the document collection. It is possible for the PMI of a term within a document to be negative. In that case, we can change the weight of this term to zero which is called Positive Pointwise Mutual Information (PPMI). It is then given by:

$$\textit{PPMI}(t,d) = \max\left(0, \log\left(\frac{P(t,d)}{P(t)P(d)}\right)\right) \tag{2.4}$$

In this present research, we weight the tag occurrences based on a variant of Positive Pointwise Mutual Information (PPMI) to associate more significance to tags that are less common and more closely correlated with a particular geographic location. This method will be explained in Section 4.2.1.

2.3.2 Term Selection

Term selection is the process of automatically selecting the terms that are most relevant (i.e. most useful) to the considered task. It is also called feature selection, variable selection, or attribute selection. The term selection process is an important component for most text mining tasks. It mostly acts as a filter for cleaning out the irrelevant or partially relevant features that can negatively impact model performance. The main objectives of term selection are [46]: (i) improving the prediction performance by providing the most relevant features, (ii) enabling faster training by minimising the number of features, and (iii) making it easier to interpret and understand the nature of the data. As term selection methods seek to reduce the number of features in the dataset, it can be used as a special case of dimensionality reduction [80]. Whereas term selection methods include and exclude terms present in the data without changing them, other dimensionality reduction approaches (such the low dimensional vector space embeddings methods that will be explained in Section 2.4) create a new combination of features.

Term selection methods often apply a statistical measure to score each term or feature. The features are ranked by their score and the most relevant features are those with the highest value. Below are some examples of widely used methods for term selection and scoring:

- Chi-Squared (χ^2) is a measure for modelling the dependency between the terms (or features) and the classes (such as gender, political preference, or land cover type). For each class $c \in C$ and each term t occurring in a document belonging to class c, the χ^2 statistic ranks terms with respect to the following quantity:

$$\chi^2(t,c) = \frac{(O_{tc} - E_{tc})^2}{E_{tc}} + \frac{(O_{t\bar{c}} - E_{t\bar{c}})^2}{E_{t\bar{c}}} + \frac{(O_{\bar{t}c} - E_{\bar{t}c})^2}{E_{\bar{t}c}} + \frac{(O_{\bar{t}\bar{c}} - E_{\bar{t}\bar{c}})^2}{E_{\bar{t}\bar{c}}} \quad (2.5)$$

where O_{tc} is the number of documents in class c where term t occurs, $O_{t\bar{c}}$ is the number of documents outside class c where term t occurs, $O_{\bar{t}c}$ is the number of documents in class c where term t does not occur, and $O_{\bar{t}\bar{c}}$ is the number of documents outside class c where term t does not occur. Moreover, E_{tc} is the expected number of occurrences of term t in documents belonging to class c, and similar for $E_{t\bar{c}}$, $E_{\bar{t}c}$, and $E_{\bar{t}\bar{c}}$ which can be computed as:

$$\begin{aligned}
E_{tc} &= N \cdot P(t) \cdot P(c) \\
E_{t\bar{c}} &= N \cdot P(t) \cdot (1 - P(c)) \\
E_{\bar{t}c} &= N \cdot (1 - P(t)) \cdot P(c) \\
E_{\bar{t}\bar{c}} &= N \cdot (1 - P(t)) \cdot (1 - P(c))
\end{aligned}$$

where N here is the total number of documents, P(t) is the probability that a document contains term t, and P(c) is the probability that a document belongs to class c. These probabilities can be estimated by:

$$P(t) = \frac{\sum_{c \in C} O_{tc}}{\sum_{t \in T} \sum_{c \in C} O_{tc}} \quad (2.6)$$

$$P(c) = \frac{|c|}{N} \quad (2.7)$$

- Correlation Coefficient (CC) is a variant of χ^2 test where $CC^2 = \chi^2$. The Correlation Coefficient has been defined in [82] as:

$$CC(t,c) = \frac{\sqrt{N} \cdot (O_{tc} \cdot O_{\bar{t}\bar{c}} - O_{t\bar{c}} \cdot O_{\bar{t}c})}{\sqrt{(O_{tc} + O_{t\bar{c}}) \cdot (O_{tc} + O_{\bar{t}c}) \cdot (O_{\bar{t}c} + O_{\bar{t}\bar{c}}) \cdot (O_{t\bar{c}} + O_{\bar{t}\bar{c}})}} \tag{2.8}$$

with O_{tc}, $O_{t\bar{c}}$, $O_{\bar{t}c}$, $O_{\bar{t}\bar{c}}$, and N as defined in χ^2. CC can be seen as a 'one-sided' version of χ^2. The correlation coefficient (CC) selects those terms that are highly related to a class, while the χ^2 may also pick out terms that are indicative of non-membership in the class.

- Log-Likelihood is an alternative to χ^2. For each term t and class $c \in C$, the log likelihood statistic is given by:

$$\begin{aligned} LL(t,c) = 2(&O_{tc} \log O_{tc} + O_{t\bar{c}} \log O_{t\bar{c}} + O_{\bar{t}c} \log O_{\bar{t}c} + O_{\bar{t}\bar{c}} \log O_{\bar{t}\bar{c}} + N \log N \\ &- (O_{tc} + O_{t\bar{c}}) \log(O_{tc} + O_{t\bar{c}}) - (O_{tc} + O_{\bar{t}c}) \log(O_{tc} + O_{\bar{t}c}) \\ &- (O_{t\bar{c}} + O_{\bar{t}\bar{c}}) \log(O_{t\bar{c}} + O_{\bar{t}\bar{c}}) - (O_{\bar{t}c} + O_{\bar{t}\bar{c}}) \log(O_{\bar{t}c} + O_{\bar{t}\bar{c}})) \end{aligned} \tag{2.9}$$

where O_{tc}, $O_{t\bar{c}}$, $O_{\bar{t}c}$, $O_{\bar{t}\bar{c}}$, and N are as defined above.

- Kullback-Leibler (KL) Divergence is a measure of how much the probability distribution of term t across documents from a given class differs from the reference probability distribution. It is given by:

$$KL(t) = \sum_{c \in C} P(c|t) \log \frac{P(c|t)}{P(c)} \tag{2.10}$$

where $P(c)$ is the probability of class c and $P(c|t)$ is is the probability of term t in class c which is estimated by:

$$P(c|t) = \frac{O_{tc}}{\sum_{c \in C} O_{tc}} \tag{2.11}$$

Note that Kullback-Leibler divergence immediately produces a single ranking for the term over all the classes, in contrast to the χ^2, Correlation Coefficient and log-likelihood, which provide a ranking per class.

2.4 Low Dimensional Vector Space Representation

A bag-of-words (BOW) representation can be seen as a very high-dimensional vector representation. An embedding is a mapping from such a high-dimensional vector representation into a relatively low-dimensional representation (e.g. 300 dimensions). Unlike in BOW representations, the individual dimensions in the vector space embedding typically have no specific meaning. They represent the overall patterns of the distance between objects by placing semantically similar objects close together in the embedding space. In this thesis, we develop two novel models for learning low-dimensional vector space embeddings by integrating the textual information derived from Flickr with the numerical and categorical information derived from structured scientific datasets. The first model, named EGEL model, generates the embedding for representing the geographic locations which will be presented in Chapter 6; The Second model, named SPATE model, generates the embedding for representing the spatiotemporal regions which will be presented in Chapter 7.

2.4.1 Vector Space Embeddings

The use of low-dimensional vector space embeddings for representing objects has already proven effective in a large number of applications, including natural language processing (NLP), image processing, and pattern recognition. In the context of NLP, the most prominent example is that of word embeddings, which represent word meaning using vectors of typically around 300 dimensions. A large number of methods for learning such word embeddings have already been proposed, including Skip-gram and the Continuous Bag-of-Words (CBOW) model [77], and GloVe [86]. They have been applied effectively in many downstream NLP tasks such as sentiment analysis [112], part of speech tagging [89, 72], and text classification [70, 38]. The model we consider in this thesis builds on GloVe, which was designed to capture linear regularities of word-word co-occurrence. In GloVe, there are two word vectors w_i and $\tilde{w}_j$ for each word in the vocabulary, which are learned by minimizing the following objective function:

$$J = \sum_{i,j=1}^{V} f(x_{ij})(w_i.\tilde{w}_j + b_i + \tilde{b}_j - \log x_{ij})^2 \tag{2.12}$$

where x_{ij} is the number of times that word i appears in the context of word j, V is the vocabulary size, b_i is the target word bias, and $\tilde{b}_j$ is the context word bias. The weighting function f is used to limit the impact of rare terms. It is defined as 1 if $x > x_{max}$ and as $(\frac{x}{x_{max}})^{\alpha}$ otherwise, where x_{max} is usually fixed to 100 and α to 0.75. Intuitively, the target word vectors w_i correspond to the actual word representations which we would like to find, while the context word vectors $\tilde{w_j}$ model how occurrences of j in the context of a given word i affect the representation of this latter word. In this thesis, we will use a similar model, specifically in Chapters 6 and 7, which will be aimed at learning spatial or spatiotemporal region vectors instead of the target word vectors.

Beyond word embeddings, various methods have been proposed for learning vector space representations from structured data such as knowledge graphs [8, 126, 115], social networks [43, 121] and taxonomies [119, 83]. The idea of combining a word embedding model with structured information has also been explored by several authors, for example, to improve the word embeddings based on information coming from knowledge graphs [124, 109]. Along similar lines, various lexicons have been used to obtain word embeddings that are better suited at modelling sentiment [112] and antonymy [84], among others. The method proposed by [71] imposes the condition that words that belong to the same semantic category are closer together than words from different categories, which is somewhat similar in spirit to how we will model categorical datasets in our embedding models.

2.4.2 Embedding Spatial or Spatiotemporal Information

The problem of representing geographic locations using embeddings has also attracted some attention. An early example is [95], which used principal component analysis and stacked autoencoders to learn low-dimensional vector representations of city neighbourhoods based on census data. They use these representations to predict attributes such as crime, which is not included in the given census data, and find that in most of the considered evaluation tasks, the low-dimensional vector representations lead to more faithful predictions than the original high-dimensional census data.

Some existing works combine word embedding models with geographic coordinates. For example, in [16] an approach is proposed to learn word embeddings based on the assumption that words which tend to be used in the same geographic locations are likely to be similar. Note that their aim is dual to our aim in this thesis: while they use geographic location to learn word vectors, we use textual descriptions to learn vectors representing geographic locations or spatiotemporal regions.

Several methods also use word embedding models to learn representations of Points-of-Interest (POIs) that can be used for predicting user visits [33, 73, 132]. These works use the machinery of existing word embedding models to learn POI representations, intuitively by letting sequences of POI visits by a user play the role of sequences of words in a sentence. In other words, despite the use of word embedding models, many of these approaches do not actually consider any textual information. For example, in [73] the Skip-gram model is utilised to create a global pattern of users' POIs. Each location was treated as a word and the other locations visited before or after were treated as context words. They then use a pair-wise ranking loss [123] which takes into account the user's location visit frequency to personalise the location recommendations. The methods of [73] were extended in [132] to use a temporal embedding and to take more account of geographic context, in particular, the distances between preferred and non-preferred neighbouring POIs, to create a "geographically hierarchical pairwise preference ranking model". Similarly, [127] developed a method for modelling places, neighbourhoods, and users from social media check-ins. They treat the check-ins as sentences to generate the embeddings which encode the geographical, temporal, and functional aspects. In [128], the CBOW model was trained with POI data. They ordered POIs spatially within the traffic-based zones of urban areas. The ordering was used to generate characteristic vectors of POI types. Zone vectors, represented by averaging the vectors of the POIs contained in them, were then used as features to predict land use types. The authors of [125] proposed a method that uses the Skip-gram model to represent POI types, based on the intuition that the vector representing a given POI type should be predictive of the POI types found in nearby places of that type. In the CrossMap method, [129] learned unsupervised embeddings for spatiotemporal hotspots obtained from social media data of locations, times and text. In one form of embedding, intended to enable reconstruction of records, neighbourhood relations in space and time were encoded by averaging hotspots in a target location's spatial and temporal neighbourhoods. They also proposed a graph-based embedding method with different nodes for modelling location, time and text. The concatenation of the location, time and text vectors were then used to predict peoples' activities in urban environments. In another work, [130] proposed the ReAct model, which is similar to CrossMap. However, while the CrossMap model is unsupervised and handles static data, ReAct is a semi-supervised model and handles continuous online data to learn the activity models.

In NLP research, embedding methods have been used to measure the language variation across geographical regions as well as over time [3, 60, 65, 87, 49]. For instance, [3] and [65] present methods to learn geographically situated word embeddings from geo-tagged tweets. They used cosine similarities between the generated embeddings to measure the spatial variation of the language across English speaking countries. The authors of [49] used the Doc2Vec method [67] to learn document embeddings from online posts in German-speaking regions. These embeddings have been used to study language variation in German. To study the temporal variation of language, [60], among others, trained the Skip-gram model on text from the Google Books corpus for the period from 1900 to 2009. They also used cosine similarity to measure the change in word meaning between the embeddings of the same words learned in different time periods. In [87], they used the CBOW model to learn spatiotemporal embeddings from geo-tagged tweets. They first split the data into 8-hour windows (i.e. the temporal granularity) for each separate country (i.e. the spatial granularity). For each time window, they then trained a joint embedding using tweets from all countries and used it to initialise the country-specific embeddings.

Despite the considerable progress that has been made on embedding social media data, the problem of embedding Flickr tags has so far received very limited attention. To the best of our knowledge, [48] is the only work that generated embeddings for Flickr tags. However, their focus was on learning embeddings that capture word meaning, which has been evaluated on word similarity tasks.

Our work in Chapters 6 and 7 is different from all these studies, as our focus is on spatial or spatiotemporal embeddings based on text descriptions (in the form of Flickr tags), along with numerical and categorical features from environmental datasets.

2.5 Machine Learning

Machine Learning (ML) is a set of general algorithms which have the ability to learn and infer based on available data [62]. Machine learning algorithms usually build a mathematical model based on a data sample, known as “training data”, in order to make predictions or decisions. They have been used in a wide variety of applications, such as natural language processing, sales and marketing, computer vision, and many others. Machine learning approaches are often categorised into supervised and unsupervised learning. Supervised learning algorithms can apply what has been learned from labelled examples to predict labels for new data. The most studied supervised learning task is classification. Unsupervised learning is usually used when the input data do not have labels. The goal of unsupervised learning is to model the underlying structure or distribution in the data to learn more about the data. The most widely studied unsupervised learning task is clustering. Our aim in this thesis is to use supervised machine learning to learn a set of models where the supervision labels will correspond to environmental features. Therefore, we will focus on supervised machine learning in the following section.

2.5.1 Supervised Machine Learning

In supervised machine learning, the algorithm builds a mathematical model from a set of data that contains both the inputs (training data) and the desired outputs (labels) [94]. It starts by analysing the labelled training dataset and then produces an inferred function to make predictions about the unlabelled examples (testing data). The learning algorithm can also compare its prediction with the correct output and find errors to modify the model accordingly. An optimal model will allow the algorithm to correctly determine the output for entries that were not a part of the training data [79].

Most supervised learning problems belong to two broad categories: regression and classification. Generally, classification aims to assign examples to predefined classes, i.e. the output variable is a category. For instance, for the problem of filtering spam emails, the output would be the prediction of either “spam” or “not spam”. However, in regression, the goal is to predict a continuous measurement for an observation, i.e. the output variable is a real value. Examples of a continuous value are the temperature, length, or weight.

Some popular examples of supervised machine learning algorithms are:

- **Support Vector Machines (SVMs)** are discriminative learning models which are based on maximizing the margins between the examples and a separation hyperplane. They are known for their strong performance in many applications including the classification of text [85, 103], images [18], and genes and proteins

[101, 68]. A similar approach to SVMs can also be used for regression problems, which is known as ‘Support Vector Regression (SVR)’.

SVMs represent the training examples as vectors, such that examples from different classes are intuitively separated by a gap which is as wide as possible. In particular, each training example is represented by (x, y), where x indicates the set of attributes and y indicates the class label (i.e. either 1 or -1 for binary classification). The separating hyperplane of linear SVM can be written as:

$$w \cdot x - b = 0 \tag{2.13}$$

where $w = (w_1, w_2, ..., w_n)$ is a weight vector (support vector) and b is a bias. The geometric intuition of the binary classification is illustrated in Figure 2.1. The testing examples are then mapped into that same space, and the label predicted based on the side of the hyperplane in which they fall.

SVMs have also been shown effective in non-linear classification using the kernel function [47]. To this end, the dot product is replaced by a kernel function. The most commonly used kernels with SVMs are (Gaussian) radial basis functions (RBF) and polynomial kernels.

- **Artificial Neural Networks (ANNs)** are inspired by how the human brain works. ANNs are based on a collection of connected nodes called artificial neurons, which loosely model the neurons in the brain. Each connection, like the synapses in a brain, can transmit a signal between neurons. An artificial neuron that receives a signal can process it and send it to other artificial neurons to which it is connected. The general architecture is illustrated in Figure 2.2. The main Neural Networks architectures used for classification are Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). They treat the text as a sequence of words, and thus implicitly assume that at least some aspects of word order are important. In particular, RNNs analyse a text word by word and store a representation of the already processed text as a fixed-dimensional vector in a hidden layer [66]. However, this is not useful for the current research where

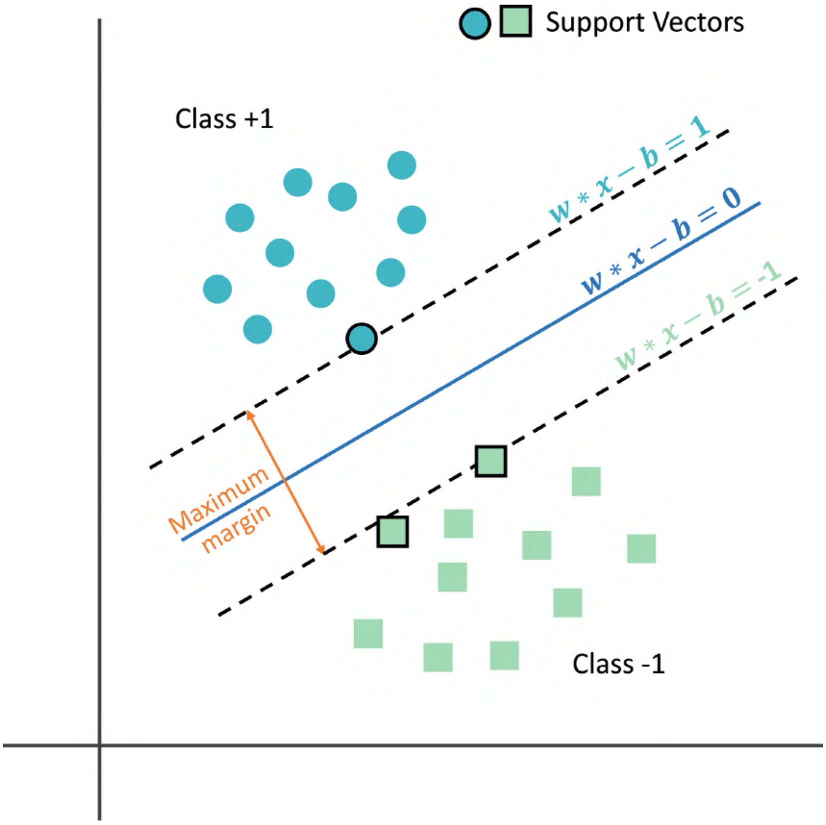


Figure 2.1: The geometric intuition of linear SVMs.

we consider Flickr tags as a source of textual data, which is a set of unrelated words. The major limitations of using Neural Networks methods are: (i) they can be computationally expensive, (ii) they tend to be difficult to configure and require careful fine-tuning of hyper-parameters, and (iii) the learned models are often difficult to interpret [99, 78].

In this thesis, we will use Support Vector Machines in all the experiment. In particular, we will learn Support Vector Machines (SVMs) for classification problems and Support Vector Regression (SVR) for regression problems. In both cases, we will use the SVM^{light} implementation[4] [57]. In addition to SVMs, we will also use methods for learning vector space embeddings of geographic locations (possibly time-dependent)

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

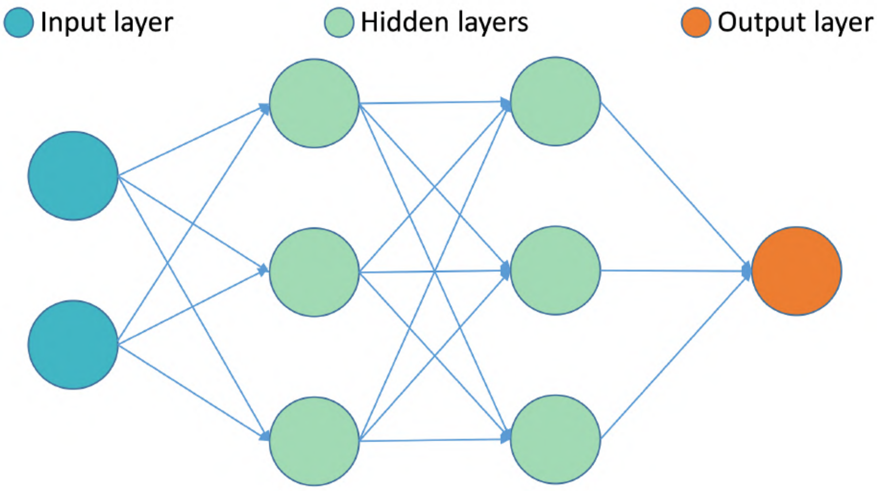


Figure 2.2: Artificial Neural Networks architecture.

which are inspired by the Artificial Neural Networks model.

2.5.2 Collective Prediction

Collective prediction plays an important role for the research presented in this thesis, particularly for the method we introduce in Chapter 5. 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 this network structure in the learning process, by exploiting information about the entities that are related to the considered one. A standard example is the problem of web page categorisation [13, 2, 26]: 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 research, we will use the Iterative Classification Algorithm (ICA) from [81], which is conceptually simple but often highly effective. Other approaches are based on inference in joint probabilistic models using Gibbs sampling [39]. However, Gibbs sampling tends to be slow [104], which is an important limitation in our setting, as we will need to make predictions about hundreds of thousands of regions.

The authors of [13] experimentally demonstrated the effectiveness of taking into account the link structure for web page categorisation. 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, in [2], they 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 [26], 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 [63] 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 [56], 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 utilises the node attributes and an iterative classification algorithm which uses 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. In [15], they 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 [74] presented an algorithm called CORENA (COllective REgression in Network dAta) which studies the transduction of collective regression in a sparsely labelled network. In particular, they iteratively augmented the descriptive and the target information of the labelled 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 feedback the reliably predicted labels only. They show that their proposed method is able to improve regression performance in the areas of social and spatial networks.

In Chapter 5 of this thesis, we focus on both collective classification and regression problems by applying 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.

2.5.3 Evaluation Measures

Classification Problems

Evaluating the performance of a classifier is usually done by measuring its effectiveness rather than its efficiency, i.e. the classifier's ability to predict the correct category, and not its computational complexity [103]. Generally, the evaluation measures in classification problems are defined from a matrix with the numbers of examples correctly and incorrectly classified for each class, called a “confusion matrix”. The confusion matrix for a binary classification problem (which has only two classes - positive and negative), is shown in Table 2.1.

Table 2.1: Confusion matrix for binary classification.

	Predicted Class	
Actual Class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

TP (true positives) is the number of examples that are correctly predicted as belonging to the positive class; TN (true negatives) is the number of examples that are correctly predicted as belonging to the negative class; FP (false positives) is the number of examples that are classified as positive while they are from the negative class; FN (false negatives) is the number of examples that are classified as negative, but their true class is positive. Common metrics for evaluating classification tasks include accuracy, precision, recall and F1-score, which are defined as:

$$Accuracy = \frac{(TP + TN)}{TP + FP + TN + FN} \tag{2.14}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{2.15}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{2.16}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{(Precision + Recall)} \tag{2.17}$$

Accuracy is the proportion of the correctly classified examples (i.e. true positive and true negative examples); precision measures the proportion of false positives; recall measures the proportion of false negatives; the F1-score is the harmonic means of precision and recall. Due to the often highly imbalanced number of positive vs. negative examples in binary classification, as negative class usually dominates the accuracy of a model, leading to miss-interpretation of the results. For example, when the positive examples of a category represent only 1% of the test set, a trivial classifier that makes negative predictions for all examples has an accuracy of 99%. However, such a system is useless. For this reason, precision, recall and F1-score are more commonly used instead of accuracy for evaluating unbalanced classification problems.

Regression Problems

Validation and evaluation of regression models is usually done by measuring the prediction error, i.e. the difference between the actual and the predicted scores. Mean Absolute Error (MAE) is one of the most widely used metrics for this purpose. It is calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a_i - p_i| \tag{2.18}$$

where a_i is the actual value for example i, p_i is the predicted value, and n is the total number of examples. Note that as a measure of error the lower value is the better.

Another important metric for evaluating the performance of a regression model is by measuring the correlation between the actual and the predicted scores [134]. Correlation can be defined as a measure of the strength of association between two variables and the direction of the relationship. In terms of the strength of the relationship, the value of the correlation varies between +1 and -1. A value of +1 indicates a perfect degree of association between the two variables. When the correlation value declines towards 0, indicates that the relationship between the two variables is weaker. The negative value correlation value indicates a negative relationship. Spearman rho is one of the most widely used correlation functions. It is computed as follows:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(a_i - p_i)^2}{n^3 - n} \tag{2.19}$$

2.6 Summary

In this chapter, we have explored the background knowledge related to mining social media platforms. To gain a full understanding of the topic, we reviewed the existing work that uses social media data for geospatial analysis and citizen science. We also describe some of the commonly used methods to generate bag-of-words representations of text documents, including the term weighting and term selection methods. These methods will be used in the following chapters for modelling locations using Flickr tags. Moreover, we reviewed the recent techniques used for learning low-dimensional vector space representations, which we will be used as an alternative way to model locations. Furthermore, we have briefly discussed some of the most commonly used supervised machine learning methods as well as the standard evaluation methods. We paid particular attention to reviewing the existing work that uses the collective prediction approach.

With the information gained from this chapter, we can move forward in the next chapter to discuss the datasets that will be used in this thesis.

Chapter 3

Data Acquisition and Preprocessing

3.1 Introduction

This chapter gives an overview of the data used in this research and explains how the datasets were collected and extracted. In particular, we consider two different data sources: textual and structured scientific data. The textual data source comes from the tags associated with the photographs in the photo-sharing platform Flickr[1]. The structured data comes from more traditional data sources such as European Environment Agency[2]. This chapter also gives brief details of each of the considered ground truth datasets which have been obtained from the ScenicOrNot website[3], Natura 2000[4] project, and NBN Atlas[5].

Section 3.2 provides an overview of the photo-sharing platform Flickr and describes our methodology for collecting, preprocessing and analysing Flickr data. In Section 3.3 we introduce the considered structured scientific data. After that, Section 3.4 presents the sources used to extract the ground truth data. Finally, Section 3.5 provides a summary of the chapter.

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

[2] https://www.eea.europa.eu/

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

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

[5] https://nbnatlas.org/

3.2 Flickr Data

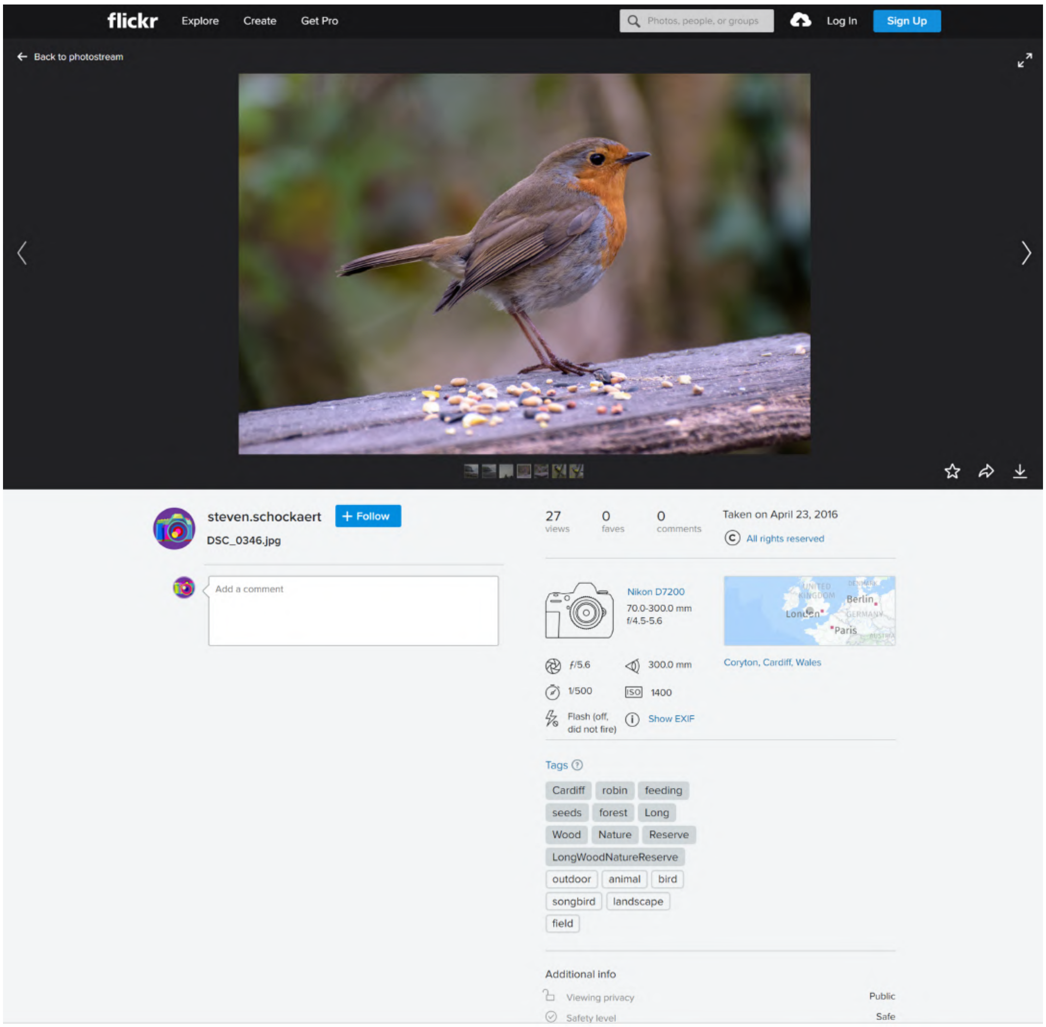
Flickr is a photo-sharing and hosting platform. It was launched by Ludicorp in 2004 and changed ownership several times[6]. Flickr at its peak hosted more than 10 billion photographs[7], many of which are associated with textual data in the form of a title, description and a set of up to 75 tags that express what is in the photographs and make it easily accessible by others. Flickr offers functionality for any users to record the timestamp and the GPS coordinates (latitude and longitude coordinates) as a meta-data attached with their photographs, which can be done manually or automatically by the camera or the smartphone. An example of a Flickr photo with the associated meta-data is shown in Figure 3.1. The benefit of adding this meta-data is to improve the search feature, as the user will be able to search for photographs taken at a specific location, time, or with a particular tag. For a large number of photographs on Flickr, these meta-data fields have been made publicly available. In the next section, we will explain our methodology for collecting those publicly available meta-data.

3.2.1 Flickr Data Collection

Some social media companies, including Flickr, make their data available through APIs. The API key (Application Programming Interface key) is a code needed as a parameter for the request, as a form of user identification to track and control the use of API. Two Types of API keys are available in Flickr: Non-Commercial and Commercial. A Non-Commercial key is suitable for the needs of this research, and usefully it does not require much validation to acquire. By providing details of the project to Flickr, an API key was given to be used for all API calls. To perform an action using the API a user needs to send a request to its endpoint specifying a method and arguments, and will then receive a formatted response in the form of the XML

[6] https://en.wikipedia.org/wiki/Flickr

[7] http://expandedramblings.com/index.php/flickr-stats



steven.schockaert
DSC_0346.jpg

27 views 0 faves 0 comments

Taken on April 23, 2016

All rights reserved

Nikon D7200
70.0-300.0 mm f/4.5-5.6

f/5.6
300.0 mm
1/500
1400
Flash (off, did not fire)

Coryton, Cardiff, Wales

Cardiff robin feeding seeds forest Long Wood Nature Reserve LongWoodNatureReserve outdoor animal bird songbird landscape field

Viewing privacy Public
Safety level Safe

Figure 3.1: Flickr photographs

file. Flickr offers many methods that can be called; the full extent can be found at `https://www.flickr.com/services/api/`, examples include *flickr.groups.browse*, *flickr.people.findByUsername*, and *flickr.photos.search* which is the method used for fetching the data. Each method has different optional and required arguments that can be applied. For the *flickr.photos.search* method, the *api_key* is the only required argument, while, for example, *user_id* , *tags*, *min_upload_date*, *max_upload_date*, and *has_geo* are optionally used to refine the search. The approach that we used to collect the meta-data associated with Flickr photographs includes the following steps:

1. Set API key using the *flickr_api.set_keys* method.

2. Use the *flickr.photos.search* method to make the request, setting the argument *has_geo* to 1 to only retrieve the georeferenced photographs. However, each API request cannot retrieve more than 4000 records. One of the ways to overcome this restriction, which we use in our implementation, is by adding time constraints (i.e. *min_upload_date* and *max_upload_date* arguments) to the request such that the crawler tries to determine time intervals in which the number of results for the query is less than 4000 but close to it. To this end, we set *max_upload_date* of the first request to the current time and *min_upload_date* by subtracting a predefined time interval.

3. Check the status of the request to determine whether the total number of photographs is acceptable (i.e. less than 4000 but close to it) and then download the detailed XML data if this is the case. Otherwise, increase or decrease the time interval to obtain a total number of photographs which is just under 4000 and re-request according to the new range.

4. Continue going back step by step until the given timestamp.

We were able to collect around 350 million georeferenced Flickr photographs worldwide, all of which were uploaded to Flickr from January 2004 to September 2015. We constructed the crawler in Python programming language; the source code is available at `https://github.com/shsabah84/Flickr_Crawler_Python`.

3.2.2 Flickr Data Preprocessing

After describing the methodology used for collecting all the publicly available georeferenced Flickr meta-data, we will now describe the methodology used for preprocessing the collected data. It includes the following steps:

1. The raw XML data are parsed to extract the required meta-data. In particular, the following set of meta-data is extracted: photo_id, owner, title, tags, accuracy, upload date, taken date, place_id, latitude, longitude, and description.

2. Photographs that do not contain any tags are removed from the collection.

3. For the experiments conducted in Chapter 7 only, which depend on spatiotemporal analysis, the photographs in which the difference between the upload date and taken date is more than six months are removed from the collection. The purpose of this step is to avoid, as much as possible, photographs with an incorrect timestamp.

3.2.3 Exploratory Data Analysis

Now we will analyse the dataset collected from Flickr in order to understand the nature of the data and preliminarily assess its potential toward predicting environmental features.

- The number of georeferenced photographs per month of each year in the collected data is calculated to evaluate the availability of the data. As shown in Figure 3.2, there is a sharp rise in Flickr's popularity, reaching a peak of over 5,250,000 new photographs in June 2013. However, as of 2013, there is also an evident decline in the number of photographs, which suggests that the popularity of the platform was reducing. Perhaps, this is related to the increased popularity of the photo-sharing platform, Instagram[8]. Nevertheless, there is still a valuable amount of data to carry on this research, with an average of almost 2,500,000 photographs per month. The plot demonstrates a trend that summer months yield larger numbers of captured photographs, while the smallest number of photographs is uploaded during winter months. The reason for this is

[8] https://www.statista.com/chart/9157/instagram-monthly-active-users/

presumed to be related to the weather conditions. As the public generates these data, it is likely that fewer photographs being taken if it is cold or raining.

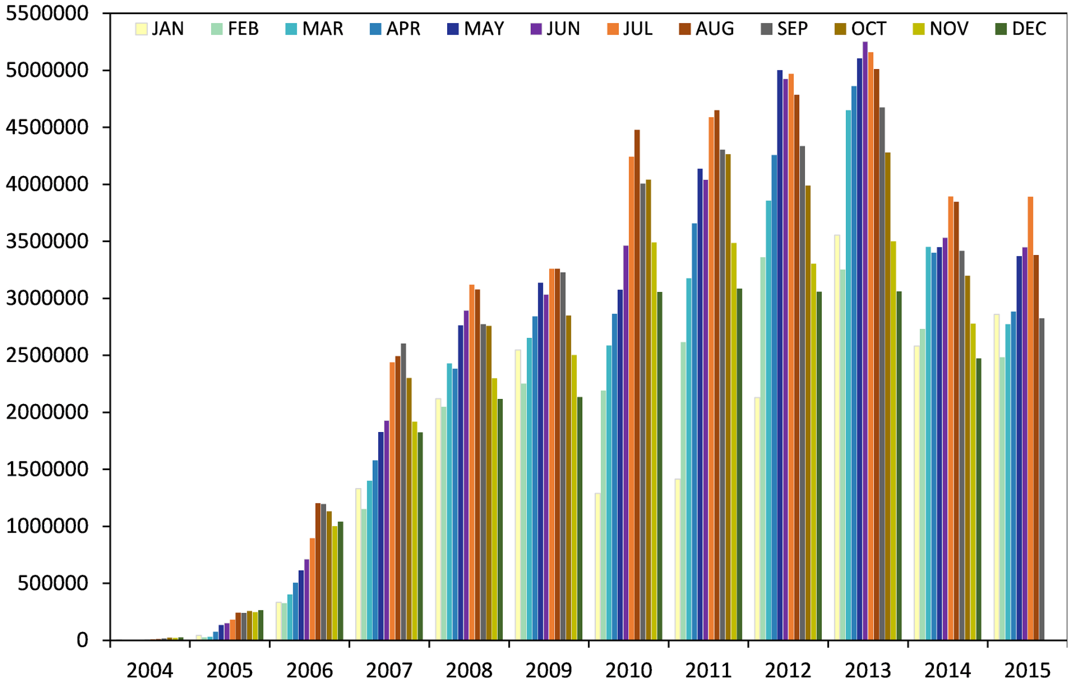


Figure 3.2: The number of georeferenced photographs in each month among the collected Flickr data.

- The frequency of scientific species names available from the European Bioinformatics Institute (EBI)[9] is calculated among the tags to evaluate the potential of the data toward modelling species distribution. Interestingly, there are more than 9500 scientific species names occurring in the collected data. Moreover, the scientific names of some types of species such as *Papilio Polyxenes* occur just under 16000 times, as can be seen from the overview in Figure 3.3. This suggests that it would be possible to extract useful information about species from Flickr meta-data, especially because only experts tend to use the scientific names to refer to specific species.

[9] https://www.ebi.ac.uk/

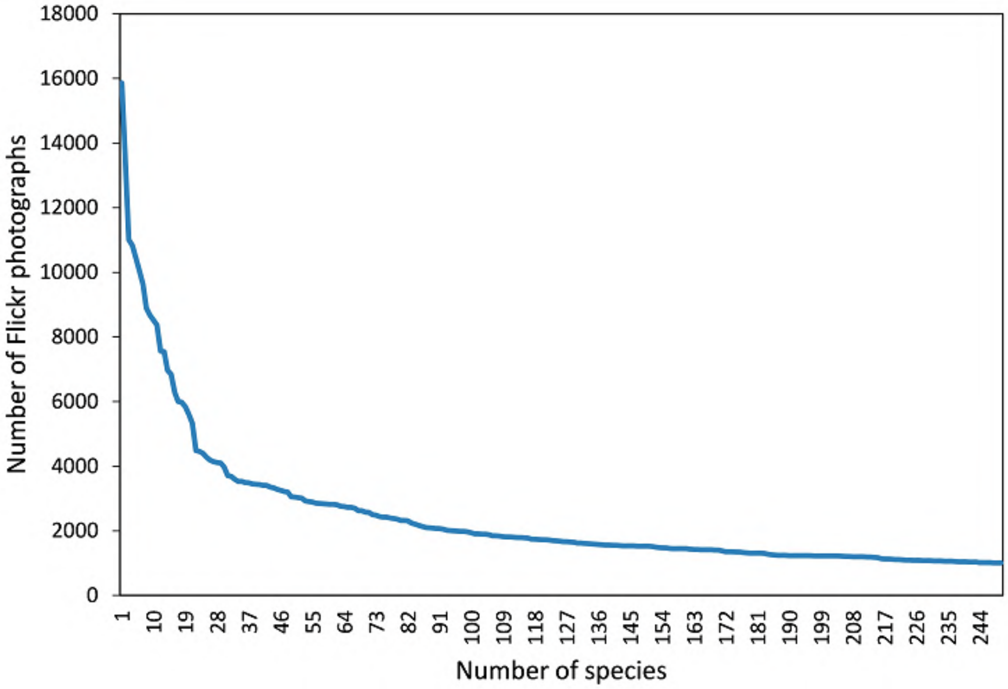


Figure 3.3: Scientific species names frequency among Flickr data.

- Kullback-Leibler (KL) divergence (Equation 2.10) between tag distributions is computed to identify tags whose occurrence is correlated with specific times of the year or with photographs of particular species. To this end, each month of the year or the species name is treated as a separate class. Samples of the top ten time related or species related tags resulting from KL divergence are shown in Table 3.1. Clearly, all the tags are informative and relevant to the considered task, which is another promising signal about the usefulness of Flickr tags for modelling the natural environment.

- The number of users and photographs in locations belonging to different CORINE land cover classes at level 3 is calculated to evaluate the availability of data in each category. The CORINE has 44 classes at the third and most detailed level (see Section 3.3 below for more information about CORINE), as shown in Figure 3.4, there is a large number of users and photographs for all these classes. This suggests that it might be possible to predict or refine the land cover type

Table 3.1: Top 10 tags with the highest KL divergence, in relation to the months of the year and the occurrence of species names.

Time-related Tags	Species-related Tags
december	pyrrhocoraxpyrrhocorax
january	casuariuscasuarius
february	alcedocristata
october	vireogriseus
july	salamandrasalamandra
christmastree	toxostomacurvirostre
thanksgiving	chordeilesminor
christmas	thryomanesbewickii
newyearseve	myiarchuscrinitus
november	nasuanasua

from Flickr data.

- Several term selection methods are applied to identify tags that occur in locations of a particular land cover type. In particular, tags that occur in photographs posted in forests (according to CORINE) are compared with all the other tags. The top ten tags resulting from the term selection methods explained in Section 2.3.2 are listed in Table 3.2. It can be clearly seen that using *Chi Squared*, *Correlation Coefficient*, and *Log Likelihood* gives similar sets of tags, which are generally related to the forest. On the other hand, using *KL divergence* gives a set of tags that describe a specific type of forest or a particular species.

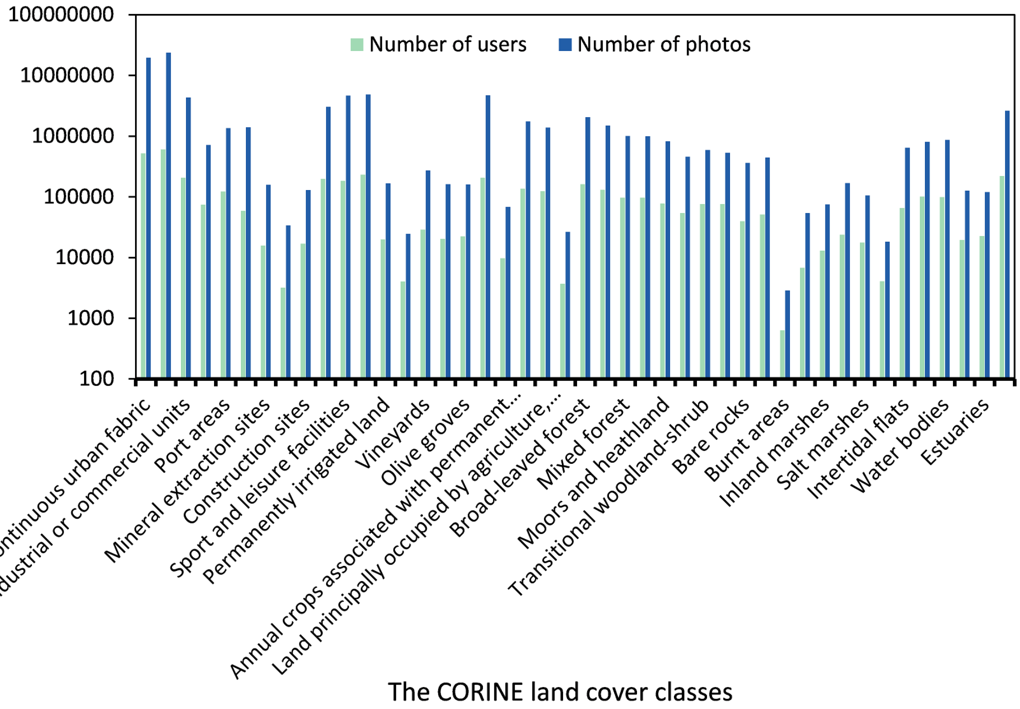


Figure 3.4: Number of photographs in each of the CORINE land cover class.

Table 3.2: Top 10 tags related to forest in different term selection methods.

Chi Squared	Correlation Coefficient	Log Likelihood	KL divergence
forest	forest	forest	beechforest
woods	woods	woods	spruce
wild	wild	trees	nadelwald
trees	trees	wild	primevalforest
mushroom	mushroom	nature	fagussylvatica
nature	nature	mushroom	silverwashedfritillary
wood	wood	wood	flyagaric
lake	lake	lake	amanitamuscaria
tree	tree	tree	lycoperdonperlatum
waterfall	waterfall	snow	apaturairis

3.3 Structured Scientific Data

There are a wide variety of structured scientific data that can be used for modelling the environment. In this section, we give an overview of the scientific datasets that we will use in the experiments in the remainder of this thesis.

- Land cover type, obtained from the CORINE Land Cover 2006 [10] dataset. CORINE Land Cover (CLC) is a European dataset which describes land cover with a 100-meter spatial resolution. It uses three levels of description: a top level with 5 classes, an intermediate level with 15 classes and a detailed level with 44 classes. A plot of the most detailed level is shown in Figure 3.5.

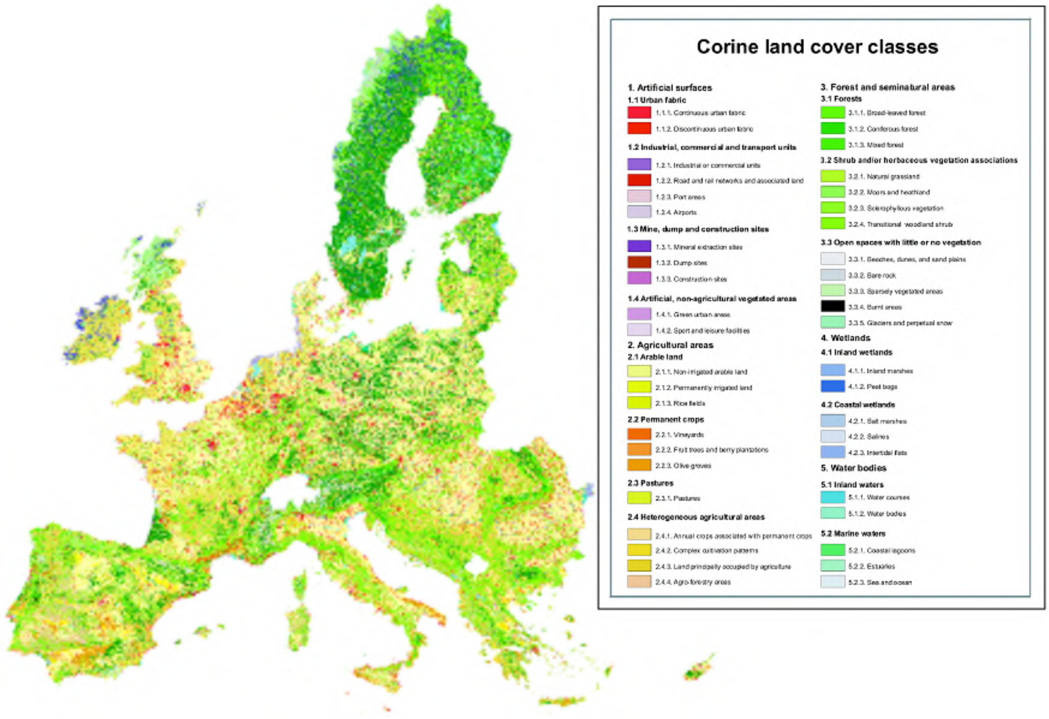

Corine land cover classes
1. Artificial surfaces
1.1 Urban fabric
1.1.1. Continuous urban fabric
1.1.2. Discontinuous urban fabric
1.2 Industrial, commercial and transport units
1.2.1. Industrial or commercial units
1.2.2. Road and rail networks and associated land
1.2.3. Port areas
1.2.4. Airports
1.3 Mine, dump and construction sites
1.3.1. Mineral extraction sites
1.3.2. Dump sites
1.3.3. Construction sites
1.4 Artificial, non-agricultural vegetated areas
1.4.1. Green urban areas
1.4.2. Sport and leisure facilities
2. Agricultural areas
2.1 Arable land
2.1.1. Non-irrigated arable land
2.1.2. Permanently irrigated land
2.1.3. Rice fields
2.2 Permanent crops
2.2.1. Vineyards
2.2.2. Fruit trees and berry plantations
2.2.3. Olive groves
2.3 Pastures
2.3.1. Pastures
2.4 Heterogeneous agricultural areas
2.4.1. Annual crops associated with permanent crops
2.4.2. Complex cultivation patterns
2.4.3. Land principally occupied by agriculture
2.4.4. Agro-forestry areas
3. Forest and seminatural areas
3.1 Forests
3.1.1. Broad-leaved forest
3.1.2. Coniferous forest
3.1.3. Mixed forest
3.2 Shrub and/or herbaceous vegetation associations
3.2.1. Natural grassland
3.2.2. Moors and heathland
3.2.3. Sclerophyllous vegetation
3.2.4. Transitional woodland shrub
3.3 Open spaces with little or no vegetation
3.3.1. Beaches, dunes, and sand plains
3.3.2. Bare rock
3.3.3. Sparsely vegetated areas
3.3.4. Burnt areas
3.3.5. Glaciers and perpetual snow
4. Wetlands
4.1 Inland wetlands
4.1.1. Inland marshes
4.1.2. Peat bogs
4.2 Coastal wetlands
4.2.1. Salt marshes
4.2.2. Salines
4.2.3. Intertidal flats
5. Water bodies
5.1 Inland waters
5.1.1. Water courses
5.1.2. Water bodies
5.2 Marine waters
5.2.1. Coastal lagoons
5.2.2. Estuaries
5.2.3. Sea and ocean

Figure 3.5: CORINE Land Cover dataset.

[10] http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-2

- Soil type, obtained from SoilGrids[11]. SoilGrids is a global raster dataset, which classifies locations into 116 types of soil, using a 250-meter spatial resolution.
- Elevation, obtained from the Digital Elevation Model over Europe (EU-DEM)[12]. EU-DEM is a Europe-wide digital surface model, encoding elevation with a spatial resolution of about 30 meters.
- Population, obtained from the European Population Map 2006[13], which is a digital raster grid that reports the number of residents (night-time population) with a 100-meter spatial resolution.
- Temperature, precipitation, solar radiation, wind speed and water vapour pressure, all of which are obtained from WorldClim[14]. The WorldClim dataset covers the monthly average values over the period from 1970 to 2000, using a 1 km spatial resolution.

All these datasets were extracted using the QGIS application by the following steps:

1. Open the raster map in the QGIS.
2. Add the locations coordinates (i.e. latitude and longitude coordinates) as a "Delimited Text Layer" to the map.
3. Use "Point Sampling Tools" to extract the correspondence value feature for each location.

[11] https://www.soilgrids.org

[12] http://www.eea.europa.eu/data-and-maps/data/eu-dem

[13] http://data.europa.eu/89h/jrc-luisa-europopmap06

[14] http://worldclim.org

3.4 Ground Truth Data

This section introduces the ground truth data that will be used in the experiments of this thesis.

3.4.1 The ScenicOrNot project

The ScenicOrNot project was initiated in 2009 by MySociety[15] and is currently hosted by the Data Science Lab at Warwick Business School[16]. It is based on an online game that allows people to evaluate places in Great Britain by rating photographs collected from the Geograph[17] photo-sharing website. The goal of the project is to crowdsource aesthetic judgements that can be used to study the impact of scenicness on human well-being in Britain. The dataset is available to use under the "Open Database Licence", so it can be downloaded directly as a TSV file. It contains ratings for 217,000 photographs 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). In particular, each record contains: ID, latitude, longitude, average rating, population variance, the votes (comma separated), and the Geograph URL for the image.

3.4.2 Natura2000 project

The European network of nature protected sites, Natura 2000[18], is an ecological network of protected areas in the European Union. It is a vital instrument to protect biodiversity, set up to ensure the survival of Europe's most valuable species and habitats. The Natura 2000 dataset consists of data submitted by national authorities with

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

[16] scenicornot.datasciencelab.co.uk

[17] http://www.geograph.org.uk/

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

an extensive description of the site and its ecology. This dataset contains information about 35,600 rare 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. The dataset is available in the Microsoft Access format and permitted to use for commercial or non-commercial purposes free of charge.

3.4.3 NBN Atlas

The National Biodiversity Network Atlas (NBN Atlas)[19] is a collaborative project committed to making biodiversity information available via the NBN Atlas. The National Biodiversity Network (NBN) registered as a charity to support the sharing of ecological data in the UK since 2000. The goal of the project is to improve the availability of high-quality species occurrence data in the UK. It is the largest collection of biodiversity information within the UK and Ireland and has revolutionised the use of biodiversity data by allowing it to be shared, downloaded, analysed, and researched by the public. NBN Atlas holds more than 220 million species occurrence records combined from individual observations and official organisations such as the "Royal Society for the Protection of Birds (RSPB)". Each species has a separate observations file in the form of a comma separated values (CSV) file. Each occurrence record in the file contains a set of meta-data including the observation's geographic location and time.

3.5 Summary

In this chapter, we introduced the Flickr data, structured scientific data, as well as the considered sources of the ground truth data that will be used in the following chapters. We also described our methodology for collecting and extracting those datasets. Now

[19] https://nbnatlas.org

that we have described the data necessary for this research, we can move on to addressing our hypotheses and research questions discussed in Chapter 1. In particular, in the next chapter, we will present new methods for modelling geographic locations using Flickr tags and structured data based on a bag-of-words (BOW) representation model. We will introduce a set of experiments to explore, compare, and assess the prediction power of using Flickr tags only, structured data only, and their combination.

Chapter 4

Modelling Locations using Bag-Of-Words Representation

4.1 Introduction

The main aim of this chapter is to provide a first exploration about the kind of scientifically useful information that can be derived from Flickr. In particular, we focus on comparing the predictive power of Flickr tags with that of structured data from more traditional sources for the task of characterising the natural environment. To this end, we introduce a new method for modelling locations using georeferenced Flickr tags. The method is based on a spatially smoothed version of positive pointwise mutual information (PPMI). We evaluate the proposed method for predicting a broad set of environmental features: scenicness, species distribution, land cover, and climate data. For the task of species distribution modelling, we will also pay particular attention to the role that is played by tags that correspond to the name of the species.

The remainder of this chapter is organised as follows. Section 4.2 presents our methodology for modelling locations based on Flickr tags and based on structured data. In Section 4.3 we then provide a detailed discussion about our experimental results. Subsequently, Section 4.4 evaluates the role of species name tags. Finally, Section 4.5 summarises our findings from this chapter.

4.2 Methodology

In this section, we will explain our methodology for representing locations using the Bag-Of-Words (BOW) model. In particular, Section 4.2.1 explains how feature vectors describing locations can be obtained from the tags associated with georeferenced Flickr photographs. After that, in Section 4.2.2 we describe how feature vectors can be derived from structured information sources. The full model is schematically summarised in Figure 4.1, which shows how these two vector representations are then combined to represent locations. These representations are then used to train a classifier and predict values at unsampled locations.

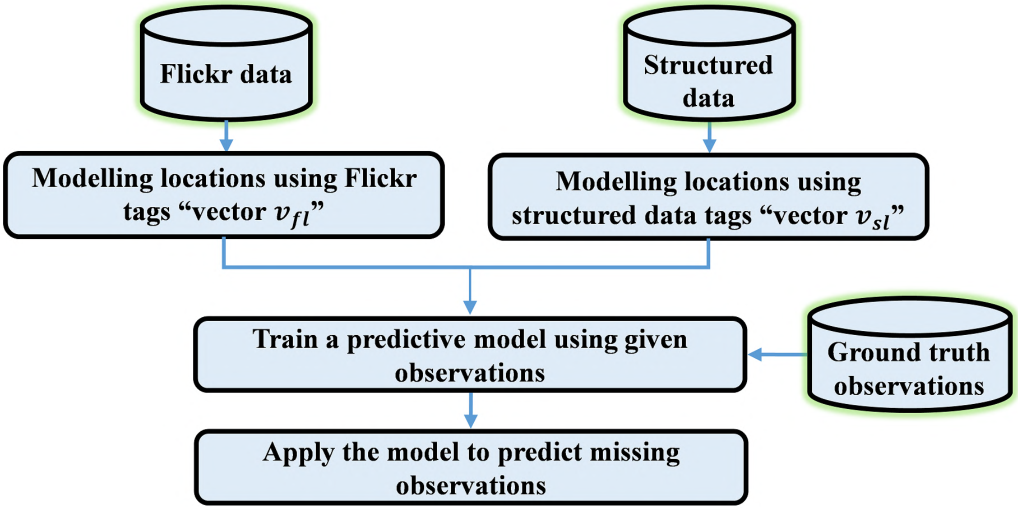

Flickr data
Structured data
Modelling locations using Flickr tags "vector v_{fl}"
Modelling locations using structured data tags "vector v_{sl}"
Train a predictive model using given observations
Ground truth observations
Apply the model to predict missing observations

Figure 4.1: Modelling locations using a Bag-Of-Words model.

4.2.1 Modelling Locations Using Flickr Tags

Many of the tags associated with Flickr photographs tell us something about the locations where these photos were taken. For example, tags might refer to 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). From the georeferenced Flickr photographs that were collected in Section 3.2, there are a total of over 70 million photographs within Europe and approximately 12 million photographs within the UK and Ireland, which are the regions our experiments in this chapter will focus on.

Let $L = \{l_1, ..., l_m\}$ be a set of locations (points), 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 often it occurs in photographs near that location. To this end, we first use a BallTree[1] to retrieve the set F_l of all Flickr photographs 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 photograph with coordinates c. Then we can 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 photographs in F_l. However, to reduce the impact of bulk uploading, we count a tag occurrence only once for all photographs 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 photographs, whereas intuitively we want photographs which are closer to l to influence our characterisation of l more than photographs which are further away. To this end, following [117], we use a Gaussian kernel to weight the tag occurrences:

$$w(t,l) = \sum_{d(c,l)\leq D} |U_{t,c}| \cdot \exp\Big(- \frac{d^2(l,c)}{2\sigma^2}\Big) \tag{4.1}$$

where σ is a bandwidth parameter.

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), as was explained in Section 2.3.1 and given by Equation 2.4. We treat the set of tags that

[1] `http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.BallTree.htm`

occur near location l as the document. In other words, we compare 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). However, here we used the weight $w(t, l)$ instead of the number of occurrences. Specifically, the weight of tag t in our bag-of-words representation of l is then given by:

$$\textit{PPMI}(t, l) = \max\left(0, \log\left(\frac{P(t, l)}{P(t)P(l)}\right)\right) \tag{4.2}$$

where:

$$P(t, l) = \frac{w(t, l)}{N} \tag{4.3}$$

$$P(t) = \frac{\sum_{l' \in L} w(t, l')}{N} \tag{4.4}$$

$$P(l) = \frac{\sum_{t' \in T} w(t', l)}{N} \tag{4.5}$$

$$N = \sum_{t' \in T} \sum_{l' \in L} w(t', l') \tag{4.6}$$

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

4.2.2 Modelling Locations Using Structured Data

The most obvious type of structured data are the coordinates of the photograph 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 structured scientific data described in Section 3.3. To encode locations, we consider a feature vector v_{sl} 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 annual average of temperature, precipitation, solar radiation, wind speed and water vapour pressure. The real-valued features have been normalised using the standard z-score. In experiments where both Flickr data and structured data are used, we simply concatenate the two corresponding feature vectors.

4.3 Experiments

In the following experiments, we evaluate how well we can predict a number of environmental features using Flickr tags and the considered structured data. For all experiments, we have set the maximum Haversine distance D (cluster radius) to 1 kilometre and the bandwidth σ to D/3. The choice of D represents a trade-off, where larger values can potentially lead to more accurate 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. To make predictions, we use Support Vector Machines (SVMs) for classification problems and Support Vector Regression (SVR) for regression problems (for more details see Section 2.5.1). For each experiment, the set of locations L was split into two-thirds for training, one-sixth for tuning the parameters of the SVM/SVR models, and one-sixth for testing.

4.3.1 Predicting Scenicness

In this first experiment, we consider the problem of predicting people's opinions of landscape beauty, using the UGC dataset from the ScenicOrNot website that was described in Section 3.4.1 as ground truth. The dataset contains 217,000 rated images at distinct locations in Great Britain. For 25,395 of the images in this dataset, our Flickr collection did not contain any georeferenced photographs within a 1km radius. Therefore, we only report results for the remaining 191,605 photographs (i.e. 88.3% of the full dataset). The number of Flickr photographs within a 1km radius of these locations varies between 1 and 397982.

Table 4.1: Results for predicting scenicness.

Dataset	Mean Absolute Error	Spearman ρ
Structured	1.031	0.556
Flickr	1.013	0.570
Structured + Flickr	**1.006**	**0.581**

For this experiment, L thus contains the locations of these 191,605 photographs. Table 4.1 shows the results for three different variants: only using structured data, only using Flickr data, and combining both. Based on the tuning data, for the SVR model, we found a Gaussian kernel to be optimal when only structured data are used, and a linear kernel to be optimal otherwise. The results in Table 4.1 show the mean absolute error between the predicted and actual scenicness scores, as well as the Spearman ρ correlation between the rankings induced by both sets of scores (for more details see Section 2.5.3). Note that the mean value of this data set is 4.372 and the standard deviation is around 1.6. While the differences are small, we find that using Flickr outperforms using structured data, and that combining both leads to the most accurate results overall. Looking at what tags most influence the regression model, among the highest weighted tags we find terms relating to natural and open-country landscape such as *scotland*, *highlands*, *mountains* and *sea*, while among the lowest weighted tags we find names of artificial and urban phenomena such as *station*, *bus*, *pub* and *railway*. This reinforces the finding from [110] that land cover categories are strongly correlated with scenicness scores.

We also tested whether the number of photographs (or users) could be used to predict scenicness, as was suggested in [12, 114, 40] for particular restricted settings. However, we actually found a negative correlation of around -0.12 (resp. -0.1) between scenicness and the number of photographs (resp. users who have posted photographs) near a given location.

4.3.2 Predicting Species Distribution

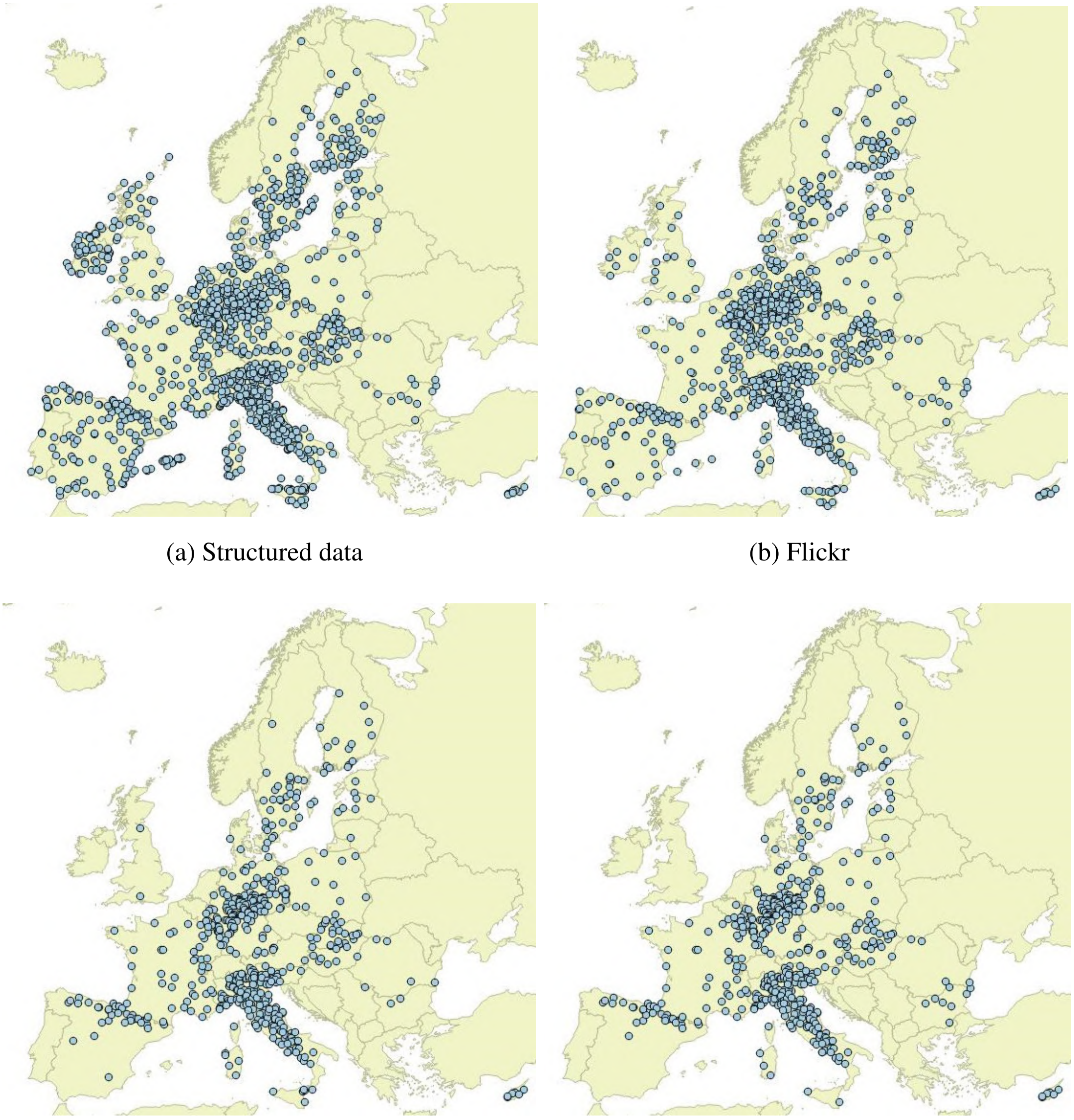
The next experiment we considered was to predict the distribution of species across Europe, using as ground truth the dataset of the European network of nature protected sites Natura 2000; see Section 3.4.2 for more details about this dataset. This dataset specifies which species occur at 26,425 different 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 are reported in Table 4.2, showing that combining structured data with Flickr data leads to substantially better results than either structured data alone or Flickr data alone. Comparing Flickr with structured data directly is more difficult, as Flickr data led to a much higher precision, whereas the structured data led to a much higher recall.

As an example, Figure 4.2 compares the predictions that were made by the different models with the ground truth for a particular species: the black woodpecker (dryocopus martius). For this species, the F1 scores were 0.594, 0.648 and 0.927 for structured data, Flickr data, and the combined data, respectively. This example shows that highly accurate distribution models can be learned for species that occur in sufficiently many sites. Interestingly, while the number of occurrences is overestimated in, e.g. Spain and the UK when only Flickr data or only structured data are used, much more accurate predictions are made for these countries using the combined model. For species that have a more restricted geographic scope (in terms of the number of sites), it is likely that better results can be obtained by looking at a wider region and by specifically counting photographs that mention the name of the species, as a separate feature. This will be discussed in detail in Section 4.4.

Table 4.2: Results for predicting species distribution.

Dataset	Precision	Recall	F1 Score
Structured	0.241	0.568	0.338
Flickr	0.577	0.112	0.188
Structured + Flickr	0.650	0.506	**0.569**



(a) Structured data

(b) Flickr

(c) Combination of structured data and Flickr

(d) Ground truth data

Figure 4.2: Prediction of the black woodpecker distribution across Europe.

4.3.3 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 sites 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 here as ground truth). We experimented with predicting CORINE land cover classification at levels 1, 2 and 3, each time treating the task as a binary classification problem. The results are presented in Table 4.3, showing again that combining structured data and Flickr data clearly leads to the most accurate results. The difference in performance between structured data alone and Flickr data alone is mixed. For example, Flickr data performing better at level 1 but worse at level 2. For level 1, we found that Flickr outperformed structured data in 4 out of the 5 classes, with the *artificial surfaces* class being the only exception. This seems related to the small number of sites for this particular class (e.g. only 4% of the training data sites belong to this class). To illustrate how Flickr tags are used to predict CORINE classes, Table 4.4 shows the five tags with the highest weight in the SVM classifier for each of the classes at level 1.

By far the largest CORINE class at level 1 is *Forest & semi natural areas*. This class has three subclasses at level 2. The predictions of the three models for these three subclasses are compared with the ground truth in Figure 4.3. Clearly, in this case, the structured data has resulted in a model that is too simplistic, essentially segmenting Europe into *Forests* and *Shrub and/or herbaceous vegetation*. Flickr data alone leads to more faithful predictions for these subclasses, but instances of *open spaces with little or no vegetation* are underreported. This issue is alleviated in the combined model.

Table 4.3: Results for predicting CORINE land cover classes, at levels 1, 2 and 3.

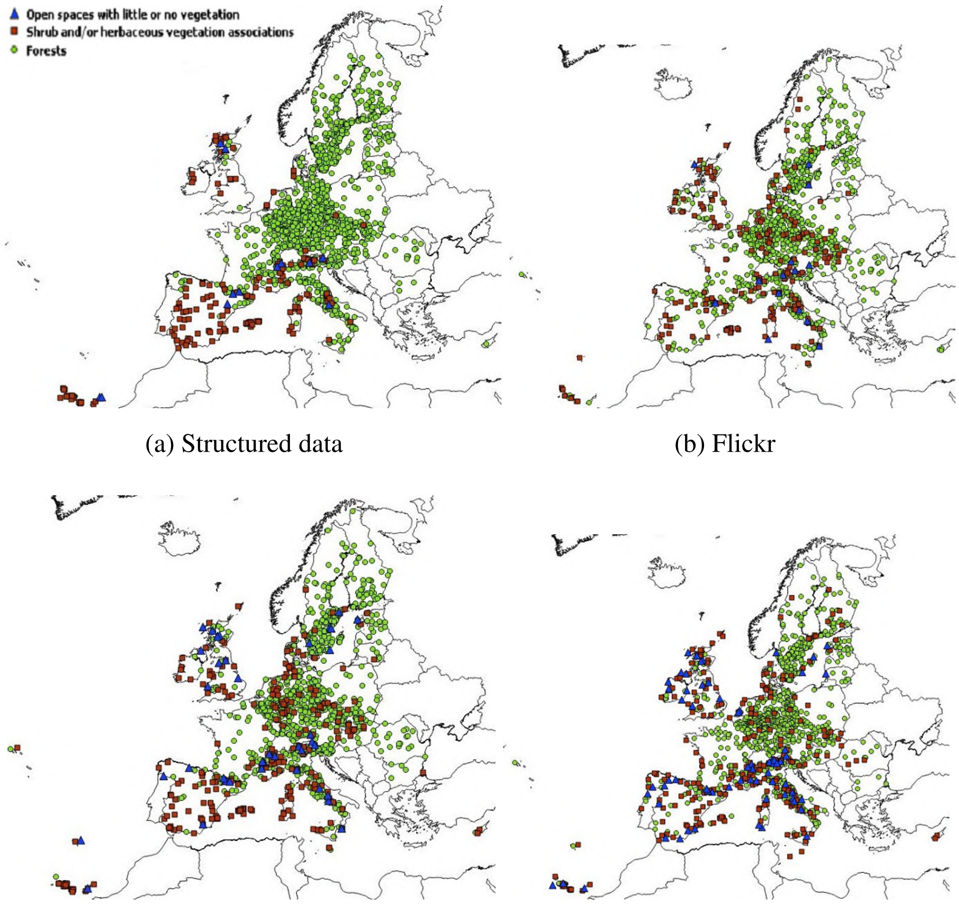
	Level 1			Level 2			Level 3		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Structured	0.437	0.363	0.397	0.346	0.160	0.219	0.207	0.070	0.105
Flickr	0.499	0.457	0.477	0.205	0.139	0.166	0.145	0.086	0.108
Structured + Flickr	0.523	0.514	**0.518**	0.270	0.199	**0.229**	0.184	0.112	**0.139**

Table 4.4: Top 5 Flickr tags for CORINE level 1 classes in the SVM models.

Artificial surfaces	Agricultural areas	Forest & semi nat. areas	Wetlands	Water bodies
Babenhausen	field	wald	bog	lake
Ceskedrahy	grass	forest	moor	island
Meppen	horse	mountains	marsh	sea
Tuplice	vineyard	woods	swamp	boat
Deutsche Reichsbahn	meadow	mountain	saline	sailing

4.3.4 Predicting Climate Data

In the last experiment in this section, we assess the usefulness of Flickr tags for 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 the feature vector representations as they constitute the ground truth. We consider five different regression problems: predicting average temperature, average precipitation, average solar radiation, average wind speed, and average water vapour pressure. The results are reported in Table 4.5, in terms of mean absolute error (MAE) and Spearman ρ. The mean and standard deviation of each of those features are shown in Table 4.6. Overall, structured data and Flickr data perform comparably. However, by far the most accurate results are obtained when combining both types of data, showing again that the information we obtain from Flickr is complementary to what is available as structured data. As an example of how Flickr tags are used by the regression model, the tag ‘sea’


▲ Open spaces with little or no vegetation
■ Shrub and/or herbaceous vegetation associations
● Forests

(a) Structured data

(b) Flickr

(c) Combination of structured data and Flickr

(d) Ground truth data

Figure 4.3: Prediction of subclasses of the CORINE class "Forest & semi natural areas".

has a very large weight in the model for predicting water vapour pressure, while the tag 'mountain' has a very low weight in this model. In Figure 4.4, we illustrate the predictions made by the different models for solar radiation. Clearly, the model based on structured data is too simplistic, mostly capturing the impact of latitude.

Table 4.5: Results for predicting average climate data.

	Structured		Flickr		Struct+ Flickr	
	MAE	ρ	MAE	ρ	MAE	ρ
Temperature	0.789	0.938	1.623	0.814	**0.728**	**0.940**
Precipitation	13.173	0.709	11.660	0.689	**10.523**	**0.755**
Solar Radiation	1726.5	0.747	926.3	0.832	**484.8**	**0.939**
Wind Speed	0.508	0.791	0.545	0.756	**0.429**	**0.846**
Water Vapor Pressure	0.060	0.903	0.083	0.719	**0.053**	**0.914**

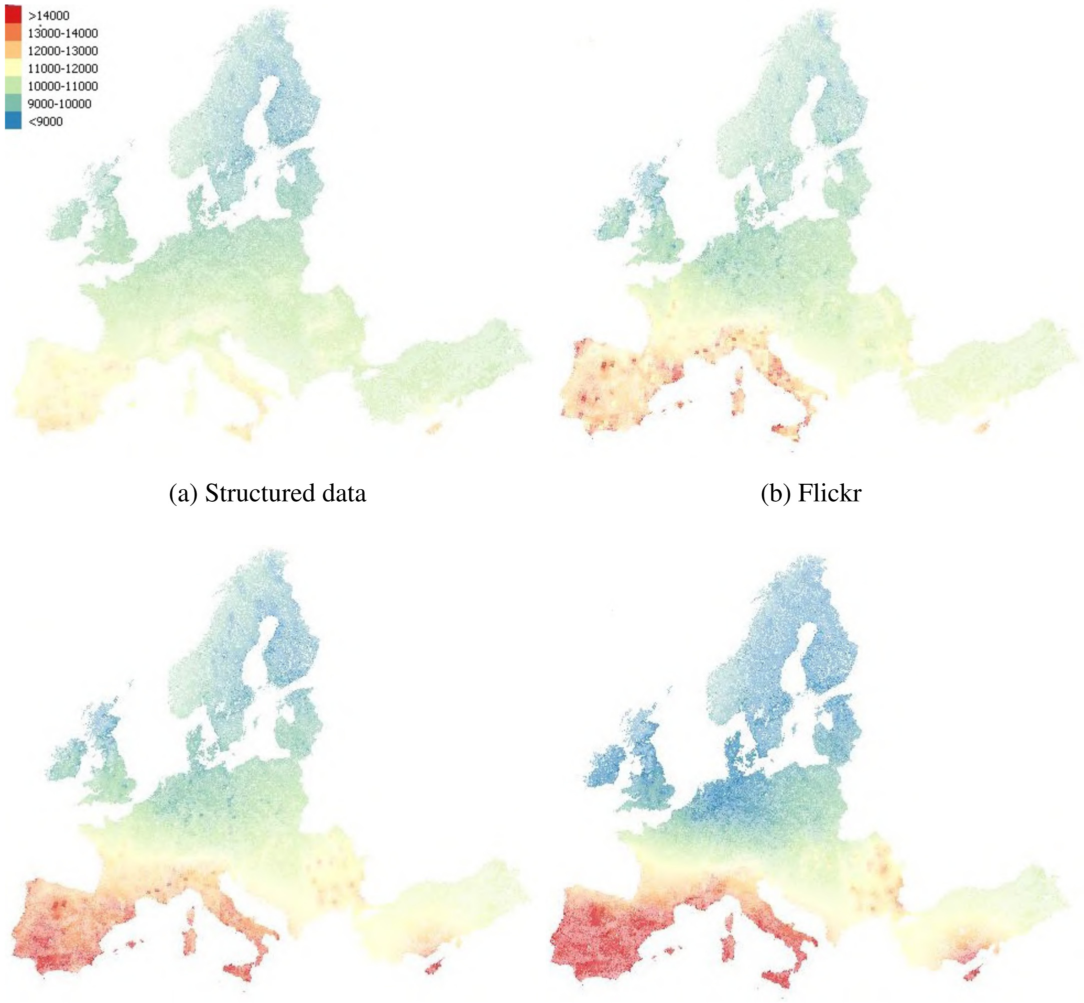
Table 4.6: Mean and Standard deviation of climate data.

	Mean	STDEV
Temperature ($^{\circ}C$)	9.268	3.490
Precipitation (mm)	66.625	24.827
Solar Radiation ($kJ\ m^{-2}day^{-1}$)	11478	2388
Wind Speed ($m\ s^{-1}$)	3.605	1.126
Water Vapor Pressure (kPa)	0.958	0.186

4.4 Evaluating the Role of Species Name Tags

In the fields of wildlife observation, there is clearly strong potential for exploiting social media, reflected in the fact that searching for named species on photo-sharing Flickr often reveals thousands of results, many of which are associated with coordinates and almost all with timestamps. 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 [4, 20].

The aim of this section is to evaluate the performance of predicting the occurrence of species in a given geographic 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



(a) Structured data

(b) Flickr

(c) Combination of structured data and Flickr

(d) Ground truth data

Figure 4.4: Prediction of solar radiation.

common name or scientific name). This method is then compared with the standard text classification method (similar to that explained above in Section 4.3.2), 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. Furthermore, we develop a meta-classifier that combines the prediction of the text classifier with information about the occurrence of the species name in or near the given region to make the final prediction.

4.4.1 Methodology and Data

Our objective here is to find a method that can use Flickr tags for predicting the occurrence of wildlife species. To this end, we use the ground truth species distribution from the National Biodiversity Network Atlas (NBN Atlas)[2] (see Section 3.4.3 for more information about NBN Atlas). The reason for choosing a different source of ground truth than in Section 4.3.2 is because Natura 2000 project deals with rare species which cannot be tagged by non-experts. 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.

To use Flickr tags for predicting the occurrence of those species, we first 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. We then use Positive Pointwise Mutual Information (PPMI), given by Equation 2.4, to weight how strongly tag t is associated with cell c. There is no need for distance weighting here because we consider grid cells instead of point locations. Consequently, each cell c is 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. Note that even species with a large number of occurrences may possibly only occur in a few cells.

Our method for 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 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

[2] NBN Atlas occurrence was download from http://nbnatlas.org. Accessed 19 April 2018.

species in the cell c or the neighbouring 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 name itself in c as a binary feature (being 1 if 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 a tag. The second classifier *SVM2* is learned using the feature vector V_m to give the final estimation. This process is illustrated below in Figure 4.5.

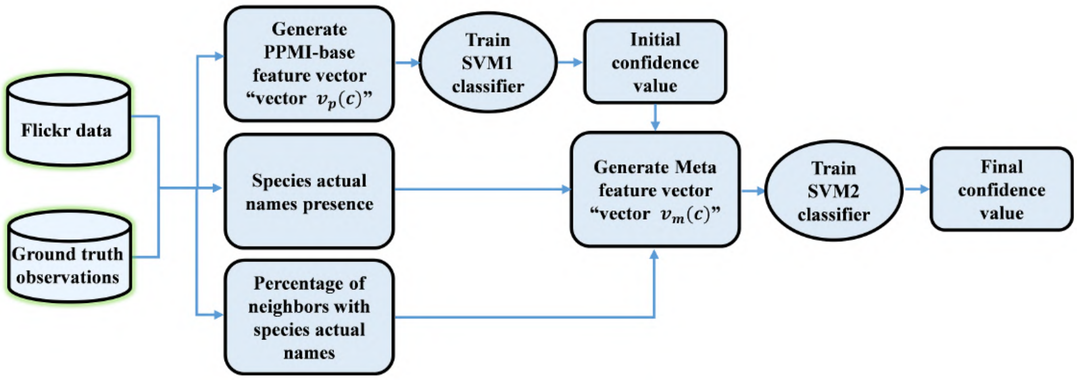

Flickr data
Ground truth observations
Generate PPMI-base feature vector "vector $v_p(c)$"
Species actual names presence
Percentage of neighbors with species actual names
Train SVM1 classifier
Initial confidence value
Generate Meta feature vector "vector $v_m(c)$"
Train SVM2 classifier
Final confidence value

Figure 4.5: The training process.

4.4.2 Evaluation

For evaluation, 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 kilometres. 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 4.6, 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

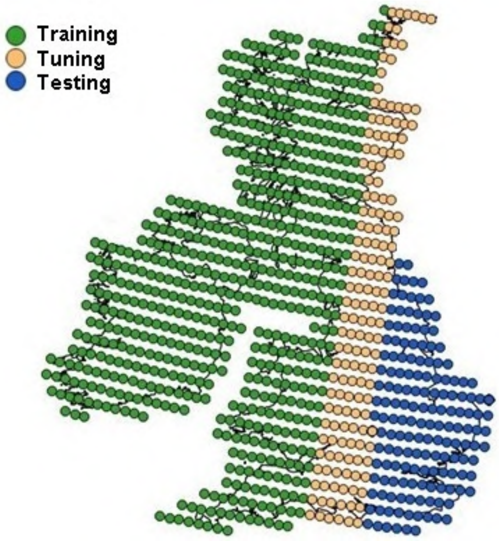


Figure 4.6: Training, Tuning, and Testing regions.

challenging than choosing the cells randomly, due to possible differences between the training and testing regions. Finally, for our evaluation, we compared the results of three different methods:

1. “Species Names” which predicts that the species occurs if its common or scientific name appears in at least one Flickr photograph in the test cell.
2. “All Flickr Tags” (*SVM1*) which uses the PPMI-based feature vector modelling all Flickr tags to train an SVM classifier.
3. “Meta features”(*SVM2*) which is the proposed method as described in Section 4.4.1.

4.4.3 Results and Discussion

The results of predicting species distribution are reported in Table 4.7 in terms of the average accuracy, average precision, average recall, and average F1 score over the 50 birds. The results clearly show that “All Flickr Tags” significantly outperforms

Table 4.7: Results for predicting the distribution of 50 species across the testing area.

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

"Species Names". However, the proposed meta-classifier leads to the most accurate 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 small, as there may not be enough training data to effectively learn an SVM classifier in such cases. To illustrate such issues, Table 4.8 shows the F1 scores of five 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 obtained better results when using the "Species name" method. Interestingly, our proposed meta-classifier, which takes account of both the species presence data and the all tags classification for nearby regions, outperforms both of the other methods for almost all the considered species.

Figures 4.7 and 4.8 visually illustrate the performance of our method. Note that these

Table 4.8: F1 scores for predicting the distribution of individual species using different methods.

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

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 the absence of any reason to predict occurrence, it is safer for a classifier to predict non-occurrence).

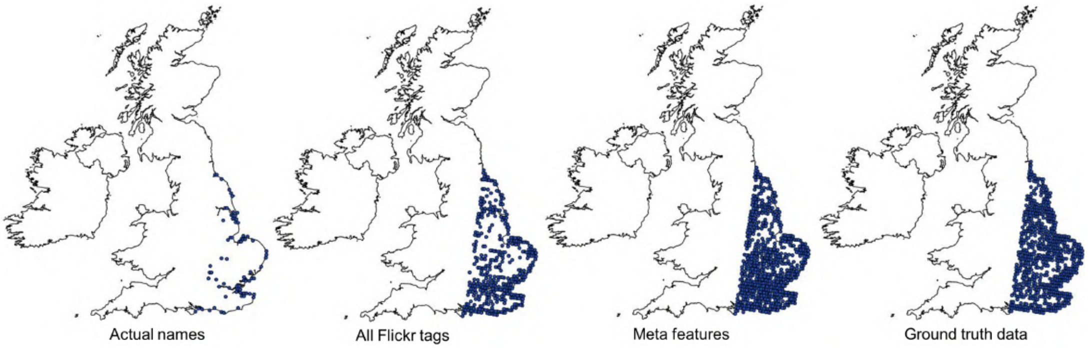


Figure 4.7: Prediction of the Dunlin distribution across the testing area with 10km grid cells.

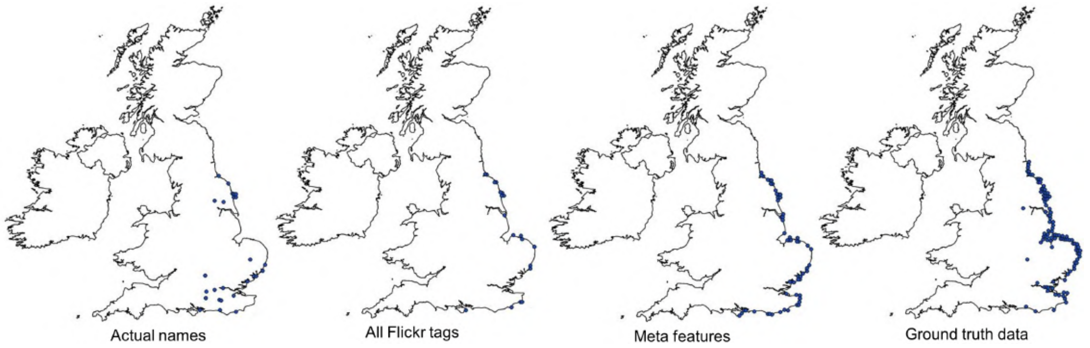


Figure 4.8: Prediction of the Atlantic Puffin distribution across the testing area with 10km grid cells.

4.5 Summary

In this chapter, we have analysed how Flickr tags can be used to supplement structured scientific data in tasks that rely on characterising 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 of predicting CORINE land cover categories. Finally, we looked at predicting five climate-related properties. Each time, we compared three different setups. In the first setup, we used features that were derived from several structured scientific datasets. In the second setup, we used a bag of words representation, capturing how strongly each tag is associated with photographs that appear near a considered location. In the final setup, we combined both data sources, concatenating the corresponding feature vectors. Our main finding from these experiments is that the combined model substantially and consistently outperformed the model that only relied on structured data sources. This strongly suggests that Flickr can indeed be valuable, as a supplement to more traditional datasets in environmental analyses. While it may be possible to reduce some of the performance gaps by considering additional scientific datasets, we found the versatility of Flickr data that was displayed in the four experiments to be remarkable.

Additionally, to get a deeper understanding of the usefulness of using Flickr tag for mapping the location of species occurrence. We have compared and combined two main strategies: (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. From these experiments, we found 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.

We have identified two directions to improve the results conducted in this chapter. First, since many of the considered features are strongly spatially autocorrelated, it may be possible to improve the predictions by formulating some of the considered tasks as collective prediction problems, where we would intuitively take into account the predictions for neighbouring sites. This improvement will be the focus of Chapter 5. Second, which will be explained in Chapter 6, we can expect to obtain more accurate predictions by improving the way we have combined structured features with bag-of-words features. In Chapter 6, this will be achieved by learning a low-dimensional vector space embedding that captures both kinds of data.

Chapter 5

Collective Prediction Model

5.1 Introduction

In this chapter, we propose a collective prediction model, which takes advantage of the fact that most environmental features are strongly spatially autocorrelated. For instance, climate features typically do not vary much between places that are just a few kilometres apart. Inspired by [2] and [26], a key feature of our approach is that the neighbourhood structure of the collective prediction model does not only depend on geographic distance but also on attribute similarity, which is estimated in our case from the Flickr tags associated with each location. 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.

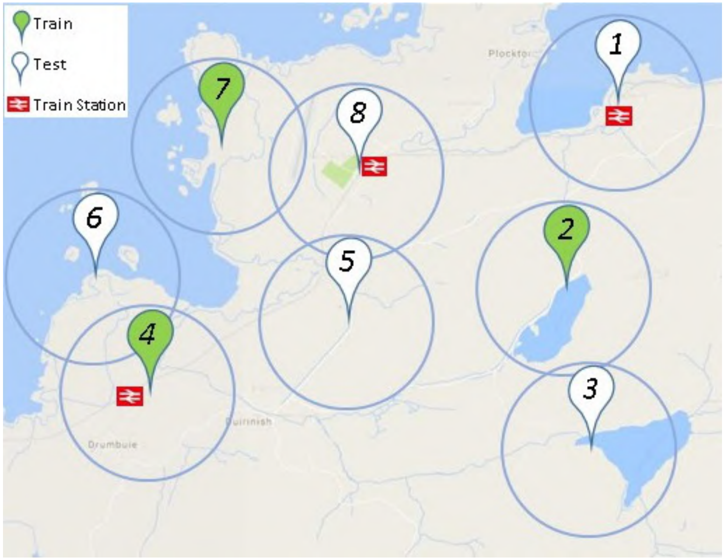
The problem we consider is to predict the value of a given feature (e.g. average temperature or land cover category) for a given set of locations, where we assume that for a subset of these locations (i.e. the training data), the correct value of the considered feature is available (e.g. temperature 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 5.1 (where, in practice, the ground truth data are 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, 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.

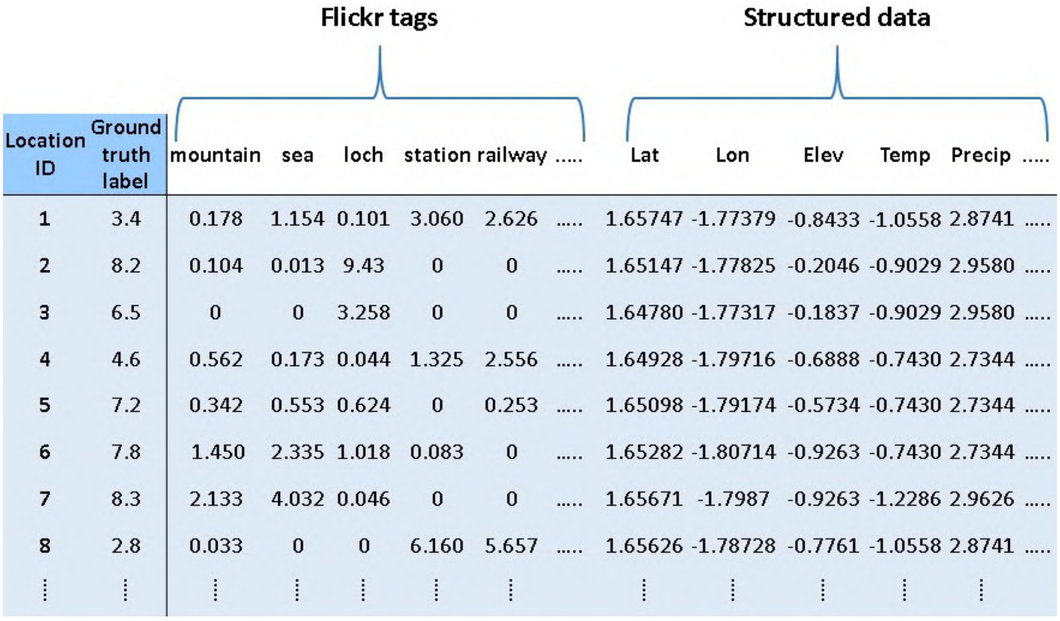
The second step crucially relies on how the neighbours are selected. As a baseline, we could choose the neighbours of a given location as those locations which are geographically closest. For example, consider the locations shown on the map in Figure 5.1 for the task of predicting scenicness. To improve the prediction for location 8, based on geographic distance, we could select locations 2, 5 and 7 as neighbours. However, locations 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 chapter is organised as follows. Section 5.2 describes our collective prediction framework. Subsequently, Section 5.3 provides a detailed discussion about our experimental results. Finally, Section 5.4 summarises the chapter.

5.2 Collective Prediction

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





		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 5.1: Modeling locations based on Flickr tags, structured features, and neighborhood structure..

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

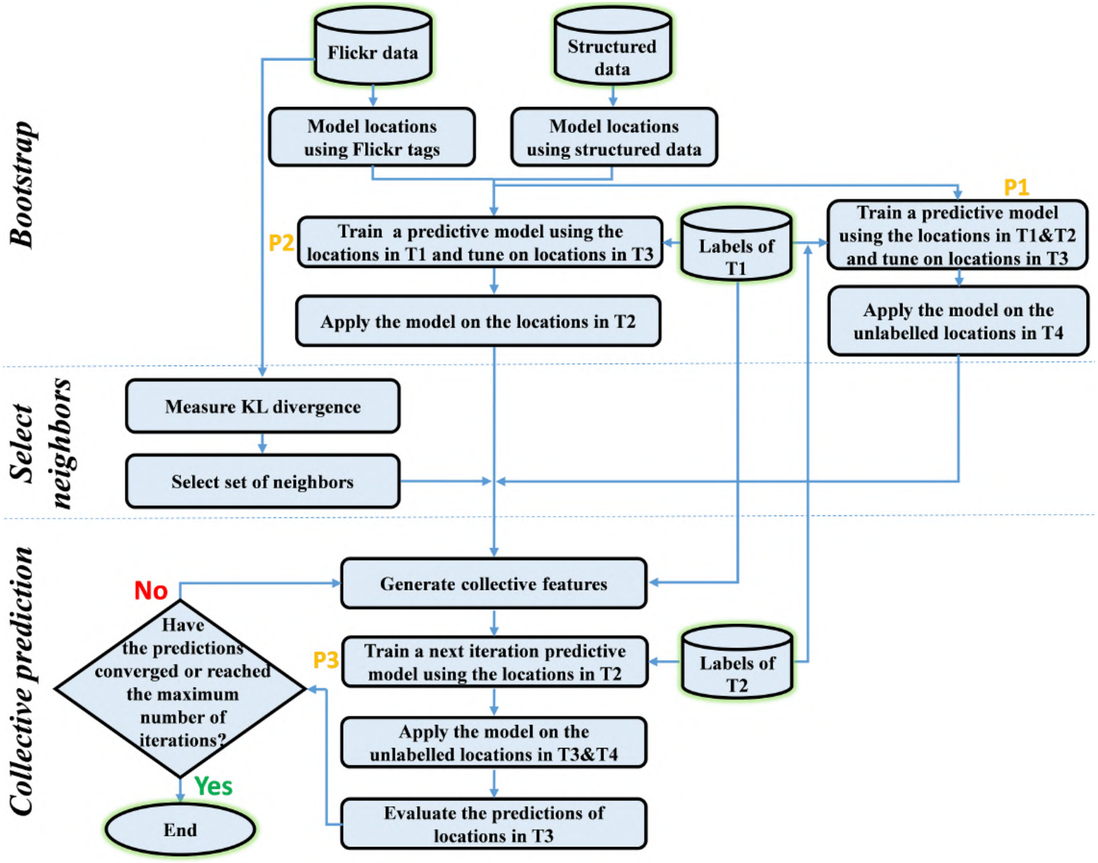

Bootstrap
Select neighbors
Collective prediction
Flickr data
Structured data
Model locations using Flickr tags
Model locations using structured data
P1
P2
Train a predictive model using the locations in T1 and tune on locations in T3
Labels of T1
Train a predictive model using the locations in T1&T2 and tune on locations in T3
Apply the model on the locations in T2
Apply the model on the unlabelled locations in T4
Measure KL divergence
Select set of neighbors
No
Generate collective features
Have the predictions converged or reached the maximum number of iterations?
P3
Train a next iteration predictive model using the locations in T2
Labels of T2
Apply the model on the unlabelled locations in T3&T4
Yes
End
Evaluate the predictions of locations in T3

Figure 5.2: The collective prediction model.

account their interrelationships. To apply the collective prediction framework to our setting, we consider each of the locations $l \in L$ 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. Besides, we consider three classifiers: $P1$, $P2$, and $P3$. The locations in $T1$ will be used for training bootstrap classifiers ($P1$ and $P2$), while those in $T2$ will be used for learning how to improve predictions based on related locations (i.e. $P3$). The locations in $T3$, finally will be used for tuning all the classifiers.

The overall method involves the following steps, which are illustrated in Figure 5.2.

Bootstrap

In this step, we use the feature vector representation (vectors v_{fl} and v_{sl}), as explained in Section 4.2, for each location in $T1 \cup T2$ to learn an SVM or SVR model. When applying our overall model, this classifier ($P1$ in Figure 5.2) 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 5.2), 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 5.1, we may want to select locations 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 that explained in Section 2.3.2. 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:

$$KL(t) = \sum_{i=1}^{n} P(C_i|t) \log \frac{P(C_i|t)}{Q(C_i)} \tag{5.1}$$

where $P(C_i|t)$ is the probability that a photograph with tag t belongs to one of the locations associated with class C_i, whereas $Q(C_i)$ is the probability that a photograph with an arbitrary tag t occurrence belongs to one of the locations associated of class C_i. We estimate $Q(C_i)$ as follows:

$$Q(C_i) = \frac{1}{N} \sum_{l \in C_i} \sum_{t' \in T} w(t', l) \tag{5.2}$$

$$N = \sum_{j=1}^{n} \sum_{l \in C_j} \sum_{t' \in T} w(t', l) \tag{5.3}$$

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

$$P(C_i|t) = \frac{\left(\sum_{l \in C_i} w(t,l)\right) + \delta \cdot Q(C_i)}{N + \delta} \tag{5.4}$$

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

For regression problems, we discretise 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 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.

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 4.2.1 but only considering the 1000 tags that were selected based on (5.1). As in Chapter 4, this vector is then concatenated with the structured feature vectors from Section 4.2.2. These feature vectors are then compared using the cosine similarity.

Iterative inference

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 5.2) 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 work, 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 neighbors 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:

$$x_i = \frac{\sum_{l \in N_i^G} \lambda \cdot \mathit{ground}(l) + \sum_{l \in N_i \setminus N_i^G} \mathit{pred}(l)}{\lambda \cdot |N_i^G| + |N_i \setminus N_i^G|} \tag{5.5}$$

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

For this iterative classification step ($P3$ in Figure 5.2), 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 reaching the maximum number of iterations. We evaluate the convergence here according to the locations in the $T3$ set. This is illustrated in Figure 5.2, which provides an overview of the whole process.

5.3 Experimental Evaluation

5.3.1 Experimental Settings

To directly evaluate the effectiveness of the proposed method, we have used the same feature vectors as in Chapter 4, i.e. the same structured datasets and same tag weighting scheme. We examined various smoothing values to select the distinctive tags in KL divergence ($\delta = 10, 100, 1000$) and chose the best value for each experiment separately based on held-out tuning data ($T3$). The thresholds used to discretise the regression problem data into C_1, C_2, and C_3, which are needed for computing the KL divergence, are listed in Table 5.1. These values were 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 neighbors, choosing k_i as $r_i + 1$, $r_i + 10$ or $r_i + 100$, again based on the held-out tuning data ($T3$). Figure 5.5 shows examples of the collective feature vectors of different locations with their ground truth labels. We set the ground truth labels weight λ to 5. Finally, we set the maximum number of iterations to 10.

For each experiment, the set of locations L was shuffled and 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 are 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 was each time used for tuning. Each training set was divided into two equal size subset $T1$ and $T2$.

5.3.2 Variants and Baseline Methods

We compared the results for seven different variants and baseline methods:

- “Structured” uses the feature vector modeling the structured scientific information from Section 4.2.2 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 4.2.1 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 two corresponding feature vectors. This process is illustrated in Figure 5.3.

- “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. The similarity is defined here as for our collective prediction method, i.e. based on a feature vector that contains the PPMI values of the top-1000 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 5.4.

- “Collective-All” uses the collective features derived from all neighbours. It is very similar to the method described in Figure 5.2 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 5.2.

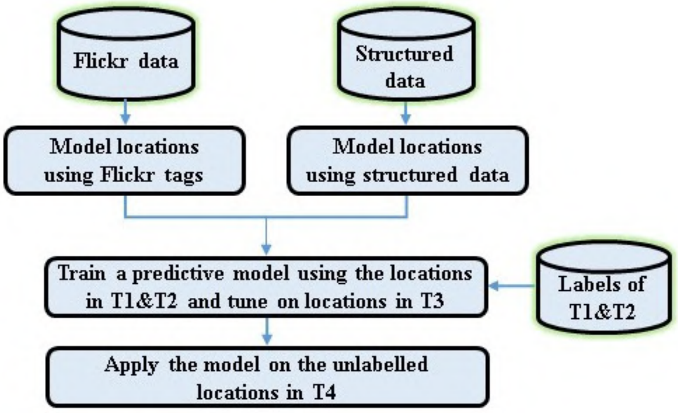
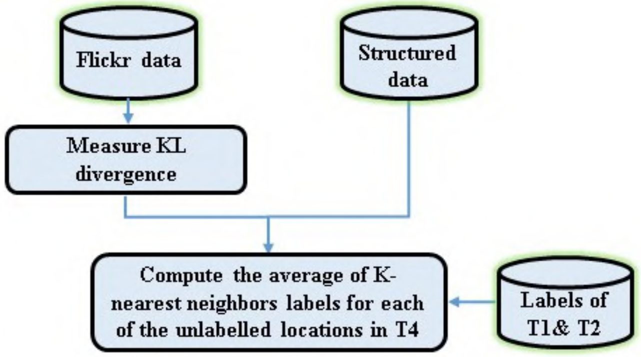
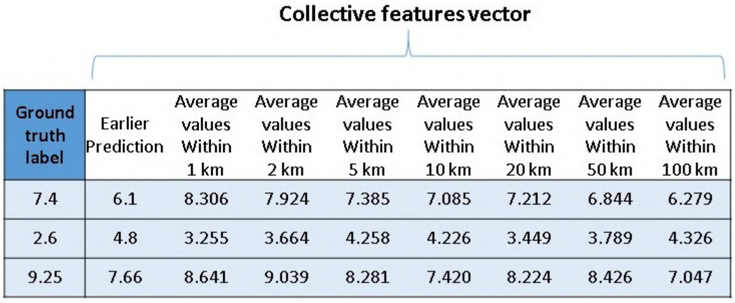

Flickr data
Structured data
Model locations using Flickr tags
Model locations using structured data
Train a predictive model using the locations in T1&T2 and tune on locations in T3
Labels of T1&T2
Apply the model on the unlabelled locations in T4

Figure 5.3: Structured+Flickr prediction model (baseline method).


Flickr data
Structured data
Measure KL divergence
Compute the average of K-nearest neighbors labels for each of the unlabelled locations in T4
Labels of T1& T2


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 5.4: K nearest neighbors prediction model (baseline method).

Figure 5.5: Modeling locations based on collective features.

Table 5.1: High and low boundaries for discretising the features from the regression problems .

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

5.3.3 Experimental Results

Here we use the same set of experiments as in Section 4.3. In particular, we will consider the following tasks.

Predicting the distribution of 100 species across Europe

This experiment was described in Section 4.3.2. The results of predicting species distribution are reported in Figure 5.6 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 “Structured + Flickr” leads to substantially better results than K Nearest Neighbors based models. 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 particular species is present as one class and all the other locations as a second class. Table 5.2 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.

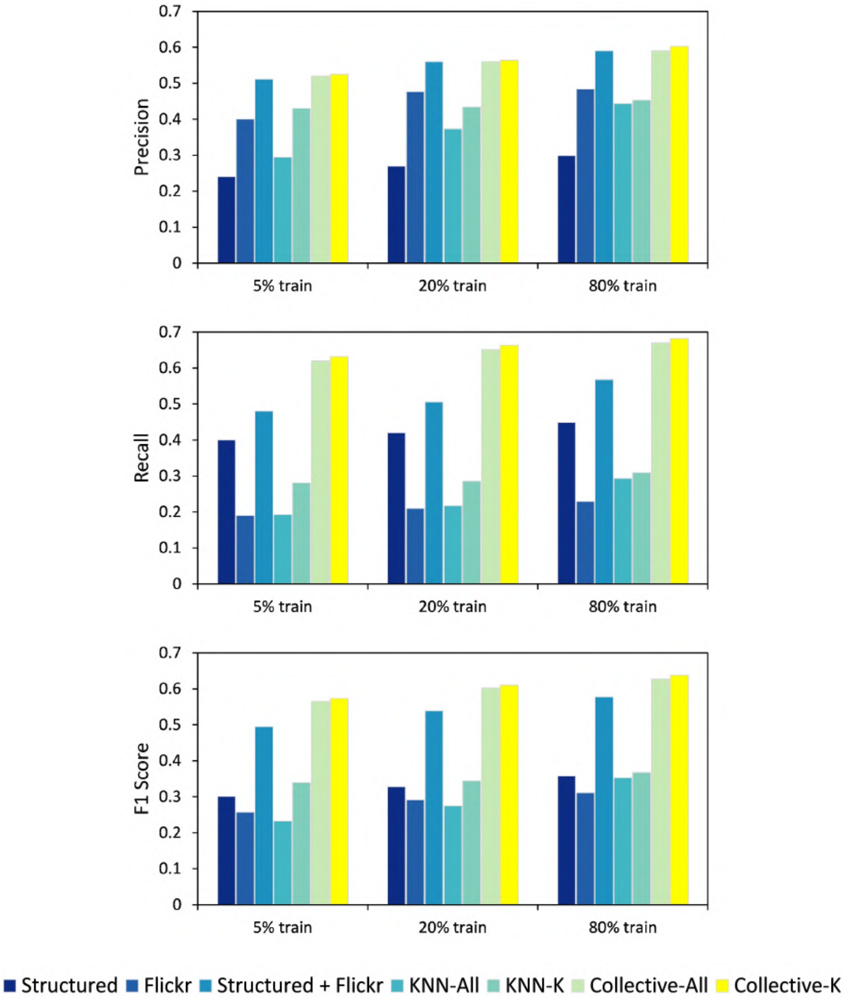


Figure 5.6: Results of predicting species distribution.

Predicting CORINE land cover classes at levels 1, 2 and 3

This experiment was explained in Section 4.3.3. The results of predicting CORINE land cover classification at levels 1, 2 and 3 are presented in Figure 5.7, Figure 5.8, and Figure 5.9 respectively in terms of the average precision, average recall and macro average F1 score. Again, the results show that the collective prediction method (Collective-K) leads to the best results overall. We compute KL divergence for each land cover class separately, where we treat the locations belonging to the target land cover type

Table 5.2: 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 5.3: 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

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 5.3 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. Indeed, it is not possible to find the optimal neighbours if only a few locations belong to that class.

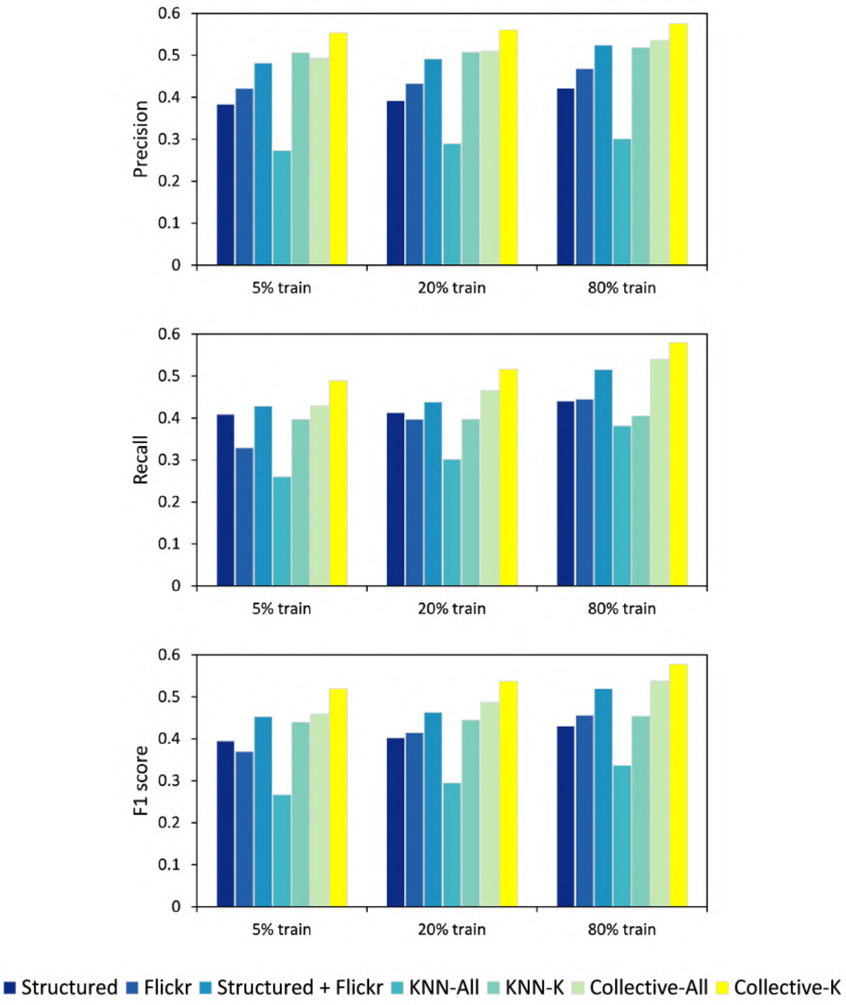


Figure 5.7: Results of predicting CORINE land cover at level 1.

Predicting people's subjective opinions of landscape beauty

This experiment was explained in Section 4.3.1. The results reported in Figure 5.10 show the mean absolute error between the predicted and actual scenicness scores, as well as the Spearman ρ correlation between the rankings induced by both sets of scores for the seven considered methods. Similar to our findings in Chapter 4, using Flickr outperforms using structured data, and combining both leads to better results than using them separately, for all the considered training/test ratios. We also find that all these setups (Structured, Flickr, and Structured+Flickr) perform better than the KNN and KNN-K methods. The collective prediction method leads to the best results overall,

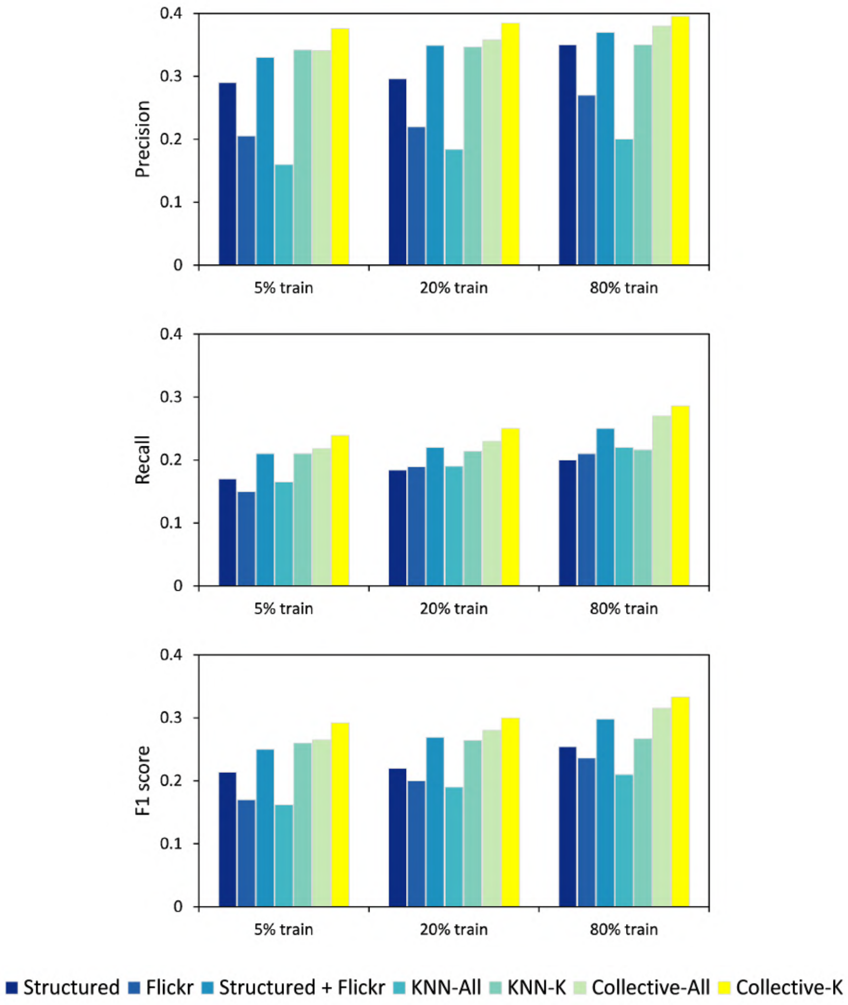


Figure 5.8: Results of predicting CORINE land cover at level 2.

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 landscapes which represent high scenicness such as *highlands*, *mountains*, and *beach* as well as names of artificial and urban phenomena which are representative of low scenicness such as *station*, *bus*, and *supermarket*.

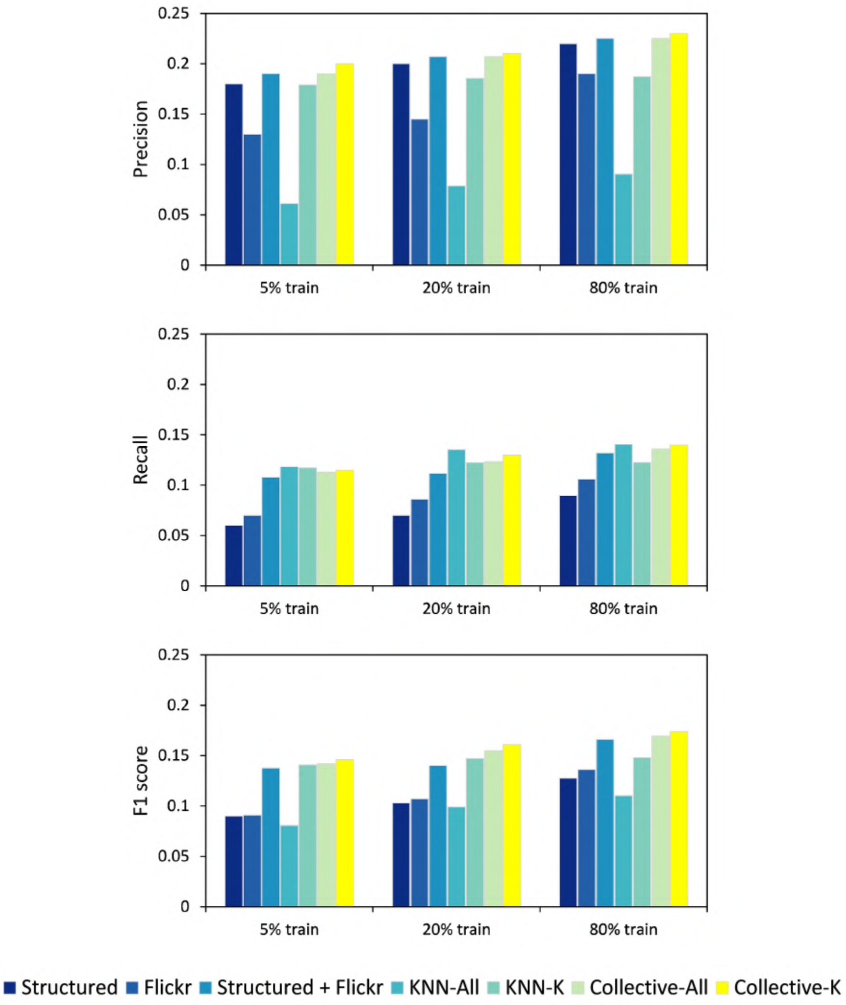


Figure 5.9: Results of predicting CORINE land cover at level 3.

Predicting climate related features

We consider the same five regression problems as in Chapter 4. The results are reported in Figure 5.11, Figure 5.12, Figure 5.13, Figure 5.14 and Figure 5.15 respectively. Overall, using collective prediction leads to an impressive improvement over the basic prediction methods, especially with the “collective-K” variant. Looking at the top selected tags in terms of KL divergence, we find names of countries, regions, and weather phenomena, which are indicative of either high or low values of the corresponding feature as shown in Table 5.4.

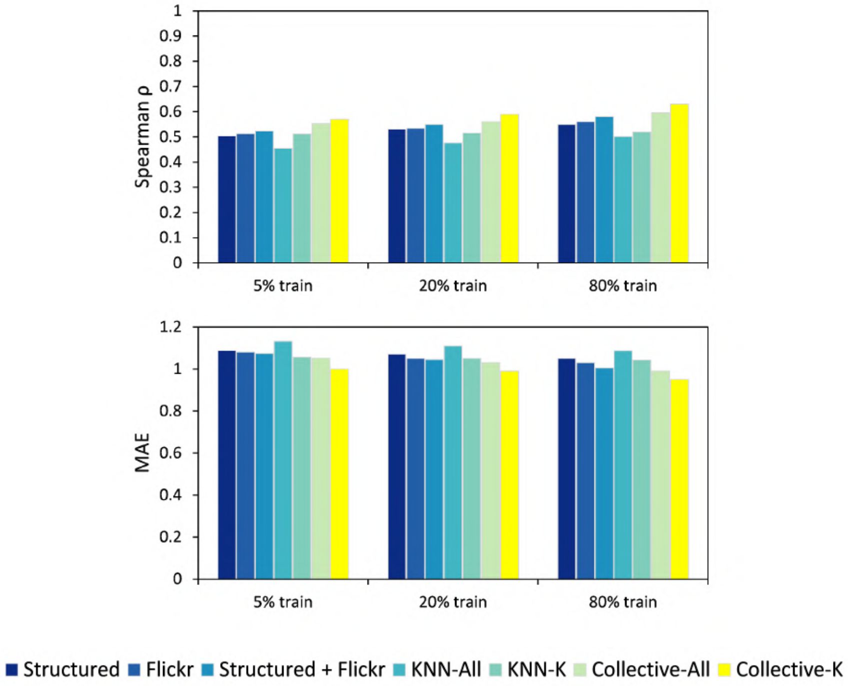


Figure 5.10: Results of predicting scenicness.

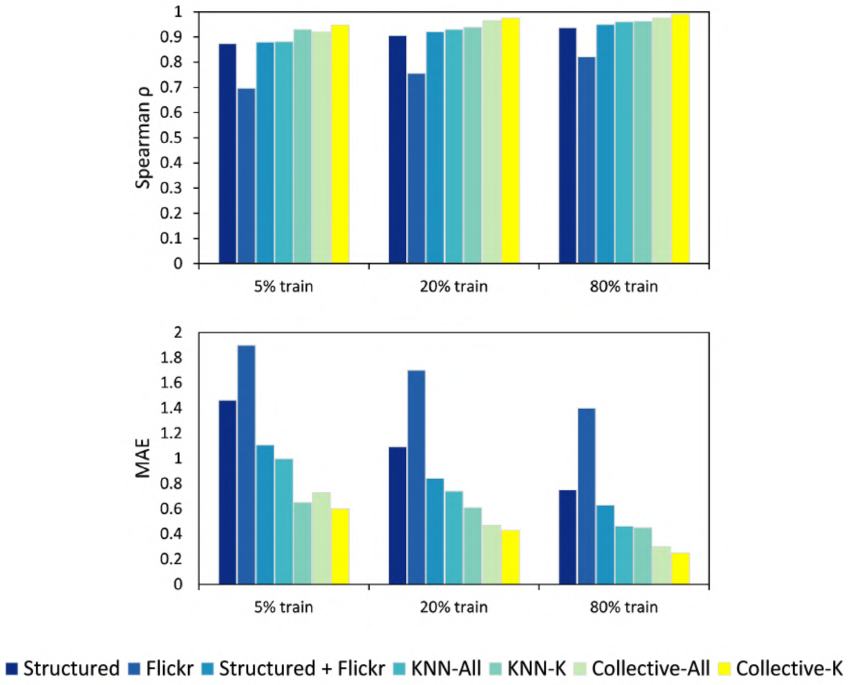


Figure 5.11: Results of predicting average annual temperature.

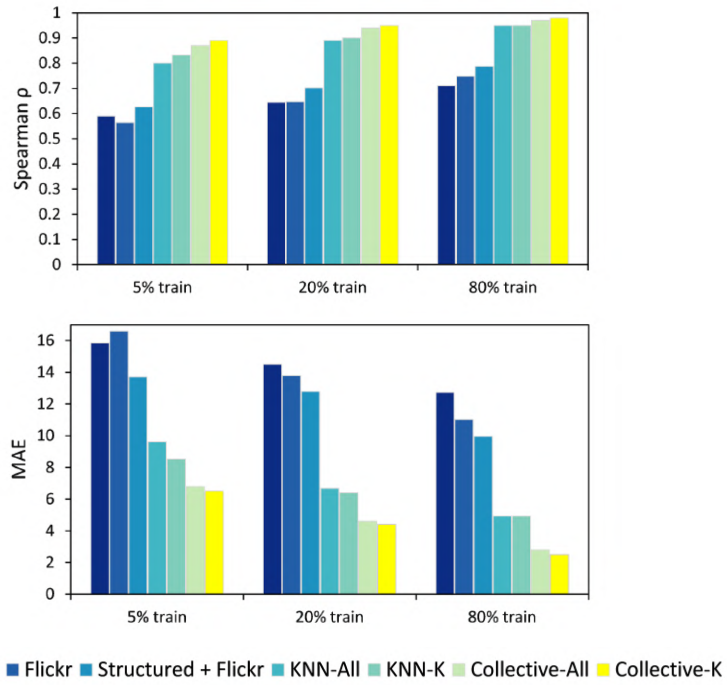


Figure 5.12: Results of predicting average annual precipitation.

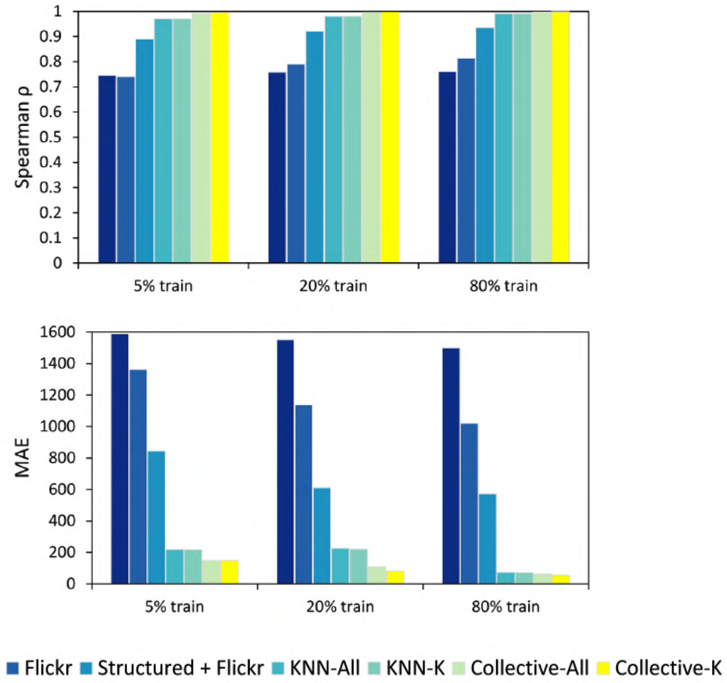


Figure 5.13: Results of predicting average annual solar radiation.

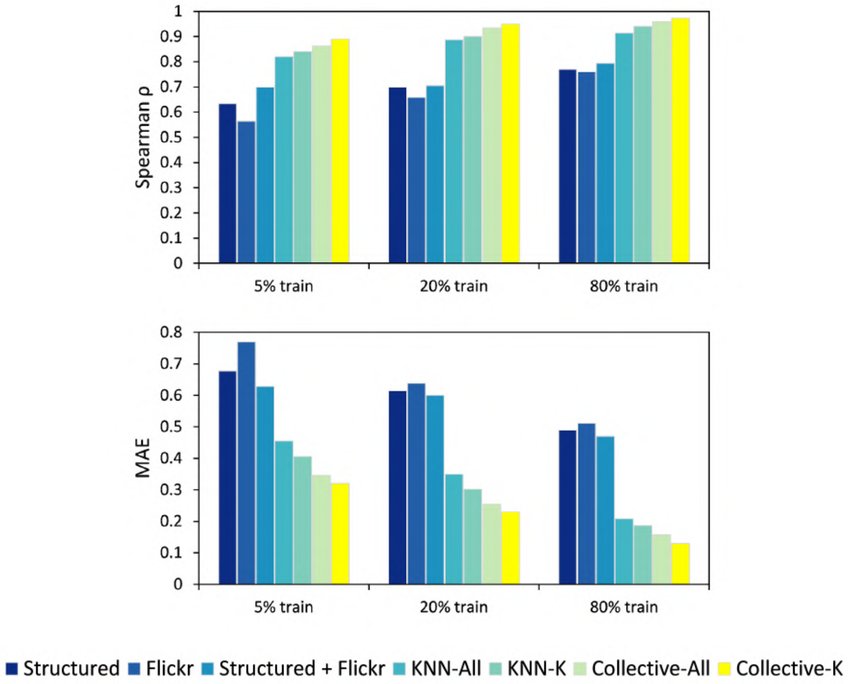


Figure 5.14: Results of predicting average annual wind speed.

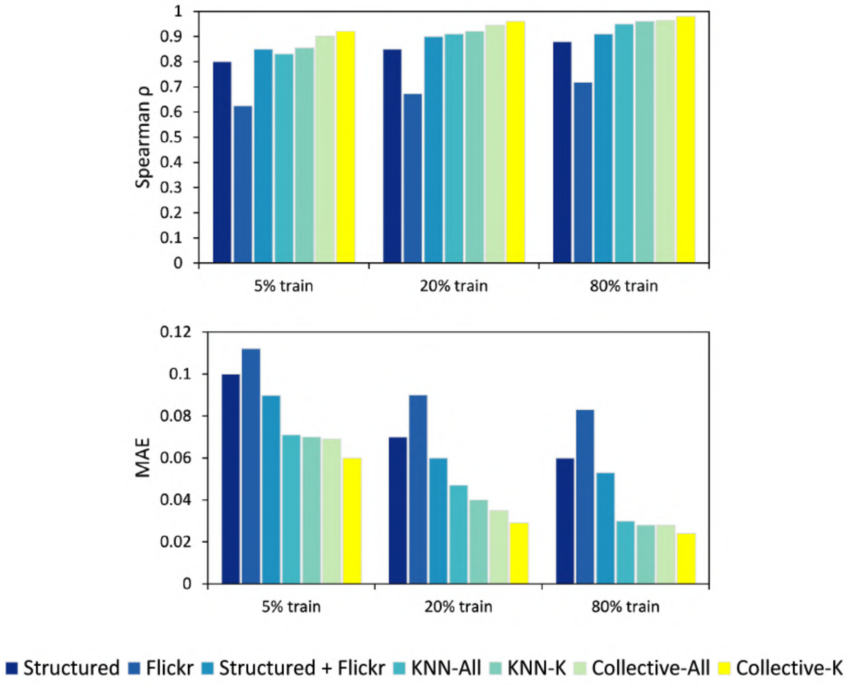


Figure 5.15: Results of predicting average annual water vapour pressure.

Table 5.4: Top 5 Flickr tags for different climate related features in terms of KL divergence.

Temperature	Precipitation	Solar Rad.	Wind Speed	Water Vapour Press.
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

5.4 Summary

In this chapter, we have proposed a collective prediction model which 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 play a key role in determining neighbourhoods, we showed that considerable gains could be made by additionally taking Flickr tags and traditional data into consideration.

In the next chapter, we will try to improve the way we have combined structured features with bag-of-words features by learning a low-dimensional vector space embedding that captures both kinds of data in an efficient way.

Chapter 6

Modelling Locations using Vector Space Embedding

6.1 Introduction

Our main hypothesis in this chapter is that by using vector space embeddings instead of bag-of-words representations, the ecological information which is implicitly captured by Flickr tags can be utilised in a more effective way. Vector space embeddings are representations in which the objects from a given domain are encoded using relatively low-dimensional vectors. They have proven useful in natural language processing, especially for encoding word meaning [77, 86], and in machine learning more generally. In this chapter, we are interested in the use of such representations for modelling geographic locations. Our main motivation for using vector space embeddings is that they allow us to integrate the textual information we get from Flickr with available structured information in a very natural way. To this end, we rely on an adaptation of the GloVe word embedding model [86] that was explained in Section 2.4.1. However, we learn vectors representing locations rather than learning word vectors. Similar to how the representation of a word in GloVe is determined by the context words surrounding it, the representation of a location in our model is determined by the tags of the photographs that have been taken near that location. To incorporate numerical features from structured environmental datasets (e.g. average temperature), we associate with each such feature a linear mapping that can be used to predict that feature from a given location vector. This is inspired by the fact that the salient properties of a given domain can often be modelled as directions in vector space embeddings [45, 23, 93]. Finally, evidence from categorical datasets (e.g. land cover types) is taken into account by requiring that locations belonging to the same category are represented using similar vectors, similar to how semantic types are sometimes modelled in the context of knowledge graph embedding [44].

The remainder of this chapter is organised as follows. The next section presents our model for embedding geographic locations from Flickr tags and structured data. Section 6.3 provides a detailed discussion about the experimental results. Finally, Section 6.4 summarizes our findings.

6.2 Embedding Geographic Location

In this section, we introduce our embedding model, which combines Flickr tags and structured scientific information to represent a set of locations L. The full model is illustrated in Figure 6.1. This figure shows how the Flickr tags representation from Section 4.2.1 are combined with the structured information from Section 4.2.2 into a low dimensional vector space embedding. The proposed embedding model aims to minimise the following objective:

$$J = \alpha J_{tags} + (1 - \alpha) J_{nf} + \beta J_{cat} \tag{6.1}$$

where $\alpha \in [0, 1]$ and $\beta \in [0, +\infty]$ are parameters to control the importance of each component in the model. Component J_{tags} will be used to constrain the representation of the locations based on their textual description (i.e. Flickr tags), J_{nf} will be used to constrain the representation of the locations based on their numerical features, and J_{cat} will impose the constraint that locations belonging to the same category should be

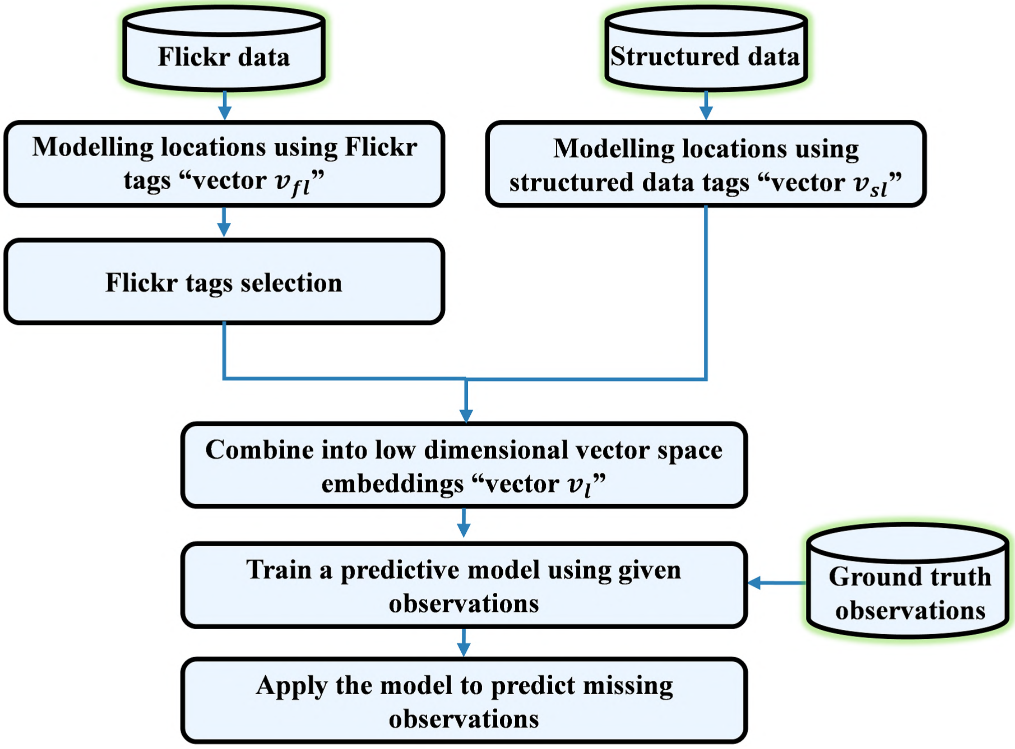

Flickr data
Structured data
Modelling locations using Flickr tags "vector v_{fl}"
Modelling locations using structured data tags "vector v_{sl}"
Flickr tags selection
Combine into low dimensional vector space embeddings "vector v_l"
Train a predictive model using given observations
Ground truth observations
Apply the model to predict missing observations

Figure 6.1: Modelling locations in vector space embedding

close together in the space. We will discuss each of these components in more detail in the following sections.

6.2.1 Tag Based Location Embedding

As illustrated in Figure 6.1, we first need to obtain weighted bag-of-words representations of locations from Flickr. Subsequently, we apply a tag selection method, which will allow us to specialise the embedding depending on which aspects of the considered locations are of interest, after which we can apply the actual embedding model.

Tag weighting

To generate a bag-of-words representation of a given location from Flickr tags, we have to weight the relevance of each tag to that location. To this end, we used the PPMI weighted feature vector v_{fl} from Section 4.2.1.

Tag selection

Inspired by [64], we use a term selection method in order to focus on the tags that are most important for the tasks that we want to consider and reduce the impact of tags that might relate only to a given individual or a group of users. In particular, we obtained good results with the set of tags that are selected with the method based on Kullback-Leibler (KL) divergence, which was explained in Section 5.2.

Location embedding

We now want to find a vector $v_{l_i} \in V$ for each location l_i such that similar locations are represented using similar vectors. To achieve this, we use a close variant of the GloVe model, where tag occurrences are treated as context words of geographic locations. In particular, with each location l we associate a vector v_l and with each tag t we associate a vector $\tilde{w_t}$ and a bias term $\tilde{b_{t_j}}$, and consider the following objective:

$$J_{tags} = \sum_{l_i \in L} \sum_{t_j \in T} (v_{l_i} \tilde{w_{t_j}} + \tilde{b_{t_j}} - \mathit{PPMI}(t_j, l_i))^2 \tag{6.2}$$

This constraint is illustrated in Figure 6.2. Note how tags play the role of the context words in the GloVe model, but instead of learning target word vectors, we now learn location vectors. In contrast to GloVe, our objective does not directly refer to co-occurrence statistics but instead uses the *PPMI* scores. One important consequence is that we can also consider pairs (l_i, t_j) for which t_j does not occur in l_i at all; such pairs are usually called *negative examples*. While they cannot be used in the standard GloVe model, some authors have already reported that introducing negative examples in variants of GloVe can lead to improvements [51]. In practice, evaluating the full objective above would not be computationally feasible, as we may need to consider millions of locations and tags. Therefore, rather than considering all tags in T for the inner summation, we only consider those tags that appear at least once near location l_i together with a sample of negative examples.

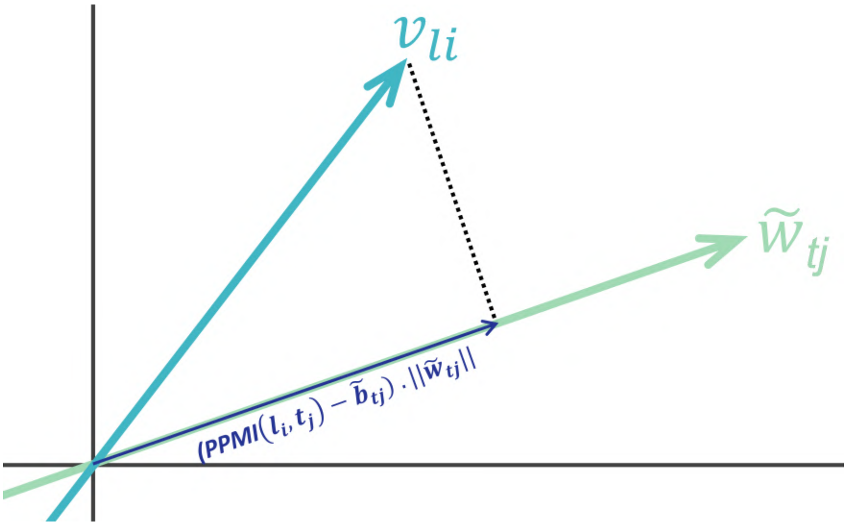


Figure 6.2: The geometric intuition of tags based embedding.

6.2.2 Structured Environmental Data

Here we have used the same datasets as in the previous chapters, where the available information about a location is encoded as the vector v_{sl}. This vector includes nine (real-valued) numerical features, which are latitude, longitude, elevation, population, and five climate-related features (avg. temperature, avg. precipitation, avg. solar radiation, avg. wind speed, and avg. water vapour pressure). In addition, 180 categorical features were used, which are the CORINE land cover classes at level 1 (5 classes), level 2 (15 classes) and level 3 (44 classes) and 116 soil types. Note that each location should belong to exactly four categories: one CORINE class at each of the three levels and a soil type. For more details about these datasets see Section 3.3.

Numerical Features Based Location Embedding

Numerical features can be treated similarly to the tag occurrences, i.e. we will assume that the value of a given numerical feature can be predicted from the location vectors using a linear mapping. In particular, for each numerical feature f_k we consider a vector $\tilde{w_{f_k}}$ and a bias term $\tilde{b}_{f_k}$, and the following objective which is illustrated in Figure 6.3:

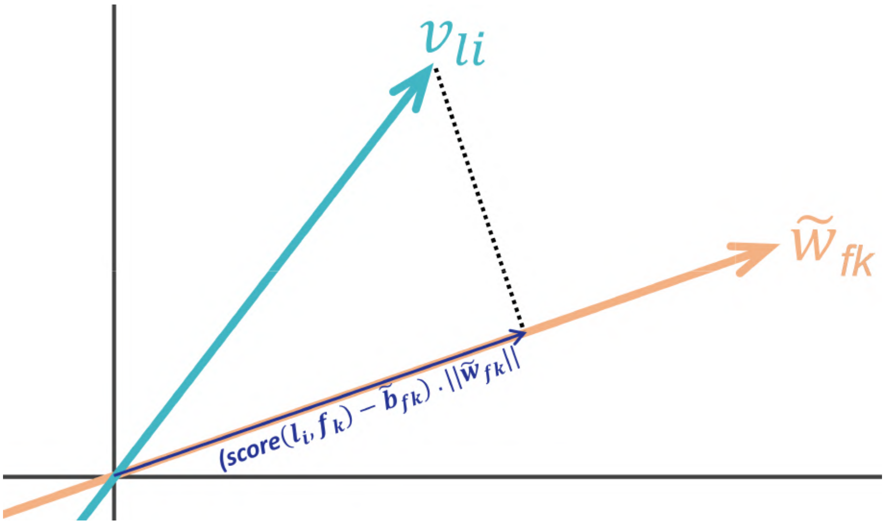

v_{li}
$\widetilde{w}_{fk}$
$(score(l_i, f_k) - \widetilde{b}_{fk}) . ||\widetilde{w}_{fk}||$

Figure 6.3: The geometric intuition of numerical features based embedding.

$$J_{nf} = \sum_{l_i \in L} \sum_{f_k \in NF} (v_{l_i}.\tilde{w_{f_k}} + \tilde{b_{f_k}} - score(f_k, l_i))^2 \tag{6.3}$$

where we write *NF* for the set of all numerical features and $score(f_k, l_i)$ is the value of feature f_k for location l_i, after z-score normalization.

Categorical Features Based Location Embedding

To take into account the categorical features, we impose the constraint that locations belonging to the same category should be close together in the space. To formalize this, we represent each category type cat_l as a vector w_{cat_l}, and consider the following objective which is illustrated in Figure 6.4:

$$J_{cat} = \sum_{l_i \in R} \sum_{cat_l \in C} (v_{l_i} - w_{cat_l})^2 \tag{6.4}$$

All the above mentioned vectors are initialized randomly and then updated iteratively using an Adagrad optimizer to minimize the considered objective function.

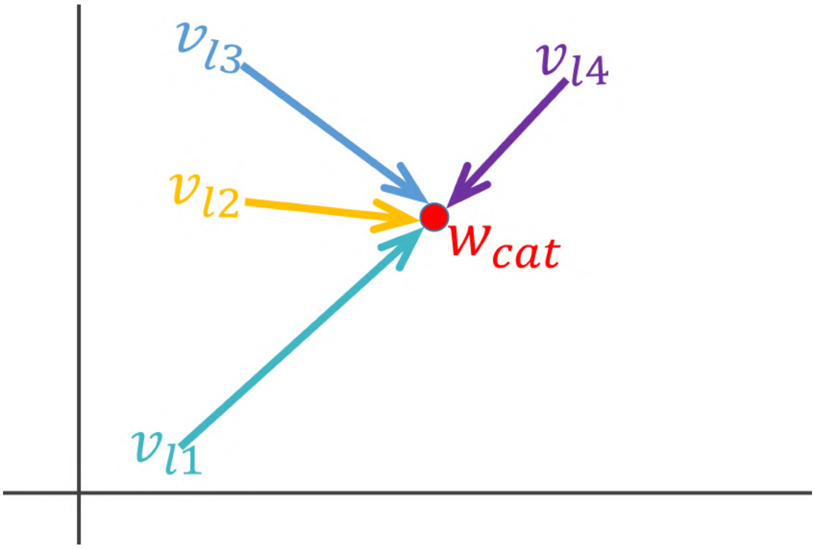


Figure 6.4: The geometric intuition of categorical features based embedding.

6.3 Experimental Evaluation

6.3.1 Experimental Settings

As in all the experiments in this thesis, we use Support Vector Machines (SVMs) for classification problems and Support Vector Regression (SVR) for regression problems to make predictions from our representations of geographic locations. For each experiment, the set of locations L was split into two-thirds for training, one-sixth for testing, and one-sixth for tuning the parameters. All embedding models are learned with Adagrad using 30 iterations. The number of dimensions is chosen for each experiment from $\{10, 50, 300\}$ based on the tuning data. For the parameters of our model in Equation 6.1, we considered values of α from {0.1, 0.01, 0.001, 0.0001} and values of β from {1, 10, 100, 1000 }. In all experiments where term selection is used, we select the top 100 000 tags. Finally, we set the number of negative examples as ten times the number of positive examples for each location, but with a cap at 1000 negative examples in each region for computational reasons. We tune all parameters with respect to the F1 score for the classification tasks, and Spearman ρ for the regression tasks.

6.3.2 Variants and Baseline Methods

We will refer to our model as EGEL (Embedding GEographic Locations). The source code is available online at `https://github.com/shsabah84/EGEL-Model.git`. For evaluation, we will consider the following variants:

- "EGEL-Tags" only uses the information from the Flickr tags (i.e. component J_{tags}), without using any negative examples and without feature selection.
- "EGEL-Tags+NS" is similar to "EGEL-Tags" but with the addition of negative examples.
- "EGEL-KL(Tags+NS)" additionally considers term selection.
- "EGEL-NF+KL(Tags+NS)" uses components J_{tags} and J_{nf} (with negative examples and term selection).
- "EGEL-Cat+KL-Tags+NS" instead uses components J_{tags} and J_{cat}.
- "EGEL-All" is our full method, i.e. it additionally uses the structured information.

We also consider the following baselines:

- "BOW-Tags" represents locations using a bag-of-words representation, with the same tag weighting as the embedding model. In particular, this is the same variant that called "Flickr" in Chapter 4 and 5.
- "BOW-KL(Tags)" uses the same representation of "BOW-Tags" but after term selection, using the same KL-based method as the embedding model.
- "BOW-All" combines the bag-of-words representation with the structured information, i.e. "Structured + Flickr" that proposed in Section 4.2.

- "GloVe" uses the objective from the original GloVe model for learning location vectors, i.e. this variant differs from "EGEL-Tags" in that instead of *PPMI*(t_j, l_i) we use the number of co-occurrences of tag t_j near location l_i.

6.3.3 Experimental Results

In this chapter, weuse the same set of experiments as in Section 4.3. In particular, we will consider the problems of predicting the distribution of 100 species across Europe, predicting soil type and predicting CORINE land cover classes at levels 1, 2 and level 3 as binary classification tasks. In addition, we considered six regression tasks: predicting five climate-related features and predicting people's subjective opinions of landscape beauty. For more information about these experiments, see Section 4.3. We present our results for the binary classification tasks in Tables 6.1 – 6.3 in terms of average precision, average recall and macro average F1 score. The results of the regression tasks are reported in Tables 6.4 and 6.5 in terms of the mean absolute error between the predicted and actual scores, as well as the Spearman ρ correlation between the rankings induced by both sets of scores. It can be clearly seen from the results that our proposed method (EGEL-All) can effectively integrate Flickr tags with the available structured information. It outperforms the baselines for all the considered tasks. Furthermore, note that the PPMI-based weighting in EGEL-Tags consistently outperforms GloVe and that both the addition of negative examples and term selection lead to further improvements. The use of term selection leads to particularly substantial improvements for the regression problems.

While our experimental results confirm the usefulness of embeddings for predicting environmental features, this is only consistently the case for the variants that use both the tags and the structured datasets. In particular, comparing BOW-Tags with EGEL-Tags, we sometimes see that the former achieves the best results. While this might seem surprising, it is in accordance with the findings in [58, 129], where it was also found that bag-of-words representations can sometimes lead to surprisingly effective

Table 6.1: Results for predicting species distribution.

	Prec	Rec	F1
BOW-Tags	0.577	0.112	0.188
BOW-KL(Tags)	0.109	0.869	0.193
GloVe	0.100	0.888	0.179
EGEL-Tags	0.102	0.884	0.182
EGEL-Tags+NS	0.124	0.827	0.215
EGEL-KL(Tags+NS)	0.157	0.644	0.252
BOW-All	0.650	0.506	0.569
EGEL-NF+KL(Tags+NS)	0.331	0.565	0.417
EGEL-Cat+KL(Tags+NS)	0.278	0.597	0.379
EGEL-All	0.563	0.601	**0.581**

baselines. Interestingly, we note that in all cases where EGEL-KL(Tags+NS) performs worse than BOW-Tags, we also find that BOW-KL(Tags) performs worse than BOW-Tags. This suggests that for these tasks there is a very large variation in the kind of tags that can inform the prediction model, possibly including user-specific tags. Some of the information captured by such highly specific but rare tags is likely to be lost in the embedding.

To further analyse the difference in performance between BoW representations and embeddings, Figure 6.5 compares the performance of the GloVe model with the bag-of-words model for predicting place scenicness, as a function of the number of tag occurrences at the considered locations. What is clearly noticeable in Figure 6.5 is that GloVe performs better than the bag-of-words model for large sets of tags and worse for smaller sets. This issue has been alleviated in our embedding method by the addition of negative examples.

Table 6.2: Results for predicting soil type.

	Prec	Rec	F1
BOW-Tags	0.170	0.445	0.246
BOW-KL(Tags)	0.309	0.439	0.362
GloVe	0.329	0.392	0.358
EGEL-Tags	0.325	0.402	0.360
EGEL-Tags+NS	0.308	0.442	0.363
EGEL-KL(Tags+NS)	0.320	0.441	0.371
BOW-All	0.398	0.438	0.417
EGEL-NF+KL(Tags+NS)	0.318	0.633	0.424
EGEL-Cat+KL(Tags+NS)	0.374	0.396	0.385
EGEL-All	0.337	0.673	**0.449**

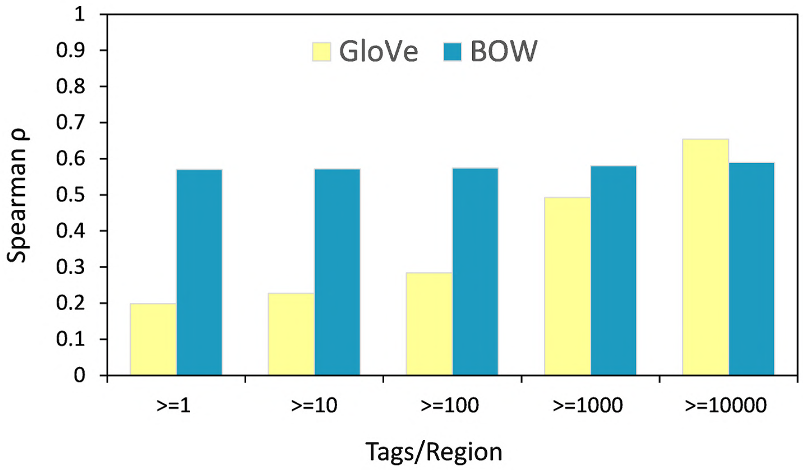


Figure 6.5: Comparison between the performance of the GloVe and bag-of-words models for predicting scenicness, as a function of the number of tag occurrences at the considered locations.

Table 6.3: **Results for predicting CORINE land cover classes, at levels 1, 2 and 3.**

	CORINE level 1			CORINE level 2			CORINE level 3		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
BOW-Tags	0.499	0.457	0.477	0.205	0.139	0.166	0.145	0.086	0.108
BOW-KL(Tags)	0.402	0.471	0.433	0.390	0.125	0.189	0.243	0.130	0.170
GloVe	0.202	0.908	0.331	0.122	0.535	0.198	0.129	0.254	0.171
EGEL-Tags	0.205	0.898	0.334	0.123	0.562	0.201	0.167	0.218	0.189
EGEL-Tags+NS	0.237	0.739	0.359	0.125	0.529	0.203	0.180	0.221	0.199
EGEL-KL(Tags+NS)	0.263	0.625	0.370	0.146	0.582	0.234	0.199	0.253	0.223
BOW-All	0.523	0.514	0.518	0.270	0.199	0.229	0.184	0.112	0.139
EGEL-NF+KL(Tags+NS)	0.452	0.673	0.541	0.266	0.493	0.346	0.182	0.360	0.242
EGEL-Cat+KL(Tags+NS)	0.268	0.616	0.373	0.188	0.437	0.263	0.202	0.253	0.225
EGEL-All	0.455	0.668	**0.542**	0.276	0.487	**0.352**	0.236	0.332	**0.276**

Table 6.4: Results for predicting average climate data.

	Temperature		Precipitation		Solar rad		Water vap		Wind speed	
	MAE	ρ	MAE	ρ	MAE	ρ	MAE	ρ	MAE	ρ
BOW-Tags	1.623	0.814	11.660	0.689	926.3	0.832	0.083	0.719	0.545	0.756
BOW-KL(Tags)	1.694	0.812	12.856	0.658	1057.1	0.756	0.084	0.715	0.531	0.737
GloVe	1.960	0.445	15.376	0.311	1507.2	0.365	0.114	0.470	0.741	0.285
EGEL-Tags	1.956	0.471	15.031	0.319	1426.6	0.411	0.107	0.461	0.734	0.328
EGEL-Tags+NS	1.979	0.448	14.937	0.321	1330.7	0.449	0.105	0.465	0.729	0.360
EGEL-KL(Tags+NS)	1.486	0.739	13.556	0.527	1008.4	0.773	0.089	0.662	0.658	0.597
BOW-All	0.728	0.940	10.523	0.755	484.8	0.939	0.053	0.914	0.429	0.846
EGEL-NF+KL(Tags+NS)	0.725	0.941	11.047	**0.799**	629.6	0.94	0.054	0.904	0.489	0.838
EGEL-Cat+KL(Tags+NS)	1.337	0.857	12.716	0.651	936.9	0.871	0.079	0.796	0.585	0.721
EGEL-All	**0.711**	**0.949**	**10.033**	0.699	**436.6**	**0.951**	**0.050**	**0.922**	**0.432**	**0.880**

Table 6.5: Results for predicting scenicness.

	MAE	ρ
BOW-Tags	1.013	0.570
BOW-KL(Tags)	1.096	0.515
GloVe	1.275	0.198
EGEL-Tags	1.121	0.376
EGEL-Tags+NS	1.145	0.407
EGEL-KL(Tags+NS)	1.058	0.537
BOW-All	1.006	0.581
EGEL-NF+KL(Tags+NS)	0.973	0.621
EGEL-Cat+KL(Tags+NS)	0.992	0.604
EGEL-All	**0.946**	**0.645**

6.4 Summary

In this chapter, we have proposed a model to learn geographic location embeddings using Flickr tags, numerical environmental features, and categorical information. While our point-of-departure is a standard word embedding model, we found that the off-the-shelf GloVe model performed surprisingly poorly, meaning that a number of modifications are needed to achieve good results. Our main findings are as follows. First, given that the number of tags associated with a given location can be quite small, it is important to apply some kind of spatial smoothing, i.e. the importance of a given tag for a given location should not only depend on the occurrences of the tag at that location, but also on its occurrences at nearby locations. To this end, we used a formulation that was introduced in Chapter 4. This method is based on a spatially smoothed version of pointwise mutual information. Second, given the wide diversity in the kind of information that is covered by Flickr tags, we find that term selection is in some cases critical to obtain vector spaces that capture the relevant aspects of geographic locations. For instance, many tags on Flickr refer to photography related terms, which we would normally not want to affect the vector representation of a given location. One exception is perhaps when we want to predict the scenicness of a given location, where e.g. terms that are related to professional landscape photography might be a strong indicator of scenicness. Finally, even with these modifications, vector space embeddings learned from Flickr tags alone are sometimes outperformed by bag-of-words representations. However, our vector space embeddings lead to substantially better predictions in cases where structured (scientific) information is also taken into account. In this sense, the main value of using vector space embeddings in this context is not so much about abstracting away from specific tag usages, but rather about the fact that such representations allow us to integrate textual, numerical and categorical features in a much more effective way than is possible with bag-of-words representations.

Given these encouraging results, in the next chapter, we will extend the proposed model by considering a spatiotemporal representation of regions.

Chapter 7

Spatiotemporal Embeddings Model

7.1 Introduction

In this chapter, we extend our approach from Chapter 6 by considering a spatiotemporal representation of regions. In particular, we learn a vector space embedding for each geographic region and each month of the year, which allows us to capture environmental phenomena that may depend on monthly or seasonal variation. Apart from extending our main model, we also introduce a new smoothing method to deal with the sparsity of Flickr tags. This is motivated by the fact that when fine-grained regions are used, and data may be sparse, the number of times that a tag is used in a particular region and month is not a reliable indicator by itself of the relevance of that tag. For evaluation, we consider the problem of predicting climate features and predicting the distribution of species in a given location and a given month. The proposed method has proven to be advantageous compared with baselines that rely only on Flickr or only on traditional sources, in particular when we have a very small training data set. We also qualitatively evaluate the proposed model by generating similarity maps for a number of selected locations.

The remainder of this chapter is organised as follows. In the next section, we provide a discussion of the related work in the area of spatiotemporal analysis and modelling. Section 7.3 describes our methodology. In particular, Section 7.3.1 and Section 7.3.2 present our methodology for spatiotemporal modelling using Flickr tags and using structured data respectively, and Section 7.3.3 then describes our spatiotemporal embeddings model. In Section 7.4, we provide a detailed discussion about the experimental results as well as the qualitative evaluations. Finally, Section 7.5 summarises our conclusions.

7.2 Related Work: Spatiotemporal Modelling

Spatiotemporal analysis and modelling have been a major interest in many research areas. Examples include environmental science [106, 76], social science [11, 50], and business [35, 36]. In particular, [36] developed a geographical and temporal weighted regression (GTWR) model to account for the location variations in time and space when modelling house prices in London from 1980 to 1998. The model is based on a spatiotemporal kernel function using a Gaussian distribution. Similar to our work, they allocated each spatial point to a time interval. However, while they model time on a linear scale, we use a circular scale since our focus is on modelling seasonality. In [11], a spatiotemporal kernel density estimation (STKDE) has been proposed which is based on multiplying the spatial kernel function by the temporal kernel function. It is a space-time cube method that extends the 2-dimensional grid used in the spatial kernel to a 3-dimensional cube and computes density values at cube centres with overlapping space-time cylinders. Time was represented on a circular scale that uses Von Mises distribution as a time kernel. STKDE has shown promising results in many applications, such as crime hotspot detection [50], and disease patterns detection [22]. In this chapter, we use the STKDE method [11] to smooth the distribution of Flickr tags over space and time, as a way of alleviating the sparsity of Flickr tags.

7.3 Methodology

Our aim in this chapter is to learn a low-dimensional vector space embedding of a set of spatiotemporal regions. Based on our findings in Chapter 6, this representation will allow us to combine the textual information derived from Flickr with the numerical, categorical, spatial, and temporal information in an efficient way. Thus the ecological information can be effectively captured by the predictive model. We will start this section by explaining how feature vectors encoding spatiotemporal regions can be obtained from the tags associated with Flickr photographs. In Section 7.3.2, we then describe how feature vectors encoding spatiotemporal regions can be obtained from the structured information sources that we will additionally consider. Finally, we will introduce our proposed spatiotemporal embedding (SPATE) model that combines both data sources.

7.3.1 Spatiotemporal Modelling Using Flickr tags

To model the spatiotemporal regions using Flickr tags, we first need to obtain weighted bag-of-words representations. We then apply a tag selection method, which will allow us to specialise tags that are related to space and/or time. Subsequently, we smooth out the tags distribution over the space and time to tackle the problem of data sparsity.

Tags Weighting

With the objective of using Flickr tags for spatiotemporal modelling, we split the target spatial area into $10km \times 10km$ grid cells. Furthermore, we discretise the timestamps with a granularity of 1 month. We thus view the overall dataset as 12 separate grid layers, each layer corresponding to a month of the year. Thus there are 12 instances for each spatial cell as illustrated in Figure 7.1. The choice of the $10km \times 10km$ spatial granularity and the one-month temporal granularity is to balance between resolution

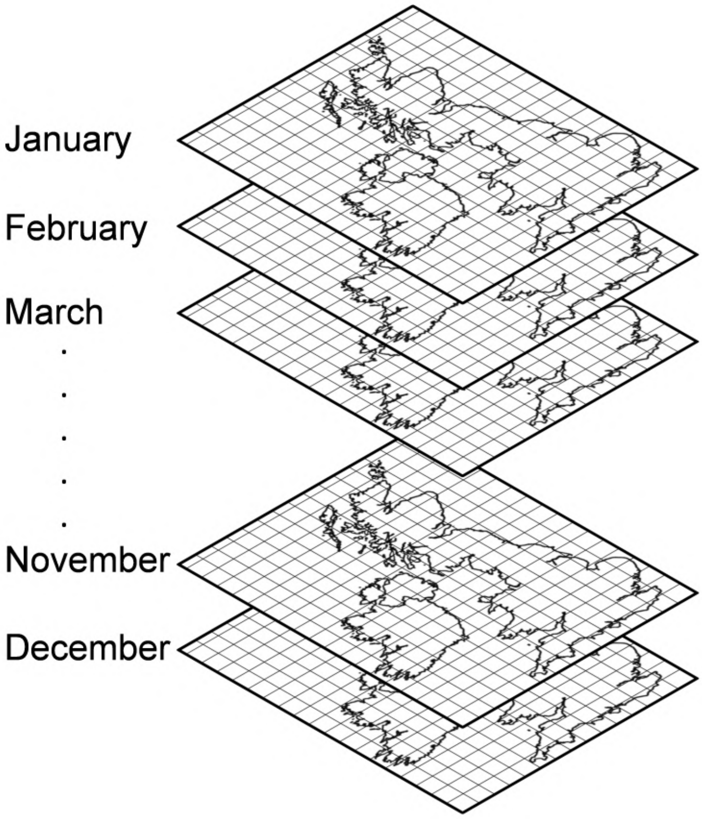


Figure 7.1: Spatiotemporal grid cells.

and computation time. Let $c_1, ..., c_n$ be the spatiotemporal grid cells, each represented by a triple (*lat*, *lon*, *m*) where *lat* is the latitude coordinate, *lon* is the longitude coordinate, and *m* is the month of the year. We associate each such cell with a histogram of Flickr tags, reflecting how many times each tag has been added to a photograph whose coordinates and time stamp fall within the cell.

Let *f(t,c)* be the number of times tag t (from the set of all tags T) occurs in the cell c. We then use Positive Pointwise Mutual Information (PPMI), which has been used in previous chapters and given by Equation 2.4, to weight how strongly tag t is associated with cell c. Each cell c can thus be represented as a sparse vector $v_f(c)$ which is defined as $(\textit{PPMI}(t_1, c), ..., \textit{PPMI}(t_k, c))$, where $t_1, ..., t_k$ is an enumeration of the tags in T.

Tag Selection

Our aim here is to select tags whose occurrence is correlated with specific times of the year (e.g. summer) or with photographs that occur in particular geographic regions (e.g. forests). When constructing the feature representation $v_f(c)$, we then only consider those tags that have been selected. Similar to Chapters 5 and 6, we will use Kullback-Leibler (KL) divergence method to select tags. However, our aim here is to determine whether a given tag is time and/or location specific. Intuitively, we assess to what extent the distribution of its occurrences across all spatiotemporal cells diverges from the overall distribution of all tag occurrences. In particular, we select those tags $T_{KL} \subseteq T$ which maximize KL divergence given by Equation 2.10. Here $P(c|t)$ is defined as the probability that a photograph with tag t has a location and time in c and $Q(c)$ is the probability that an arbitrary tag occurrence is assigned to a photograph in c, both are estimated in the same way as in Chapter 5 (Equations 5.2 and 5.4).

Now we will use the notation $v_{KL}(c)$ for the sparse vector representation of cell c encoding the *PPMI* weight of those tags in T_{KL} only.

Spatiotemporal Smoothing

The vector representation $v_{KL}(c)$ encodes which tags are most strongly correlated with the spatiotemporal grid cell c. However, these scores are computed from sometimes very limited amounts of data, and for some cells we may not have any photographs at all. This is because we are only looking at the tags for a particular month, so on average, for a given geographic region, we have 1/12th of the tags that we had in the previous chapter. To tackle this problem, we used kernel density estimation to smooth the *PPMI* weight of each tag in T_{KL} over a larger region. For this purpose, we used the spatiotemporal kernel density estimation method that was introduced in [11]. In particular, we define the smoothed weight of tag t in cell c as follows:

$$\textit{KDE}(t, c) = \frac{\hat{s}(t, c)}{\max_c (\hat{s}(t, c))} \tag{7.1}$$

The reason for normalising the *KDE* value is to keep the weight of all the tags within the same range and avoid the impact of the dominant tags. The $\hat{s}(t,c)$ value is computed as:

$$\hat{s}(t,c) = \sum_{i=1}^{n_t} \mathit{PPMI}(t,c_i) \cdot K_s\left(\Lambda_{\mathit{lat}^i}, \Lambda_{\mathit{lon}^i}\right) \cdot K_m\left(\Lambda_{m^i}\right) \tag{7.2}$$

where n_t is the number of cells c with tag t, $\Lambda_{\mathit{lat}^i} = \frac{c_{\mathit{lat}} - c_{\mathit{lat}^i}}{h_s}$, $\Lambda_{\mathit{lon}^i} = \frac{c_{\mathit{lon}} - c_{\mathit{lon}^i}}{h_s}$, and $\Lambda_{m^i} = \frac{c_m - c_{m^i}}{h_m}$. Here c_{lat}, c_{lon} and c_m are respectively the latitude, longitude and the month of cell c, while c_{lat^i}, c_{lon^i} and c_{m^i} are respectively the latitude, longitude and the month of cell c_i. With h_s the spatial smoothing bandwidth of tag t, and h_m the temporal smoothing bandwidth of tag t. For the spatial kernel function K_s, we use a Gaussian distribution [108] given by:

$$K_s(c_{\mathit{lat}}, c_{\mathit{lon}}) = \frac{1}{2\pi} \exp\left(-\frac{(c_{\mathit{lat}} - c_{\mathit{lat}^i})^2 + (c_{\mathit{lon}} - c_{\mathit{lon}^i})^2}{2h_s^2}\right) \tag{7.3}$$

As the temporal kernel K_m, we use a von Mises distribution [113] which is a continuous probability distribution on the circle. The Von Mises distribution was chosen because of its wrap-around property (it is sometimes called circular Gaussian) which is well suited to the cyclic nature of the months of the year representation. Here, we first encode months using values in {0,...,11}, then the month value is mapped to its corresponding point on the circle by:

$$\theta(c_m) = \frac{2\pi c_m}{12} \tag{7.4}$$

Now ‘January’ is represented as $\frac{\pi}{6}$, ’February’ is represented as $\frac{\pi}{3}$ and so on, as explained in Figure 7.2.

The von Mises distribution is defined by:

$$K_m(\theta(c_m)) = \frac{1}{2\pi I_0(h_m)} \exp\left(h_m \cos(\theta - \Theta)\right) \tag{7.5}$$

where I_0 is the modified Bessel function of order 0.

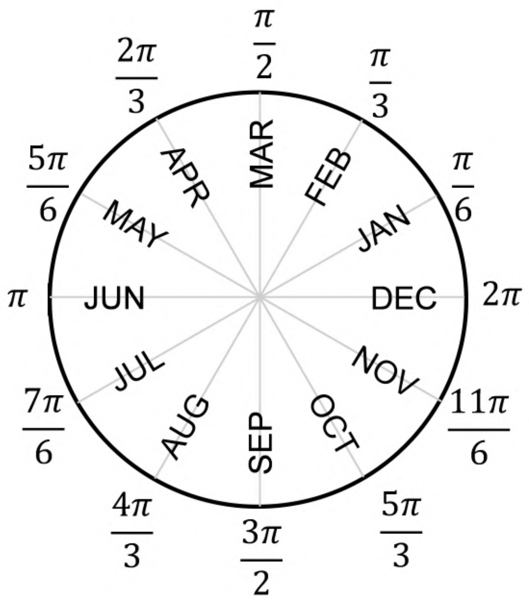


Figure 7.2: The representation of the months as circular data.

Finally, to model the spatiotemporal grid cell c using Flickr tags, we consider a vector $v_{KDE}(c)$ encoding the smoothed weight of all the tags $\{t_1...t_{nt}\} \in T_{KL}$ which is defined as $(KDE(t_1, c), ..., KDE(t_{n_t}, c))$. This vector will be used to train the proposed embeddings model in Section 7.3.3.

Bandwidth Selection

The critical parameter in any kernel-based method is the selection of the optimal bandwidth. The variables h_s and h_m are of key importance, and their values are generally considered to be more important than the type of the kernel itself. In general, large values lead to over-smoothing, while small values lead to under-smoothing. Various methods have been developed for selecting the optimal kernel bandwidth. In this chapter, we experimentally compare the performance of three of the most widely used methods.

1. The rule of thumb [108] is a simple and fast method. It estimates a fixed kernel

bandwidth based on the data driven scale of the distribution which is defined as:

$$h = \hat{\alpha} \left(\frac{n * (d+2)}{4} \right)^{-1/(d+4)} \tag{7.6}$$

where n is the size of the data, d is the number of dimensions, and $\hat{\alpha}$ is the data standard deviation. Here we need to estimate two different bandwidths (the spatial and the temporal bandwidths). They are both estimated using Equation 7.6. However, for estimating the temporal bandwidth h_m, d is equal to 1 and $\hat{\alpha}$ is the circular standard deviation. For estimating the spatial bandwidth h_s, d is equal to 2 and $\hat{\alpha} = \frac{s_1+s_2}{2}$ where s_1 and s_2 are the standard deviations of the latitude and the longitude coordinates respectively.

2. The adaptive kernel bandwidth [1, 10] is based on the idea of making the value of h vary between different regions according to the local density. In particular, a wider bandwidth is selected for regions with low density while a narrower bandwidth is selected for regions with high density. It is usually achieved by the following steps. Firstly, compute a pilot estimate of $\hat{s}(t,c)$ (Equation 7.2) using the fixed bandwidth as described above in Equation 7.6. This estimate is used to give an overall approximation of the smoothed value of the data. Secondly, compute a local bandwidth scalar, which is computed by:

$$b_c = \sqrt{\frac{g}{\hat{s}(t,c)}} \tag{7.7}$$

where g is the geometric mean of $\hat{s}(t,c_1), ..., \hat{s}(t,c_n)$, which is given by:

$$g = \left(\prod_{i=1}^{n} \hat{s}(t,c_i) \right)^{1/n} \tag{7.8}$$

Finally, the adaptive local bandwidths are given by $h_{s(c)} = h_s \cdot b_c$ and $h_{m(c)} = h_m \cdot b_c$ which can be used in Equation 7.2 to make the final estimation for tag t.

3. The leave-one-out kernel estimator [9] is based on the idea of selecting the kernel bandwidth estimator that minimizes the mean integrated square error (MISE)

[102] given by:

$$\mathit{MISE} = \frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{s}(t,c_i) - p(t,c_i))^2}{p(t,c_i)} \tag{7.9}$$

where $\hat{s}(t,c_i)$ is the estimated density of tag t at the grid cell c_i after removing the cell c_i from the data. Furthermore, $p(t,c_i)$ is the probability of the *PPMI* weight of tag t at the grid cell c_i (i.e. the true density), which is computed as:

$$p(t,c_i) = \frac{\mathit{PPMI}(t,c_i)}{\sum_{c' \in C}\mathit{PPMI}(t,c')} \tag{7.10}$$

And $\hat{s}_{-i}(t,c_i)$ is computed here as:

$$\frac{\sum_{j=1 j\neq i}^{n_t}\mathit{PPMI}(t,c_j)K_s(\Lambda_{\mathit{lat}^{ij}},\Lambda_{\mathit{lon}^{ij}})K_m(\Lambda_{m^{ij}})}{\sum_{j=1 j\neq i}^{n}K_s(\Lambda_{\mathit{lat}^{ij}},\Lambda_{\mathit{lon}^{ij}})K_m(\Lambda_{m^{ij}})} \tag{7.11}$$

The optimal bandwidths h_s and h_m that minimize Equation 7.9 can be used to smooth the tag t distribution over all the spatiotemporal grid cells in Equation 7.2.

7.3.2 Spatiotemporal Modelling Using Structured Environmental Data

In this section, we used the following external datasets as sources of numerical features (see Section 3.3 for details):

- Monthly average of temperature, precipitation, solar radiation, wind speed and water vapour pressure.
- Elevation.
- Population.

Several of the considered datasets have a resolution which is finer than our $10km \times 10km$ grid cells. To this end, we look up the feature values at 100 locations, distributed uniformly within the grid cell. To obtain a feature vector for the spatiotemporal grid cell c representing these numerical features, we first average these 100 values for each numerical feature across the grid cell. Then we normalise these features values using the standard z-score.

In addition, we used the following datasets as sources of categorical features:

- CORINE land cover type at level 1, 2, and level 3.
- Soil type.

The categorical features are represented as a vector, encoding for each of the categories what percentage of the grid cell (i.e. the average of the 100 locations) belongs to that category.

Apart from the features from these external datasets, the geographic coordinates and time stamp of the cell c are clearly also important structured features, which should be included in the feature vector describing a spatiotemporal cell. For the spatial features, each grid cell c has been represented by the normalised coordinate values which computed as:

$$norm(lat, c) = \frac{lat - min(latitude)}{max(latitude) - min(latitude)} \tag{7.12}$$

$$norm(lon, c) = \frac{lon - min(longitude)}{max(longitude) - min(longitude)} \tag{7.13}$$

where lat and lon are respectively the latitude and longitude coordinates of the centre of the grid cell c. And $max(latitude)$, $max(longitude)$, $min(latitude)$, and $min(longitude)$ are the maximum and minimum latitude and longitude over the study area. The reason for normalising these features is to ensure that they are within the same range as the other features. Note that we have also tried projecting the latitude and longitude coordinates into three-dimensional geographic coordinates, but that gave worse results. Finally, the month m corresponding to the cell c is represented as the coordinates $(cos(\theta(m)), sin(\theta(m)))$ of that month, as before (see Figure 7.2).

We will use the notation $v_s(c)$ for the feature vector representation of cell c encoding all the above mentioned structured features.

7.3.3 Spatiotemporal Embeddings

Our aim in this Section is to learn a low-dimensional vector space embedding of a set of spatiotemporal cells C. This representation will allow us to combine the textual information derived from Flickr with the corresponding numerical, categorical, spatial, and temporal information in an efficient way. The proposed embedding model has the following objective function:

$$J = (1 - 2\alpha - 2\beta)J_{\mathit{tags}} + \alpha(J_{\mathit{nf}} + J_{\mathit{cat}}) + \beta(J_{\mathit{spatial}} + J_{\mathit{temp}}) \tag{7.14}$$

where $\alpha, \beta \in [0, 1]$ are parameters to control the importance of each component in the model with $2\alpha + 2\beta < 1$. The components J_{tags}, J_{nf}, J_{cat}, J_{spatial} and J_{temp} intuitively encode the information we have about the spatiotemporal cells from the different sources. The objective function J thus encodes the available information in the form of an optimization problem. In particular, our goal is to learn vector representations for the spatiotemporal cells which minimize J.

Component J_{tags} will be used to constrain the representation of the cells based on their textual description (i.e. Flickr tags), J_{nf} will be used to constrain the representation of the cells based on their numerical features, J_{cat} will impose the constraint that cells belonging to the same category should be close together in the space, J_{spatial} will be used to constrain the representation of the cells based on their spatial feature (i.e. the latitude and longitude coordinates), and J_{temp} will be used to constrain the representation of the cells based on their temporal feature (i.e. month of the year). The components J_{nf} and J_{cat} share the same weight (α) as they have the same key importance in our model and a relatively similar number of features (i.e. similar impact on the embeddings model). The components J_{spatial} and J_{temp} share the same weight (β) for the same reasons. However, the component J_{tags} has a different weight as it involves a larger number of features, these features are of a different nature and their relative importance may also be quite different (e.g. the number of occurrences of a single tag is likely to be less important than the land cover class).

Tags Based Embedding

We now want to find a vector $v_{emb}(c) \in V$ for each spatiotemporal grid cell c. The component J_{tags} intuitively encodes the requirement that we want spatiotemporal cells whose associated Flickr tag distributions are similar to be represented by similar vectors. This is achieved by requiring that the scores $\mathit{KDE}(t_j, c)$ for each tag t_j can be predicted from the vector representation of the cell c. To this end, we use the same tag based objective function as in Chapter 6 (*EGEL* model) that is given by Equation 6.2.

Numerical Features Based Embedding

Numerical features have been treated similarly to the $\mathit{KDE}(t_j, c)$ scores. In particular, we consider the same objective function as is given by Equation 6.3.

Categorical Features Based Embedding

For the categorical features, we again consider the same objective function as in Chapter 6, which is given by Equation 6.4.

Spatial Features Based Embedding

Latitude and longitude coordinates can be incorporated in the same way as the numerical features. However, we treat them as a separate constraint because this allows us to tune the importance of the geographic location of a grid cell c, relative to the numerical and categorical features, based on how we choose the parameters α and β. Therefore, for $s_c \in \{lat, lon\}$, we consider a vector $\tilde{w_{s_c}}$ and a bias term $\tilde{b_{s_c}}$, and the same objective function given by Equation 6.3.

Temporal Features Based Embedding

We represent the temporal features, specifically the months of the year, as equidistant points on the unit circle (as shown in Figure 7.2). To encode temporal information in the embedding, we assume that there is a linear transformation that maps the vector representations of the spatiotemporal cells onto a 2-dimensional plane, such that all cells from a given month are (approximately) projected onto the vector representation of that month. This is similar to how we handle the spatial features, where the two linear lat/lon constraints could also be seen together as mapping the grid cells onto a 2-dimensional plane such that the projection reflects their geographic location. To formalize this constraint, we encode each month m_i by a 2-dimensional vector $\tilde{w_m}$ representing the coordinates $(\mathit{cos}(\theta(m)), \mathit{sin}(\theta(m)))$ of m_i on the temporal circle. We define a projection matrix P as a $2 \times n$ matrix that maps the spatiotemporal cell vector $v_{\mathit{emb}}(c)$ into 2-dimensional space and a 2-dimensional bias term $\tilde{b_m}$, and consider the following objective:

$$J_{\mathit{temp}} = \sum_{c \in C} ||v_{\mathit{emb}}(c).P + \tilde{b_m} - \tilde{w_m}||^2 \tag{7.15}$$

7.4 Experimental Evaluation

In this section, we will formally evaluate our proposed SPAtioTemporal Embeddings (SPATE) model. The SPATE source code is available online at `https://github.com/shsabah84/SPATE-model.git`. The full model is illustrated in Figure 7.3. This figure shows how the Flickr tags representation from Section 7.3.1 is combined with the structured information from Section 7.3.2 to represent the spatiotemporal cells C that can be used to predict values at un-sampled regions.

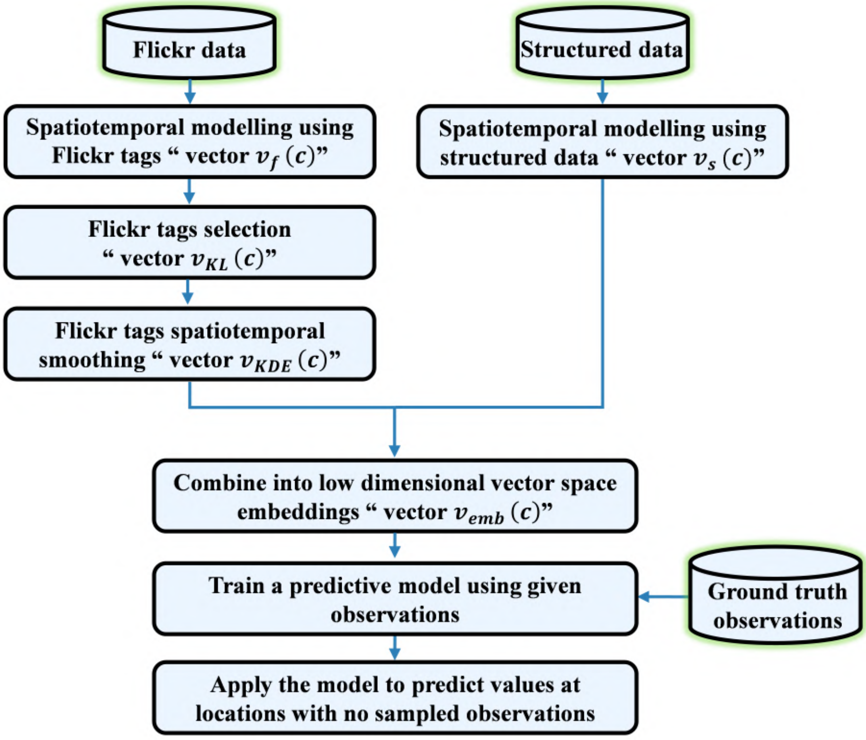

Flickr data
Structured data
Spatiotemporal modelling using Flickr tags " vector $v_f(c)$"
Spatiotemporal modelling using structured data " vector $v_s(c)$"
Flickr tags selection " vector $v_{KL}(c)$"
Flickr tags spatiotemporal smoothing " vector $v_{KDE}(c)$"
Combine into low dimensional vector space embeddings " vector $v_{emb}(c)$"
Train a predictive model using given observations
Ground truth observations
Apply the model to predict values at locations with no sampled observations

Figure 7.3: The spatiotemporal embeddings (SPATE) model.

We will start this section by evaluating the bandwidth selection methods that were described in Section 7.3.1 and choose the best method for our problem. Then we will define our experimental setting and the proposed baseline methods. Subsequently, we will introduce our experiments and provide a detailed discussion about the results. Finally, we will qualitatively evaluate our generated vectors.

7.4.1 Selecting the Optimal Bandwidth for Each Tag

We evaluate the performance of the considered bandwidth selection methods from Section 7.3.1 in term of MISE (see Equation 7.9) on a randomly selected sample of 100 cells for each tag in T_{KL}. For the leave-one-out kernel estimator method, we considered the range {2, 1, 0.5, 0.25, 0.125, 0.05, 0.025, 0.0125, 0} in latitude/longitude degrees for the spatial bandwidth h_s value and the range $\{2\pi, \pi, \pi/2, \pi/6, 0\}$ for the

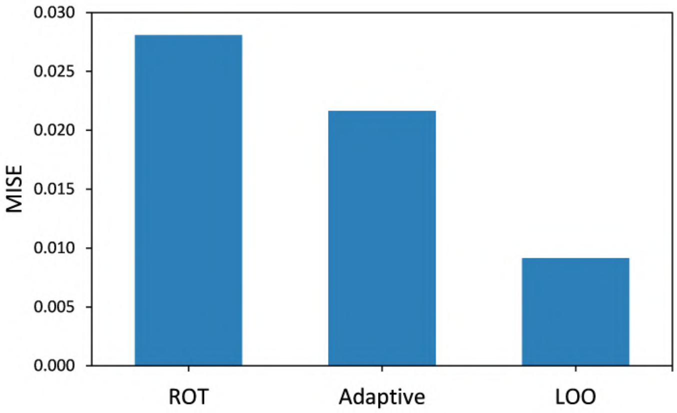


Figure 7.4: The average MISE of all the considered tags when using the rule of thumb (ROT), the adaptive kernel bandwidth (Adaptive), and the leave-one-out kernel estimator (LOO).

temporal bandwidth h_m value. The choice of these two ranges was found to be reasonable for most of the tags based on a small set of initial experiments. Note that a spatial bandwidth of value 0 would mean only temporal smoothing is applied, and vice versa if the temporal bandwidth is set to 0.

The results are summarised in Figure 7.4. We found that the fixed bandwidth selected by the rule-of-thumb method works reasonably well for tags with a uni-modal distribution (e.g. the name of a city). However, for tags with a multi-modal distribution (e.g. supermarket, beach and rain), it leads to a significant over-estimation of the bandwidth. The adaptive kernel bandwidth method performs better than the fixed bandwidth estimator in many cases, especially those with a multi-modal distribution, but it is computationally expensive. However, we found that the leave-one-out kernel estimator method outperforms both of them. Therefore, in the remaining experiments, we will use the spatial and temporal bandwidths (h_s and h_m) estimated from the leave-one-out kernel estimator method as the optimal bandwidths. In particular, when applying KDE (see Equation 7.1), for each tag we use the specific bandwidth parameters that were selected with this method.

7.4.2 Experimental Settings

In all experiments, we use Support Vector Machines (SVMs) for classification problems and Support Vector Regression (SVR) for regression problems. We randomly split the set of spatiotemporal grid cells C into one-third for testing and two-thirds for training and tuning. To evaluate the impact of the training data size on the model performance, we experimented with 1%, 10% and 100% of the training and tuning set. Each time we hold out 10% of the considered set for tuning the parameters and use the rest for training. In fact, the setting with a small amount of training data makes the problem more challenging and provides additional insight into the performance of our proposed model.

To compute KL divergence, the smoothing parameter δ was selected from $\{100, 1000, 10000\}$ based on the tuning data. Table 7.1 shows the ten tags with highest KL divergence weight resulting from these smoothing values. Clearly, using $\delta = 100$ gives a set of tags that are specifically related to small geographic regions and/or particular times, while using $\delta = 1000$ gives a set of tags describing larger regions. Using $\delta = 10000$ gives a set of more general tags or even more general regions, as well as names of well-known cities. We select the top 100 000 tags from the ranking with $\delta = 1000$, where it gave us the best results based on initial experiments. However, for a grid cell c, we only consider those tags t for which $\mathit{KDE}(t|c) > \frac{1}{3}$ for computational reasons.

All embedding models are learned with an Adagrad optimiser, which is used to minimise the objective function using 30 iterations and an initial learning rate of 0.5. The number of dimensions is chosen for each experiment from $\{10, 50, 300\}$ based on the tuning data. For the parameters of our model in Equation 7.14, we considered values of α from {0.01, 0.02, 0.04, 0.06, 0.08, 0.1} and we considered values of β between 0 and 1 with an increment of 0.05. While we chose the best values of the parameters for each experiment separately, based on the tuning data, we noticed that consistently good results were obtained when using $\alpha = 0.04$ and $\beta = 0.45$. Note that we tune all parameters with respect to the F1 score for the classification tasks and Spearman ρ for

Table 7.1: Top 10 Flickr tags in terms of KL divergence.

$\delta = 100$	$\delta = 1000$	$\delta = 10000$
struy	islay	cambridge
may	gairloch	bournemouth
tiree	ashford	chester
strathglass	longleat	york
march	orkney	cornwall
waterfordhalf2009	sywell	cardiff
stmelliongolfclub	braintree	sheffield
bawdeswell	snetterton	lakedistrict
stkilda	popham	oxford
helmsdale	dungeness	norfolk

the regression tasks.

7.4.3 Variants and Baseline Methods

For the formal evaluation, we will compare our proposed "SPATE" model with the following main baseline representations:

- "Structured" uses the feature vector $v_s(c)$ modelling the structured information from Section 7.3.2.
- "Flickr" uses the KDE-based feature vector $v_{KDE}(c)$ modelling Flickr tags from Section 7.3.1.
- "Structured + Flickr" uses the combination of both structured data and Flickr data by concatenating the vectors $v_s(c)$ and $v_{KDE}(c)$.

To evaluate the impact of the spatiotemporal smoothing on Flickr tag representation, we will consider the following variants:

- “Flickr-noKDE” uses the PPMI-based feature vector $v_f(c)$ modelling Flickr tags from Section 7.3.1 (i.e. without including the tag selection and spatiotemporal smoothing steps).

- “Flickr(1BW)” uses the KDE-based feature vector modelling Flickr tags from Section 7.3.1. However, here we select the value of the bandwidths h_s and h_m that minimise the average MISE over all the considered tags when computing KDE weight (i.e. using the same bandwidths for all the tags). This variant will thus allow us to assess the effectiveness of using tag-specific bandwidth values.

7.4.4 Experimental Results

We consider two tasks to evaluate our proposed SPATE model: predicting species distribution and predicting climate-related features.

Predicting Species Distribution

For this task, we use ground truth data from the National Biodiversity Network Atlas (NBN Atlas); see Section 3.4.3 for more information about NBN Atlas. We focused our evaluation on the same 50 birds sample that has been considered in Section 4.4.1. These birds have at least 1000 observations in the NBN Atlas. This restriction to species with a sufficient number of observations is necessary to ensure that the ground truth is sufficiently reliable. Note that even species with a large number of observations may sometimes only occur in a few spatiotemporal cells. In NBN Atlas, each species record contains a set of meta-data including the observation’s latitude, longitude and month, which is the information that we need in our experiments. For each of these 50 birds, we consider a binary classification problem, i.e. predicting whether or not the bird occurs in a particular cell (i.e. whether a grid cell contains at least one observation in the NBN Atlas data).

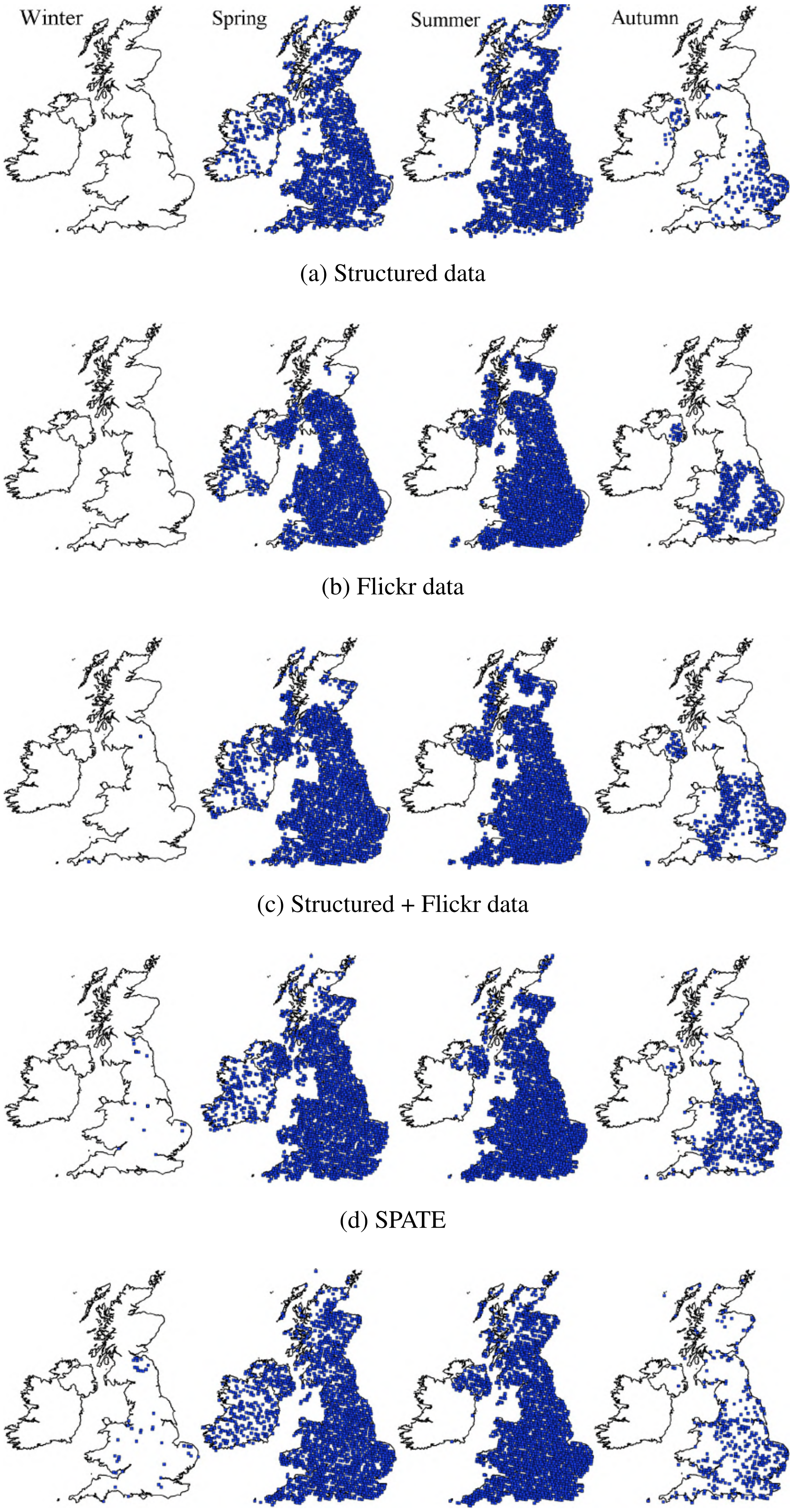
The results are reported in Table 7.2 in terms of macro-average precision, recall, and F1 score over the 50 birds. Note that we have tuned all the parameters with respect to F1 score. The results clearly show that combining Flickr tags with the available structured data leads to better results than using them separately. Moreover, combining them in our proposed spatiotemporal embeddings (SPATE) model leads to the best results. It significantly outperforms all the considered baselines, especially for the setting with the least amount of training data. Furthermore, note that the proposed KDE based spatiotemporal smoothing of Flickr tags leads to substantial improvements over the non-smoothed version in "Flickr-noKDE" and smoothing each tag with different bandwidths in Flickr consistently outperforms the method of smoothing all the tags with the same bandwidth in "Flickr-1BW". We also found the normalisation of the spatiotemporal KDE in Equation 7.1 to be critical to obtain good results. Based on the tuning data, for the SVM model, we found a linear kernel to be optimal when using Flickr data only and the combination of "Structured + Flickr", and a Gaussian kernel to be optimal for the "Structured", and "SPATE" models. For the embedding model, we found that the best results were obtained for 300 dimensions.

As an example, Figure 7.5 visually compares the predictions that were made by the different models with the ground truth for a particular bird: the Swift (Apus apus). The seasons in Figure 7.5 are defined as winter (December, January, February), spring (March, April, May), summer (June, July, August) and autumn (September, October, November). It can be clearly seen from Figure 7.5 that the predictions made by using "Structured" only, "Flickr" only, or "Structured + Flickr" are under-reported for winter and imbalanced (i.e. overestimated in some regions and underestimated in another) for the other seasons. However, "SPATE" leads to superior predictions over all the seasons. To get further insight into the performance of the considered models, Figure 7.6 shows the monthly average F1 score for the predictions made for this particular species. Although using "Flickr" outperforms using "Structured", and "Structured + Flickr" further improves the results, "SPATE" leads to the best results over all months. Interestingly, for the months with low numbers of occurrences, such as January and

Table 7.2: Results for predicting the monthly distribution of 50 species across the UK and Ireland.

	1%			10 %			100 %		
	Prec	Recall	**F1**	Prec	Recall	**F1**	Prec	Recall	**F1**
Structured	0.424	0.246	0.311	0.501	0.345	0.409	0.525	0.422	0.468
Flickr-noKDE	0.091	0.005	0.010	0.141	0.022	0.038	0.388	0.034	0.063
Flickr-1BW	0.400	0.342	0.369	0.469	0.406	0.435	0.494	0.415	0.451
Flickr	0.436	0.373	0.402	0.529	0.454	0.489	0.631	0.466	0.536
Struc + Flickr	0.448	0.384	0.414	0.536	0.465	0.498	0.629	0.474	0.541
SPATE	0.485	0.423	**0.452**	0.540	0.476	**0.506**	0.610	0.487	**0.542**

November, “SPATE” is the only model that made positive predictions while other models predicted all negatives. This example suggests that highly accurate distribution models can be learned using any of the considered models when we have sufficiently large numbers of occurrences as in the spring and summer months. However, our proposed “SPATE” model still performs better in the months with very low numbers of occurrences, as in the winter and autumn months. Additionally, when we look at the prediction confidence score of this species over the spatiotemporal grid cells, we found that our proposed SPATE model makes much higher confidence predictions than the other proposed baselines. As an example, Figure 7.7 shows the prediction confidence score obtained from different models for a particular location (latitude= 54.81503616 and longitude= -2.086120293) over all the testing months. Clearly, as can be seen in Figure 7.7 the predictions made by the SPATE model have very high confidence for the correct predictions and very low confidence for the incorrect predictions (see the incorrect prediction in August) which further illustrates the strong performance of our proposed model. Note that all the results reported in Figure 7.5, 7.6, 7.7 are for the setting where only 1% of the training/tuning data was used.



(a) Structured data

(b) Flickr data

(c) Structured + Flickr data

(d) SPATE

(e) Ground truth data

Figure 7.5: Prediction of the seasonal distribution of Swift across the UK and Ireland using 1% of the data for training/tuning.

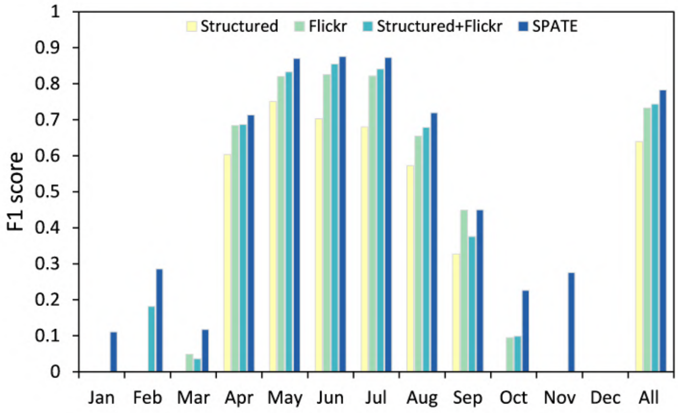


Figure 7.6: F1 score of predicting the monthly distribution of Swift.

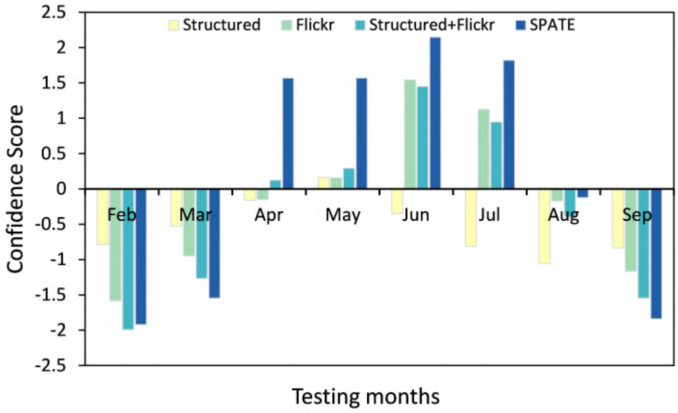


Figure 7.7: The prediction confidence score for location coordinate (54.81503616, -2.086120293) over the testing months. Jan, Oct, Nov and Dec are not shown in the figure because the corresponding cells are in the training set for that location. Note that Swift has positive ground truth observations in that location in April, May, June, July and August, and negative ground truth observations in February, March and September.

Predicting Climate Features

For this task, we consider five different regression problems: predicting the monthly average of precipitation, solar radiation, temperature, wind speed, and water vapour pressure. For these experiments, we do not include any of these climate features in the structured representations (and embeddings derived from them) as they serve here as ground truth. The results of these experiments are reported in Table 7.3 in terms of mean absolute error (MAE) and Spearman ρ correlation between the predicted and actual values for all spatiotemporal cells in the testing set. The mean and standard deviation of each of those features are shown in Table 7.4. Note that we tune all the parameters with respect to Spearman ρ. We can see from the results that combining structured and Flickr data outperforms using them separately. However, combining them using our proposed spatiotemporal embeddings (SPATE) model leads to a substantial improvement over the baseline methods, especially when we consider only 1% of the training/tuning data. Note that, for settings with more training data, it is not surprising that all methods perform well as climate features are strongly autocorrelated in time and space. For these experiments, based on the tuning data, we found that the best results for the SPATE model were obtained for 10 dimensions.

In Figure 7.8, we visually illustrate the predictions made by the different models for seasonal precipitation. The model based on structured data performs worst while the "SPATE" model is the best for all the seasons. While the overall differences between the results for precipitation (especially in term of Spearman ρ in Table 7.3) are small, clear differences between their performance are still noticeable in Figure 7.8. To get a clearer picture about the performance of each model, Figure 7.9 shows the monthly average MAE and Spearman ρ for predicting the precipitation. Although "Flickr" performs better than "Structured" in terms of MAE, it performs worse in term of Spearman ρ. The combination of "Structured + Flickr" performs in between them. Interestingly, our proposed "SPATE" model has the best performance in terms of MAE and Spearman ρ for all the months. Looking at the prediction of a particular location (latitude= 55.26469626 and longitude= -4.784080876) over all the testing months, we can see that the "Structured" model predictions do not deviate too far from the mean value, which have not affected the Spearman ρ score as much as MAE. The "Flickr" model makes more varied predictions, although they are still far from the ground truth. The combination of “Structured + Flickr” leads to more faithful predictions. However, the “SPATE” model performs significantly better. Again, all the results reported in Figure 7.8, 7.9, 7.10 are when using only 1% of the training/tuning data.

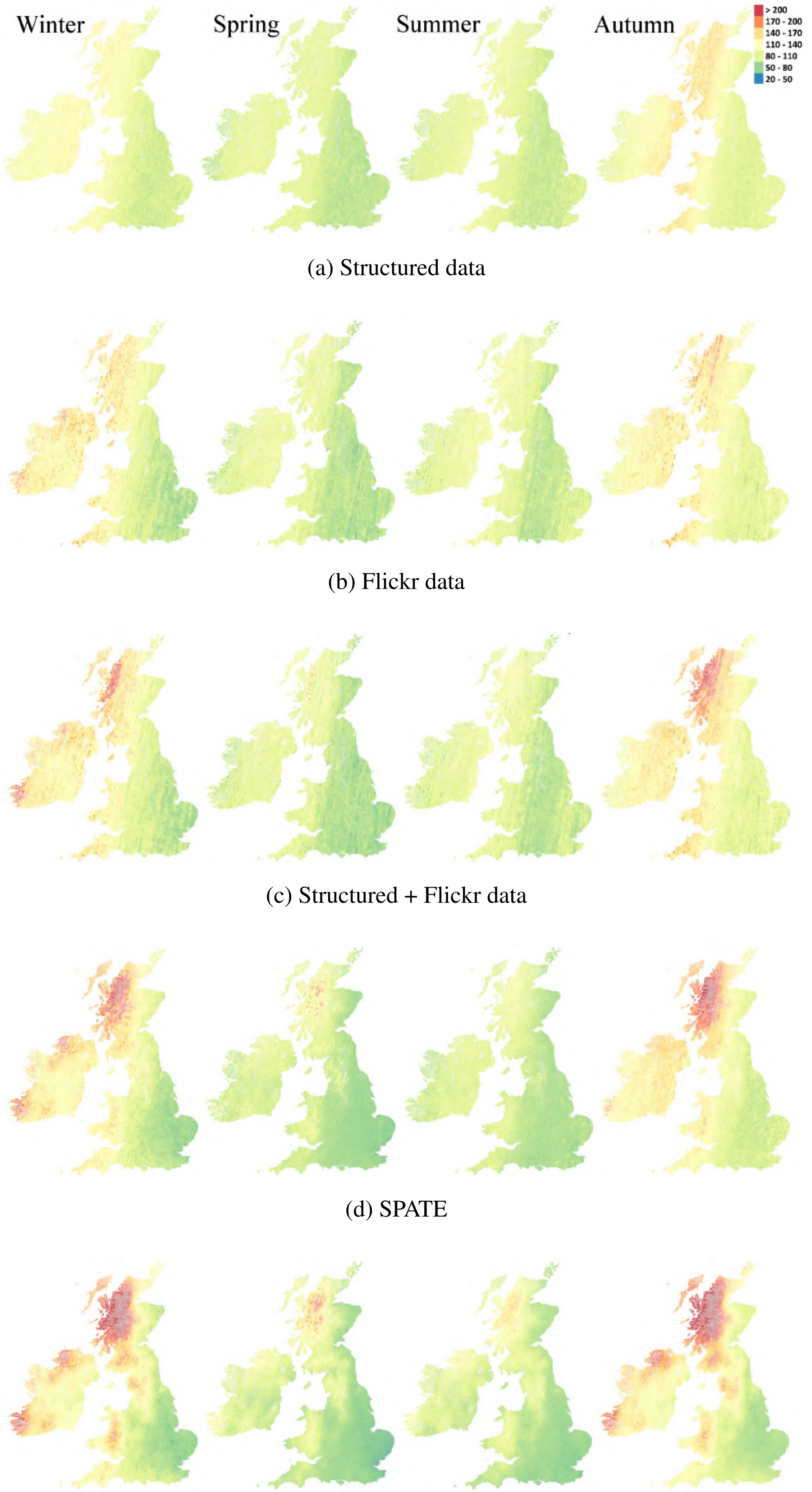
7.4.5 Location Similarity

In this section, we qualitatively evaluate the nature of the vectors generated by the SPATE model. Figure 7.11 and Figure 7.12 show the similarity maps of a number of selected locations in July and January, respectively. The selected locations include the cities of London, Dublin and Hull, the sparsely populated but popular tourist areas of Snowdonia and Skye, which are mountainous, and the tourist area of Roseland Heritage Coast which is coastal and scenic, non-intensive agricultural land with small villages. The similarity has been measured according to the Euclidean distance between the vector representation of the cell, which the considered location belongs to and the other cells using 300 vector dimensions.

As a general observation, in all cases, the maps do succeed in highlighting regions that are very similar in several respects to the respective selected location. Thus London and Dublin are both similar to other major urban conurbations such as Birmingham, Manchester, Glasgow, Newcastle upon Tyne, Bristol, Cardiff and Belfast. They are least similar to sparsely populated, mountainous rural areas such as the Highlands of Scotland and, in the case of London, the west of Ireland. Note that Dublin, the capital of Ireland is more similar to the rural west of Ireland, to which it is culturally related, than is London, just as London, the capital of England, is more similar, than is Dublin, to the geographically much closer rural areas of East Anglia in England. This latter distinction can be attributed to the general vocabulary of Flickr which is more similar, in references to places and activities, between regionally adjacent places. Hull is an industrial seaport and city. Similar locations in Summer and Winter are other commercial and industrial coastal locations such as Liverpool, Newcastle upon Tyne, Bristol, Cardiff and Southampton, along with other relatively highly populated industrial in-

Table 7.3: Results for predicting the monthly average climate features.

		1%		10 %		100 %	
		MAE	ρ	MAE	ρ	MAE	ρ
Precipitation	Structured	31.492	0.509	26.758	0.683	22.354	0.742
	Flickr-noKDE	32.214	0.125	31.724	0.202	30.808	0.268
	Flickr-1BW	28.750	0.538	23.492	0.697	22.865	0.725
	Flickr	28.240	0.549	23.601	0.698	22.562	0.741
	Structured + Flickr	27.385	0.562	22.999	0.711	20.780	**0.773**
	SPATE	24.509	**0.669**	22.971	**0.714**	21.402	0.767
Solar Radiation	Structured	4867.2	0.776	2476.0	0.895	1083.1	0.947
	Flickr-noKDE	5266.1	0.333	4603.9	0.386	4440.6	0.419
	Flickr-1BW	2434.5	0.829	1621.6	0.895	1534.4	0.914
	Flickr	2359.4	0.841	1575.9	0.901	1480.3	0.928
	Structured + Flickr	2045.2	0.884	1076.4	0.950	936.5	**0.973**
	SPATE	1415.3	**0.907**	1041.4	**0.955**	1030.6	0.960
Wind Speed	Structured	1.072	0.246	0.956	0.429	0.901	0.492
	Flickr-noKDE	1.081	0.082	1.070	0.130	1.063	0.170
	Flickr-1BW	1.099	0.217	0.963	0.418	0.897	0.493
	Flickr	1.084	0.251	0.959	0.421	0.874	0.512
	Structured + Flickr	1.001	0.347	0.938	0.456	0.873	0.522
	SPATE	0.953	**0.442**	0.930	**0.467**	0.848	**0.523**
Water Vap Press.	Structured	0.193	0.586	0.154	0.699	0.126	0.760
	Flickr-noKDE	0.234	0.110	0.226	0.250	0.225	0.279
	Flickr-1BW	0.187	0.607	0.155	0.698	0.136	0.748
	Flickr	0.186	0.612	0.152	0.707	0.134	0.752
	Structured + Flickr	0.176	0.661	0.143	0.738	0.126	0.777
	SPATE	0.135	**0.752**	0.122	**0.771**	0.119	**0.779**
Temperature	Structured	2.060	0.826	1.063	0.929	0.837	0.953
	Flickr-noKDE	3.415	0.228	3.142	0.350	2.979	0.397
	Flickr-1BW	1.653	0.849	1.372	0.888	1.074	0.919
	Flickr	1.636	0.845	1.306	0.891	1.034	0.931
	Structured + Flickr	1.302	0.907	1.054	0.932	0.823	**0.961**
	SPATE	1.164	**0.920**	1.010	**0.939**	0.935	0.946



(a) Structured data

(b) Flickr data

(c) Structured + Flickr data

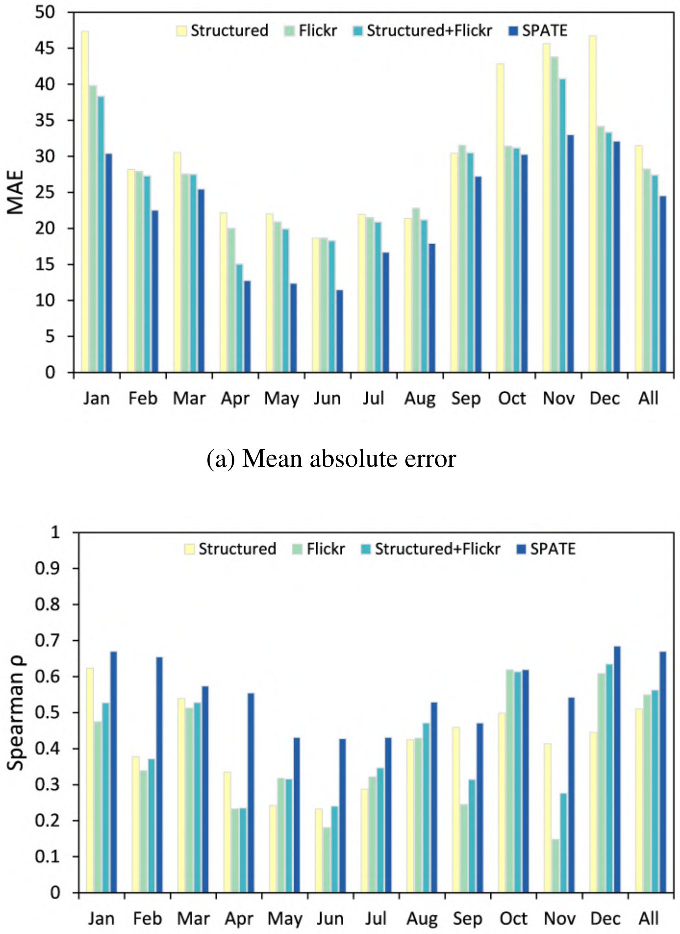
(d) SPATE

(e) Ground truth data

Figure 7.8: Prediction of the seasonal precipitation across the UK and Ireland using 1% of the data for training/tuning..

Table 7.4: Mean and Standard deviation of the monthly average climate data.

	Mean	STDEV
Precipitation (mm)	94.750	44.037
Solar Radiation (kJ $m^{-2}day^{-1}$)	9243.9	5847.7
Wind Speed (m s^{-1})	4.750	1.454
Water Vapor Press (kPa)	0.897	0.302
Temperature (°C)	9.021	3.970



(a) Mean absolute error

(b) Spearman ρ

Figure 7.9: The monthly prediction results of precipitation.

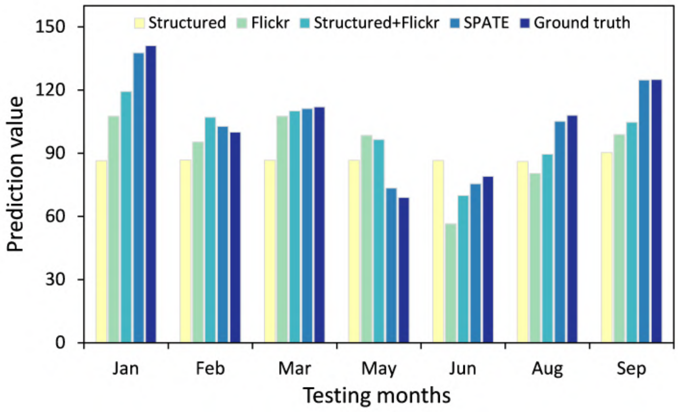
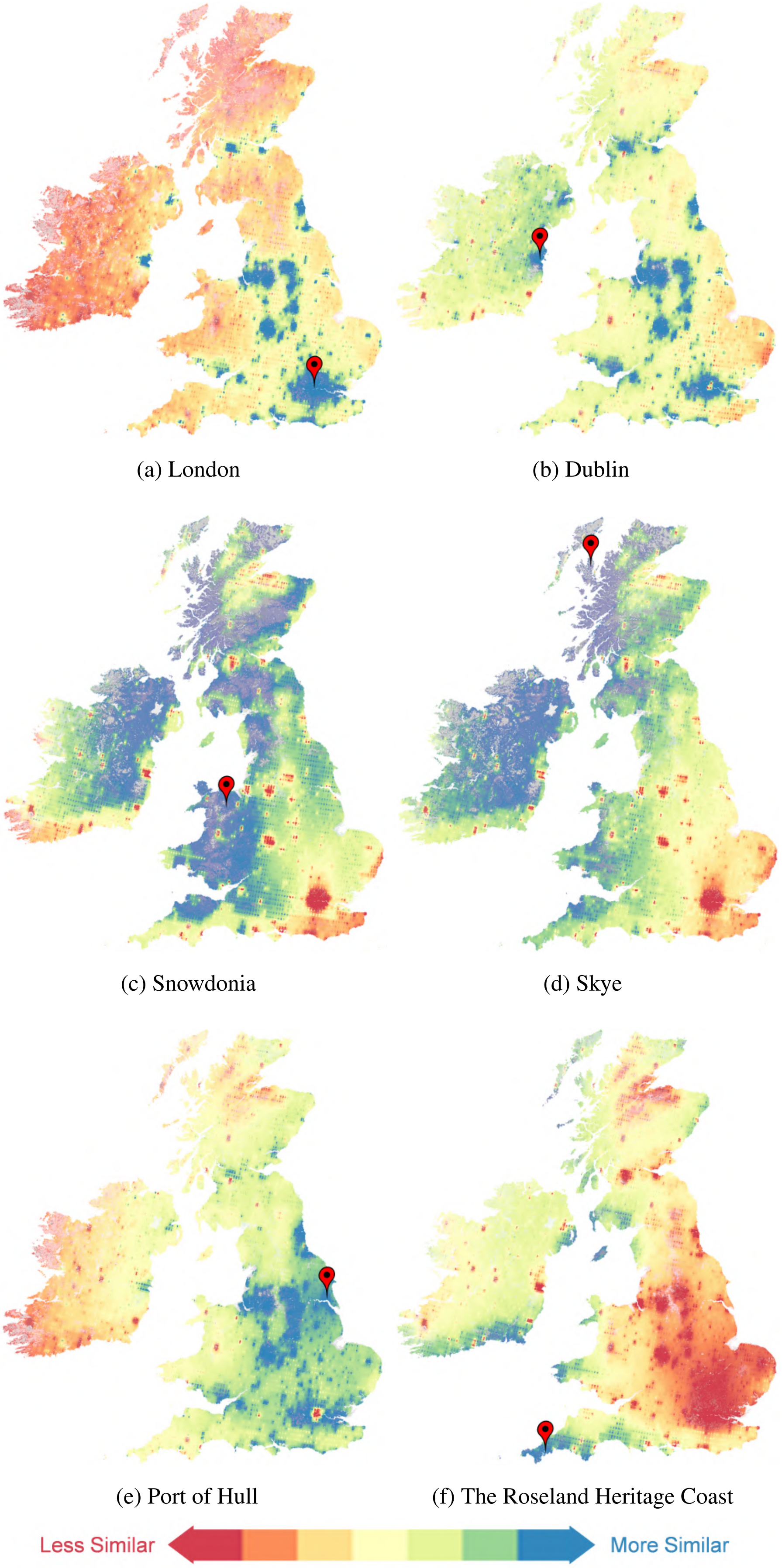


Figure 7.10: The monthly average value of predicting the amount of precipitation for location coordinate (latitude= 55.26469626 and longitude= -4.784080876) over the testing months. April, July, October, November, and December are not shown in the figure because the corresponding cells are in the training set for that location.

land regions such as Birmingham, Leeds and Manchester. It is most different from the west of Ireland and the highlands of Scotland, which are mountainous regions with low population and pastoral agriculture.

Differences between summer and winter are much less marked than the differences between regions at the same times of the year, particularly for the cities. However, an example of a seasonal city difference can be observed for London, which is more different in the summer (July) from relatively remote rural areas such as parts of Wales and Cornwall. In the latter regions (Wales and Cornwall) there might be higher levels of observations in Summer of the natural environment and outdoor leisure activities when there are more tourists than in winter. The nature of different types of tourist activity might also explain the pronounced differences in summer between the mountainous but popular tourist area of Snowdonia and the also popular coastal tourism areas of south-west Ireland and south-east England. The Isle of Skye, while generally similar in summer and winter to other relatively low populated rural areas, has a more significant



(a) London

(b) Dublin

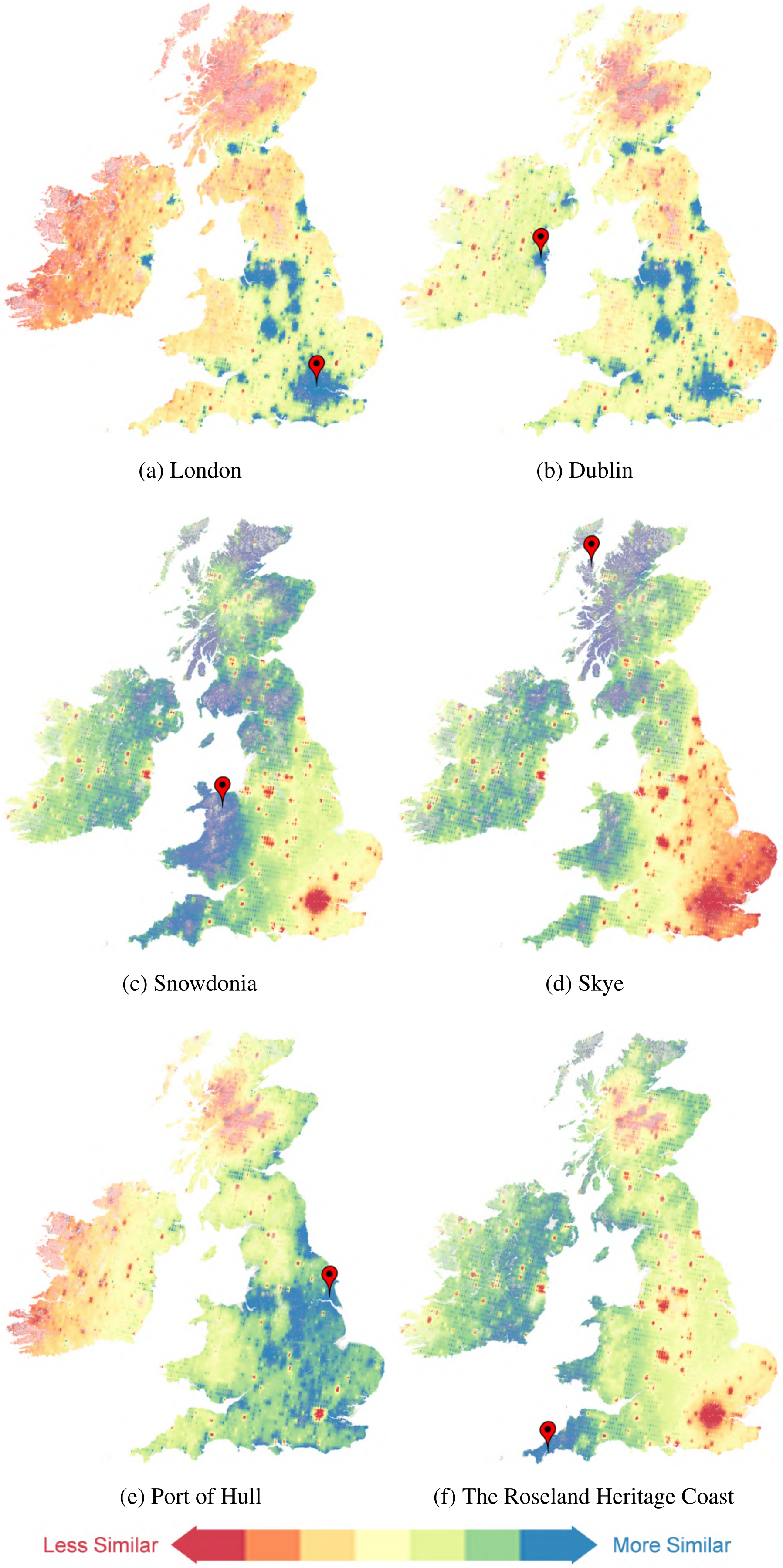
(c) Snowdonia

(d) Skye

(e) Port of Hull

(f) The Roseland Heritage Coast

Figure 7.11: Location's similarity maps in July.



(a) London

(b) Dublin

(c) Snowdonia

(d) Skye

(e) Port of Hull

(f) The Roseland Heritage Coast

Figure 7.12: Location's similarity maps in January

difference from the south-east of England in winter than in summer. Speculatively, this might reflect the fact that, in winter, Skye with its low indigenous population and much lower levels of tourism (in winter) will have relatively low levels of contribution to social media than the more populated areas of south-east England.

7.5 Summary

In this chapter, we have proposed a novel model for learning vector space embeddings of spatiotemporal entities which is able to integrate structured environmental information and textual information from Flickr tags. Furthermore, to handle the problem of Flickr data sparsity, we presented a method based on kernel density estimation to smooth the distribution of Flickr tags over space and time. For evaluation, we have considered two experimental tasks. The first experiment aimed to predict the monthly distribution of species across the UK and Ireland, using observations from the National Biodiversity Network Atlas as ground truth. In the second experiment, we looked at predicting five climate-related features.

The experimental results show that smoothing the distribution of Flickr tags leads to substantial improvements in comparison with the non-smoothed version. Moreover, combining Flickr tags with structured data consistently outperformed using them separately. This strongly suggests that Flickr can be a valuable supplement to more traditional datasets. Notably, our proposed spatiotemporal embeddings (SPATE) model provides an efficient integration of Flickr tags with structured information that outperforms all the considered baselines, especially when we considered very small training datasets.

Chapter 8

Conclusions and Future Work

8.1 Introduction

This final chapter provides a summary of the research conducted in the thesis. First, it relates the contributions to the thesis hypothesis and summarises the main findings. Subsequently, we address each of the considered research questions. It ends by highlighting some possible directions for future work.

8.2 Thesis Summary and Contributions

The main point of departure for this thesis was the observation that with the popularity of social media, a large amount of user-generated textual data that is grounded in time and space has become available. In particular, the photo-sharing platform Flickr hosts more than 10 billion photographs[1], most of which are associated with short textual descriptions in the form of tags to describe what is depicted in the photograph. In addition, the time at which these photographs were taken and their geographical coordinates are available as meta-data for many photographs. The tags associated with such georeferenced photographs often describe the location where they were taken, and Flickr can thus be regarded as a source of environmental information. Several

[1] http://expandedramblings.com/index.php/flickr-stats

previous works have shown the potential of Flickr tags for characterising the environment, which can complement more traditional sources. However, most of these studies are based on manual analysis, with little automated exploitation of the associated tags [92, 30]. This motivates us to automate methods that can utilise Flickr as an additional source of environmental information.

The research hypothesis for this thesis was presented in Chapter 1. To remind the reader, the hypothesis is: “Social media can be used as a valuable source of ecological information. In particular, we can use the meta-data associated with the photographs on the photo-sharing platform Flickr as a complementary source to the publicly available scientific datasets in order to predict spatially and temporally grounded information about the natural environment. This meta-data allows us to improve the prediction of features such as the scenicness of a place, species distribution, land cover categories, and several climate-related features”. We have developed several methods to test this hypothesis to the point where it is possible to say that it is true. In fact, all the research presented in this thesis, mainly Chapters 4, 5, 6, and 7, supports this hypothesis.

After summarising the related work in Chapter 2 and introducing the considered datasets and preprocessing strategies in Chapter 3, Chapter 4 presented a new method that uses georeferenced Flickr tags for modelling locations and predicting environmental features. This method represents each location as a concatenation of a bag-of-words representation derived from Flickr and a feature vector encoding the numerical and categorical features obtained from the structured datasets. The main aim was to compare the predictive power of Flickr tags with that of structured environmental data from more traditional sources for the task of predicting environmental phenomena. 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 of predicting CORINE land cover categories.

Finally, we looked at predicting five climate-related properties. We have found that combining Flickr tags with available environmental data substantially improves the predictions. This indicates that Flickr can be considered as a complementary source of ecological information. Furthermore, in Section 4.4 of the same chapter, we presented a method for mapping the location of wildlife species occurrence using Flickr tags. We have shown that while a method based simply on the presence or absence of the species name provides good precision, higher recall with similar precision can be achieved with a meta-classifier that combines the presence-absence data with predictors based on all the tags.

In Chapter 5, we proposed a novel collective prediction model, which takes advantage of the fact that most environmental features are strongly spatially autocorrelated. For example, climate features typically do not vary much between places that are just a few kilometres apart. While this indicates that geographic distance should play a significant role in determining neighbourhoods, we found that substantial gains can be made by considering Flickr and traditional data. 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.

To improve the way of combining Flickr tags with structured environmental features, Chapter 6 developed a novel model, named EGEL (*Embedding GEographic Locations*). This model integrates both Flickr and environmental data into low-dimensional vector space embeddings. We found that this approach led to more accurate predictions than the previous approach from Chapter 4 that concatenated the bag-of-words data with the structured data.

Following on from that approach, in Chapter 7 we considered the SPATE (*SPAtioTemporal Embeddings*) model. In particular, SPATE learns a vector space embedding for each geographic region and each month of the year. Besides the added consideration of time, we also introduced a new smoothing method to deal with the sparsity of Flickr tags. This was motivated by the fact that when fine-grained regions are used, and data may be sparse, the number of times that a tag is used in a particular region and month is not a reliable indicator by itself of the relevance of that tag. For evaluation, we have considered two experimental tasks. The first experiment aimed to predict the monthly distribution of species across the UK and Ireland, using observations from the National Biodiversity Network Atlas as ground truth. In the second experiment, we looked at predicting five climate-related features. The experimental results show that smoothing the distribution of Flickr tags leads to substantial improvements in comparison with the non-smoothed version. Furthermore, combining Flickr tags with structured data consistently outperformed using them separately, especially when our embedding model is used as the considered representation.

In conclusion, the experimental results obtained from this thesis support our hypothesis. Flickr tags have been used successfully for (a) modelling geographic locations, (b) building neighbourhood structure, and (c) spatiotemporal modelling. This thesis proposed the use of user-generated content as a rich source of information to identify a wide variety of environmental phenomena. We have proposed and evaluated methods and techniques for developing algorithms that can efficiently utilise Flickr tags as an additional source of ecological information. Overall, the research conducted in this thesis has made significant advances in web mining and geographic information retrieval and analysis, particularly with respect to knowledge discovery and data mining in social media.

8.3 Research Questions

In this section, the research questions previously identified in Section 1.2 will be discussed in relation to the research undertaken in this thesis. Each research question will be repeated, and the relevant research will be discussed, including any related analysis, evaluation approaches and new knowledge that has been acquired.

Research Question 1: *Is it possible to extract large amounts of high-quality environ-*

mental information from Flickr, and if so, how complementary is this information to publicly available scientific datasets?

Our main finding in Chapters 4, 5, 6, and 7 is that the combined model substantially and consistently outperformed the model that only relied on structured data sources. This strongly suggests that Flickr can be valuable as a supplement to more traditional datasets in environmental analyses. While we have not been able to precisely identify the nature of the information contained in Flickr tags, we found them to be consistently helpful in a variety of different ways.

Research Question 2: *How can we deal with the sparsity of Flickr tags for location (and possibly time-dependent) representation?*

To handle the problem of Flickr data sparsity, we proposed two methods. The first method, which was presented in Chapter 4, is aimed at the representation of geographic locations. Given that the number of tags associated with a given location can be quite small, we applied a kind of spatial smoothing, i.e. the importance of a given tag for a given location should not only depend on the occurrences of the tag at that location but also on its occurrences at nearby locations. To this end, we use a formulation in Equation 4.2 which is based on a spatially smoothed version of pointwise mutual information. The second method, which was presented in Chapter 7, is based on kernel density estimation to smooth the distribution of Flickr tags over space and time. This second method is more sophisticated because the smoothing is more important in the considered spatiotemporal setting. The experimental results show that smoothing the distribution of Flickr tags leads to substantial improvements in comparison with the non-smoothed version. This confirms that smoothing the distribution of Flickr tags over space (and possibly time-dependent) is an effective way of alleviating the sparsity of Flickr tags, especially when fine-grained regions are considered.

Research Question 3: *How can we best integrate these representations with the available structured environmental data to improve the predictive power?*

To answer this question, Chapter 6 introduced a model for representing the geographic location by combining the feature vector derived from Flickr datasets with that derived from the numerical and categorical datasets into low dimensional vector space embeddings. The method has been extended in Chapter 7 for modelling spatiotemporal regions. The experimental results obtained from both chapters prove that vector space embeddings provide more effective integration than bag-of-words (BOW) representation.

8.4 Future Work

In this section, we discuss some of the ways in which the research in this thesis can be extended further in future.

Identifying photographs of sightings. We can develop a classifier to identify when a Flickr photograph tagged with the name of a particular organism actually corresponds to a sighting. To this end, we could use Amazon Mechanical Turk (AMT) workers to obtain an initial training set, which could then be used to obtain a much larger amount of training data in an automated way. The classification would be based on the tags associated with the photograph, the tags associated with other photographs from that user, and the tags associated with other photographs near the corresponding location. We could then consider the use of visual features to improve the classification of borderline cases. Relevant instances could be obtained by retrieving all photographs which have been tagged with either the scientific name or a known common name for an organism listed in, for example, the encyclopedia of life (EOL)[2].

Georeferencing photographs. Only about 0.03% of Flickr photographs have coordinates. Using methods developed in [118], we can accurately estimate the coordinates for many of the remaining photographs. This would allow us to increase the number of

[2] `http://eol.org`

sighting records for a particular species or phenomena. It would also be useful to get more photographs for learning the embedding models.

Users validation. Few of the users who tag photographs of organisms are experts, which means that the classification we obtain from Flickr may not always be accurate. To cope with this, we can estimate a degree of confidence we have in the classifications we obtained, for example by putting higher confidence in users who use scientific names.

Embedding species. We could learn a low dimensional vector space embedding for each species. This can be done by encoding the available ecological and habitat information about the considered species as well as all Flickr tags that occur in photographs tagged by the species name. We can also consider the textual and structured data about the considered species from other resources such as the encyclopedia of life. All these features would be integrated into a low dimensional vector space embedding representing this species which can be used to predict or confirm species observation.

Integrating data sources. Flickr is just one possible source of data. Extending the same analysis to data collected from other social media platforms such as Twitter, Instagram, and Wikipedia may alleviate the problem of data sparsity and improve the quality of the prediction. We could also consider additional scientific data sources, for example, remote sensing and earth observation data. Any new dataset can be added as an additional constraint in our embedding model.

8.5 Summary

In this thesis, we have shown that it is possible to extract scientifically useful information from unstructured and noisy social media platforms like Flickr. We have proposed several novel methods which have the potential to make an impact in the area of social media mining and geographic information retrieval.

Bibliography

[1] Ian S Abramson. On bandwidth variation in kernel estimates-a square root law. *The annals of Statistics*, pages 1217–1223, 1982.

[2] Ralitsa Angelova and Gerhard Weikum. 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*, pages 485–492, 2006.

[3] David Bamman, Chris Dyer, and Noah A Smith. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 828–834, 2014.

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

[5] Vijay V Barve. *Discovering and developing primary biodiversity data from social networking sites*. PhD thesis, University of Kansas, 2015.

[6] Ian D Bishop and David W Hulse. Prediction of scenic beauty using mapped data and geographic information systems. *Landscape and urban planning*, 30(1-2):59–70, 1994.

[7] Marianna Bolognesi. Flickr distributional tagspace: evaluating the semantic spaces emerging from flickr tags distributions. *Big Data in Cognitive Science*, pages 144–173, 2016.

[8] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795. 2013.

[9] Adrian W Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360, 1984.

[10] Chris Brunsdon. Estimating probability surfaces for geographical point data: an adaptive kernel algorithm. *Computers and Geosciences*, 21(7):877–894, 1995.

[11] Chris Brunsdon, Jonathan Corcoran, and Gary Higgs. Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems*, 31(1):52–75, 2007.

[12] Stefano Casalegno, Richard Inger, Caitlin DeSilvey, and Kevin J Gaston. Spatial covariance between aesthetic value & other ecosystem services. *PloS one*, 8(6):e68437, 2013.

[13] Soumen Chakrabarti, Byron Dom, and Piotr Indyk. Enhanced hypertext categorization using hyperlinks. In *ACM SIGMOD Record*, volume 27, pages 307–318, 1998.

[14] Olga Chesnokova, Mario Nowak, and Ross S Purves. A crowdsourced model of landscape preference. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 86, pages 19:1–19:13, 2017.

[15] Sumit Prakash Chopra. *Factor graphs for relational regression*. PhD thesis, New York University, 2008.

[16] Anne Cocos and Chris Callison-Burch. The language of place: Semantic value from geospatial context. In *Proceedings of the 15th Conference of the European*

Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, volume 2, pages 99–104, 2017.

[17] Rion Brattig Correia, Lang Li, and Luis M Rocha. Monitoring potential drug interactions and reactions via network analysis of instagram user timelines. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 492–503, 2016.

[18] Rémi Cuingnet, Charlotte Rosso, Marie Chupin, Stéphane Lehéricy, Didier Dormont, Habib Benali, Yves Samson, and Olivier Colliot. Spatial regularization of svm for the detection of diffusion alterations associated with stroke outcome. *Medical image analysis*, 15(5):729–737, 2011.

[19] Eduardo Cunha and Bruno Martins. Using one-class classifiers and multiple kernel learning for defining imprecise geographic regions. *International Journal of Geographical Information Science*, 28(11):2220–2241, 2014.

[20] Stefan Daume. Mining twitter to monitor invasive alien species - an analytical framework and sample information topologies. *Ecological Informatics*, 31:70–82, 2016.

[21] Munmun De Choudhury, Moran Feldman, Sihem Amer-Yahia, Nadav Golbandi, Ronny Lempel, and Cong Yu. Constructing travel itineraries from tagged geo-temporal breadcrumbs. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1083–1084, 2010.

[22] Eric Delmelle, Coline Dony, Irene Casas, Meijuan Jia, and Wenwu Tang. Visualizing the impact of space-time uncertainties on dengue fever patterns. *International Journal of Geographical Information Science*, 28(5):1107–1127, 2014.

[23] J. Derrac and S. Schockaert. Inducing semantic relations from conceptual spaces: a data-driven approach to plausible reasoning. *Artificial Intelligence*, pages 74–105, 2015.

[24] 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.

[25] 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, Jan-12-2010 2010.

[26] Shiri Dori-Hacohen, David Jensen, and James Allan. Controversy detection in wikipedia using collective classification. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 797–800. ACM, 2016.

[27] Cathal Doyle, Rodreck David, Jane Li, Markus Luczak-Roesch, Dayle Anderson, and Cameron M Pierson. Using the web for science in the classroom: Online citizen science participation in teaching and learning. *OSF Preprints*, 2019.

[28] Paul S Earle, Daniel C Bowden, and Michelle Guy. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6), 2012.

[29] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, 2010.

[30] 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.

[31] Jacinto Estima, Cidália C Fonte, and Marco Painho. 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*, pages 1–6, 2014.

[32] Jacinto Estima and Marco Painho. 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, 2014.

[33] Shanshan Feng, Gao Cong, Bo An, and Yeow Meng Chee. Poi2vec: Geographical latent representation for predicting future visitors. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 102–108, 2017.

[34] Jazmine A Maldonado Flores, Jheser Guzman, and Barbara Poblete. A lightweight and real-time worldwide earthquake detection and monitoring system based on citizen sensors. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*, pages 137–146, 2017.

[35] A Stewart Fotheringham, Ricardo Crespo, and Jing Yao. Exploring, modelling and predicting spatiotemporal variations in house prices. *The Annals of Regional Science*, 54(2):417–436, 2015.

[36] A Stewart Fotheringham, Ricardo Crespo, and Jing Yao. Geographical and temporal weighted regression (gtwr). *Geographical Analysis*, 47(4):431–452, 2015.

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

[38] Lihao Ge and Teng-Sheng Moh. Improving text classification with word embedding. In *IEEE International Conference on Big Data*, pages 1796–1805, 2017.

[39] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6:721–741, 1984.

[40] Gianfranco Gliozzo, Nathalie Pettorelli, and Mordechai Muki Haklay. Using crowdsourced imagery to detect cultural ecosystem services: a case study in south wales, uk. *Ecology and Society*, 21(3), 2016.

[41] Michael F Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.

[42] Christian Grothe and Jochen Schaab. Automated footprint generation from geotags with kernel density estimation and support vector machines. *Spatial Cognition & Computation*, 9(3):195–211, 2009.

[43] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864, 2016.

[44] Shu Guo, Quan Wang, Bin Wang, Lihong Wang, and Li Guo. Semantically smooth knowledge graph embedding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 84–94, 2015.

[45] Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21, 2015.

[46] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[47] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[48] Mika Hasegawa, Tetsunori Kobayashi, and Yoshihiko Hayashi. Social image tags as a source of word embeddings: A task-oriented evaluation. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC-2018)*, pages 969–973, 2018.

[49] Dirk Hovy and Christoph Purschke. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, 2018.

[50] Yujie Hu, Fahui Wang, Cecile Guin, and Haojie Zhu. A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation. *Applied geography*, 99:89–97, 2018.

[51] Shoaib Jameel and Steven Schockaert. D-glove: A feasible least squares model for estimating word embedding densities. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1849–1860, 2016.

[52] Shelan Jeawak, Christopher Jones, and Steven Schockaert. Mapping wildlife species distribution with social media: Augmenting text classification with species names. *Liebniz International Proceedings in Informatics*, 2018.

[53] Shelan Jeawak, Christopher Jones, and Steven Schockaert. Predicting environmental features by learning spatiotemporal embeddings from social media. *Ecological Informatics, Elsevier*, 2019.

[54] 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*, volume 86, pages 21:1–21:13, 2017.

[55] Shelan S Jeawak, Christopher B Jones, and Steven Schockaert. Embedding geographic locations for modelling the natural environment using flickr tags

and structured data. In *European Conference on Information Retrieval*, pages 51–66, 2019.

[56] Yasong Jiang, Taisong Li, Yan Zhang, and Yonghong Yan. Collective prediction based on community structure. *Physica A: Statistical Mechanics and its Applications*, 465:587–598, 2017.

[57] Thorsten Joachims. Making large-scale svm learning practical. Technical report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1998.

[58] Armand Joulin, Edouard Grave, Piotr Bojanowski, Maximilian Nickel, and Tomas Mikolov. Fast linear model for knowledge graph embeddings. *arXiv preprint arXiv:1710.10881*, 2017.

[59] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[60] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*, 2014.

[61] Ioannis Korkontzelos, Azadeh Nikfarjam, Matthew Shardlow, Abeed Sarker, Sophia Ananiadou, and Graciela H Gonzalez. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of biomedical informatics*, 62:148–158, 2016.

[62] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[63] Zhenzhen Kou and William W Cohen. Stacked graphical models for efficient inference in markov random fields. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 533–538, 2007.

[64] Sicong Kuang and Brian D Davison. Learning word embeddings with chi-square weights for healthcare tweet classification. *Applied Sciences*, 7(8):846, 2017.

[65] Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. Freshman or fresher? quantifying the geographic variation of language in online social media. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[66] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[67] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[68] Christina Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: A string kernel for svm protein classification. In *Biocomputing 2002*, pages 564–575. World Scientific, 2001.

[69] Daniel Leung and Shawn Newsam. Exploring geotagged images for land-use classification. In *Proceedings of the ACM multimedia 2012 workshop on Geotagging and its applications in multimedia*, pages 3–8, 2012.

[70] Joseph Lilleberg, Yun Zhu, and Yanqing Zhang. Support vector machines and word2vec for text classification with semantic features. In *Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on*, pages 136–140, 2015.

[71] Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1501–1511, 2015.

[72] Quan Liu, Zhen-Hua Ling, Hui Jiang, and Yu Hu. Part-of-speech relevance weights for learning word embeddings. *arXiv preprint arXiv:1603.07695*, 2016.

[73] Xin Liu, Yong Liu, and Xiaoli Li. Exploring the context of locations for personalized location recommendations. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 1188–1194, 2016.

[74] Corrado Loglisci, Annalisa Appice, and Donato Malerba. Collective regression for handling autocorrelation of network data in a transductive setting. *Journal of Intelligent Information Systems*, 46(3):447–472, 2016.

[75] Christopher S. Lowry and Michael N. Fienen. Crowdhydrology: Crowdsourcing hydrologic data and engaging citizen scientists. *Ground Water*, 51(1):151–156, 2013.

[76] Marnie Isla McLean. *Spatio-temporal models for the analysis and optimisation of groundwater quality monitoring networks*. PhD thesis, University of Glasgow, 2018.

[77] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[78] Luis Miralles-Pechuán, Dafne Rosso, Fernando Jiménez, and Jose M García. A methodology based on deep learning for advert value calculation in cpm, cpc and cpa networks. *Soft Computing*, 21(3):651–665, 2017.

[79] Tom M. Mitchell. *Machine Learning*, volume 45. 1997.

[80] Dunja Mladenić. Feature selection for dimensionality reduction. In *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*, pages 84–102, 2005.

[81] Jennifer Neville and David Jensen. Iterative classification in relational data. In *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pages 13–20, 2000.

[82] Hwee Tou Ng, Wei Boon Goh, and Kok Leong Low. Feature selection, perceptron learning, and a usability case study for text categorization. In *SIGIR*, volume 97, pages 67–73, 1997.

[83] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, pages 6341–6350, 2017.

[84] Masataka Ono, Makoto Miwa, and Yutaka Sasaki. Word embedding-based antonym detection using thesauri and distributional information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 984–989, 2015.

[85] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86, 2002.

[86] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 1532–1543, 2014.

[87] Lawrence Phillips, Kyle Shaffer, Dustin Arendt, Nathan Hodas, and Svitlana Volkova. Intrinsic and extrinsic evaluation of spatiotemporal text representations in twitter streams. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 201–210, 2017.

[88] Chad D Pierskalla, Jinyang Deng, and Jason M Siniscalchi. 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, 2016.

[89] Lin Qiu, Yong Cao, Zaiqing Nie, Yong Yu, and Yong Rui. Learning word representation considering proximity and ambiguity. In *Twenty-eighth AAAI conference on artificial intelligence*, pages 1572–1578, 2014.

[90] Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 116–125, 2014.

[91] Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110, 2007.

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

[93] Sascha Rothe and Hinrich Schütze. Word embedding calculus in meaningful ultradense subspaces. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–517, 2016.

[94] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited, 2016.

[95] Marzieh Saeidi, Sebastian Riedel, and Licia Capra. Lower dimensional representations of city neighbourhoods. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[96] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.

[97] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[98] Gerard Salton and Michael J McGill. Introduction to modern information retrieval. *McGraw-Hill, Inc.*, 1986.

[99] Kl Saravanan and S Sasithra. Review on classification based on artificial neural networks. *International Journal of Ambient Systems and Applications (IJASA)*, 2(4):11–18, 2014.

[100] Uta Schirpke, Erich Tasser, and Ulrike Tappeiner. Predicting scenic beauty of mountain regions. *Landscape and Urban Planning*, 111:1–12, 2013.

[101] Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert. *Support Vector Machine Applications in Computational Biology*. 2004.

[102] D Erran Seaman and Roger A Powell. An evaluation of the accuracy of kernel density estimators for home range analysis. *Ecology*, 77(7):2075–2085, 1996.

[103] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[104] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93, 2008.

[105] Pavel Serdyukov, Vanessa Murdock, and Roelof Van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491, 2009.

[106] Gavin Shaddick and James V Zidek. *Spatio-temporal methods in environmental epidemiology*. 2015.

[107] S Andrew Sheppard, Andrea Wiggins, and Loren Terveen. Capturing quality: retaining provenance for curated volunteer monitoring data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1234–1245, 2014.

[108] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. 1986.

[109] Robert Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451, 2017.

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

[111] Stefan Steiniger, M Ebrahim Poorazizi, and Andrew JS Hunter. Planning with citizens: Implementation of an e-planning platform and analysis of research needs. *Urban Planning*, 1(2):46–64, 2016.

[112] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1555–1565, 2014.

[113] Charles C Taylor. Automatic bandwidth selection for circular density estimation. *Computational Statistics & Data Analysis*, 52(7):3493–3500, 2008.

[114] Patrizia Tenerelli, Urška Demšar, and Sandra Luque. Crowdsourcing indicators for cultural ecosystem services: a geographically weighted approach for mountain landscapes. *Ecological Indicators*, 64:237–248, 2016.

[115] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080, 2016.

[116] George Valkanas and Dimitrios Gunopulos. Event detection from social media data. *IEEE Data Eng. Bull.*, 36(3):51–58, 2013.

[117] Steven Van Canneyt, Steven Schockaert, and Bart Dhoedt. Discovering and characterizing places of interest using flickr and twitter. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 9(3):77–104, 2013.

[118] Olivier Van Laere, Steven Schockaert, and Bart Dhoedt. Finding locations of flickr resources using language models and similarity search. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 48. ACM, 2011.

[119] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

[120] Jingya Wang, Mohammed Korayem, and David Crandall. Observing the natural world with flickr. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 452–459, 2013.

[121] Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. Community preserving network embedding. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 203–209, 2017.

[122] Xiaofeng Wang, Donald E Brown, and Matthew S Gerber. Spatio-temporal modeling of criminal incidents using geographic, demographic, and twitter-derived information. In *2012 IEEE International Conference on Intelligence and Security Informatics*, pages 36–41, 2012.

[123] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: Learning to rank with joint word-image embeddings. *Machine. Learning*, 81(1):21–35, October 2010.

[124] C. Xu, Y. Bai, J. Bian, B. Gao, G. Wang, X. Liu, and T.-Y. Liu. Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1219–1228, 2014.

[125] Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 35:1–35:10, 2017.

[126] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. 2014.

[127] Jing Yang and Carsten Eickhoff. Unsupervised learning of parsimonious general-purpose embeddings for user and location modeling. *ACM Transactions on Information Systems (TOIS)*, 36(3):32, 2018.

[128] Yao Yao, Xia Li, Xiaoping Liu, Penghua Liu, Zhaotang Liang, Jinbao Zhang, and Ke Mai. Sensing spatial distribution of urban land use by integrating points-of-interest and google word2vec model. *International Journal of Geographical Information Science*, 31(4):825–848, 2017.

[129] Chao Zhang, Keyang Zhang, Quan Yuan, Haoruo Peng, Yu Zheng, Tim Hanratty, Shaowen Wang, and Jiawei Han. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *Proceedings of the 26th International Conference on World Wide Web*, pages 361–370, 2017.

[130] Chao Zhang, Keyang Zhang, Quan Yuan, Fangbo Tao, Luming Zhang, Tim Hanratty, and Jiawei Han. React: Online multimodal embedding for recency-

aware spatiotemporal activity modeling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 245–254, 2017.

[131] Haipeng Zhang, Mohammed Korayem, David J Crandall, and Gretchen LeBuhn. Mining photo-sharing websites to study ecological phenomena. In *Proceedings of the 21st international conference on World Wide Web*, pages 749–758, 2012.

[132] Shenglin Zhao, Tong Zhao, Irwin King, and Michael R. Lyu. Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 153–162, 2017.

[133] Yu Zheng. Tutorial on location-based social networks. In *Proceedings of the 21st international conference on World wide web, WWW*, volume 12, 2012.

[134] Daniel Zwillinger and Stephen Kokoska. *CRC standard probability and statistics tables and formulae*. Crc Press, 1999.