DISSERTATION

submitted

to the

Combined Faculty for the Natural Sciences and Mathematics

of

Heidelberg University, Germany

for the degree of

Doctor of Natural Sciences

Put forward by

Diplom-Geograph Christian Sengstock
Born in Bad-Soden am Taunus, Germany

Oral examination: ....

Geographic Feature Mining: Framework and Fundamental Tasks for Geographic Knowledge Discovery from User-generated Data

Advisor: Prof. Dr. Michael Gertz

Abstract

We live in a data-rich environment where massive amounts of data such as text messages, articles, images, and search queries are continuously generated by users. In this environment, new opportunities to discover and utilize knowledge about the real-world arise, such as the extraction and description of places and events from social media records, the organization of documents by spatio-temporal topics, and the prediction of epidemics by search engine queries. Major challenges addressed in these data- and application-specific works arise from the unstructured and complex nature of the data, and the high level of uncertainty and sparsity of the attributes.

Despite the evident progress in utilizing specific data sources for different applications, there remains a lack of common concepts and techniques on how to exploit the data as high-quality sensors of geographic space in a general manner. However, such a general point of view allows to address the common challenges and to define fundamental building blocks to deal with problems in fields like information retrieval, recommender systems, market research, health surveillance, and social sciences.

In this thesis, we develop concepts and techniques to utilize various kinds of user-generated data as a steady source of information about geographic processes and entities (together called geographic phenomena). For this, we introduce a novel conceptual data mining framework, called *geographic feature mining*, that provides the foundation to discover and extract highly informative and discriminative dimensions of geographic space in a unifying and systematic fashion. This is achieved by representing the qualitative and geographic information in the records as *geographic feature signals*, each constituting a potential dimensions to describe geographic space. The mining process then determines highly informative features or feature combinations from the candidate sets that can be used as a steady source of auxiliary information for domain-specific applications.

In developing the framework, we make contributions to several fundamental problems: (1) We introduce a novel *probabilistic model to extract high-quality geographic feature signals*. The signals are robust to noise and background distributions, and the model allows to exploit diverse kinds of qualitative and geographic information in the records. This flexibility is achieved by utilizing a Bayesian network model and the robustness by choosing appropriate prior distributions. (2) We address the problem of *categorizing and selecting geographic features based on their spatio-temporal type*, such as feature signals having landmark, regional, or global semantics. For this, we introduce representations of the signals by interaction characteristics and evaluate their performance in clustering and data summarization tasks. (3) To *extract a small number of highly informative feature combinations* that reflect geographic phenomena, we introduce a model that extracts latent geographic features from the candidate signals using dimensionality reduction. We show that this model outperforms document-centric topic models with respect to the informativeness of the extracted phenomena, and we exhaustively evaluate how different statistical properties of the approaches affect the characteristics of the resulting feature combinations.

Zusammenfassung

Heute wird permanent eine Vielzahl unterschiedlicher Daten von Benutzern erzeugt, wie Textnachrichten, Artikel, Bilder oder Suchanfragen. Hierdurch ergeben sich neuartige Möglichkeiten, um geographische Phänomene zu erkennen und dieses Wissen für Anwendungen nutzbar zu machen. Dazu gehören etwa die Extraktion von interessanten Orten und Ereignissen anhand von Informationen in sozialen Medien, die Organisation von Dokumenten auf Basis von geographischen Themen oder die Vorhersage von Epidemien mittels Suchanfragen. Grundlegende Herausforderungen in diesen oft daten- und anwendungsspezifischen Arbeiten liegen in der unstrukturierten und komplexen Natur der Daten und in den großen Unsicherheiten bezüglich der Aussagekraft der Attribute.

Trotz zahlreicher Fortschritte bei der Analyse von benutzergenerierten Daten fehlt es an grundlegenden Konzepten und Techniken, um diese als Sensoren für geographische Phänomene zu verstehen und nutzbar zu machen. Solch ein grundlegender Ansatz würde es jedoch erlauben, elementare Probleme zu identifizieren und hierdurch fundamentale Bausteine zur Lösung von Forschungsproblemen im Bereich des Information Retrieval, der Empfehlungssysteme, der Marktforschung, des Gesundheitswesens und der Sozialwissenschaften zu entwickeln.

Diese Dissertation entwickelt Techniken und Konzepte zur Nutzung von benutzergenerierten Daten als eine ständige Informationsquelle über geographische Phänomene. Wir präsentieren ein neuartiges konzeptionelles Data Mining-Rahmenwerk, genannt *Geographic Feature Mining*. Dieses erlaubt es, geographische Phänomene aus verschiedenartigen Datensätze in einer einheitlichen und systematischen Art und Weise zu extrahieren, indem die jeweiligen qualitativen und geographischen Information als *geographische Feature-Signale* beschrieben werden. Hierbei bildet jedes Signal eine potentielle Dimension, um den geographischen Raum zu beschreiben. Die Aufgabe des Mining-Prozesses ist es dann, hoch-informative Signale oder Signal-Kombinationen zu extrahieren und diese als geographisches Wissen für domänenspezifische Analysen und Anwendungen verfügbar zu machen.

Durch die Entwicklung des Rahmenwerks leisten wir zudem mehrere Beiträge zu fundamentalen Forschungsproblemen. (1) Wir präsentieren einen neuartigen probabilistischen Ansatz, um hochwertige geographische Feature-Signale zu extrahieren. Die extrahierten Signale sind robust gegenüber einer Vielzahl von Unsicherheiten in den Daten. Zudem erlaubt es das Modell, eine Vielzahl an qualitativen und geographischen Informationen in den Daten auszunutzen. (2) Wir befassen uns mit dem Problem, geographische Feature-Signale auf Basis ihrer semantischen Ähnlichkeit zu kategorisieren und zu selektieren, wie etwa Signale, welche einen einzelnen Ort, mehrere Orte oder eine Region beschreiben. Hierfür führen wir Repräsentationen der Signale basierend auf ihrer *Interaktions-Charakteristik* ein. (3) Um eine kleine Anzahl informativer Signal-Kombinationen aufzudecken, präsentieren wir ein Modell mit dem *latente geographische Dimensionen* aus einer Vielzahl von Feature-Signalen mittels Dimensionalitäts-Reduktion extrahiert werden können. Alle vorgestellten Methoden werden in umfangreichen und vergleichenden Experimenten hinsichtlich ihrer Effektivität evaluiert. Hierzu verwenden wir reale Daten aus Photo-Communities, Microblogs und von Wikipedia.

Acknowledgements

First of all, thanks to my advisor, Prof. Dr. Michael Gertz, for giving me the chance to step into the world of knowledge discovery research, his invaluable support and advice over the last 5 years, and for creating an exceptionally productive and positive working environment. Thanks also to Prof. Dr. Alexander Zipf for supporting this research and giving me various opportunities to present and discuss my work within the Geographic Information Science group at the Institute of Geography, Heidelberg University. In particular, thanks to the Institute of Computer Science, Heidelberg University for providing a great research environment and making this interdisciplinary work happen. I also want to thank my former boss at the European Media Laboratory and Heidelberg Mobil, Matthias Jöst: You have been a great mentor and set the course to bring my latent interest in computer science to life.

Special thanks to my colleague and friend Jannik Strötgen for tolerating me as a roommate and for being the NLP master vis-à-vis for the last years. Thanks to Conny Junghans for being a great PhD candidate mentor in the first year and to Florian Flatow for inevitable advice on math problems and notations. Also, thanks Anh and Hamed for being able to work with you on research projects and of course thanks to the rest of the group (Ayser, Canh, Hui, Katharina, Thomas) for all the fun time and discussions we had. Of course, thanks to Natalia for the support in all of the administration tasks, and to the Hiwis for helping us running our systems. Thanks to all of you in the combinatorics group and the distributed systems group for great lunch hours and discussions.

Thanks also to my dear friends for forming such a close group and helping each other over the years, in particular Ilia, Matthias, and Axel for long discussions on science, life and all of the rest that matters. Moreover, thanks to all my flatmates-friends, ”Bergheimer 1B“ rocks!

Deepest thanks to my parents for their constant love, support, and advice over all the years. You set the course for all of this and I am deeply grateful for it. Also, thanks to my sister Tini and my brother Phillip for sticking together and supporting each other unquestioning.

Finally, thanks Elise for the last two years. Our relationship has become my new foundation to plan for the future and made me feel great when leaving my office in the evening.

Contents

Chapter 1

Introduction

1.1 Motivation

Today, a massive and steadily increasing amount of textual data, photos, and videos is generated by users on a daily basis. Such data includes records of various social network services (e.g., Facebook, Twitter, Foursquare), communication platforms (e.g., Blogger, Tubmblr, Whatsapp), media sharing platforms (e.g., Flickr, Panoramio, Youtube), or collaborative repositories (e.g., Wikipedia). In addition to user-generated records, user actions like search queries, click-through sequences, votes, check-ins, user profile data, and social network information (friends) are available. This data is of course at hand to the service providers, but may also be crawled from the Web or is made available by dedicated data APIs.

This data-rich environment provides new opportunities to discover and utilize knowledge about the real-world. For example, records often contain geographic information of varying granularity and precision that describes where and when users generated them or about what place and time the content refers to. Consequently, the past decade has witnessed a tremendous amount of research in the fields of information retrieval, text mining, computer vision, and geographic information science with the aim of exploiting the data to realize innovative applications and to address novel research questions. Such applications include the extraction of places and events [Rattenbury et al., 2007a], the prediction of land cover [Leung and Newsam, 2010], forecasting the spread of epidemics [Ginsberg et al., 2009], or recommending points-of-interest (POIs) and trajectories [Zheng et al., 2009], among others.

1.1.1 Applications and Challenges

Since a lot of different fields of research cope with the ubiquitously available user-generated data, a heterogeneous set of application-driven techniques has been developed. Problems addressed in the different fields are, for example:

- *Information retrieval*: Improving browsing and data organization of photo collections by automatically discovering places and events [Ahern et al., 2007; Rattenbury et al., 2007a; Serdyukov et al., 2009], the extraction of geographical topics in blogs and tags sets [Mei, 2006; Yin et al., 2011], and the retrieval of context-aware search results [Sengstock and Gertz, 2011a].

- *Disaster Monitoring and Health Surveillance*: Tracking hurricanes [Sakaki et al., 2010] and predicting the distribution of car-animal accidents [Xu et al., 2012] by using Twitter data. Using search engine queries and Twitter data to predict the spread of epidemics [Ginsberg et al., 2009; Johnson et al., 2004; Paul and Dredze, 2011].
- *Recommender Systems*: Recommending POIs and routes using user positions extracted from photo series and travel blogs [Ye et al., 2011; Zheng et al., 2011, 2009].
- *Market Research*: Predicting polls and forecasting opinions and interests on the basis of clicks on online photo collections and search queries [Gallagher, 2010; Radinsky et al., 2012].
- *Social Sciences*: Describing and identifying city cores using Flickr tags [Hollenstein and Purves, 2010] and analyzing geographic conceptualization of space [Deng and Lemmens, 2009].

Apart from the very distinct techniques and models employed to address the problems, these works share a number of common challenges. These are summarized as:

- Extracting qualitative and geographic information from unstructured and complex data records.
- Coping with a high level of uncertainty of the considered attributes.
- Handling the sparse distribution of the geographic information, even when a huge number of records is available for processing.
- Coping with geographic heterogeneity and trends, such as a strong and biased background population and areas with a small number of contributions.
- Appropriate handling and utilization of spatial dependency and auto-correlation.

These common challenges are treated in the above works in very different and specific ways. These include exploiting different types of data features in complex data sources, such as user-links in social networks [Zheng et al., 2011] or linking to external knowledge sources [Yuan et al., 2012], appropriate regularization in statistical models [Xu et al., 2012], use of prior distributions and smoothing techniques [Cheng et al., 2010], and domain-specific pre-processing routines [Paul and Dredze, 2011].

One can argue that applications need their own techniques and models to appropriately leverage specific data sources. In this thesis, however, we want to explore how to utilize user-generated data in a more general fashion in order to develop fundamental building blocks to realize domain-specific applications and to state well-defined research questions.

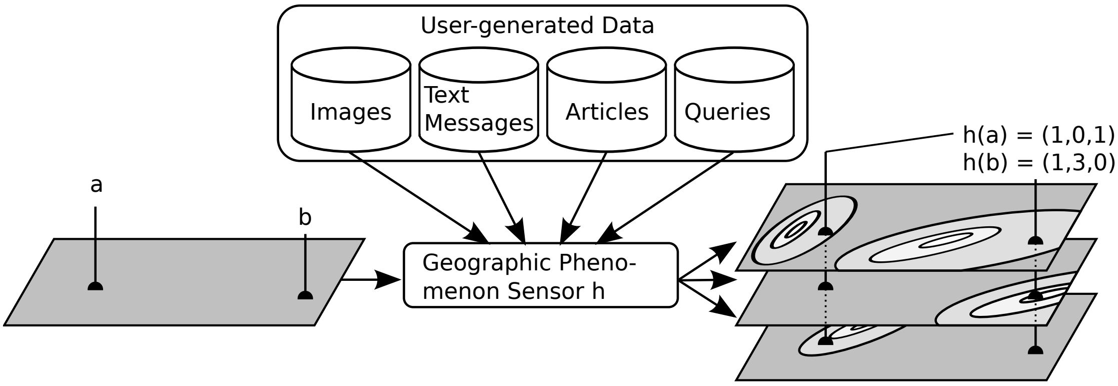


Figure 1.1: Illustration of the utilization of user-generated data as a sensor h for geographic phenomena. The sensor extracts spatio-temporal signals of geographic phenomena, allowing to compare the locations a and b by the vectors $h(a)$ and $h(b)$ in which each dimension describes a particular semantics of geographic space.

1.1.2 Sensor and Signals

The underlying idea driving this research is to exploit user-generated data as a general sensor to extract informative dimensions of geographic space. This assumes that geographic processes and entities (together called *geographic phenomena*) are treated as semantic dimensions of geographic space. For this, a phenomenon is viewed as a spatio-temporal intensity signal, describing where, when, and to what degree a social or physical process or entity occurred.

Understanding users as sensors of geographic information has been introduced in the field of geographic information science [Goodchild, 2007]. There, the data gathered by the users is called *volunteered geographic information* (VGI). The data is collected by collaborative services allowing users to add information in a more-or-less structured manner (such as point-of-interest, areas-of-interest, or roads). Examples of such services are OpenStreetMap[1], WikiMapia[2], or mash-ups build on top of map interfaces such as Google Maps[3] .

Here, we focus on data that has not been generated by users with the intend of providing accurate and structured geographic information. Instead, this data is generated on a massive scale and is mainly unstructured, such as text messages, images, or search queries. However, this data is more and more understood as a steady *observational data source* for geographic phenomena, and has been used to address specific research questions in the social sciences [Crandall and Snavely, 2012; Sheth, 2009].

Different from existing works, we abstract from particular applications and claim that information about geographic phenomena is essential to differentiate between locations

[1] www.openstreetmap.org

[2] www.wikimapia.org

[3] developers.google.com/maps

and/or areas in space and time. By understanding location semantics through processes and entities, we are able to develop fundamental tasks to find similar locations, extract interesting places, and to describe space and time in general.

Figure 1.1 illustrates the idea of extracting a set of geographic phenomena from user-generated data, with each phenomenon constituting a semantic dimension. Without any information about the distribution of processes and entities, two points a and b in geographic space can only be compared by the distance to each other. To decide if the two locations are similar (or distinct), one needs to know something about the semantics of these locations. One source of such qualitative information are surveys or geographic databases (such as provided by VGI services). Today, however, we are given a steadily increasing and accessible amount of data in the form of articles, photos, text messages, and search queries.

1.1.3 Unifying Mining Framework

In this thesis, we develop a conceptual data mining framework to extract informative geographic dimensions from heterogeneous user-generated data sources, called *geographic feature mining*. We call the dimensions *geographic features*, each of which described by a spatio-temporal intensity distribution. This is achieved by transforming the qualitative and geographic information in the records to *geographic feature signals*. The aim of the mining process is then to discover and extract highly informative features or feature combinations from the candidate set. These signals can afterwards be used as covariate information in subsequent spatio-temporal analysis task and applications.

The idea of proposing such a general approach, however, only makes sense if it helps to realize of domain-specific applications by addressing the fundamental challenges of user-generated data. We see the framework as an abstraction between the user-generated data and particular applications. A schematic view on the different layers in shown in Figure 1.2. It differentiates between (1) the data selection/pre-processing level that extracts raw qualitative and geographic information from the data sources, (2) the framework level that discovers and extracts a set of informative geographic feature signals, and (3) the application level that utilizes the feature signals. In the remainder of this thesis, the connection between feature signals and application-specific problems is exhaustively discussed. Specific applications considered in this thesis are:

- *Point-of-interest (POI), place, and event discovery and summarization*: These applications extract geographic entities and textual descriptions from the data in order to use them for information organization or monitoring [Ahern et al., 2007; Rattenbury et al., 2007a; Serdyukov et al., 2009].

- *Area-of-interest (AOI) discovery and segmentation*: These applications segment space into coherent semantic regions, e.g., to analyze functional parts of cities [Yuan et al., 2012]

- *Spatio-temporal prediction*: These applications extract a spatio-temporal distribution from a number of direct or indirect observations (such as hurricanes, earth-

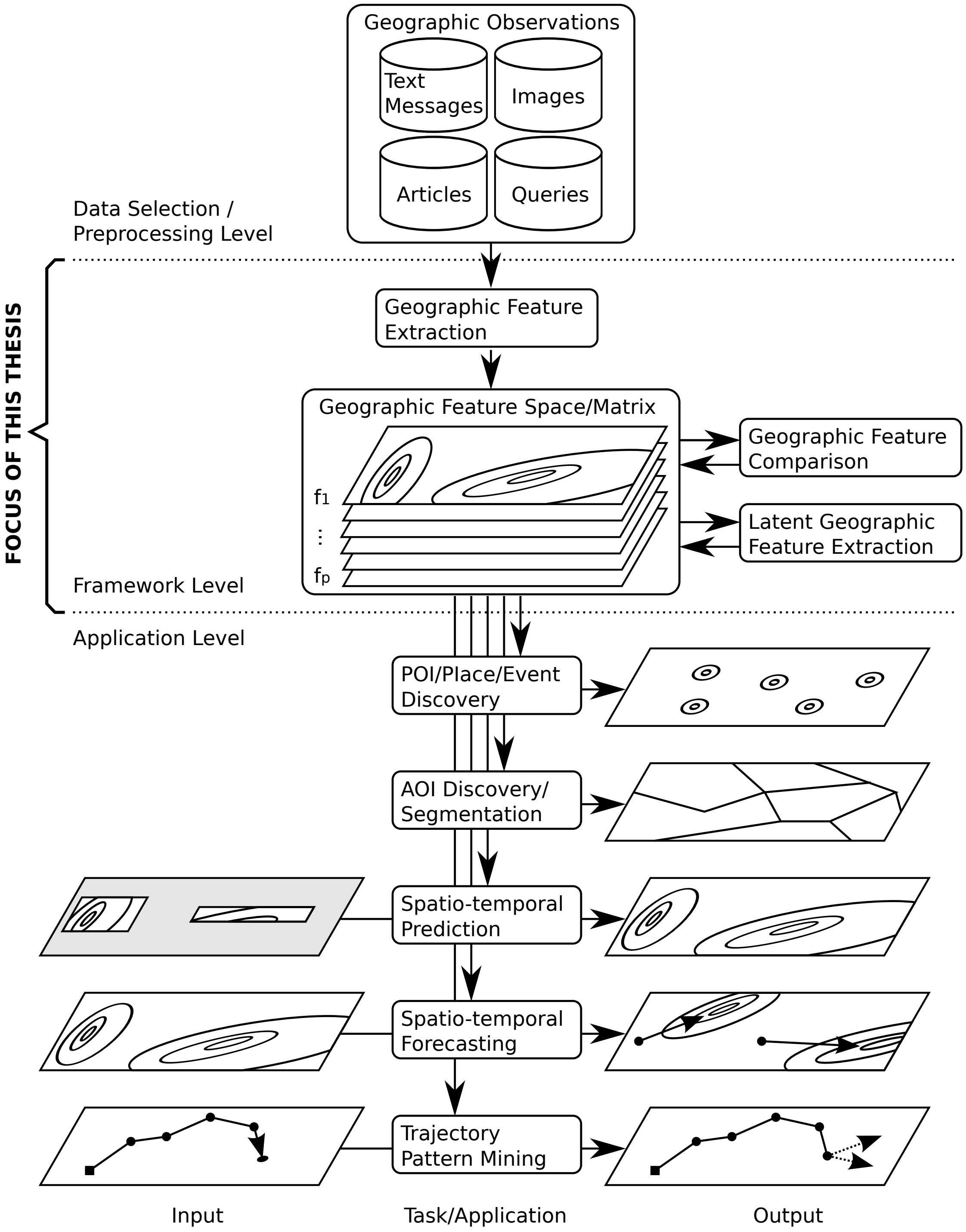

Geographic Observations
Text Messages
Images
Articles
Queries
Data Selection / Preprocessing Level
FOCUS OF THIS THESIS
Geographic Feature Extraction
Geographic Feature Space/Matrix
f1
⋮
fp
Geographic Feature Comparison
Latent Geographic Feature Extraction
Framework Level
Application Level
POI/Place/Event Discovery
AOI Discovery/ Segmentation
Spatio-temporal Prediction
Spatio-temporal Forecasting
Trajectory Pattern Mining
Input
Task/Application
Output

Figure 1.2: Schematic view on different levels of geographic knowledge discovery from user-generated data. The data selection/pre-processing level is concerned with data-specific selection and information extraction problems. The framework level is concerned with the extraction of informative geographic features. The application level makes use of the geographic feature signals to realize domain-specific tasks.

quakes, accidents, or crime areas) [Sakaki et al., 2010; Xu et al., 2012].

- *Spatio-temporal forecasting*: These applications predict the spatio-temporal distribution of a phenomenon in the future [Gallagher, 2010].
- *Trajectory pattern mining*: These applications extract interesting paths or forecast the next location of a user given past user location data [Ye et al., 2011; Zheng et al., 2011, 2009].

The concepts and techniques developed in this thesis should help to address the common challenges in these works. For this, we summarize the desiderata for a framework to utilize user-generated data as follows:

- *Unifying*: The framework should allow to process a variety of user-generated data sources and be applicable to different domain-specific tasks.
- *Systematic*: The framework should have clearly defined sub-tasks and an iterative mining process to guide the analyst in the discovery of geographic knowledge. This includes general input and output representations, interestingness measures, and criteria to judge about the quality of the results.

1.2 Contributions

This thesis makes several contributions in the field of geographic knowledge discovery (GKD). The unique contributions are as follows:

- We develop a conceptual data mining framework for the discovery of geographic phenomena on the basis of geographic feature signals. In order to define the subproblems and their unique challenges, concepts and methods to formalize the input data, the output patterns, and intermediate data representations are introduced. We show that the framework allows to describe various application and research-oriented tasks in a unifying and systematic fashion. Preliminary ideas of this research have been published in: *Reliable Spatio-temporal Signal Extraction from Human Activity Records (SSTD 2013)* [Sengstock et al., 2013a].
- We propose a model to extract robust geographic feature signals from user-generated data on the basis of a probabilistic Bayesian network. The approach is able to model diverse kinds of input data and to cope with huge levels of noise and uncertainty. We show that the model subsumes a number of basic signal extraction techniques, and is able to extract more robust and meaningful features. Parts of this research have been published in: *A Probabilistic Model for Spatio-temporal Signal Extraction from Social Media (ACMGIS 2013)* [Sengstock et al., 2013b]. This publication presents the Bayesian network model for the robust extraction of geographic feature signals.

- We propose a novel technique to compare geographic features by their spatio-temporal signals. Existing applications compare spatio-temporal signals by their intensity distribution, resulting in geographic features being similar if they happen at the same points in space and time. Our technique allows to find features that are of the same type (event, place, trajectory, etc.). This feature type similarity allows to filter and select feature candidates or to compare observational data sources on the basis of their covered spatio-temporal information. Parts of this research have been published in: *Exploration and Comparison of Geographic Information Sources using Distance Statistics (ACMGIS 2011)* [Sengstock and Gertz, 2011b]. This publications covers the problem of finding similar geographic features by using different representations of the interaction characteristics of the signals.

- We propose a technique to discover geographic phenomena from user-generated data using dimensionality reduction. For this, we define a latent factor model of geographic features. We propose several realizations of the model by employing different dimensionality reduction techniques (KMeans, PCA, SPCA, ICA). Our experiments show that enforcing sparsity in the latent factor model parametrization allows to discover more meaningful phenomena. Moreover, our experiments show how signal transformations can be used as a parameter to discover phenomena of different spatio-temporal types. Parts of this research have been published in *Latent Geographic Feature Extraction (ACMGIS 2012)*[Sengstock and Gertz, 2012b]. This publication deals with the comparison of the different dimensionality reduction approaches. A proof-of-concept application that utilizes the technique for information organization is published in: *Latent Contextual Indexing of Annotated Documents (WWW 2012)* [Sengstock and Gertz, 2012a].

1.3 Overview

The remainder of this thesis is structured as follows:

- In Chapter 2, we first detail the research context of this work. Then, we define essential concepts, such as spatio-temporal variables and spatio-temporal lattices, and review essential techniques to process spatio-temporal data such as spatio-temporal discretization and density estimation.

- In Chapter 3, we develope a conceptual data mining framework to extract geographic features from user-generated data. For this, we first review various works that utilize user-generated data and reveal commonalities in their underlying models. Then, we introduce a general representation of user-generated data as geographic observations, and introduce fundamental sub-tasks to extract informative dimensions, namely (1) geographic feature extraction, (2) geographic feature comparison, (3) and latent geographic feature extraction. The chapter can be seen as an introduction to the research problems addressed in the subsequent Chapters 4, 5, and 6.

- In Chapter 4, we propose a novel probabilistic approach to extract geographic feature signals from user-generated data. We first develop a model that encodes qualitative and geographic information in the data by conditional probability distributions in a Bayesian network. Then, we discuss its generality in comparison to existing approaches and evaluate its robustness and flexibility against its competitors.

- In Chapter 5, we propose a novel technique to compare geographic feature signals on the basis of their spatio-temporal type. For this, we first review spatial point patterns to develop the concept of interaction characteristics of spatio-temporal signals, and then introduce and evaluate different representations by using clustering and data summarization tasks.

- In Chapter 6, we propose a technique to discover a small number of informative geographic features from a huge set of candidate features. We first review existing approaches and then introduce the latent geographic feature model based on dimensionality reduction of the candidate signals. We exhaustively evaluate the statistical properties of different dimensionality reduction approaches and show that the latent geographic feature model outperforms document-centric topics models with respect to the informativeness of the extracted phenomena.

- In Chapter 7, we summarize the thesis, provide a general discussion, and give an outlook on future research directions.

1.4 Notations

The following notations are used throughout this thesis:

$[a, b], (a, b), [a, b), (a, b]$	Real-valued intervals, $a, b \in \mathbb{R}$
$[a; b], (a; b), [a; b), (a; b]$	Intervals in the natural numbers, $a, b \in \mathbb{N}$
D_S	Spatial domain
D_T	Temporal domain
D_C	Spatio-temporal/context domain
$s \in D_S$	Spatial point
$t \in D_T$	Temporal point
$c = (s, t) \in D_{S,T}$	Spatio-temporal point
$z(s)$	Spatial variable
$z(s, t), z(c)$	Spatio-temporal variable
$z_f(c)$	Spatio-temporal feature signal of feature f
$\mathbf{W}$	Neighborhood matrix
F	Set of features
L	Discrete spatial or spatio-temporal lattice
$\mathbf{Z}_{L,F}$	Geographic feature matrix
$\mathcal{N}(x; \mu, \sigma)$	(Univariate) Gaussian density function
$\mathcal{N}(x; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate Gaussian density function
μ	mean
$\boldsymbol{\mu}$	multivariate mean
σ	standard deviation
$\boldsymbol{\Sigma}$	covariance matrix
$\mathbf{I}_k$	$k \times k$ identity matrix
$\mathrm{p}_X(x) = \mathrm{p}(x)$	Probability density function or probability mass function of random variable X
$\mathbf{1}\{pred\}$	Identity function (1 if *pred* is true, 0 otherwise)

We always use bold-face/upper-case variables to denote matrices (e.g., $\mathbf{A}$). If necessary we use bold-face/lower-case variables to denote vectors (e.g., $\mathbf{a}$). However, we also use normal-face/lower-case variables (e.g., a) to denote variables that might be vector-valued, but that are not used in vector-algebraic statements. For example, we use both, $\mathbf{s} \in D_S$ and $s \in D_S$ to denote a spatial point in 2-dimensional spatial space.

Chapter 2

Background and Definitions

2.1 Research Context

This thesis deals with the discovery of geographic phenomena from user-generated data. In computer science, the domain coping with advanced data analysis techniques on huge amounts of data is called *Knowledge Discovery in Databases* (KDD). The subdomain focusing on spatial and spatio-temporal data and knowledge is called *Geographic Knowledge Discovery* (GDK). Other subdomains cope specifically with textual or image data, which is common input to our problems. In the following we give a broad overview of common concepts and terminology in the domain of data mining and its subdomains. For a detailed introduction see [Han et al., 2012; Leskovec et al., 2014; Tan et al., 2006].

2.1.1 Knowledge Discovery in Databases

Data mining is the process of discovering knowledge from data and is used synonymously to KDD. As Han states [Han et al., 2012, p. 5], data mining is actually a misnomer, since knowledge is to be mined, not data. Still the term is accepted and commonly used. Definitions of data mining include:

- "Data mining is the process of discovering interesting patterns from massive amounts of data" [Han et al., 2012, p. 33].
- "Data mining is the process of discovering interesting and potentially useful patterns of information embedded in large databases" [Shekhar and Chawla, 2003, p. 182].
- "Data mining is the process of extracting useful models [from] data" [Leskovec et al., 2014, p. 17].

Other phrases have a similar meaning, such as knowledge extraction, pattern mining, data archeology, data dredging, or big data analysis.

The domain of data mining has a huge overlap with traditional statistics and machine learning. Compared to traditional data analysis, the unique aspects of data mining can be identified as (1) having complex input data, (2) not a strict hypothesis to test, and (3) using massive amounts of data [Tan et al., 2006, p. 1].

The term *pattern* is used to describe a particular type of information. Patterns can be diverse kinds of things, such as itemsets, sequences, labels, text summaries, distributions, or parameters describing a model. The aim of data mining is to find interesting patterns among a possibly huge number of candidates. Patterns are the output of data mining and represent the mined knowledge.

The input to data mining is any kind of data collection, such as relational databases, data warehouses, transactional data, textual data, multimedia data, data streams, and sensor measurements. Transforming a data collection into a valuable format such that data analysis routines can be used, is an essential part of KDD.

Discovery is the process of finding interesting patterns (and hence new knowledge) without searching for them explicitly. Rather, data mining algorithms should allow to explore data and discover new knowledge in an automated fashion with only a small amount of prior assumptions. In this sense, data mining is similar to exploratory data analysis in statistics.

A data mining task transforms the input data into a particular type of pattern. Han uses the term data mining functionalities to refer to the following set of technical low-level tasks (as opposed to domain-specific high-level tasks) [Han et al., 2012, p. 15]:

- *Class characterization and discrimination*: Summarization and description of data having an certain class (characterization) and comparison of different classes (discrimination) [Han et al., 2012, p. 16].
- *Frequent pattern and association rule mining*: Determination of patterns that frequently occur in the data. The frequent patterns allow to discover association rules and correlations between attributes.
- *Classification and regression*: Classification and regression is the process of finding a model that describes and distinguishes data classes or value distributions [Han et al., 2012, p. 18]. This task is similar to supervised learning, predictive modeling, and statistical inference.
- *Cluster analysis*: Clustering is the process of generating class labels from a group of data and can be used to derive a taxonomy from the records [Han et al., 2012, p. 20]. From a statistical perspective it is similar to finding peaks in density distributions.
- *Outlier analysis*: Outlier analysis aims to detect and discover highly irregular data records to remove noise or to discover rare events (e.g., fraud detection) [Han et al., 2012, p. 21].

Knowledge Discovery Process

Discovering knowledge from data can be seen as an iterative sequence, called the *knowledge discovery process.* Two such sequences are described in [Han et al., 2012, p. 6] and

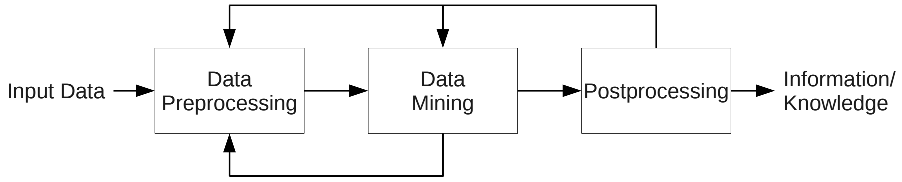


Figure 2.1: Knowledge discovery process.

[Tan et al., 2006, p. 3]. Figure 2.1 shows a knowledge discovery process slightly adapted from [Tan et al., 2006, p. 3]. The process can be understood as follows:

- *Data Pre-processing*: The input data is processed by one or several steps including data cleaning, integration, selection, transformation, and information extraction. The last step refers to the extraction of information items (features, concepts) from unstructured data such as text or images (see Section 2.4.1).

- *Data mining*: Extraction of patterns from the pre-processed data. A pattern might itself be the input to other data mining tasks.

- *Post-processing*: Representation and filtering of patterns using interesting measures, statistical tests, and visualization. The post-processed patterns might be input to other data mining tasks.

In this work we use the term data mining to refer to the whole process (synonymous to KDD) and use the term *data mining task* to refer to techniques that allow to mine particular types of output patterns. If a particular knowledge discovery tasks includes several loops between data mining tasks, we use the term data mining sub-task to refer to them.

Conceptual Data Mining Framework

Data mining is a highly application-driven domain [Han et al., 2012, p. 23] and the input and patterns often have domain specific semantics.

In this thesis we define concepts, patterns, mining tasks, interesting measures, and a general process to mine geographic feature signals from data. We use the term *data mining framework* to refer to such a abstract and conceptual view on the problem. A data mining framework is assumed to define:

- A set of concepts and methods to define input data, intermediate data, output patterns, and interestingness measures.

- A set of data mining (sub-)tasks to discover and filter interesting patterns.

- A user-driven process of pre-processing, data mining tasks, and post-processing to transform the input data into knowledge.

2.1.2 Geographic Data Mining

Geographic data mining is a subdomain of data mining and refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases [Han et al., 2012, p. 405]. In this thesis we treat time as an inherent dimension of geographic space and use the term spatio-temporal synonymously to geographic.

In [Han et al., 2012, p. 405] the unique challenge of geographic data mining is identified as extending traditional spatio-temporal data analysis methods by placing emphasis on efficiency, scalability, cooperations with database systems, improved interaction with the user, and the discovery of new types of knowledge.

Geographic data mining has a number of unique aspects (see [Miller and Han, 2009, p. 10-13]). The following aspects have important impact on this work:

- *Measurement framework*: Space and time are highly interrelated and provide a measurement framework for the other dimensions [Miller and Han, 2009, p. 10]. Hence, patterns not only occur on a attribute level (e.g., a signal value), but also in space and time. Spatio-temporal patterns are important to distinguish different type of geographic information (e.g., events, places, trends). In this work we will deal with spatio-temporal patterns of a huge number of attributes (later called features).

- *Spatio-temporal dependency*: From a spatial perspective, dependency is the property of phenomena to be similar in close spatial proximity. This is often called the law of geography, meaning, near things are more similar to each other than distant things. This law extends naturally to the spatio-temporal case. Spatial dependency is also called small-scale variation in spatial statistics. In this work, spatio-temporal dependency is used as viable prior knowledge to extract robust spatio-temporal signals from noisy observations.

- *Spatio-temporal heterogeneity*: From a spatial perspective, spatial heterogeneity is the property of spatial phenomena to vary over space. For example, the population varies from region to region, and it is far from uniformly distributed on Earth. Spatial heterogeneity can be observed on a medium to large scale. The medium to large scale variation of a temporal signal is called temporal trend. In this work, heterogeneity of a (possibly unknown) background population must be taken into account when extracting signals from observations.

- *Scale*: The spatial and temporal resolution directly influences the semantics of the analyzed processes. Hence, other than in scale-independent data domains, model parameters that affect the resolution need to be set carefully. We will discuss the influence of the bandwidth in non-parametric models in detail in Section 2.3.2.

- *Diverse spatio-temporal input data*: Spatio-temporal data might be given in diverse representations, e.g., raster data describing a distribution, spatio-temporal points of measurements, vector geometries, or spatio-temporal relationships between objects. All of these representations can be found in user-generated data. To be able to process diverse type of spatio-temporal data in a unifying fashion, we will make use spatio-temporal influence variables to describe spatio-temporal information.

Important geographic data mining tasks (which could also be named geographic data mining frameworks according to our definition in Section 2.1.1) are co-location pattern mining and spatial association rule mining.

- Co-location pattern mining finds subsets of boolean spatial features frequently located together [Shekar and Huang, 2001]. The focus of this task is on efficient mining of such subsets [Yoo et al., 2005] and on measures to select interesting patterns [Sengstock et al., 2012]. The geographic aspect lies in the concept of a co-location, defined by a distance measure and a threshold in spatial or spatio-temporal space.

- Spatial association rule mining finds itemsets and rules between items with spatial predicates (nearby, north, contains, etc.). The input of spatial association rule mining are transactions with spatial and temporal attributes. The focus is on the efficient extraction of itemsets and derived rules [Koperski and Han, 1995]. Recently, also the spatial distribution characteristics of rules and itemsets have been studied in order to develop context-aware interesting measures [Sengstock et al., 2012].

In this thesis we develop a novel data mining framework that finds informative spatio-temporal signals from qualitative and geographic influence signals. As for the frameworks mentioned above, to be eligible to be treated as a separate framework, we will clearly define the input and output data representations and a mining process to discover interesting patterns.

2.1.3 Other Fields

This thesis is concerned with user-generated data often given as text or images. In the following we will use the term *unstructured data* to refer to such data records. To transform unstructured data into an appropriate input format to data mining, domain-specific techniques from natural language processing (NLP) and computer vision are needed as pre-processing tasks.

In both domains the term information extraction is used to describe techniques and models to extract information form data records. [Davies, 2012; Feldman and Sanger, 2007]. Information extraction is an important pre-processing task of geographic feature mining to extract attributes and spatio-temporal information from unstructured data. We will discuss information extraction concepts in Section 2.4.1.

Apart from the importance of information extraction, other aspects of NLP and computer vision are important in this work (see [Feldman and Sanger, 2007, p. 3-8] for a detailed discussion on textual data):

- *Feature-based representation*: Since the records in NLP (text documents) and computer vision (images, videos) have no explicit structure of their semantics (e.g., mentioned people, locations, objects, etc), they need to be transformed into a structured representation. The attributes that represent information items of a record are usually called features.
- *High-dimensional problems*: To represent documents or images often a huge number of low-level features is used. This forces the data analysis techniques to work with high-dimensional data. Often, special techniques (such as dimensionality reduction) or assumptions (sparsity) are needed to obtain meaningful results.
- *High level of uncertainty*: Different from attributes in a database, features are highly noisy, cover redundant information, and are often meaningless when analyzed on their own.

In this work we use the term geographic feature with the same meaning as in NLP and computer vision. We give a precise definition of geographic features in Section 3.4.1.

2.2 Spatio-temporal Analysis

The primary pattern of this thesis are spatio-temporal signals. In the following, we introduce basic concepts and methods to model and analyze such signals.

2.2.1 Spatio-Temporal Data

Spatio-temporal data has a spatial and a temporal dimension. The space in which spatio-temporal points are described is called the *spatio-temporal space*, or the *(geographic) context space*. We denote it as

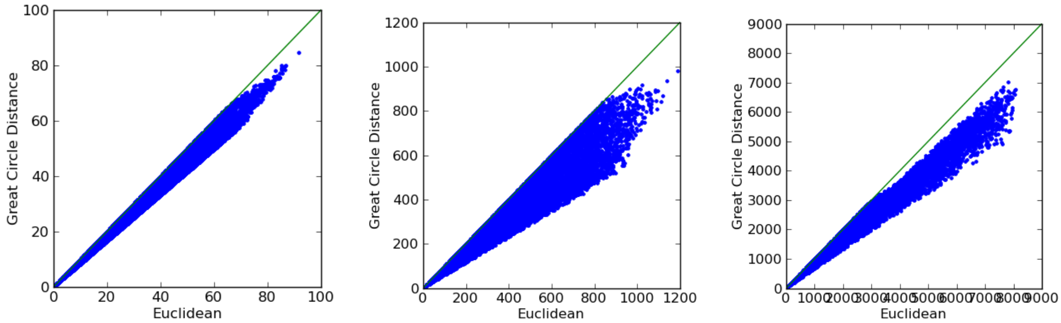
$$D_C = (D_S \times D_T) \in \mathbb{R}^{d+1} \tag{2.1}$$

with $D_S \subseteq \mathbb{R}^d$ being the spatial space and $D_T \subseteq \mathbb{R}$ being the temporal space. The term context space derives from the intuition that it is the spatio-temporal context of an observation. The spatial space defines an *area of interest*, the temporal space defines an *interval of interest*, and we call D_C a *(spatio-temporal) window of interest*. A point in context space $\mathbf{c} = (\mathbf{s}, t) \in D_C$ is described by a spatial point $\mathbf{s}$ and a temporal point t.

The context space need to be equipped with a structure to allow for distance computations between points. Usually we assume a 2-dimensional Euclidean space with its Euclidean norm used as the distance. If a separate spatial distance $d_S(\mathbf{s}_i, \mathbf{s}_j)$ and temporal distance $d_T(t_i, t_j)$ are defined, we assume that a combined spatio-temporal distance exists, e.g., by a mixture of distances

$$d_C(\mathbf{c}_i, \mathbf{c}_j) = \lambda d_S(\mathbf{s}_i, \mathbf{s}_j) + (1 - \lambda) d_T(t_i, t_j), \lambda \in [0, 1]. \tag{2.2}$$

Possible non-euclidean spatial candidates are distances on a road network or the distances between points on a sphere.



(a) Los Angeles

(b) Germany

(c) USA

Figure 2.2: Scatter plot of Euclidean distance versus great circle distance for $n = 10000$ randomly generated pairs of points inside the respective bounding boxes.

Geographic Longitude and Latitude

Spatial data points are often given as geographic longitude and latitude (note that here geographic refers only to spatial space)

$$\mathbf{s} = (lng, lat) \in D_S \subseteq [-180, 180] \times [-90, 90]. \tag{2.3}$$

The proper distance between two points on Earth is the great circle distance (gcd)

$$d_{gcd}(\mathbf{s}_i, \mathbf{s}_j) = r \arccos(\sin lat_i \sin lat_j + \cos lat_i \cos lat_j \cos |lng_i - lng_j|) \tag{2.4}$$

with radius $r = 6371$ km (average radius of the Earth).

Using geographic coordinates as axes in Euclidean space is called the equi-rectangular geographic projection. The projection is not conformal, i.e., distances between points in the projected Euclidean space do not match the distances of the points on the sphere. For small distances, however, the distances are similar to each other. We make use of the Euclidean distance $||\mathbf{s}_i - \mathbf{s}_j||$ to compute distances between close geographic coordinates using

$$d_{eucl-km}(\mathbf{s}_i, \mathbf{s}_j) = ||\mathbf{s}_i - \mathbf{s}_j|| \frac{180.0}{\pi r}. \tag{2.5}$$

Figure 2.2 shows a scatter plot for $n = 10000$ point pairs randomly generated in the bounding boxes of the USA, Germany, and Los Angeles. The scatter plots show that the distances are close to equal for small distances and are highly correlated within the proposed regions. Table 2.1 shows the average errors between the gcd and the Euclidean distance defined as

$$\Delta d(\mathbf{s}_i, \mathbf{s}_j) = \frac{d_{eucl-km}(\mathbf{s}_i, \mathbf{s}_j)}{d_{gcd}(\mathbf{s}_i, \mathbf{s}_j)}. \tag{2.6}$$

We consider the errors for the areas of interest to be small compared to the uncertainty of the input data, which justifies the general Euclidean calculations on longitude and latitude data points. For accurate distance computations in Euclidean space, a conformal projection of the data points needs to be applied as a pre-processing step.

	Los Angeles	Germany	USA
$\min(\Delta d)$	1.000	1.000	1.000
$\max(\Delta d)$	1.210	1.748	1.553
$\text{avg}(\Delta d)$	1.097	1.262	1.173
$\text{sd}(\Delta d)$	0.073	0.213	0.114

Table 2.1: Statistics of Δd for given areas of interest.

Spatio-temporal Variable

A spatio-temporal signal describes the distribution of a variable in space and time. We make use of a concept similar to a geostatistical spatial variable to describe the variation of a variable in continuous space [Cressie and Wikle, 2011]. We only consider spatio-temporal signals that are positive, and we use a positive-valued spatio-temporal variable to describe a signal

$$z(\mathbf{c}) \in \mathbb{R}_+, \mathbf{c} \in D_C. \tag{2.7}$$

A spatial variable normalized by integrating to 1 is denoted

$$\dot{z}(\mathbf{c}) = \frac{z(\mathbf{c})}{\int_{D_C} z(\mathbf{c}) d\mathbf{c}}. \tag{2.8}$$

A normalized variable can be interpreted as a probability density function that describes the probability to find the signal at a point in context space

$$\mathrm{p}(\mathbf{c}) = \dot{z}(\mathbf{c}). \tag{2.9}$$

The distribution of $\mathrm{p}(\mathbf{c})$ is used to describe the spatio-temporal characteristics of a variable.

Spatio-temporal Influence

The primary type of spatial information in user-generated data records are vector geometries, such as points, polygons, and bounding boxes. To handle those spatial data representations in a uniform fashion, we represent them as continuous or piecewise constant spatial variables. Similarly, a temporal interval or a timestamp is represented by a continuous or piecewise constant temporal variable. We assume that the joint spatial and temporal information of a record can always be represented by a spatio-temporal signal, as defined in (2.7).

The spatio-temporal signal of the input data represents a positive and additive influence over space and time. A high influence at a spatio-temporal points means that the spatial and temporal information of the record is relevant. A low influence means that the information is not relevant.

In the following we define spatial influence signals for basic spatial geometries. Similar influence functions can be defined for points and intervals in temporal space.

A point $\mathbf{s}' \in D_S$ can be represented as a piecewise constant step function

$$z_{\mathbf{s}'}(s) = \begin{cases} 1 & \text{if } \mathbf{s}' = \mathbf{s} \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

A polygon or bounding box covering area $A \subseteq D_S$ can be represented by

$$z_A(\mathbf{s}) = \begin{cases} 1/|A| & \text{if } \mathbf{s} \in A \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

A more natural representation of the influence of a geometry in space can be obtained by using a continuous spatial variable. We will make use of a Gaussian distribution to model a smooth influence of points. The Gaussian distribution is defined as

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^2|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{2.12}$$

A spatial point $\mathbf{s}' \in D_S$ can now be defined by using the point as the center and a covariance matrix to specify the shape of the influence. A symmetric influence of the point $\mathbf{s}'$ with distance α can be modeled by the following covariance matrix

$$\boldsymbol{\Sigma}_\alpha = \mathbf{I}_d \alpha, \tag{2.13}$$

where $\mathbf{I}_d$ is the d-dimensional identity matrix. The points' influence is then defined as

$$z_{\mathbf{s}',\sigma}(\mathbf{s}) = \mathcal{N}(\mathbf{s}; \mathbf{s}', \boldsymbol{\Sigma}_\alpha). \tag{2.14}$$

Similarly, we can represent the influence of a bounding box using a multivariate Gaussian with the covariance matrix $\boldsymbol{\Sigma}$ being an estimate on the basis of the boxes' $n = 4$ corner points

$$bbox = (\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \mathbf{s}_4). \tag{2.15}$$

The sample mean of the Gaussian is

$$\bar{\mathbf{s}}_{bbox} = \frac{1}{n} \sum_{\mathbf{s}_i \in bbox} \mathbf{s}_i. \tag{2.16}$$

The sample covariance matrix is

$$\boldsymbol{\Sigma}_{bbox} = \frac{1}{n} \sum_{\mathbf{s}_i \in bbox} (\mathbf{s}_i - \bar{\mathbf{s}}_{bbox})(\mathbf{s}_i - \bar{\mathbf{s}}_{bbox})^\top. \tag{2.17}$$

The Gaussian representation of a bounding box then is

$$z_{bbox}(\mathbf{s}) = \mathcal{N}(\mathbf{s}; \bar{\mathbf{s}}_{bbox}, \boldsymbol{\Sigma}_{bbox}). \tag{2.18}$$

The influence of several geometries can be mixed using the spatial variable representation. Given a set of spatial variable geometries $z_1, \ldots, z_n$ and mixing weights $\alpha_1, \ldots, \alpha_n$, a mixed spatial variable is

$$z_{mixed}(\mathbf{s}) = \sum_{i=1}^{n} \alpha_i z_i(\mathbf{s}). \tag{2.19}$$

2.2.2 Lattice Data

The geographic feature signals (patterns) are represented by discrete spatio-temporal distributions. The underlying discrete representation of space and time is called a spatio-temporal lattice [Rue and Held, 2010]. A spatio-temporal lattice is defined as a number of disjoint spatio-temporal windows

$$\begin{aligned} L = \{l_1, \ldots, l_n\}, l_i, l_j \subseteq D_C, \\ l_i \neq l_j, l_i \cap l_j = \emptyset. \end{aligned} \tag{2.20}$$

A spatio-temporal window l_i is also called a cell of the lattice. The spatial areas represented in a lattice are denoted L^S, the temporal intervals L^T and the respective dimensions of a cell

$$l = (l^S, l^T) \in L^S \times L^T. \tag{2.21}$$

The center of a cell l_i is denoted

$$\mathbf{c}_i^* = (\mathbf{s}_i^*, t_i^*) \in D_C. \tag{2.22}$$

Different from continuous context space, spatio-temporal structure between the cells is not given by a distance function but by a neighborhood function. We define a neighborhood function as a vector-valued function

$$\mathrm{nbg}(l) \in \mathbb{R}_+^{n_L}, l \in L. \tag{2.23}$$

The neighborhood function returns the cells' dependencies to all cells in the lattice (including itself). A higher value means a higher dependency. The dependency of a cell to itself is defined to be 0. An alternative representation of the neighborhood is by using a neighborhood matrix

$$\mathbf{W} \in \mathbb{R}_+^{n_L \times n_L}, w_{ij} = \mathrm{nbg}(l_i)_j. \tag{2.24}$$

In a regular 2-dimensional lattice, the neighborhood can also be described by a discrete kernel

$$K(i, j) \in \mathbb{R}_+, i, j \in [-m; m], \tag{2.25}$$

where $K(0, 0)$ represents the dependency of the cell to itself and $K(i'+i, j'+j)$ represents the dependency of cell (i', j') to the cell with offset i and j.

Similar to the continuous case, we define a spatio-temporal variable over the cells in a lattice as

$$z(l) \in \mathbb{R}_+, l \in L. \tag{2.26}$$

We use $z^C(\mathbf{c})$ and $z^L(l)$ to differentiate between continuous and discrete spatio-temporal variables if not clear from the context.

2.2.3 Spatio-temporal Grids

A spatio-temporal grid G defines a lattice L over a given space D_C. For this, the grid G partitions the space D_C into disjoint windows $L = \{l_1, \ldots, l_n\}$. We define a spatio-temporal grid on the basis of a function that maps the points $\mathbf{c} \in D_C$ to a window $l \in L$,

$$l = G(\mathbf{c}) \in L, \mathbf{c} \in D_C. \tag{2.27}$$

A spatial grid maps a spatial point $\mathbf{s}$ to an area l_i^S

$$l^S = G_S(\mathbf{s}) \in L^S. \tag{2.28}$$

A temporal grid maps a temporal point t to an interval l_i^T

$$l^T = G_T(t) \in L^T. \tag{2.29}$$

A spatio-temporal grid can be constructed from a spatial and a temporal grid by

$$l = (l^S, l^T) = G(\mathbf{c}) = (G_S(\mathbf{s}), G_T(t)). \tag{2.30}$$

The neighborhood of a regular grid is usually given by kernel K.

Temporal Grids

In this work we use temporal grids that partition the temporal domain D_T by a basis resolution δ_T,

$$\delta_T \in \{\text{second, minute, hour, day, week, month, year}\}. \tag{2.31}$$

We denote a grid that partitions D_T using δ_T as

$$l^T = G_{\delta_T, D_T}(t) \in L_{\delta_T}^T, t \in D_T. \tag{2.32}$$

By convention, the intervals intersecting the boundaries of D_T are not considered to belong to $L_{\delta_T}^T$. Intervals of the types given above are almost equi-length (with small variations in year and month). An equi-length interval grid can easily be defined based on multiples of second, minute, hour, day, or week intervals.

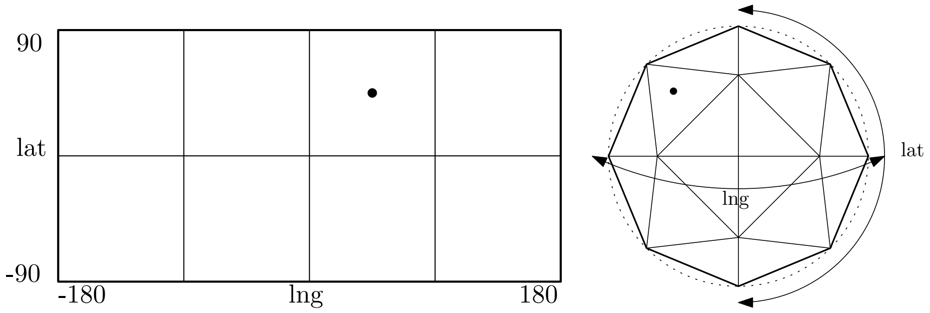
Equi-rectangular Spatial Grid

Different from the temporal domain, defining an equi-area spatial grid on spherical space is non-trivial. We shortly discuss the equi-rectangular grid to partition longitude/latitude pairs onto a regular grid, since this is heavily used in this thesis.

The equi-rectangular projection maps a point

$$\mathbf{s} = (lng, lat) \in D_S \subseteq [-180, 180) \times [-90, 90) \tag{2.33}$$

into 2-dimensional Euclidean space by just using the longitude and latitude as x and y coordinates (see Figure 2.3(a)) The equi-rectangular projection partitions the Euclidean



(a) Equi-rectangular grid

(b) Polyhedron-based grid

Figure 2.3: Mapped point on a equi-rectangular and a geodesic grid.

space into rectangles of equi-distant intervals on the x- and y-axis. Let δ_S be the interval for both, the x- and y-axis. $l_i^{S,x}, l_i^{S,y}$ are the indexes for cell l_i in a $n_x \times n_y$ grid defined on the area of interest D_S. The bounding box is defined by the corner points

$$(lng_{min}, lat_{min}, lng_{max}, lat_{max}). \tag{2.34}$$

The zero indexed grid has the dimensions

$$n_x = \lfloor (lon_{max} - lon_{min})/\delta \rfloor, n_y = \lfloor (lat_{max} - lat_{min})/\delta \rfloor. \tag{2.35}$$

The grid is then defined by

$$l^S = (l^{S,x}, l^{S,y}) = G_{\delta_S, D_S}(\mathbf{s}) = \left(\left\lfloor \frac{lng - lng_{min}}{\delta_S} \right\rfloor, \left\lfloor \frac{lat - lat_{min}}{\delta_S} \right\rfloor \right). \tag{2.36}$$

The equi-rectangular grid is not equi-area. This means, the areas on the sphere represented by cells $l^S \in L^S$ have different sizes. The missing equi-area property of the equi-rectangular grid must be considered when aggregating data. Let the bounding box of cell l^S be

$$bbox_l = (lon_{min}, lat_{min}, lon_{min} + \delta, lat_{min} + \delta)$$

The size of the area A_l of cell l on a sphere with radius $r = 6371$ km is

$$|A_l| = \frac{2\pi r^2 \delta}{360} \left(\sin(lat_{min} + \delta) - \sin(lat_{min}) \right). \tag{2.37}$$

It decreases with higher absolute latitude $|lat_{min}|$. Hence, when aggregating data into cells of an equi-rectangular grid, a cell with lower absolute latitude will have a larger area and hence will cover more points. In this thesis, the largest study area are the United States. As shown in Table 2.2, the standard deviation of the area variation using a resolution of $\delta = 1.0$ degrees is 1010.59 km, for a mean area of 9963.91 km. The cell area variation will hence contribute to an average error of 0.101. Since this error is small

	USA	Germany	Los Angeles
BBox	(-125.25, 21.52), (-52.78, 49.31)	(5.84, 47.39), (14.35, 55.16)	(-118.53, 33.69), (-117.88, 34.30)
δ	1.0	0.1	0.01
$L_y \times L_x = \|L\|$	28 x 73 = 2044	78 x 86 = 6708	61 x 65 = 3965
A_{min} km^2	8108.72	70.68	1.02
A_{max} km^2	11462.41	83.63	1.03
A_{min}/A_{max}	0.7074	0.8451	0.9930
$\overline{A}(SD)$ km^2	9963.91 (1010.59)	77.27 (3.79)	1.03 (0.00)

Table 2.2: Typical equi-rectangular grids used in this thesis.

compared to the error in the input data, we neglect the area normalization in this work. However, for larger study areas (e.g. a world wide data analysis) normalization needs to be considered.

Even without the equi-area property, the equi-rectangular grid has a number of benefits

- *Efficiency*: No pre-computation is needed to use a grid G_{δ_S, D_S}. Mapping of a point to a cell is a $O(1)$ operation by using Equation (2.36). Neighbors of a cell can directly be found in $O(1)$ time by index manipulation.

- *Fast Convolution*: The regular 2-dimensional representation allows for fast computation of a convolution (e.g., for Gaussian smoothing) using the Fast Fourier Transformation (FFT).

- *Hierarchical Structure*: 2-dimensional Euclidean space can easily be indexed hierarchically by subsuming cells into higher-level cells, e.g., by using a quadtree. This allows for an efficient aggregation along several scale levels.

Equal-area Spatial Grids

Several approaches to construct equi-area grids exist, for example, by recursively tiling the faces of a polyhedron (see Figure 2.3(b)). An equi-area grid used in astronomy is the Hierarchical Equal Area isoLatitude Pixelization grid (HEALPix) [Gorski et al., 2005]. Other geodesic grids are described in [Sahr et al., 2003].

2.2.4 Discretization

Since the spatio-temporal input data is represented as a continuous (or piecewise constant) spatio-temporal variable, it needs to be discretized onto a given discrete lattice for subsequent modeling (e.g., aggregation). Given a continuous signal $z^C(\mathbf{c})$ and a lattice

L. Let $\mathbf{c} \in l$ denote the spatio-temporal points in the window of cell l. The discretized signal $z^L(l)$ is obtained by

$$z^L(l) = \int_l z^C(\mathbf{c})d\mathbf{c}. \tag{2.38}$$

Assuming that the size $|l|$ of the cells is small compared to the support of the continuous function $z^C(\mathbf{c})$, an efficient approximation is obtained by just evaluating the center points $\mathbf{c}^*$ of the lattice windows

$$\tilde{z}^L(l) = z^C(\mathbf{c}^*). \tag{2.39}$$

In the case where the input data is a point $\mathbf{c}' \in D_C$ represented by a piecewise constant step function on $z^C_{\mathbf{c}'}(c)$ the discretization is obtained by

$$z^L(l) = z^C_{\mathbf{c}'}(\mathbf{c})\mathbf{1}\{\mathbf{c}' \in l\}. \tag{2.40}$$

2.2.5 Distribution Characteristics

By interpreting the normalized signal $\dot{z}(\mathbf{c})$ as a probability density function $\mathrm{p}(\mathbf{c})$, we can use information theoretic measures to describe characteristics of the signal. In the following, we only consider normalized discrete spatio-temporal variables $\dot{z}^L(l)$, where $\mathrm{p}(l)$ is a probability mass function with $\sum_{i=1}^{n_L} \mathrm{p}(l) = 1$.

Let $z_q(l)$ be the signal of phenomenon q and $\mathrm{p}_q(l)$ be the normalized signal. The entropy of a signal describes the uncertainty of the distribution as the average number of bits needed to encode the signal of phenomenon q,

$$E[q] = \sum_{l \in L} \mathrm{p}_q(l) \log \mathrm{p}_q(l). \tag{2.41}$$

The entropy will be zero if the whole probability mass is in one cell. Hence, the signal exhibits zero uncertainty since it will always exist at the same location. The entropy will be $\log n_L$ if the mass is uniformly distributed over all cells. In this case the signal exhibits total uncertainty, since the phenomenon q can occur at each location with equal probability.

Given two phenomena q_i and q_j and their normalized signals $\mathrm{p}_{q_i}(l)$ and $\mathrm{p}_{q_j}(l)$, respectively. The Kullback Leibler divergence represents how much additional information is needed to express the probability distribution of q_i on basis of q_j,

$$KL[q_i||q_j] = \sum_{l \in L} \mathrm{p}_{q_i}(l) \log \frac{\mathrm{p}_{q_i}(l)}{\mathrm{p}_{q_j}(l)}. \tag{2.42}$$

It is an information theoretic measure to compare how similar the distributions p_{q_i} and p_{q_j} are. A practical example to use the KL-div is to compare a phenomenon q_i against a background distribution q_j. A phenomenon having a similar distribution like the background distribution is likely to be non-informative. The entropy equals the Kullback Leibler divergence of q_i and q_j if $\mathrm{p}_{q_j}(l)$ is a uniform distribution.

2.3 Density Estimation

The problem of density estimation is to model a distribution $\mathrm{p}(\mathbf{x})$ given a finite set of observations $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ [Bishop, 2006, p. 67]. We write $X \sim \mathrm{p}(\mathbf{x})$ to denote a random variable distributed according to the distribution $\mathrm{p}(\mathbf{x})$. A function of a variable $g(X)$ is itself a random variable, since it depends on the distribution of X. We overload the upper case variable X to denote the set of observations and the random variable, so $x \in X$ is a sample (observation) of the random variable $X \sim \mathrm{p}(x)$.

Density estimation is an elementary technique in this work to compute the parameters of a model and to represent a set of observations (points) by an underlying smooth distribution. In the following, we review parametric and non-parametric approaches of density estimation and we discuss the influence of the bandwidth (smoothing) on the semantics of a distribution in the context of spatio-temporal data.

Finding the optimal distribution for a given model is called estimation. Here we assume the points $X \subseteq D_C$ to be spatio-temporal observations of an underlying continuous process (the geographic phenomenon). Estimating the density distribution is hence a way to recover the phenomenon from the observations.

2.3.1 Parametric

Parametric density estimation is performed using a parametric model. A parametric model is a set that can be parametrized by a finite number of parameters [Wasserman, 2004, p. 105]. Generally it can be written

$$\mathcal{M} = \{f(x; \theta) | \theta \in \Theta\}, \tag{2.43}$$

where $f(x; \theta)$ is called a (parametric) density function.

Multivariate Gaussian

A simple parametric model is the family of Gaussian distributions parametrized by the parameters

$$\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{2.44}$$

The density function is defined as

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^2 |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right). \tag{2.45}$$

A typical estimation approach is maximum likelihood estimation. There, the observations are assumed to be independently and identically distributed (i.i.d.). The likelihood of the data given a model parametrized by parameters θ is then

$$L(X; \theta) = \prod_{i=1}^{n} f(\mathbf{x}_i; \theta). \tag{2.46}$$

Algorithm 2.1 EM algorithm for Gaussian mixture models.

```
Initialize $\hat{\mathbf{y}}$, $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ randomly.
  (1) For each $\mathbf{x}_i$ set its $\hat{y}_i$ to index $j$ of the Gaussian where $\mathcal{N}(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$ is maximal.
  (2) Estimate $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$ on basis of the points $X_j$ using Equations (2.47) and (2.48).
  (3) Go to step (1) until convergence of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$.
```

Maximum likelihood estimation finds the optimum of the likelihood function with respect to the parameters θ [Bishop, 2006, p. 113].

For the multivariate Gaussian, the closed form solution of the maximum likelihood estimates are given as

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \tag{2.47}$$

and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top. \tag{2.48}$$

Using the parameterized multivariate Gaussian to model a number of spatio-temporal points $X \subseteq D_C$ assumes that the underlying process has a single center (unimodal). After estimating the parameters the center is given by $\hat{\boldsymbol{\mu}}$, and the shape of the distribution around the center is given by $\hat{\boldsymbol{\Sigma}}$.

Gaussian Mixture Model

A Gaussian mixture model (GMM) assumes the distribution to consist of a number of k Gaussian distributions. The parameters of the model are

$$\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}), \tag{2.49}$$

where each parameter is a k-dimensional vector

$$\begin{aligned} \boldsymbol{\mu} &= (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k), \\ \boldsymbol{\Sigma} &= (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k), \\ \boldsymbol{\pi} &= (\pi_1, \ldots, \pi_k). \end{aligned} \tag{2.50}$$

The parameter π_i is called the mixing coefficient for the Gaussian i. The density function is given as

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \pi_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \tag{2.51}$$

The number of Gaussians, k, can also be seen as a parameter in the Gaussian mixture model. However, it is usually assumed to be given the user.

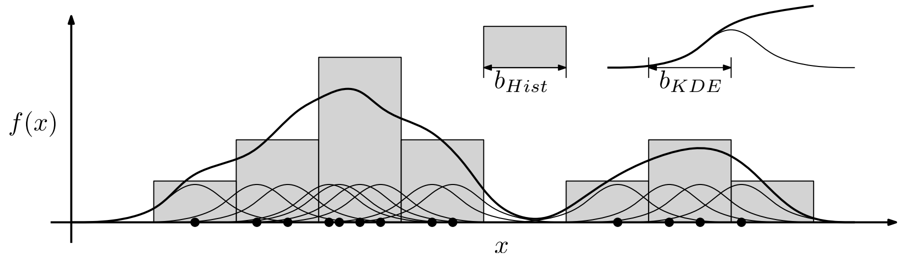


Figure 2.4: Discrete and continuous intensity distributions for $n = 13$ points using a histogram estimator with bandwidth b_{Hist} and an KDE estimator with a Gaussian kernel with bandwidth b_{KDE}.

Different from a multivariate Gaussian, there exists no closed form solution for the optimal parameters using maximum likelihood estimation. Instead, an iterative estimation procedure is used to find the optimal parameters $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$, and $\hat{\boldsymbol{\pi}}$ called expectation maximization (EM) [Bishop, 2006, p. 435].

For this, a set of latent (hidden) variables $\mathbf{y}$ is introduced. Let $y_i \in [1; k], i \in [1; n]$ be the index of a Gaussian to which a point i has been assigned. Given the points assigned to Gaussian i as

$$X_i = \{\mathbf{x}_j | y_j = i\}. \tag{2.52}$$

Then, the respective parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ can be estimated as shown in Section 2.3.1. The EM algorithm estimates the parameters $\hat{\mathbf{y}}$ and $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ iteratively until convergence as shown in Algorithm 2.1.

GMMs allow to model multi-modal distributions, e.g., a distribution of popular places. The parametrization allows to identify the centers and the shapes of k peaks, which might represent particular locations. GMMs can be seen as a statistically founded form of the KMeans clustering algorithm [Bishop, 2006, p. 424].

2.3.2 Non-parametric

Non-parametric density estimation estimates a density for a given set of observations without assuming a parametric distribution [Bishop, 2006, p. 435]. Without any assumptions on the functional form of the distributions, non-parametric models generally include all possible distributions

$$\mathcal{M} = \{\text{all possible distributions}\}. \tag{2.53}$$

However, in order to estimate a distribution consistently, smoothness assumptions need to be made [Wasserman, 2004, p. 359]. A possible smooth non-parametric model is

$$\mathcal{M} = \{f | \int f''(x)^2 dx < \infty\} \tag{2.54}$$

which is the set of functions that are not “too wiggly” [Wasserman, 2004, p. 107].

Non-parametric models are needed if the distribution of the observations is not well explained in a functional form. This is clearly the case for spatio-temporal distributions that explain processes happening on Earth. In the following we review Kernel density estimation (KDE) and histograms. Histograms are used extensively in this work. We also introduce KDE, since this non-parametric density estimation technique motivates an adapted form of the histogram estimator. Figure 2.4 shows density estimates using continuous KDE and a discrete histogram.

Kernel Density Estimation

Given a set of observations

$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}n\} \subseteq D_C. \tag{2.55}$$

A kernel

$$K_b(\mathbf{x}, \mathbf{x}_i) \in \mathbb{R}_+, \mathbf{x}, \mathbf{x}_i \in D_C \tag{2.56}$$

measures the influence of observation $\mathbf{x}_i$ at point $\mathbf{x}$. The parameter h is called the bandwidth of the Kernel, controlling the support of the influence around $\mathbf{x}_i$. A commonly used kernel is the Gaussian distribution such that

$$K_b(\mathbf{x}, \mathbf{x}_i) = \mathcal{N}(\mathbf{x}; \mathbf{x}_i, \mathbf{\Sigma}_b), \tag{2.57}$$

where the parameter b is the standard deviation of a symmetric covariance in the d-dimensional space

$$\mathbf{\Sigma}_b = \mathbf{I}_d \; b. \tag{2.58}$$

Kernel density estimation works by summing up the influences of all points at a given point $\mathbf{x}$ and normalizing the function accordingly by integrating to 1. The bandwidth b is now treated as a parameter of the model

$$f(\mathbf{x}; b) = \frac{1}{C_b} \sum_{i=1}^{n} K_b(\mathbf{x}, \mathbf{x}_i), \tag{2.59}$$

where C_b is a normalizing constant such that $f(\mathbf{x}; b)$ integrates to 1. Given the Gaussian kernel, KDE simplifies to

$$f(\mathbf{x}; b) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(2\pi b^2)^{1/2}} \exp\left\{ -\frac{||\mathbf{x} - \mathbf{x}_i||^2}{2b^2} \right\}. \tag{2.60}$$

Assume we want to evaluate $f(\mathbf{x}; b)$ at all center points of a lattice L. If the kernel K_b has infinite support (such as the Gaussian kernel), the complexity will be $O(n_L \times n)$, where n is the number of points and n_L is the number of cells. Given a kernel K_b with finite support proportional to b such that n_b cells are affected for an observation $\mathbf{x}_i$, the complexity reduces to $O(n_b \times n)$, which is likely to be much smaller than the first case.

Histogram

A histogram is the simplest non-parametric density estimation approach [Wasserman, 2004, p. 359]. Given a set of observations X and a lattice of equi-area spatio-temporal cells partitioning space D_C,

$$L = \{l_1, \ldots, l_m\}. \tag{2.61}$$

Let $b = \delta_C$ be the bandwidth of the histogram model. Choosing a larger bandwidth results in less cells and a coarser resolution, and vice versa. Let the resulting lattice be L_b. The histogram is a piecewise constant probability mass function $f(\mathbf{x}; b)$ defined as

$$f(\mathbf{x}; b) = \sum_{l_i \in L_b} \mathbf{1}\{\mathbf{x} \in l_i\} \frac{n_i}{n}, \tag{2.62}$$

where $n_i = \sum_{\mathbf{x}_j \in X} \mathbf{1}\{\mathbf{x}_j \in l_i\}$ is the number of points falling in cell l_i. By building a (possibly sparse) matrix of observation counts over the lattice cells, the density can be obtained in $O(1)$. Evaluating the density over the lattice will take $O(n)$ time.

Kernel-convoluted Histogram

Different from KDE the histogram is not continuous and the result is a less smooth distribution [Wasserman, 2004, p. 361]. We can, however, choose a histogram bandwidth $b_{hist} < b_{KDE}$ and smooth the histogram by convolution with a discrete kernel K. For the 2-dimensional case let $f(x, y)$ be a count matrix with $x, y \in [1; m]$. The convolution of f using K is defined as

$$(f * K)(x, y) = \sum_{x'=-k}^{k} \sum_{y'=-k}^{k} f(x - x', y - y')K(x', y'). \tag{2.63}$$

Histograms for Spatio-temporal Influence Signals

Given a number of records $r_1, \ldots, r_n$ and their spatio-temporal influence signals $z_1(\mathbf{c}), \ldots, z_n(\mathbf{c})$. The influence signals can be seen as a record-specific Kernel, with the same meaning as in KDE. This justifies to use KDE on the influence signals as

$$f(\mathbf{c}; b) = \frac{1}{C_b} \sum_{i=1}^{n} z_i(\mathbf{c}). \tag{2.64}$$

A histogram estimator can be used by first discretizing the continuous signals on a lattice L_b and then summing up and normalizing the influences along the cells

$$f(l; b) = \frac{1}{C_b} \sum_{i=1}^{n} z_i(l) \tag{2.65}$$

with

$$C_b = \sum_{l_j \in L_b} \sum_{i=1}^{n} z_i(l_j). \tag{2.66}$$

2.3.3 Scale and Bandwidth

Scale defines the resolution at which a geographic phenomena is analyzed. This is true for the spatial scale and a temporal scale. We use the term *scale level* to refer to a certain spatio-temporal, spatial, or temporal resolution b. In the following, we only discuss the influence of the spatial scale level on the semantics of geographic phenomena. However, the discussion is similarly valid for the spatio-temporal and temporal scale.

Scale Level

A scale level defines the resolution of analysis. For example, the following scale levels can be defined:

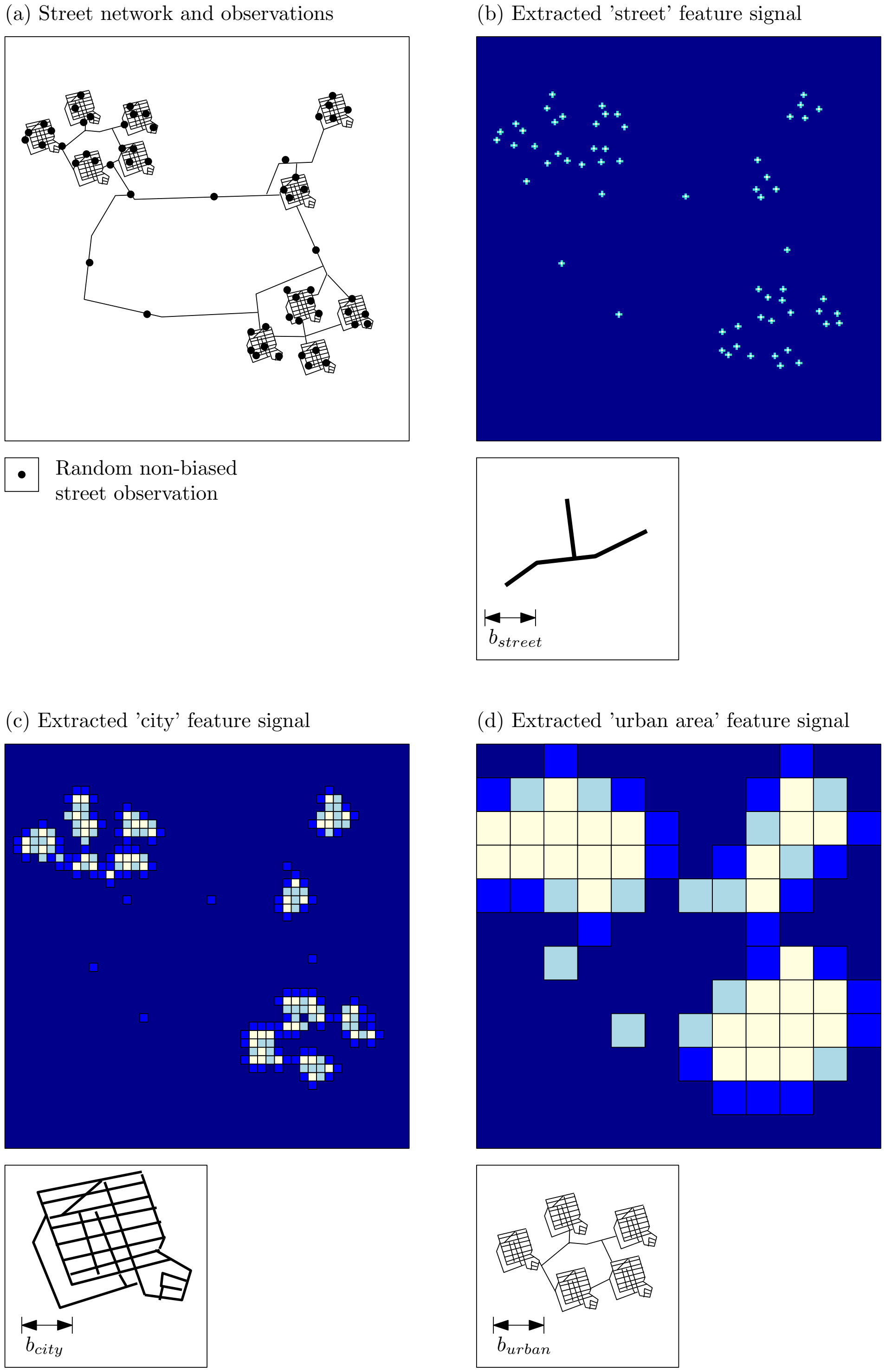
- *Street level*: Phenomena such as roads or the number of cars on a road. The resolution in which this phenomena changes is in the order of meters.
- *City level*: Phenomena such as city districts with high crime or a lot of tourists. The resolution in which this phenomena changes is in the order of kilometers.
- *Country level*: Phenomena such as regions with high unemployment rates or physical entities such as mountains. The resolution in which this phenomena changes is in the order of 10-100 of kilometers.
- *Global level*: Phenomena such as continents or countries with high number of educated people or high usage of iPhones. The resolution in which this phenomena changes is in the order of 100-1000 of kilometers.

We say a low scale level has a fine/high resolution, a high scale level has a coarse/low resolution.

The semantics of a geographic phenomena changes depending on the scale level of analysis. Figure 2.5 shows a street network acting as a physical phenomena. On a street level the concrete shape of the streets describes the network itself. On a higher level, clusters of streets and the shape of the cluster might describe districts or small towns. On an even higher level, clusters of districts might describe city regions.

Scale and Non-Parametric Estimators

The bandwidth of non-parametric estimators is directly related to the resolution of analysis, and hence to the scale level of interest. We refer to the example provided in Figure 2.5. Given a number of observations X over the space D_S that happen proportional to the number of street segments in their surrounding. A density estimation on different scale levels b_{street}, b_{city}, b_{urban} will result in signals having different semantics. Using b_{street} we may are able to recover the street network. Using b_{city} we may are able to recover populated districts (e.g., towns). Using b_{urban} we may be able to recover urbanized regions. Hence, the parameter b changes the semantics of the estimated phenomena and must be treated as a user-defined parameter in the mining process: Choosing a particular bandwidth influences what types of phenomena are to be mined.


(a) Street network and observations

(b) Extracted 'street' feature signal

(c) Extracted 'city' feature signal

(d) Extracted 'urban area' feature signal

Figure 2.5: Density estimates of spatial points using different smoothing bandwidths. The sample points represent observations of streets (such as photos showing streets or cars). The bandwidths correspond to scale levels and result in signals of different semantics (streets, cities, urban areas). More observations are needed to extract a meaningful signal representing the street network. However, enough observations exist to extract a signal describing the distinct cities and the urban areas.

Scale and Parametric Estimators

Parametric models such as Gaussian distributions or Gaussian mixture models do not have a bandwidth parameter. However, their parametrization is also related to the scale level of interest.

In a Gaussian distribution the covariance matrix describes the shape of the signal. Scale level information can be given as prior knowledge, e.g., by forcing the shape to be symmetric and having a given standard deviation as shown in Eq. (2.13). Without restrictions on the covariance matrix the estimated Gaussian can be seen as choosing the best scale for the provided data.

For GMMs and KMeans, the number of clusters k affects the scale level of interest. If a number of points is assumed to be described by a multimodal distribution, a small number of centers will induce a high scale level (low resolution), while a large k will induce a low scale level (high resolution).

2.3.4 Robustness

As discussed in the previous section, the semantics of a signal obtained by model $\mathcal{M}$ depends on the bandwidth b. The bandwidth can, however, not be chosen arbitrarily. It depends on:

(a) The area of interest D_C. The bandwidth must be chosen small enough to be able to describe the phenomenon structure within D_C.

(b) There is a lower bound b_{min} on the bandwidth that depends on the number and the spatio-temporal accuracy of the observations X. An estimated signal on a low scale level with only a few observations will sufficiently describe the phenomenon.

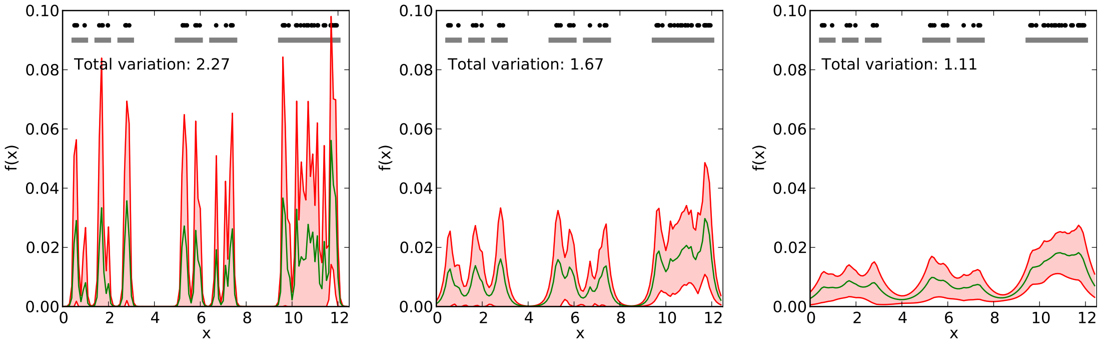
In the following, we review a technique to asses the robustness of a model. We call a signal that only changes slightly for small variations in the input data a *robust signal.* A signal that varies heavily even for small variations of the input data is called a *non-robust signal* (detailed later). This will allows us to filter signals in which we cannot be confident, e.g., because it is based on a too small sample of observations. We describe how to determine confidence bands for an estimated signal. The confidence band can be visualized together with the signal to support the discovery process. An aggregation on the total variability of the signal can also be used to filter non-robust distributions.

Bootstrap

We employ a technique from statistics to compute confidence intervals for arbitrary models $\mathcal{M}$, called the bootstrap [Wasserman, 2004].

Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subseteq D_C$ be the input observations. A set X_* is a random sample with replacement of the input observations X of size n (hence, of the same size as the input data). In X_* some observations are missing and some are duplicated. We denote k sampled sets as

$$X_*^{(1)}, \ldots, X_*^{(k)}. \tag{2.67}$$



(a) $b = 0.05$ (b) $b = 0.2$ (c) $b = 0.6$

Figure 2.6: Confidence intervals based on the bootstrapped standard error estimate for different bandwidths.

Given a model $\mathcal{M}$ and its density function $f(\mathbf{x};\theta)$. Let $f(\mathbf{x};\theta)^{(i)}$ be the estimated density on the basis of set $X_*^{(i)}$. The sample mean of the signal at point $\mathbf{x}$ over the sampled densities is

$$mean[f(\mathbf{x};\theta)] = \frac{1}{k}\sum_{j=1}^{k} f(\mathbf{x};\theta)^{(j)}. \tag{2.68}$$

The sample variance of the density at point $\mathbf{x}$ is

$$svar[f(\mathbf{x};\theta)] = \frac{1}{k}\sum_{i=1}^{k}\left(f(\mathbf{x};\theta)^{(i)} - \frac{1}{k}\sum_{j=1}^{k} f(\mathbf{x};\theta)^{(j)}\right)^2. \tag{2.69}$$

The standard deviation describes the variation of the signal based on small variations of the input data.

By assuming the variation to be the standard error of the signal, we can construct a confidence band by a normally distributed error. The 95%confidence band is then,

$$(mean[f(\mathbf{x};\theta)] - 2svar[f(\mathbf{x};\theta)], mean[f(\mathbf{x};\theta)] + 2svar[f(\mathbf{x};\theta)]). \tag{2.70}$$

This confidence band describes the range of the estimated signal values in 95% of the time. The bootstrap can be used on every model since it only depends on the input data. Let the runtime complexity of the model be $O(M)$, then the complexity of the bootstrap is $O(kO(M))$.

Bandwidth Selection

Figure 2.6 shows a KDE-estimated signal using different bandwidths and their 0.95 percent confidence bands computed by the bootstrap. The input data consists of points randomly sampled over the context domain (here 1-dimensional) and only the points overlapping with a set of intervals are kept as observations. The data can be seen as observations of intervals of different lengths in the context domain. The intervals are arranged in three clusters, where the sub-intervals in each cluster become larger from (a) to (c), with (c) consisting of only a single interval. The different interval lengths correspond to different semantics expressed on different scale levels.

Figure (a) shows the resulting signal and the confidence bands using a bandwidth that allows to recover the short intervals on the left. The resulting signal shows several peaks at the interval locations, however, the confidence band indicates that we cannot be sure that they are just showing up by chance.

Using larger bandwidths, corresponding to the scale levels of the middle and the right interval lengths, the confidence band of the signals becomes more narrow. In Figure (b) peaks within the left and center intervals exists, however, they are still rather noisy. In particular, the lower band often hits the x-axis. Finally, in Figure (c), the three interval clusters are estimated with a relatively narrow confidence band. Even the lower band (worst-case) shows the three interval clusters.

The increasing density from the left to the right clusters in Figure (c) reflects the intensity of the phenomena measured by the signal. As gaps exist in the left and the center cluster, the density is lower. As a result, the signal (c) is robust with respect to the input data. Robust estimations for (a) and (b) will need more observations to obtain a robust estimate.

2.4 User-generated Data

The input to the data mining framework and the problems proposed in this thesis comes from user-generated data sources. Such data sources are often unstructured, such as textual messages, text documents, images, and videos. Moreover, they often comprise complex relationships between different kinds of information, e.g., between users, documents, text, links, locations, check-ins, among others. In the following we introduce concepts and methods to process and model such data sources.

2.4.1 Unstructured Data

User-generated data sources often include text documents, images, and videos. These content types contain attributes and spatio-temporal information in a non-structured form, i.e., implicitly covered in language or image pixels. An important pre-processing step is hence to extract basic information items from the records. In the following, we review techniques and terminology from the fields of natural language processing and computer vision related to feature representation and information extraction.

Feature Representation

Often the term *unstructured data* is used to refer to records whose information is not represented by an explicit data schema or attributes [Weiss et al., 2005, p. 2]. The term unstructured data is slightly misleading, since the records of course have an implicit structure to convey the information [Feldman and Sanger, 2007, p. 3]. Such implicit structure is exhibited by language and image pixels. An important task in unstructured data analysis is to leverage different kinds of elements in the records in order to transform it from an irregular and implicitly structured representation into an explicitly structured representation. For this, pre-processing operations are used to transform the raw, unstructured, original-format content into a carefully structured, intermediate representation [Feldman and Sanger, 2007, p. 3-4].

The extracted information items are often called features. The term stems from its usage in machine learning, where a feature is an attribute of an observation [Hastie et al., 2009, p. 9]. In text mining the term *document feature* is used to describe basic information items such as characters, words, or terms [Feldman and Sanger, 2007, p. 5-6]. In computer vision features include extracted corners, edges, textures, or shapes [Davies, 2012]. The above features are used to represent parts of a record or a record as a whole and is called the *feature representation* of a record.

Feature representations of the records are then used for subsequent tasks in text mining and computer vision. One such task is the extraction of high-level features, such as concepts and entities in text mining [Feldman and Sanger, 2007] or objects in images and videos [Davies, 2012]. Such high-level feature extraction tasks use low-level features as underlying data representations. To extract high-level features even other data sources can be exploited, called background or domain knowledge [Feldman and Sanger, 2007, p. 42]. For example, attributes of extracted entities in a text messages can be used as features to represent the document.

We denote a collection of unstructured data records such as text documents, images, or videos as

$$R = \{r_1, \ldots, r_n\}. \tag{2.71}$$

We use the term *information extraction* to refer to any task that extracts low- or high-level features from unstructured data records. A set of features is denoted

$$F = \{f_1, \ldots, f_p\}. \tag{2.72}$$

The function that represents a given record by a vector of features is called a *feature extraction function*

$$\psi_F(r) \in \mathbb{R}^p_+, r \in R. \tag{2.73}$$

The function returns a positive real value for each of the p features. The value of a single feature f in record r is

$$\psi_f(r) \in \mathbb{R}_+, r \in R. \tag{2.74}$$

It is also called the influence or the signal of feature f in r.

Feature Types

Depending on the data type, different kinds of features can be extracted. In the following we shortly list features types for different types of unstructured data that are meaningful for geographic feature mining.

Textual Features. Low-level textual features include words and terms. Removing those features that exhibit only a small amount of information, since they occur more or less equally common in the documents, is called stop-word removal. Words and terms can also be stemmed to reduce verbs and nouns to a common basis. Using such low-level textual features can be seen as a domain-independent representation. If no prior assumptions on the semantics of the features is proposed, such a representation can be used as a starting point.

High-level features are often domain dependent. For example, if phenomena related to products or persons are of interest. In this case, high-level features need to be extracted from the documents first. According to [Feldman and Sanger, 2007, p. 96], the following high-level features can be extracted from text:

- *Entities*: Basic building blocks of documents, e.g., people, companies, locations, genes, drugs.
- *Attributes*: Specifications of entities, e.g., age of a person, type of organization.
- *Facts*: Static relations between entities.
- *Events*: Dynamic relations between entities.

Given textual data, spatio-temporal information can be extracted on the basis of mentioned places and times [Lieberman, 2010; Strötgen and Gertz, 2013]. These entities constitute a geographic scope of the content and can be used as spatio-temporal information, similar to associated GPS coordinates or timestamps. We say this spatio-temporal information has been extracted using geographic and/or temporal expressions.

Image and Video Features. Low-level image features, such as lines, corners, textures and shapes can be used as a domain independent feature set. A common set of features (not necessarily positive-valued) is the set of MPEG features [Le Gall, 1991]. Another set of low-level features are SIFT features [Lowe, 2004].

In the context of geo-referenced social media, low-level image features have been used to predict photo locations [Hays and Efros, 2008] and to find representative images for places and windows in space and time [Chen and Roy, 2009; Crandall et al., 2009; Rattenbury et al., 2007b].

However, for the task of geographic feature mining we assume that low-level features are first transformed into meaningful high-level ones. Today, efficient algorithms exist to recognize objects in images and videos [Viola and Jones, 2001]. The techniques use a training collection of images/videos with annotated objects and are able to detect objects in images and videos with high precision. A very valuable set of features is hence a set of objects. Each image can then be represented by a set of objects shown in it.

Social Media Tags. Tags are user-defined terms associated with records. They are initially meant to index the records by a user defined vocabulary [Davis, 2006]. In today's social media services tagging of records is very common. Thereby, a tag of an image can be seen as a weak high-level feature, explaining the content of the image. Hashtags in Twitter messages play the same role in assigning a topic to a short message.

In this work, we use tags as high-level user-provided record features. The value of tags as high-level record features has been shown in a variety of works. Different semantics of tags have been analyzed in [Davis, 2006]. The semantics of different temporal distributions has been analyzed in [Dubinko et al., 2007]. Together with low-level image features, they have also been used to detect places and events [Crandall et al., 2009] and for place recommendation [Shepitsen et al., 2008].

2.4.2 Complex Data

User-generated data is not only often unstructured, but also comprised of relationships between information items, such as between users, links, documents, words, and locations. Here we assume that basic information items have already been extracted from the data using techniques explained in the previous section.

Current research in data mining deals with knowledge discovery in such *complex data sources*. As described in [Han et al., 2012, p. 585], complex data includes sequence data, graph and network data, spatio-temporal data, multimedia data, and text data, among others. In this thesis we use the term complex data for data sources that contain (different types of) information items that are connected to each other.

Generative Models

A recently popular technique to mine knowledge from complex data are generative models [Bishop, 2006, p. 365]. Let X, Y, Z be a number of variables describing some type of information in the data source (e.g. documents, words, users). A generative model allows to generate samples from the joint distribution

$$\mathrm{p}(x, y, z; \Theta), \tag{2.75}$$

where Θ represent the parameters of the distribution. To achieve this, a generative model needs describe the joint distribution of the variables. In contrast, a discriminative model only describes a conditional distribution, e.g.

$$\mathrm{p}(z|x, y; \Theta). \tag{2.76}$$

Note that a generative model is more general than a discriminative one, since a conditional distribution can always be obtained from a joint distribution by integration and Bayes rule.

Direct modeling of the joint distribution as a single function is often not possible since the dependencies between variables and parameters are too complex [Kollar and Friedman, 2009, p. 3]. In addition, estimates of a function with a huge number of variables are often weak, since not enough data is available to fit the huge number of parameters needed to describe the functional relationship [Kollar and Friedman, 2009,

p. 5]. To overcome this problem, a joint model is defined based on independence assumptions among the variables. For this, the joint distribution is assumed to factor in a model-specific way. For example,

$$\mathrm{p}(x, y, z) = \mathrm{p}(x|y, z; \Theta_1)\,\mathrm{p}(y|z; \Theta_2)\,\mathrm{p}(z; \Theta_3). \tag{2.77}$$

Each factor is represented by a distinct (parametrized) function. Since now the interaction between some variables does not need to be modeled anymore, the joint model needs a smaller number of parameters. The joint distribution can be described as a graph of factors, where the conditional dependence between the factors is represented as directed edges. Each factor is a conditional probability distribution (CPD), and in a generative model, the product of all CPDs resembles the joint distribution.

The graph represents the hierarchical relationship between variables. Because of this, these models are also called *hierarchical models*. Another used phrase which refers to the factor graph is *directed graphical model*. Finally, a synonym to directed graphical model is *Bayesian network*. We use the latter term later in Chapter 4.

Using a graphical model (often a generative one) to describe the statistical structure of complex data involves:

- *Definition of conditional probability distributions (CPDs) and network structure*: During this step, often hidden (latent) variables are introduced to realize assumed but unknown relationships between variables. The CPDs and the network structure encode knowledge and assumption about how the different types of information interact with each other in a graphical model.

- *Estimation of parameters*: This task is generally model independent, however, often models are built in a way such that parameter learning is possible for a huge amount of data [Kollar and Friedman, 2009, p. 5]. Prominent learning techniques are sampling methods (Markov Chain Monte Carlo; MCMC), expectation maximization (EM), and variational inference.

We refer the reader to [Bishop, 2006] and [Hastie et al., 2009] for an overview of different types of graphical models and estimation techniques, and to [Kollar and Friedman, 2009] for a detailed introduction into the topic.

Topic Models

Topic models are a specific type of generative model that are aimed to extract topics occurring in text documents. A topic is similar to a category of a document. A topic is, however, not a unique label, but a distribution over words. Since the topics are not known in advance, the task of topic modeling is to infer the topics from the data.

Let D, W, Q be the variables describing the documents, words, and topics, respectively. Since Q is not known but expected to exists as a variable in the model, it is called a latent variable. Two well-known topic models are Probabilistic Latent Semantic Analysis (PLSA) [Hofmann, 1999] and Latent Dirichlet Allocation (LDA) [Blei et al., 2003]. The models allow to predict the distributions p($q|d$) and p($w|q$). These describe the probability that a topic belongs to a document and that a word belongs to a topic, respectively. Both models have been extended to incorporate other types of information in complex data sources (e.g., timestamps, location, links, users). We will detail several of these approaches in the next chapter.

2.5 Summary

In this chapter, we detailed the research context of this thesis and introduced necessary definitions and techniques to process spatio-temporal and user-generated data.

We focused on a signal-based representation of geographic information, which is an essential concept for the development of the geographic feature mining framework in the following chapter. In particular, we showed how spatial data types like points, polygons, and bounding boxes can be understood as and transformed to spatio-temporal intensity signals. Moreover, we reviewed density estimation of spatio-temporal data, and discussed its challenges and peculiarities. In this context, we discussed scale, estimation bandwidth, and robustness, which are essential when dealing with noisy and uncertain data but are often neglected in related works. Finally, we introduced concepts and methods to process common types of user-generated data, namely unstructured and complex data. There, we reviewed feature-based data representations and introduced information extraction techniques.

Chapter 3

Geographic Feature Mining: A Unifying Framework

3.1 Introduction

A huge amount of research emerged in the last decade that addressed the utilization of various forms of geographic information in user-generated data to realize novel applications and analysis methods. Since the data sources, the types of spatio-temporal information, and the addressed problems are very different from each other, this body of research is highly heterogeneous, application-, and data-specific.

In this chapter, we propose a data mining framework to systematically discover geographic knowledge from diverse types of user-generated data sources in the form of geographic feature signals. In short, a geographic feature is defined as a dimension of space and time and is described by a spatio-temporal signal. The aim of geographic feature mining is to extract and discover interesting features that convey knowledge about geographic phenomena, e.g., social and cultural habits, physical processes, or places and events. We show in the remainder of this chapter that this simple framework conception allows to define a huge number of different applications and tasks found in related work. Herewith, we also provide a big picture on the problem of mining geographic knowledge from user-generated data in general.

We first give an high-level overview of a heterogeneous body of research that deals with the extraction of geographic knowledge from user-generated data and related applications. Thereby, we focus on identifying common problems and applications and describe various underlying models in a unifying probabilistic fashion to uncover their similarities.

Then, on the basis of a fundamental set of problems, applications, and insights into the underlying models, we present a conceptual data mining framework (as defined in Section 2.1.1) that allows to systematically extract geographic knowledge in a highly data independent manner. We summarize the key features of the proposed framework as:

- *Unifying*: This allows to use different kinds of data sources and to define various applications and mining tasks.
- *Systematic*: This allows to represent the information in the data and the output patterns by simple primitives. By this, we are able to focus on the essential problems and questions (what is to be measured, what is the semantics of a pattern), and to use the right tools and models to address them. Moreover, a mining process allows to extract knowledge in an iterative manner.

The remainder of this chapter is organized as follows. In Section 3.2, we start with reviewing related work and identifying common applications and models. Then, we introduce concepts and data representations to model the input data in Section 3.3 and the output patterns in Section 3.4. In Section 3.5, we present interestingness measures of geographic feature signals. In Section 3.6, we propose a process and (sub-)tasks to extract interesting geographic knowledge from the data.

3.2 Comparison of Models and Applications

This section gives an overview of related work dealing with applications and models to extract geographic knowledge from user-generated data. We present works from a diverse range of fields, such as multimedia, computer vision, information retrieval, Web technology, and data mining.

The particular works are often very application- and data-specific. Our aim is to identify common concepts and methods to extract geographic knowledge from user-generated data for particular tasks and applications. For this, we focus on the underlying models that describe the spatio-temporal distribution of information contained in the data. Such models (and the resulting spatio-temporal distributions) are needed for a variety of applications:

(1) Location prediction of records and users on the basis of the record content or users' friends.

(2) Detection and description (summarization, labeling) of popular places and events from photos, documents, and text messages.

(3) Detection, description and ranking of user trajectories from recorded user locations.

(4) Recovery of geographic phenomena such as accidents/crimes, natural disasters, and epidemics from photos, text, and search queries.

(5) Analysis of the spatio-temporal variation of textual topics (spatio-temporal topic models).

(6) Supervised learning of spatio-temporal models based on user-generated data for prediction and forecasting.

(7) Spatio-temporal visualization and indexing of user-generated records.

We organize the related work by their underlying spatio-temporal models and use a unifying notation to describe them in a probabilistic fashion. The user-generated data sources contain different kinds of information, which we represented by the following variables:

- $r \in R$: A document, record, query, or user action in the data source.
- $f \in F$: A discrete (categorical) feature such as a word, object, characteristic image element, or normalized query string extracted from the records. The records might be represented by a large number of features (e.g., the words in a document).
- $q \in Q$: A high-level feature or concept that is aimed to be found by using a model. For example, spatial clusters of records, a cluster/subset of features, or a latent variable describing a distribution over features.
- $u \in U$: A user who created a record.
- $s \in D_S$: A spatial point in continuous spatial space or an area in discrete spatial space.
- $t \in D_T$: A timestamp or interval in temporal space.
- $c = (s, t) \in D_C$: A point or a window in spatio-temporal space.

This set of variables describes information in a user-generated data source, and we can think of an underlying generative process that generates the data (see Section 2.4.2). The process is described by the joint distribution over all variables (if they exist)

$$\mathrm{p}(r, f, q, u, s, t). \tag{3.1}$$

To describe geographic information in the data, the relationship between features $f \in F$ or high-level features $q \in Q$ and spatio-temporal space $(s, t) \in D_C$ are particularly important. This relationship is expressed in the distributions

$$\mathrm{p}(s, t), \mathrm{p}(s, t, f), \mathrm{p}(s, t|f), \mathrm{p}(f|s, t), \mathrm{p}(s, t, q), \mathrm{p}(s, t|q), \mathrm{p}(q|s, t). \tag{3.2}$$

We show that these probabilities are in fact often used to realize particular applications, such as extracting places and events, estimating locations of records, or predicting geographic phenomena.

Most of the works do not present their ideas in a probabilistic fashion, for example, by describing their approaches on the basis of clustering, heat-maps, or count vectors. A major effort of the following review is to reduce these approaches to a simple probabilistic description. Uncovering and classifying the underlying models allows to identify common concepts and methods in a heterogeneous body of research, and, as a consequence, to formulate a unifying framework to mine geographic knowledge for a variety of tasks.

3.2.1 Place and Event Models

A large body of research deals with the identification of places and events from geo-referenced records, such as photos or text messages. Usually, places and events are defined as popular phenomena that happen in a small spatial area and/or time interval.

Clustering-based Approaches

We first review works that cluster the records spatially or spatio-temporally based on their geographic information. The resulting clusters are supposed to represent popular places or events and describe an area or spatio-temporal window. The clusters are used for various applications, such as the organization and visualization of photo collections, extraction of points-of-interests (POIs), or the prediction of record location on the basis of the record content.

In [Ahern et al., 2007] the authors propose a techniques to extract labels for interesting places and events to organize photo collections and to visualize landmarks on a map. The input are geo-referenced and tagged photos. The authors extend their approach to extract representative images for discovered places and events in [Kennedy et al., 2007] and [Kennedy, 2008]. In the following, we detail their approach to extract places and place descriptions.

As a pre-processing step a spatial grid is used to partition the photos into medium-scale areas of interest. A number of places is then extracted in each partition separately. K-means is used to identify spatial clusters inside each partition. The number of clusters is chosen by the number of photos in the partition or by choosing a fix number (around 20) that allows to present a reasonable number of places on a map. Spatial clustering finds the modes of the underlying density distribution of the photos. Since K-means can be described by a Gaussian Mixture Model, the underlying distribution of the photos is

$$\mathrm{p}(s) = \mathrm{GMM}(s; \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{3.3}$$

Given a cluster q (representing a place) with centroid s_q and covariance matrix Σ_q, the cluster-conditional spatial distribution can be described as

$$\mathrm{p}(s|q) = \mathcal{N}(s; s_q, \Sigma_q). \tag{3.4}$$

Hence, the photos inside a partition are supposed to represent a number of places or events with each instance having a single center s_q and a particular shape Σ_q.

Even if not mentioned in the papers, we note that the partitioning scheme normalizes the medium-scale background density (see spatio-temporal heterogeneity in Section 2.1.2). By clustering the whole spatial space, regions with high background density would consume most of the clusters, while less populated areas will have no clusters at all. By prior partitioning each area is assumed to have a fixed number of clusters, which homogenizes the number of clusters in space.

To extract representative tags for each cluster a TF-IDF-like scoring function of the tags is used. Given a tag f inside a cluster q, the score is defined as

$$\mathrm{score}_{Ken}(f,q) = \overbrace{\mathrm{count}_r(f,q)}^{\mathrm{TF}} \times \overbrace{\frac{\mathrm{count}_r(\cdot)}{\mathrm{count}_r(f)}}^{\mathrm{IDF}}, \tag{3.5}$$

where count_r counts the number of records supporting the given arguments (e.g., having tag f and/or being assigned to cluster q), and $\mathrm{count}_r(\cdot)$ is the number of all records. To limit the influence of users contributing large amounts of photos, the score can also be computed based on the user counts

$$\mathrm{score}'_{Ken}(f,q) = \mathrm{count}_u(f,q) \times \frac{\mathrm{count}_u(\cdot)}{\mathrm{count}_u(f)}. \tag{3.6}$$

Here, count_u counts the users supporting the given arguments. Such an adapted counting scheme is used in several papers and is not mentioned for the following works anymore.

In [Crandall et al., 2009] a very similar approach is used to extract places and place summaries. Here, the mean-shift clustering algorithm is used to find dense photos locations, and no prior partitioning into regions is assumed as a pre-processing step. Using mean-shift allows to determine the number of clusters from the data by choosing a bandwidth b [Comaniciu and Meer, 2002]. For each cluster, the representative tags are extracted using

$$\mathrm{score}_{Cran,R}(f,q) = \frac{\mathrm{count}_r(f,q)}{\mathrm{count}_r(f)}. \tag{3.7}$$

We note that this score is just the tag-conditional cluster probability

$$\mathrm{p}(q|f) = \frac{\mathrm{p}(f,q)}{\mathrm{p}(f))} = \frac{\frac{\mathrm{count}_r(f,q)}{\mathrm{count}_r(\cdot)}}{\frac{\mathrm{count}_r(f)}{\mathrm{count}_r(\cdot)}} = \frac{\mathrm{count}_r(f,q)}{\mathrm{count}_r(f)}. \tag{3.8}$$

Note also that $\mathrm{score}_{Ken}(f,q)$ and $\mathrm{score}_{Cran}(f,q)$ behave very similar. In fact, they are the same if $\mathrm{score}_{Ken}(f,q)$ is multiplied by the constant $\mathrm{count}_r(\cdot)$. Other TF-IDF-like scores (e.g., using logged IDF) will also behave similar. Another similar work about place extraction and summarization on the basis of clustering and scoring is presented in [Jaffe, 2006].

In [Yin and Cao, 2011] the clustering-based approach is used as a pre-processing step. Here, a number of places have been extracted from a photo collections and used as points-of-interest (POIs). The identified POIs are then used to mine and rank interesting POI trajectories of tourists.

In [Deng and Lemmens, 2009] the photos are spatially clustered using DBSCAN. These clusters cannot be described by a single center, but by a density distribution

$$\mathrm{p}(s|q) = \mathrm{KDE}_q(s;\varepsilon), \tag{3.9}$$

where $\mathrm{KDE}_q(s;\varepsilon)$ is the Kernel-density estimate of the points assigned to cluster q, and ε is the distance parameter of DBSCAN. Areas of the clusters can be extracted by a convex hull of the cluster points or based on a confidence region of the extracted KDE estimate. The aim of the authors is to use the tag-tag correlation matrix of the photos in a cluster to analyze different conceptualizations of space. In [Kisilevich et al., 2010] the authors propose a modified DBSCAN algorithm to extract more equally sized regions, which are then summarized by representative tags as shown above.

A slightly different approach to extract events from text messages in proposed in [Becker and Gravano, 2010]. There, the records are clustered not only on the basis of the geographic information, but jointly based on features (tags, title, description), spatial, and temporal information. This results in clusters describing sets of records with similar content in spatio-temporal proximity. To achieve a joint clustering the authors use an ensemble-clustering method. Distance functions for feature-space (bag-of-words representation of the records), spatial space (geodesic distance), and temporal space (distance between timestamps) are defined separately. Then, the cluster-ensemble method combines the separate clusterings on each of the distances into a single joint clustering result. By this approach the clusters might overlap in a region and/or time interval if their content is different but their spatio-temporal information is not. This work is less focused on the spatio-temporal distribution of events. However, the distribution could easily be estimated on the basis of the records assigned to event clusters (as shown in (3.9)), or on basis of the spatio-temporal centroid. To determine the most representative label of overlapping events, the clusters would have be scored first (e.g., by choosing the cluster with the higher number of records, equally to $\mathrm{p}(q)$).

All of the approaches above allow to extract representative features for each cluster q by choosing the features with highest probability $\mathrm{p}(q|f)$. We generalize this idea and describe the representativeness of features for an arbitrary point or area in space s. Let A_q be the region that contains all records assigned to cluster q and lets assume that the regions are non-overlapping. In this case $\mathrm{p}(q|f)$ and $\mathrm{p}(A_q|f)$ are the same, since A_q just represents the fraction of space that is covered by cluster q. If we do not know about the clusters we can use any disjoint set of regions $A_1, \ldots, A_n$ to compute $\mathrm{p}(A|f)$ and by shrinking the areas we can assume to end up with a single point $\mathrm{p}(s|f)$. Note that $\mathrm{p}(s|f)$ can be computed by using Kernel density estimation and a given bandwidth b,

$$\mathrm{p}(s|f) = \mathrm{KDE}_f(s; b), \tag{3.10}$$

where $\mathrm{KDE}_f(s; b)$ denotes KDE of all records having feature f. Hence, the clustering approach can be seen as a segmentation of spatio-temporal space into dense regions $A_{q_1}, \ldots, A_{q_k}$ and the representativeness is determined by

$$\mathrm{p}(q|f) = \mathrm{p}(A_q|f) = \int_{A_q} \mathrm{p}(s|f) ds. \tag{3.11}$$

As a result, we see that $\mathrm{p}(s|f)$ is an important probability to determine the representativeness of a feature at a point s.

Another application of the spatial clustering model is proposed in [Crandall et al., 2009] and [Cao et al., 2009]. There, it is used to predict the location of records by the record content. In both approaches a set of photos with associated spatial GPS information are spatially clustered to obtain important places (modes in the density distribution $\mathrm{p}(s)$). Hence, they use the clustering model to extract locations with a high likelihood of occurring photos. Then, for each cluster, a classifier (SVM, canonical correlation analysis) is learned using a feature representation of the records based on the image features or the tags. The classifier is trained using the photos in the cluster as positive instances, and all other photos as negative instances.

To predict the location of an unseen record, the classifier's confidence values for a positive labeled prediction are determined, and the location of the classifier having the highest confidence is used as the records location estimate.

For this task, the extraction of places is used as a pre-processing task. The idea can be generalized by assuming a discrete distribution $\mathrm{p}(c)$ over a given spatio-temporal grid and training a classifier for each cell.

All of the clustering-based models are working on continuous spatial or spatio-temporal data, hence no discretization takes place. Only in [Ahern et al., 2007; Kennedy, 2008; Kennedy et al., 2007] the background distribution of photos has been taken into account. The other approaches are biased by spatio-temporal heterogeneity. This might be justified if only the most popular places and events in the whole area are important for the task.

Unimodal Distributions

The above works assume the records to define a multimodal spatio-temporal distribution, where each peak can be represented by a unimodal instance. Other works assume that the spatio-temporal information of records and/or features describes a single place or event instead. Such models can be used to decide if a set of records or a feature exhibits a place or event semantics.

In [Liang et al., 2010] the authors are interested in extracting tags from a photo collection that are highly predictive to identify a unique place. For this, they fit a symmetric Gaussian distribution to the spatial record coordinates of all records that have a particular tag. The resulting tag-conditional spatial distribution can be described as

$$\mathrm{p}(s|f) = \mathcal{N}(s; s_f, \sigma_f). \tag{3.12}$$

To filter out tags that describe a particular place, they choose those tags that have a small estimated standard deviation $\hat{\sigma}_f$.

A more complex unimodal model is described in [Backstrom et al., 2008]. In this work the authors aim is to describe the spatial focus of queries in a search engine on the basis of the user locations. For this, the authors propose a model to extract the center and the shape of user location distributions for a given set of queries. The locations of users is obtained by the IP-address, the scale level of interest is on a country to global level.

They assume that a query f has a single center s_f (unimodal) and the likelihood of a user at location s to submit this query is described by

$$g(s; f) = g(s; \gamma_f, s_f, \alpha_f) = \gamma_f ||s - s_f||^{-\alpha_f}, \tag{3.13}$$

where γ_f is a frequency parameter of the query (height of the bump), and α_f a shape parameter (called dispersion in their paper) describing the form of the bump. A small α_f will result in an almost uniform distribution, while a large α_f results in a single peak with exponential characteristics. After normalization this function can be described as the probability that a location s belongs to a query f

$$\mathrm{p}(s|f) \propto g(s; \gamma_f, s_f, \alpha_f). \tag{3.14}$$

Given this functional form of the distribution, the parameters γ_f, s_f and α_f can be obtained by maximum likelihood estimation using an optimization routine presented in the paper. Notably, in [Cheng et al., 2010] their approach has been used to filter tags $f \in F$ that have a small amount of locality, by thresholding the dispersion parameter α_f.

The same model has been used by the authors to predict the location of Facebook users [Backstrom et al., 2010]. Based on empirical analysis they find that the distance of a user to its friends can be described by

$$g(||s_u, s_v||) = 0.0019\ (||s_u - s_v|| + 0.196)^{-1.05}, \tag{3.15}$$

which is of a similar functional form as (3.13). Given friends with known location as training data, the location of a user can be found using maximum likelihood optimization.

The above unimodal models can be used if $\mathrm{p}(s|f)$ or $\mathrm{p}(s|q)$ are assumed to have a single peak, such as a place or event. The clustering models can be seen as mixture models of such unimodal distributions, i.e., the GMM is the mixture model of Gaussian distributions. In [Backstrom et al., 2008] also a mixture model for their exponential distribution is proposed to model queries with several centers. No comparison between these two models has been performed, however.

3.2.2 Heat-Map Visualizations and KDE Models

The former models estimate spatio-temporal densities based on continuous spatio-temporal points and by assuming particular distribution characteristics. In the clustering-based models distributions are expected to have a number of bumps, with each bump representing a certain phenomenon (e.g., a place or an event). In the unimodal models the phenomena are expected to be described by a single bump, e.g., a single place or event.

The most general category of geographic phenomena are processes having an arbitrary distribution. Given continuous spatio-temporal data, such distributions can be estimated using non-parametric density estimation. Such models make no assumptions on the shape of the distribution and are not described by a parametrized function. Instead, they expect a parameter describing the smoothness of the distribution, called the bandwidth (see Section 2.3). Phenomena that lack a functional form are, for example, the amount of traffic accidents, or natural phenomena (beach, coast).

A number of works copes with the spatial visualization of such phenomena by using heat-maps. A heat-map is just a non-parametrized density estimate. Given continuous data a heat-map can be described by a Kernel-density estimate with bandwidth b,

$$\mathrm{p}(s) = \mathrm{KDE}(s; b). \tag{3.16}$$

Spatial heat-map visualizations can be found in many works, e.g., in the visualization of news in Twitter [Earle et al., 2010; Sankaranarayanan et al., 2009], extracted topics in blogs [Adams and McKenzie, 2012], dense regions of geo-referenced photos [Toyama et al., 2003], and tags describing vernacular geographies in cities [Hollenstein and Purves, 2010]. These works all focus on different kinds of applications. However, they all need accurate and robust density estimates of the spatio-temporal information in the records.

3.2.3 Discrete Spatio-temporal Distributions

The previous models all work on a continuous spatio-temporal space. We now introduce a number of works that represent the spatio-temporal density distribution by a discrete distribution over a lattice. For this, note that the number of counts $k_1, \ldots, k_n, k \in \mathbb{N}$ over the cells $c_1, \ldots, c_n$ can be described as a discrete distribution

$$\mathrm{p}(c_i) = \frac{k_i}{\sum_{j=1}^{n} k_j}. \tag{3.17}$$

In fact, a distribution over a discrete spatio-temporal lattice can be obtained for any positive, additive, real-valued signal (see Section 2.2.1). Estimating densities in a given set of bins (given by a spatio-temporal grid defining a lattice) is also called the non-parametric histogram estimator (see Section 2.3.2).

In [Rattenbury et al., 2007a] the authors propose a way to determine if a tag is representative for a place or an event. Extended versions of their work are described in [Rattenbury et al., 2007b] and [Rattenbury and Naaman, 2009]. For this, they first extract a discrete spatio-temporal distribution over a lattice L for each tag f by counting the number of corresponding photos falling in a cell c

$$\mathrm{p}(c|f) = \frac{\mathrm{count}_r(f, c)}{\mathrm{count}_r(f)}. \tag{3.18}$$

They build discrete distributions for several scale levels by defining the lattices $L_1, \ldots, L_k$ along a quadtree. In each scale level the tags are scored by their spatio-temporal clustering characteristic. For this, they use the entropy of the distribution $\mathrm{p}(c|f)$ (see Section 2.2.5)

$$E_f^j = \sum_{c_i \in L_j}^{n} \mathrm{p}(c_i|f) \log \mathrm{p}(c_i|f), \tag{3.19}$$

where j denotes the scale level. A small entropy indicates that most of the probability mass is in a few cells. An entropy of $\log n$ (with $n = |L_j|$ being the number of cells) means the the probability mass is uniformly distributed. A major focus on their work is to find tags that show a clustering behavior at multiple levels (scale-structure identification, SSI). For this, they sum up the entropy scores at each level to obtain a global clustering score over all scale levels. The global score is

$$E^*_f = \sum_{j=1}^{k} E^j_f. \tag{3.20}$$

Extracting tags that have a consistent clustering behavior among all the levels is a reasonable assumption to find scale-independent phenomena. However, as discussed in Section 2.3.3, the semantics of a phenomena might change at different scales. This means, if the analyst is interested in a certain scale level then not the global clustering score E^*_f but the local score E_f is of interest. The result of their approach is a set of tags and associated spatio-temporal distributions, each having a score stating how likely it is a place or event. In their works the authors do not discuss what tag is the best label for a certain place or window in space and time (representativeness).

In [Emily et al., 2009] the same SSI approach as in [Rattenbury and Naaman, 2009] has been used to select tags that have a landmark semantics. They propose an adapted global clustering score by exploiting the co-occurrence of tags on the basis of their average Jaccard distance. It is left unclear if this smoothed score improves the selection of tags with scale-independent clustering characteristics.

Both of the works above use the discrete spatio-temporal distribution of tags $\mathrm{p}(c|f)$ to determine spatio-temporal cluster characteristics on basis the entropy. Other characteristics might be determined as well, e.g., if a spatio-temporal distribution is described by a particular number of clusters or represents a trajectory. We will introduce a novel approach to compare tags on the basis of their spatio-temporal characteristics in Chapter 5.

In [Chen and Roy, 2009] the authors present an approach to determine place or event related tags from photo collections. These tags are then used to extract places and events in a subsequent step. They first extract a continuous density distribution for each tag based on a wavelet transform. Wavelets are a non-parametric form of density estimation (see Section 2.3 and [Wasserman, 2004]). The resulting continuous density distributions of each tag $\mathrm{p}(c|f)$ are then transformed into a discrete spatio-temporal space using a discretization over a lattice. This results in smoothed and de-noised discrete spatio-temporal distributions compared to a direct count-based discretization.

The authors separate periodic and non-periodic distributions into separate groups by using heuristics (peaks every 7 days, etc.). Then, each of the two groups of tags is clustered separately by their discrete spatio-temporal distributions to find clusters of tags with similar distribution $\mathrm{p}(c|q_1), \ldots, \mathrm{p}(c|q_k)$, where $\mathrm{p}(c|q)$ is the normalized centroid of the cluster (detailed below).

A similar approach has been proposed in [Zhang et al., 2012b]. The authors use georeferenced photos to find clusters of tags that have a similar spatio-temporal distribution. For this, they first extract discrete spatio-temporal distributions for each tag $f \in F$ on the basis of the number of corresponding photos falling into a cell of a lattice L. The resulting distributions are represented by a matrix

$$\mathbf{Z} \in \mathbb{N}^{n_L \times n_F}. \tag{3.21}$$

To obtain groups of tags with similar spatio-temporal semantics the columns of the matrix are clustered using K-means. The resulting clusters contain tags that have a similar spatio-temporal distribution, and the centroid of the cluster represents the average spatio-temporal distribution of its tags. Before clustering the authors normalize the columns by the L2-norm of the column. The authors state that this will result in less clusters having only a single tag. However, no comparison or evaluation is performed.

Let the centroid of cluster q be $\boldsymbol{\mu}_q \in \mathbb{R}_+^{n_L}$. We can represent the distribution of a cluster as a discrete probability distribution by normalizing accordingly

$$\mathrm{p}(s|q) = \frac{\boldsymbol{\mu}_q}{\sum_{i=1}^{n_L} \mu_{qi}}. \tag{3.22}$$

The aim of their research is mainly a comparison of their proposed form of spatio-temporal similarity to other similarity measures, e.g., co-occurrences of tags in the photo tag-sets. The evaluation involves a number of users judging whether or not the resulting clusters are spatio-temporally relevant. Since a co-occurrence similarity will clearly result in clusters that are less spatio-temporally specific, the goal of the evaluation is thus somewhat unclear. We interpret their work as an approach to extract dominant spatio-temporal distributions occurring in the data (similar to the work proposed by [Chen and Roy, 2009]).

In [Leung and Newsam, 2010] the authors propose a model to extract land cover types from geo-referenced images. For this, they build a classifier based on manually labeled image data, to predict a 'developed' and an 'undeveloped' label for each image. The labels can be seen as two features f_d and f_u. They extract discrete spatial distributions for each tag

$$\mathrm{p}(c|f_d) \text{ and } \mathrm{p}(c|f_u) \tag{3.23}$$

and assign the label to those cells that with a higher feature-conditional cell-probability. Note that this is exactly the same model as is used to extract representative tags for clusters. Their preliminary results show that there is a correlation between the distribution obtained by ground-truth and the estimated land cover types obtained by their model.

In [Hays and Efros, 2008] the authors use discrete spatial distributions to visualize and compare scenes that are shown on an images (such as mountains, desert, etc.). For this, they assume a training collection of geo-referenced images and a similarity function $sim(r_i, r_j)$ between images based on low-level image features. Given a new image, they compute the similarity to all other images in the training set and extract a discrete spatial distribution by a weighted histogram over the lattice. Given a set of photos $r_1, \ldots, r_m$ with coordinates $e_1, \ldots, e_m$, and a spatial lattice $L = \{s_1, \ldots, s_n\}$. The unnormalized discrete spatial distribution of a photo r' is defined as

$$\tilde{\mathrm{p}}(s|r') \propto \sum_{i=1}^{n} \mathbf{1}\{e_i \in s\} sim(r', r_i), \tag{3.24}$$

which can easily be normalized by

$$\mathrm{p}(s|r') = \frac{\tilde{\mathrm{p}}(s|r')}{\sum_{i=1}^{n} \tilde{\mathrm{p}}(s_i|r')}. \tag{3.25}$$

The authors show that by using a reasonable large photo collection the distributions of natural phenomena can be extracted with good precision using this simple technique.

They also use their idea to predict the location of a photo. For this, they either return the mode of $\mathrm{p}(s|r')$ or the location of the most similar image r' on the basis of $sim(r, r'), \forall r' \in R$ as the result.

As an additional application they extract the spatio-temporal distribution for a huge number of images. Then, they cluster the distributions to obtain spatio-temporal signals of dominant image distributions similar to the idea of [Zhang et al., 2012b] and [Chen and Roy, 2009]. Let q be a cluster and $\boldsymbol{\mu}_q$ be the cluster centroid, then the spatio-temporal distribution of the scene (phenomenon) represented by this cluster is

$$\mathrm{p}(s|q) = \boldsymbol{\mu}_q \tag{3.26}$$

and can be seen as a dominant geographic feature in the data.

Finally, we explain the use of discrete spatio-temporal distribution for geographic phenomenon recovery. The authors in [Xu et al., 2012] propose a statistical model to extract the intensity of car accidents in the United States from Twitter messages. First, they extract a set of accident related tweets by using a set of keywords, and use this set of tweets as positive observations. Then, they use a spatio-temporal lattice defined over the US-states and days to aggregate the counts of positive observations. The number of positive counts is modeled by a Poisson distribution in each cell. Let $x_1, \ldots, x_n$ be the counts of positive tweets in the cells $c_1, \ldots, c_n$. The number of positive tweets is assumed to follow

$$\mathrm{p}(x|c_i) = \mathrm{Poisson}(x; \lambda_i), \tag{3.27}$$

where λ_c is the intensity of the Poisson distribution at cell c. Given a background distribution over the cells $b_1, \ldots, b_n$ the model can be extended by the link function

$$\gamma(x, b) = x \cdot b, \quad \mathrm{p}(x|c_i) = \mathrm{Poisson}(\gamma(x, b_i); \lambda_i). \tag{3.28}$$

The maximum likelihood estimate of λ_i then is

$$\hat{\lambda}_i = \frac{x_i}{b_i}. \tag{3.29}$$

Note that λ is not a probability but a positive real-valued intensity parameter. The discrete distribution of positive tweets (those having feature f_p) over the lattice can be obtained from the intensity by

$$\mathrm{p}(c_i|f_p) = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}. \tag{3.30}$$

If the background distribution $b_1, \ldots, b_n$ is the count of all tweets falling into a cell (positive and negative events), then $\mathrm{p}(c_i|f_p) = \lambda_i$.

Such Poisson models on spatio-temporal lattices are a primary tool in statistical epidemiology [Lawson, 2001]. Since the counts are represented by a Poisson distribution conditional in each cell, the model can easily be extended to take covariate information into account in a well-defined manner (such as shown for the background distribution).

Note that in all models presented so far no spatial-interaction between the cells has been modeled explicitly. Instead, spatial interaction is included by choosing appropriate bandwidth parameters or number of clusters. For the linked Poisson model, the authors in [Xu et al., 2012] propose explicit spatio-temporal regularization on basis of the neighborhood matrix $\mathbf{W}$ of the lattice by reformulating the estimation of λ_i as a regularized maximum likelihood estimation problem.

3.2.4 Language Models

In the following, we present discrete spatio-temporal models that are used to describe the distribution of words in textual data. From a spatio-temporal point of view, these models are equivalent to some of the models described before. However, since their focus is the modeling of language, specific smoothing and estimation techniques have been used to cope with word sparsity.

In [Cheng et al., 2010] the authors propose a model to predict the city of a user by his tweets. For this, the authors assume a set of geo-referenced tweets as training data. Their problem can be formulated as finding the location (city) s that maximizes

$$\mathrm{p}(s|f_1, \ldots, f_k), \tag{3.31}$$

where $f_1, \ldots, f_k$ are the words in the tweets of the user. Given a discrete lattice over city locations, they compute a distribution for each word in the training corpus on the basis of the tweets user location (GPS coordinate)

$$\mathrm{p}(s|f) = \frac{\mathrm{count}_r(f, s)}{\mathrm{count}_r(f)}. \tag{3.32}$$

Then, by using a naive base assumption they compute

$$\mathrm{p}(s|f_1, \ldots, f_k) = \prod_{i=1}^{k} \mathrm{p}(s|f_i) \tag{3.33}$$

to predict the most likely city. To tune their model they propose to use only predictive local words and different smoothing techniques. To filter local words they use the model proposed by [Backstrom et al., 2008] by using a dispersion threshold. They also make use of spatio-temporal smoothing. For this, they use another coarser spatial state lattice L_{state} and mix the city and state level probabilities. Another spatio-temporal smoothing is based on a lattice-based convolution given a neighborhood matrix $\mathbf{W}$ (see Section 2.3.2).

Exactly the same idea to predict the locations of text documents has been used in [Wing and Baldridge, 2011] and [Serdyukov et al., 2009]. In [Wing and Baldridge, 2011] sophisticated smoothing techniques have been used to limit the influence of sparsity (back-off smoothing), which shows to greatly improve the prediction performance.

Note that these models are inherently based on the distributions $\mathrm{p}(c|f)$. We see that estimating this distribution robustly is a major task for a huge number of applications.

3.2.5 Temporal Models

Works that focus on temporal information in user-generated data often employ techniques from time-series analysis [Gallagher, 2010; Guralnik and Srivastava, 1999; Ma and Perkins, 2003]. In this work we will not study these models and techniques in detail, however, in the following we mention works that have a considerable overlap with spatial and spatio-temporal modeling.

From a conceptual point of view temporal models are not different from spatial and spatio-temporal models. Given a set of features or clusters the temporal distribution can be represented by a temporal distribution

$$\mathrm{p}(t|f) \text{ or } \mathrm{p}(t|q). \tag{3.34}$$

The models, however, often make use of a Markov assumption. Let the density at a point in time be $\mathrm{p}(t)$. By the Markov assumption this probability depends on a number of prior time intervals $\mathrm{p}(t|t-1,\ldots,t-k)$. The number k is the order of the Markov chain, denoting the size of the temporal dependence. Using the Markov assumption allows to model special characteristics of temporal data, such as trends and periodicities. The generalization of a Markov model to spatial and spatio-temporal space is a Markov Random Field (which is not detailed in this thesis).

A particular focus on temporal distributions of tweets can be found in [Chae et al., 2012] where the authors use the seasonal-trend decomposition (STD) technique to detect unusual temporal patterns. For this, the authors first select a subset of records on the basis of tweet keywords or the tweets locations. Then, they extract a number of textual topics from the tweets using LDA as a pre-processing step. By using the document topic weights $\mathrm{p}(q|r)$ they extract a temporal distribution by a weighted discrete temporal density distribution, as shown in (3.24). This temporal distribution is then decomposed using STD to find unusual temporal patterns in each topic. The authors do not focus on the spatial dimension of the events but extract a discrete spatial distribution for selected LDA topics as a post-processing step.

Without considering the spatial domain at all, the extraction of spatio-temporal events shares a number of research goals with topic detection and tracking (TDT) [Allan et al., 1998; Rajaraman and Tan, 2001]. There, a stream of documents is to be clustered into a set of textual topics. In TDT, the focus is on the content similarity of documents and their temporal frequency. Given that a huge number of documents is similar in a short time interval, this is regarded as a topic, and new records are accordingly labeled by this topic if their content is similar. Primary challenges are, however, not the modeling of temporal peaks, but the identification of documents that have similar content [Allan et al., 1998]. See also [Kleinberg, 2002].

In the context of social media, different type of temporal events have been analyzed in [Lehmann et al., 2012] on the basis of the temporal hashtag distributions. There, the focus is on a characterization of distribution types by the amount of tweets occurring before and after the peak.

3.2.6 Trajectory Models

Another type of geographic phenomena can be described by trajectories. A trajectory is a temporal sequence of spatial points, forming a polyline in spatio-temporal space. Given user-generated data, some works cope with the discovery of trajectory phenomena.

In [Sakaki et al., 2010] the authors propose a system to detect trajectories of geographic phenomena in Twitter. For this, they first train a classifier (SVM) on the basis of a bag-of-word representation of the tweets. The training data consists of manually labeled tweets that describe a particular phenomenon (e.g., a hurricane). Through this each incoming tweet can be labeled as belonging to a phenomenon or not.

A temporal model is used to detect if the number of positive tweets is higher than expected. In this case the spatio-temporal trajectory is estimated using a hidden Markov model (HMM). The HMM uses a latent spatial variable $s_f^{(i)}$ to describe the hurricane center at time step i. The parameter is then updated at every time step by the incoming positive observations and by the former location $s_f^{(i-1)}$. They propose a Kalman filter and a particle filter approach to predict the HMM parameters, with the particle filter showing better results.

Using a HMM to model the spatio-temporal density distribution is a valuable approach if the phenomenon of interest is assumed to have a single moving center. We can describe the spatial distribution of the phenomenon at time t as

$$\mathrm{p}(s|f,t) = \mathrm{p}(s|s_f^{(t-1)}, R^{(t)}), \tag{3.35}$$

where $s^{(t)}$ is the location at time step t, $R^{(t)}$ are the positive observations in the respective time interval, and $s^{(t-1)}$ is the location at the former interval.

In [Yin and Cao, 2011] the authors propose a model to mine user trajectories in discrete spatio-temporal space from geo-referenced photo collection data. They first spatially cluster the photos to extract points-of-interest and then describe them using representative tags. Then, they extract trajectories over point-of-interest on the basis of the users' photo sequences. Those trajectories are then mined for frequent sequential patterns using standard techniques [Zhu et al., 2006]. Using the frequent sequential trajectory patterns, the points-of-interest descriptions, and user importance features, they rank the trajectories by an interestingness score.

3.2.7 Spatio-temporal Topic Models

Currently a great body of research copes with the extraction of textual topics from complex data (text with associated information) using generative models (see Section 2.4.2)

A topic $q \in Q$ can be seen as hidden information in the documents that describes a weighted distribution of topics over documents $\mathrm{p}(q|r)$, and each topic is described by a distribution over words $\mathrm{p}(f|q)$. This allows to organize documents with respect to their topics and to analyze the content in the documents by the topics' word distributions. Particularly interesting for this thesis are topic models that aim to model the spatio-temporal variation of topics. In the following we propose basic modeling strategies to describe the spatio-temporal distribution of topics.

Extracted topics can be seen as high-level features of documents (similar to recognized objects in images). Hence, we can use them directly to compute a weighted spatio-temporal distribution $\mathrm{p}(c|q)$ as shown in (3.24). Exactly this scheme is used in [Chae et al., 2012] and [Adams and McKenzie, 2012]. There, first the LDA topics of tweets and blogs are determined and the spatio-temporal distribution of the topics is extracted in a subsequent step.

In the following, we describe extensions of probabilistic latent semantic analysis (PLSA) and LDA that model the distribution $\mathrm{p}(s, t, f, q)$ explicitly. By this, they do not compute the topic information $\mathrm{p}(q|r)$ and $\mathrm{p}(f|q)$ as a pre-processing step (independently of the spatio-temporal record information) but assume that feature, topics, space, and time depend on each other. Anyhow, the models need to make assumptions on how the information items are interrelated.

In [Sizov, 2010] the author models the distribution of tags in space and time in geo-referenced photo collections. They propose an extension of LDA such that each topic is described by two univariate Gaussian distributions that independently represent the latitude and the longitude information. The resulting distribution is a symmetric multivariate Gaussian that can be tied on the longitude and latitude axis. The spatial distribution can be represented by a constraint multivariate Gaussian

$$\mathrm{p}(s|q) = \mathcal{N}(s; \mu_q, \Sigma'_q), \tag{3.36}$$

where Σ'_q is a constraint covariance matrix adhering to the above assumptions. Since the model is based on a multivariate distribution of each topic, it assumes that records with similar topic cluster around a single location, i.e., the underlying topic phenomenon is a landmark or place. The authors describe different applications of their model: Content organization (visualization) of photos by topics and location by using $\mathrm{p}(s|q)$, location-aware keyword search of photos by using $\mathrm{p}(r|w, s)$, location suggestion on the basis of keyword queries by using $\mathrm{p}(s|w)$, and location-aware tag recommendation by using $\mathrm{p}(w|s)$. Note that these applications can be achieved by all of the topic models below.

In [Yin et al., 2011] the authors propose a similar model to extract spatial topics from geo-referenced photo collections. They extend PLSA such that each topic has not only a distribution over words, but also a distribution over a number of Gaussian distributions. The spatial distribution of a topic can be described as

$$\mathrm{p}(s|q) = \mathrm{GMM}(s; \boldsymbol{\alpha}_q, \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3.37}$$

where $\boldsymbol{\alpha}_q$ is a topic specific vector of weights for each of the Gaussian distributions. The authors use their model to extract and visualize keyword-depended spatial topics in the photo collection. For this, they choose the topic q that most likely describes a set of keywords $w_1, \ldots, w_k$ using $\mathrm{p}(q|w)$, and draw its spatial distribution $\mathrm{p}(s|q)$. The model is learned by using the EM algorithm.

In [Mei et al., 2006] the authors propose another PLSA-based mixture model to extract spatio-temporal topics from Web-logs. For this, they assume a set of documents with associated timestamps and discrete locations (a spatio-temporal lattice). The model distinguishes between two types of topics, global and spatio-temporal ones. A spatio-temporal topic has an associated discrete spatio-temporal distribution $\mathrm{p}(s, t|q)$ over the cells in the lattice. The model is learned using an EM algorithm. The authors propose generalized versions on arbitrary graphs (with the lattice being a particular instance) in [Mei, 2006; Mei et al., 2008].

In [Hao et al., 2010] the authors propose an extension of PLSA to extract location relevant topics from travelogues (blogs). For this, they also assume that a document is represented by a mixture of global and local topics. The spatial information is obtained by geo-tagged terms in the documents (e.g., place names). Other than in [Mei et al., 2006], they do not assume that the locations are associated with the whole document, but to a text segments (sentence or paragraph). The local topics have an associated discrete distribution over the locations $\mathrm{p}(s|q)$. They use the model to (1) find similar locations on the basis of spatio-temporal topic and word distributions $\mathrm{p}(s|q)$ and $\mathrm{p}(s|w)$, respectively (2) to recommend locations by keyword queries, and (3) to extend existing blogs by location-relevant content from other blogs.

Similar to [Mei et al., 2006] and [Hao et al., 2010], the authors in [Eisenstein et al., 2010] present an LDA extension that allows to distinguishes between local and global topics. They assume geo-referenced tweets as input data. They use their model to (1) analyze lexical variation by topic and location, (2) segmentation of geographic space in coherent linguistic communities, and (3) prediction of author locations. As a specific model assumption, they assume local topics to be specializations of global topics. This allows to analyze the lexical variation of a topic in space.

In [Hong et al., 2012] the authors propose a generative model to describe the relationships between users, terms, topics, locations, and language. They assume that words in a tweet depend on the location and the topic, geographic regions have different language variations, topics distributions over regions are topic dependent, and users tend to occur in a handful of locations. Their model is a primary example of modeling complex data such as given in Twitter using a rich generative model. Their primary evaluation consists of using the model to predict the location of tweets and comparing the predicted location with that of the real GPS coordinate. Regarding the spatio-temporal model, they assume a discrete set of regions and timestamp. The regions have been obtain in a pre-processing step by spatial clustering the photo GPS coordinates.

Finally, an approach that exploits location information in news and blogs is proposed in [Wang et al., 2007]. The authors extend the LDA topic model by a distribution over location entities. For this, they extract location names in the text and use them to describe topics in addition to the word distribution. Since they do not model the spatial coordinates and/or extents of the locations, their work can better be understood as an extension of LDA to handle entities found in the text.

3.2.8 Supervised Models

Finally, we review works that focus on predictive spatio-temporal models by using user-generated data as covariate information. A supervised spatio-temporal model takes a number of measurements in space and time

$$X = \{(c_1, y_1), \ldots, (c_m, y_m)\}, \tag{3.38}$$

where c_i is the measurement location and y_i is the measurement value. The measurements are assumed to be samples from a spatio-temporal process that is described by a distribution $\mathrm{p}(c, v)$. Prediction is the process of estimating the measurement at an arbitrary point in space and time

$$\mathrm{p}(v|c). \tag{3.39}$$

In [Zhang et al., 2012a] the authors propose a model to predict natural phenomena (such as snow cover) by spatio-temporal distributions of geographic features extracted from geo-referenced photos. For this, they extract a number of discrete spatio-temporal distribution $\mathrm{p}(s|f)$ for a set of keywords $f \in F$ (e.g., 'white','snow','mountains'). This allows to represent a location s by a feature-vector over keyword intensities. Then, they use remote images as training data to learn a model $\mathrm{p}(v|s)$.

In [Ginsberg et al., 2009] the authors propose a model to predict phenomena on the basis of search engine queries. For this, they use the user locations and timestamps of a large number of queries $f \in F$, where each f is a particular query string. Their model first obtains the counts of queries in space (states) and time (weeks), resulting in a number of discrete distributions $\mathrm{p}(c|f)$. As in [Zhang et al., 2012a], they use these distributions to extract geographic-feature vectors to represent a location s by a feature-vector over the query intensities. The phenomenon to predict (the outcome variable) is the number of physician visits in space and time $X = \{(c_1, y_1), \ldots, (c_m, y_m)\}$. Given a training set made of the physician visits and the associated feature vectors, they train a linear model to predict the number of visits $\mathrm{p}(v|c)$.

The term forecasting is used if a measurement is to be predicted in a future point in time where also no covariate information is available (such as the vectors extracted from image data or the queries above). Such forecasting models exploit characteristics of the past temporal distribution (see Section 3.2.5)

In [Gallagher, 2010] the authors use the clicks of users on photos, the click timestamp, and user location to model the user interest in particular objects shown on a photo. For this they use the tags of the photos as weak high-level features, e.g, people or products. In their approach they focus on forecasting models of spatio-temporal distributions of click counts. They use a discrete temporal distribution $\mathrm{p}(t|f)$ on the basis of the number of clicks that fall into day intervals (temporal lattice). Then, an auto-regressive temporal model is used to predict the interest in a future time period, given the past temporal distribution. Their work shows that spatio-temporal distributions of user interest from user-generated data have a predictive temporal component.

3.2.9 Application Commonalities

The former sections presented a variety of different works with respect to their underlying models. We saw that most applications and tasks are based on an appropriate modeling of features, feature combinations (high-level features), and records in space and time.

Given a user-generated data source, the essential spatio-temporal information can be described probabilistically by the distributions

$$\mathrm{p}(s,t), \mathrm{p}(s,t,f), \mathrm{p}(s,t,q), \mathrm{p}(s,t|f), \mathrm{p}(s,t|q), \mathrm{p}(s|t,f), \mathrm{p}(q|s,t). \tag{3.40}$$

Given these distributions the applications rely on them like follows:

- Finding regions, hotspots, events, or places relies on $\mathrm{p}(s,t)$ to find dense windows in space and time.
- Extracting representative features relies on selecting the most likely features at an area/interval/window according to $\mathrm{p}(s,t|f)$.
- Comparison of features by their spatio-temporal semantics is based on $\mathrm{p}(s,t|f)$ or $\mathrm{p}(s,t,f)$.
- Finding the most likely location of a record relies on $\mathrm{p}(s,t|f)$ (feature-dependent, e.g., text prediction) or $\mathrm{p}(s,t)$ (feature-independent).
- Describing the variation of textual topics in space and time relies on $\mathrm{p}(s,t|q)$, where q is a latent topic.
- Extracting covariates for supervised spatio-temporal models relies on $\mathrm{p}(s,t|f)$ or $\mathrm{p}(s,t|q)$ to extract feature vectors for areas/intervals/windows.

In the following, we propose a conceptual geographic data mining framework that extracts the above distributions from the data using simple and general primitives, namely, geographic observations and geographic feature signals. Based on these primitives, we define a process that allows to realize different tasks and applications easily, and to clearly define and address fundamental sub-problems.

3.2.10 Other Approaches

Finally, we mention ideas in related work that overlaps with the aim of building a data- and application-independent framework to mine geographic knowledge from user-generated data.

The authors in [Kennedy et al., 2007] and [Crandall and Snavely, 2012] propose the idea of using geo-referenced photo collections as a source to identify places and events, and to use them to browse and visualize the data. Similar to our framework, they consider tags or visual features of geo-referenced photos as potential features to represent geographic semantics, and they state that this idea can be used to realize and extend a variety of applications in information retrieval. However, since both works use a clustering-based approach, their approaches are limited to extract phenomena as points or areas that represent popular places and events. In this thesis, we develop a framework that is more general with respect to the phenomena semantics, by modeling them as arbitrary spatio-temporal distributions.

In [Serdyukov et al., 2009] and [O'Hare and Murdock, 2012] the authors propose a language model to describe the distribution of features (words, tags) in geographic space. They state that such a model allows to estimate the geographic focus of text and to enable new location-based services. The primary focus on the distribution $\mathrm{p}(f, s, t)$ is similar to our proposed framework. However, their model is specifically developed for geo-referenced textual data, e.g., text snippets or tag sets with an associated point coordinate. Moreover, their work is not focused on discovering interesting dimensions of geographic space, but to associate the textual records with locations.

In [Singh, 2010] the authors introduce the idea of *social pixels* as a common abstraction to represent spatio-temporal information contained in social media. A social pixel describes a vector of social influences at a particular point in space and time. The authors also focus on representing user-generated data by the distribution $\mathrm{p}(f, s, t)$, where a discrete bin $c = (s, t)$ is called a social pixel. Their focus, however, is on a declarative algebra to query information contained in social pixels and not on data mining tasks. Moreover, they do not discuss how a meaningful social pixel space is to be extracted, other than from counting the number of information items in the cells.

Finally, in [Naaman et al., 2010] an abstraction of social media data sources as *social awareness streams* is introduced. Their aim is to understand social activity and communication patterns by analyzing how the message content varies on the basis of user characteristics and personal networks. Hence, the focus is not on utilizing social media as spatio-temporal observations to discover geographic phenomena, but on a unifying abstraction to analyze communication patterns and topics in the data.

3.3 Input Data Representation

In order to abstract from particular types of user-generated data, we now propose a general representation of the input-data in the form of geographic observations.

3.3.1 Geographic Observations

The main idea behind the applications and geographic knowledge discovery tasks presented so far is to exploit relationships between qualitative information and spatio-temporal information found in user-generated data to describe and discover geographic patterns. We use the term qualitative information to denote low- or high-level features that describe the content of records (words in documents, tags of photos, query strings), and spatio-temporal information to describe all kinds of spatial and/or temporal evidence available in the data (GPS coordinates, IP-addresses, geographic/temporal expressions).

By focusing on qualitative and spatio-temporal information the data can be seen as a set of measurements observing relationships between these two information items.

We call these measurements *geographic observations*. Data sources like photo collections, text messages, or Web sites provide rich sources to extract such geographic observations. For example, by building relationships between user locations and text snippets, or GPS coordinates and photo content.

Extracting an accurate set of features is already a highly application-specific process. In the following, we assume that observations are generated using a (possibly large) number of low-level candidate features as qualitative information. Finding features or features combinations that have geographic semantics is then part of the knowledge discovery process.

Often a measurement is associated with one or several users, e.g., a user clicking or uploading a record, or users writing a document. We make use of this ternary relationship between features, spatio-temporal information, and users, if available.

A geographic observations is hence defined as

$$o = (E_F, E_C[, E_U]), \tag{3.41}$$

with E_F representing information about qualitative features, E_C representing information about the spatio-temporal context, and E_U representing (optional) information about the (observing) users.

3.3.2 Uncertainty and Influence

Measurements are supposed to be a highly uncertain kind of information. A joint observation of a feature and some spatio-temporal information provides only a small amount of evidence about their semantic relationship. Hence, to discover significant patterns, a huge number of measurements is needed.

Uncertainty also exists in all of the three information items of an observation themselves. About the qualitative features E_F (what is measured), about the spatio-temporal information E_C (where and when is it measured), and about who did the measurement E_U. We use the term *influence* to describe the amount of evidence an observation contains about features, the spatio-temporal context, and the observing users.

3.3.3 Measurements

As shown in the related work a lot of different data sources have been used to extract geographic knowledge. Furthermore, even for the same data, different kinds of qualitative and geographic information can be extracted.

To generate a set of geographic observations we differentiate between records (providing the qualitative features) and measurement events (defining the geographic context, the observing user, and constituting the relationship). We denote a set of records

$$R = \{r_1, \ldots, r_n\}. \tag{3.42}$$

The content of the records is the primary source to extract qualitative features. In some cases, features might not be extracted from the records directly but from associated resources, e.g., from the URLs shown in a query result or from associated user profiles.

A measurement builds a relationship between features, spatio-temporal context, and users. We denote measurement as

$$H = \{h_1, \dots, h_m\}. \tag{3.43}$$

Often the measurement are defined on the basis of the records themselves. For example, in a photo collection the photos constitute measurements that relate features (tags, image objects), a spatio-temporal context (GPS coordinate, timestamp), and user information (uploading user). Measurements can, however, also be defined by of users clicking a photo [Gallagher, 2010], submitting a query [Backstrom et al., 2008], or checking-in at a location in a location-based social network [Cheng et al., 2011].

A particular interesting type of measurement are co-locations in textual data. For example, one can define a measurement as the features (terms, entities) and the spatio-temporal information (geographic and/or temporal expressions) co-occurring in a text segment (sentence or paragraph) or in close proximity. In NLP, such co-occurrences are analyzed to uncover semantic relatedness between terms [Weeds et al., 2004], e.g., to find synonyms or different semantics of a word [Turney and Pantel, 2010]. In this work we assume co-occurrences between features and spatio-temporal information to represent geographic observations, and we are particularly interested in the spatio-temporal feature distributions.

The extraction of geographic observations from user-generated data can be described as applying the information extraction functions ϕ_F, ϕ_C, and ϕ_U on the measurements

$$o = (E_F, E_C, E_U) = (\phi_F(h), \phi_C(h), \phi_U(h)). \tag{3.44}$$

Information extraction functions describe the influence a measurement has in its respective domain (features, spatio-temporal space, users) and will be described in the following sections. Each measurement represents a geographic observation such that the set of geographic observations can be described as

$$O = \{(\phi_F(h), \phi_C(h), \phi_U(h)) | h \in H\}. \tag{3.45}$$

3.3.4 Feature Influence

A (qualitative) feature extraction function describes the influence of a given set of categorical features $F = \{f_1, \dots, f_p\}$ on a measurement h

$$\phi_F(h) = \psi_F(\omega(h, R)) \in \mathbb{R}^p_+. \tag{3.46}$$

We make use of a feature extraction function ψ for unstructured data defined in (2.73). Since a measurement might refer to partial records (e.g., a sentence in a document), we use

$$r' = \omega(h, R), h \in H \tag{3.47}$$

to denote the content that is relevant to a measurement. Information that is not given as categorical data (e.g., ratings, real values) first needs to be transformed by discretization and a 1-of-K feature representation [Bishop, 2006, p. 74].

3.3.5 Spatio-temporal Influence

The spatio-temporal context of a measurement is described by a spatio-temporal variable (2.7) describing the influence of a measurement in space and time

$$\phi_C(h)(c) \in \mathbb{R}_+, h \in H, c \in D_C. \tag{3.48}$$

We use the notation of a functional to clarify that the resulting spatio-temporal influence is a signal $z_h(c)$ in space and time (see Section 2.2.1). Hence, even if the spatio-temporal information is given as a point and/or a timestamp, we assume that it can be represented as a continuous signal in space and time.

3.3.6 User Influence

Often an observation is only influenced by a single user. However, it is reasonable to assume that several users have influence on a record, e.g., if a records has been created by several users. Furthermore, we might treat friends of a user as having influence on the users' observation. As discussed in [Anagnostopoulos et al., 2008], the fact that actions of a user can induce his/her friends to behave in a similar way can be measured in a social network. This social influence is also a valuable input data to describe the user influence of geographic observations. For example, close friends of a user (as measured in a social network) will more likely travel together and make similar observations. Their measurement are hence less independent, which reduces the amount of additional evidence in the extracted relationship between features and spatial-context. By handling the mutual influence these users have on their records (such as counting on the basis of users), we can extract more robust and meaningful feature signals.

The set of all users is denoted

$$U = \{u_1, \ldots, u_k\}. \tag{3.49}$$

To model this multi-user influence of the observations we use a user influence function similar to the feature and the spatio-temporal influence functions

$$\phi_U(h)(u) \in \mathbb{R}_+, h \in H, u \in U. \tag{3.50}$$

Let u_i be the single user creating observation h_i. In the case of a single user influence, the influence function will be

$$\phi_U(h_i)(u) = \mathbf{1}\{u_i = u\}. \tag{3.51}$$

More complex user influence functions can be built based on the distance of users in social networks (e.g., the number of hops), or by computing the similarity of users on the basis of similar behavior by using collaborative filtering approaches [Schafer et al., 2007].

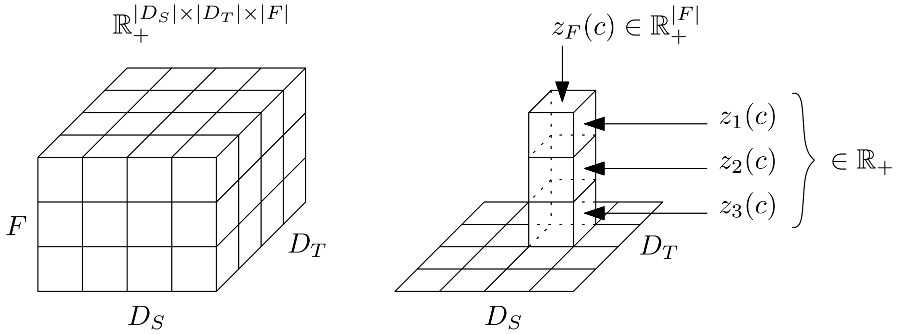

$\mathbb{R}_+^{|D_S|\times|D_T|\times|F|}$
F
D_T
D_S
$z_F(c) \in \mathbb{R}_+^{|F|}$
$z_1(c)$
$z_2(c)$
$z_3(c)$
$\in \mathbb{R}_+$
D_T
D_S

Figure 3.1: Geographic Features and Geographic Feature Space.

3.4 Output Patterns

Given a set of geographic observations a crucial step is to transform them into patterns that allow to describe and discover geographic knowledge. The main patterns of interest in this work are spatio-temporal distributions of features in space and time. We call them geographic feature signals. In the following we describe this representation and clarify its semantics.

3.4.1 Geographic Features

Generally, we define a *geographic feature* to be a property of space and time. This is similar to the definition of a document feature in text mining, which represents a property of the documents. The influence of a geographic feature f at a spatio-temporal point is represented by a positive, real-valued spatio-temporal variable called the *geographic feature signal*

$$z_f(c) \in \mathbb{R}_+, c \in D_C. \tag{3.52}$$

A set of geographic candidate features $F = \{f_1, \ldots, f_p\}$ allows to construct a (possible high-dimensional) multivariate geographic feature signal

$$z_F(c) = (z_{f_1}(c), \ldots, z_{f_p}(c))^\top. \tag{3.53}$$

As shown in Figure 3.4.1, a multivariate geographic feature signal allows to represent a spatio-temporal point $c \in D_C$ by a vector of geographic feature influences, called a *geographic feature vector*. We say that a spatio-temporal point is represented in geographic feature space. As it is the case for document or image features (see Section 2.4.1), a geographic feature vector allows to represent the semantics of a spatio-temporal point or window on the basis of a numeric vector in geographic feature space.

The geographic feature vectors of the cells in a lattice $c_i \in L$ describe a cell-feature matrix

$$\mathbf{Z}_{L,F} := (z_F(c_1), \ldots, z_F(c_n))^\top \in \mathbb{R}_+^{n \times p}. \tag{3.54}$$

We call this also the *geographic feature matrix* of a lattice. Note that the rows are cells and the columns are the features. Other than in a document-term matrix the rows have a dependency, defined by the neighborhood matrix $\mathbf{W}$.

Probabilistic Interpretation

Given the geographic feature matrix

$$\mathbf{Z} \in \mathbb{R}_{+}^{n_F \times n_L}. \tag{3.55}$$

Since the signals represent positive and additive influences of the features $f_1, \ldots, f_p$ over the cells $c_1, \ldots c_n$, we can describe the matrix by the joint distribution

$$\mathrm{p}(f_i, c_j) = z_{ij} / \sum_{i=1}^{n_F} \sum_{j=1}^{n_L} z_{ij}. \tag{3.56}$$

From this distribution we can derive the conditional probabilities $\mathrm{p}(c|f)$, $\mathrm{p}(f|c)$, $\mathrm{p}(f)$, and $\mathrm{p}(c)$. As we have shown in the related work, a lot of applications and discovery tasks can be reduced to use these distributions.

3.4.2 Geographic Phenomena

Geographic features and their respective spatio-temporal signals are assumed to be the primary patterns to discover geographic phenomena. We define a *geographic phenomenon* q to be any social or physical process or entity that can be identified in space and time. This means, we can identify where, when, and at what intensity a phenomenon q occurs. Given this definition, a geographic phenomenon is just the same as a geographic feature. Actually, we say that a geographic feature is a *low-level feature* and a geographic phenomenon is a *high-level feature.* For the low-level features, we assume that they represent a property of space and time. We don't know yet what kind of phenomena they represent, however. The aim of extracting high-level features (phenomena) is to find features or feature combinations that are interesting and convey knowledge about underlying phenomena. The relation of (low-level) geographic features and (high-level) geographic phenomena is hence similar to low-level image features (patches, lines) and high-level image features (objects). We use the notation

$$z_q(c) \in \mathbb{R}_{+}, c \in D_C \tag{3.57}$$

to refer to the spatio-temporal distribution of a phenomenon q.

One way to discover geographic phenomena is to select geographic feature signals that are interesting. Hereby, a phenomenon q is represented by a single geographic feature f, and we assume

$$z_q(c) \simeq z_f(c), \forall c \in D_C. \tag{3.58}$$

However, sometimes a phenomenon is represented by a combination of features. A phenomenon signal is then described as

$$z_q(c) = \eta(z_{f_1}(c), \ldots, z_{f_p}(c)) = \eta_F(c). \tag{3.59}$$

For example, a linear combination of geographic feature signals is

$$\eta_F(c) = \alpha_0 + \sum_{i=1}^{p} \alpha_i z_{f_i}(c), \tag{3.60}$$

where α is a vector of feature weights and α_0 is the intercept. Given that the features F are descriptive (such as words, tags, or color names) a description of a phenomenon q can be extracted on the basis of its top-k features as determined by α,

$$\text{description}(q) = (\text{label}(f_i) | f_i \in F \wedge \alpha_1 > \alpha_i > \alpha_p)_{i=1}^{k}. \tag{3.61}$$

Probabilistic Interpretation

We assume that a number of unknown geographic phenomena $q_1, \ldots, q_k$ can be discovered in a geographic feature matrix $\mathbf{Z}_{L,F}$. The combination of features is represented by $\mathrm{p}(f|q)$ and the spatio-temporal distribution of a geographic feature by $\mathrm{p}(c|q)$.

Phenomena can be seen to exists in an unknown joint distribution $\mathrm{p}(f, c, q)$, where q is a latent variable. Geographic phenomenon discovery tasks then try to estimate the joint distribution based on the given distribution $\mathrm{p}(f, c)$ defined by $\mathbf{Z}$.

3.5 Interestingness

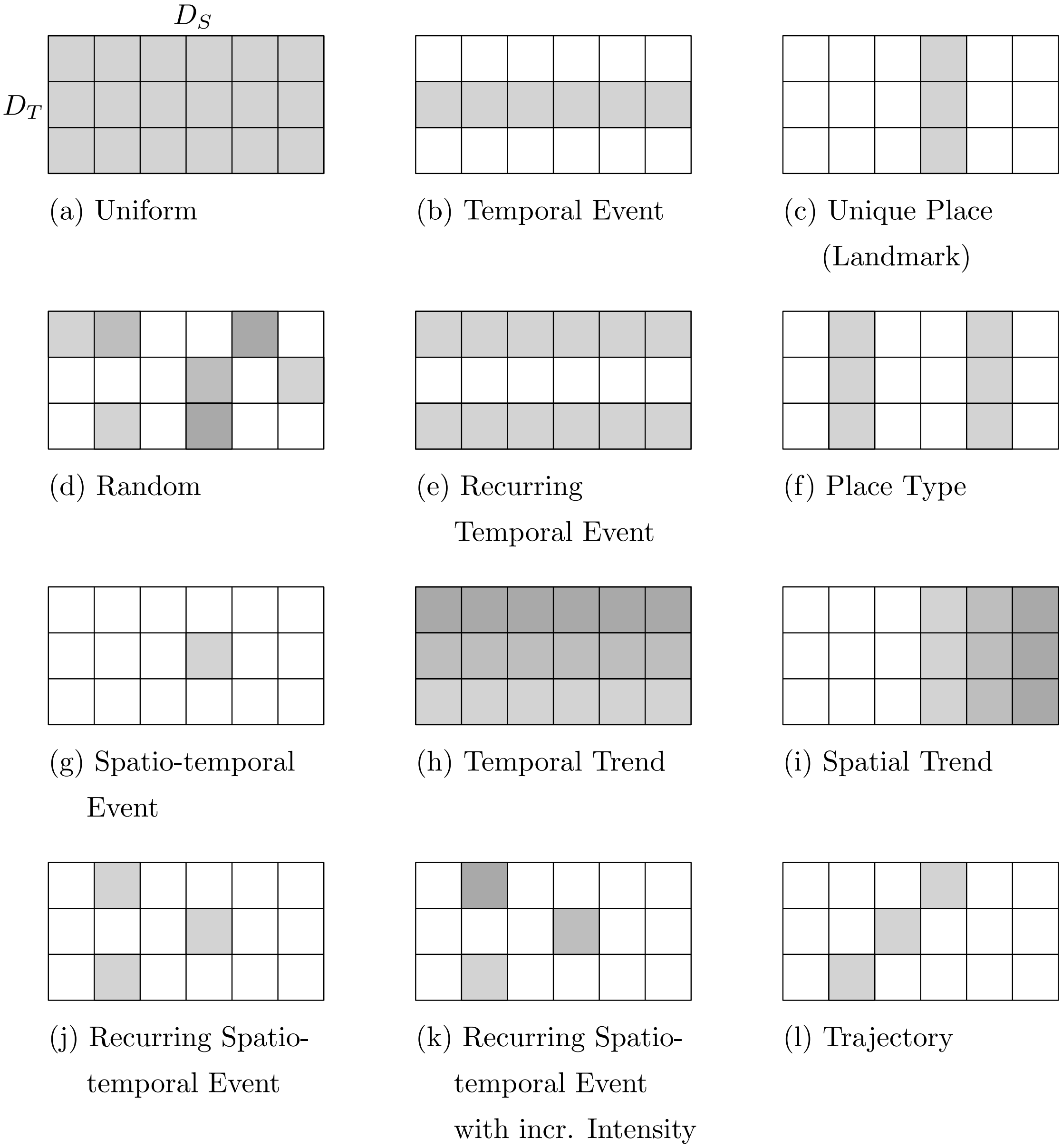
An essential part of knowledge discovery is the selection of interesting patterns among a possibly huge number of candidates. In this section, we introduce different types of interestingness of spatio-temporal distributions.

3.5.1 Spatio-temporal Distribution Type

Geographic feature distributions $\mathrm{p}(c|f)$ or feature combinations $\mathrm{p}(c|q)$ can describe several types of geographic phenomena (places, events, trajectories). Depending on the knowledge discovery task or the applications, only specific types might be of interest. In the following we characterize the spatio-temporal distributions of particular phenomenon types as shown in Figure 3.5.1:

- *Uniform*: A signal that is distributed equally likely among space and time. Such a signal is likely to be non-interesting as a geographic feature since it does not allow to distinguish between points in space and time. A uniform signal can be identified by a high entropy $E[\mathrm{p}(c|f)]$. Feature with such distribution types have a similar meaning like stop-word features in NLP (see Section 2.4.1).

- *Places (landmarks), temporal events, spatio-temporal events*: Signals that have a single peak in spatial space, temporal space, or spatio-temporal space. As shown in [Rattenbury et al., 2007a] such features can be used to extract popular places or events or to extract highly predictive features. Signals with a single peak can be identified by a low entropy $E[\mathrm{p}(c|f)]$ or by fitting the distribution to a unimodal function (see Section 3.2.1) and evaluating the fitting error.



(a) Uniform

(b) Temporal Event

(c) Unique Place (Landmark)

(d) Random

(e) Recurring Temporal Event

(f) Place Type

(g) Spatio-temporal Event

(h) Temporal Trend

(i) Spatial Trend

(j) Recurring Spatio-temporal Event

(k) Recurring Spatio-temporal Event with incr. Intensity

(l) Trajectory

Figure 3.2: Different types of spatio-temporal distributions. The x-axis of the grids represent the spatial domain and the y-axis the temporal domain. Each cell is a spatio-temporal window in geographic space and the color encodes the intensity of the signal.

- *Place type, recurring temporal event, recurring spatio-temporal event*: A signal that has several spatial peaks, temporal peaks, or spatio-temporal peaks (possibly in a regular interval). Such phenomena can be seen as attributes of places or (temporal/spatio-temporal) events. For example, a feature that describes a certain type of place or event. Such features can be identified by analyzing if the distribution is multi-modal, for example, by clustering of the signal (see Section 3.2.1) and evaluating the error of the clustering result.

- *Spatial trend, temporal trend, spatio-temporal trend*: A signal with medium to large-scale variation withing spatial, temporal, or spatio-temporal space. A spatial trend is similar to spatial heterogeneity. For instance, in some areas there might be a higher background population resulting in a higher signal intensity. Temporal trends are often monotonically increasing, representing a growing background population.

- *Trajectory*: A signal forming a line in spatio-temporal space. Hence, it has high signal at close temporal intervals and close spatial areas. Identifying if a signal has trajectory semantics can be achieved by fitting a Hidden Markov Model on the basis of the feature distribution (see Section 3.2.6) and evaluating the error of the fit.

3.5.2 Dominance

The spatio-temporal distribution type depends on $\mathrm{p}(c|f)$ and is independent of the amount of signal of a feature f in the geographic feature matrix $\mathbf{Z}_{L,F}$. Describing geographic feature by their amount of signal, allows instead to select highly dominant features. The dominance can be described by the probability that the feature occurs in the geographic feature matrix

$$\mathrm{p}(f) = \sum_{c \in L} \mathrm{p}(f, c). \tag{3.62}$$

The dominance in a particular region, interval, or window $A \subseteq L$ can be obtained as

$$\mathrm{p}(f, A) = \sum_{c \in A} \mathrm{p}(f, c). \tag{3.63}$$

3.5.3 Representativeness

In [Ahern et al., 2007] and [Crandall et al., 2009], among others, the selection of representative features for an area A_q described by a spatio-temporal cluster q is of particular interest. As we stated in Section 3.2.1 the representativeness of a feature f in cluster q covering window A_q can be generalized by

$$\mathrm{p}(q|f) = \mathrm{p}(A_q|f) = \sum_{c \in A_q} \mathrm{p}(c|f). \tag{3.64}$$

Hence, for any area A or any cell c, the features with the highest value for $\mathrm{p}(A|f)$ or $\mathrm{p}(c|f)$ are the most representative. Note that the representativeness as defined here does not depend on the dominance $\mathrm{p}(f)$. The product of dominance and representativeness is just the joint distribution $\mathrm{p}(c, f)$.

3.6 Mining Process

We now give a high-level overview of the geographic feature mining framework by defining a set of fundamental data mining (sub-)tasks and describing how they are connected to each other in the mining process.

A schematic overview of the process is shown in Figure 3.3. The framework consists of data mining (sub-)tasks that allow to extract, filter and combine an initially huge number of geographic candidate features into a small number of meaningful geographic phenomenon signals. The process contains loops to iteratively improve a set of signals by updating assumptions and parameters in the mining tasks.

3.6.1 Geographic Observation Generation

The first step in the framework is the generation of geographic observations from user-generated data as proposed in Section 3.3. This tasks assumes a user-generated data source R, a defined set of measurement events H, and appropriate influence functions ϕ_F, ϕ_C, and ϕ_U.

Geographic observation generation is a pre-processing step to transform the implicit relations between features, spatio-temporal information and users into a set of geographic observations O. This task is the most application- and data-specific. The underlying assumptions in selecting candidate features, spatio-temporal information, users, and their ternary relation determine the semantics the observations, and thus the semantics of the output patterns. Choosing an initial set of feature extraction functions and measurement events hence highly depends on the data and what kind of geographic knowledge should be mined (social or cultural habits, natural phenomena, etc.).

Once a set of geographic observations has been generated, the following data mining tasks are much more automated and can be seen as tools to help an analyst to discover geographic knowledge.

3.6.2 Geographic Feature Extraction

Given a set of observations O, the first task in the framework is to extract the spatio-temporal signals of the candidate features. The resulting signals are represented in a geographic feature matrix $\mathbf{Z}_{L,F}$.

This step needs to cope with the high amount of uncertainty in the geographic observations to extract robust and meaningful signals for a possibly huge number of features. This data mining task is heavily concerned with density estimation of feature influences in space and time, and an appropriate modeling of user redundancy and a possible background distribution. In existing works, fundamentally different kinds of

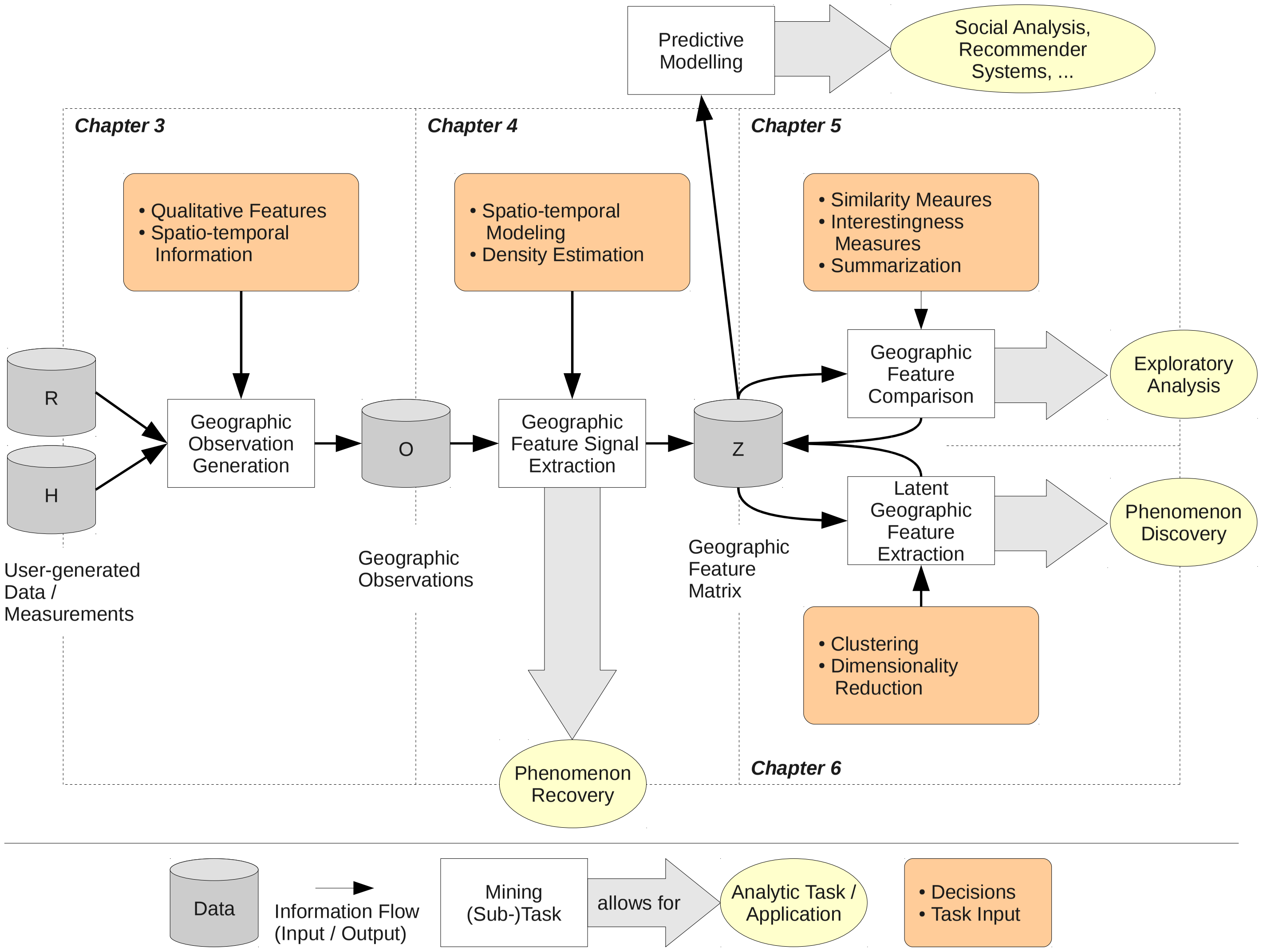

Predictive Modelling
Social Analysis, Recommender Systems, ...
Chapter 3
Chapter 4
Chapter 5
• Qualitative Features
• Spatio-temporal Information
• Spatio-temporal Modeling
• Density Estimation
• Similarity Meaures
• Interestingness Measures
• Summarization
R
H
Geographic Observation Generation
O
Geographic Feature Signal Extraction
Z
Geographic Feature Comparison
Exploratory Analysis
Latent Geographic Feature Extraction
Phenomenon Discovery
User-generated Data / Measurements
Geographic Observations
Geographic Feature Matrix
• Clustering
• Dimensionality Reduction
Phenomenon Recovery
Chapter 6
Data
Information Flow (Input / Output)
Mining (Sub-)Task
allows for
Analytic Task / Application
• Decisions
• Task Input

Figure 3.3: Schematic overview of the geographic feature mining framework.

models have been used to extract geographic feature signals (if such a concept has been defined at all) according to the particular application needs. We assume geographic feature extraction to be an application-independent tasks and formulate applications on the basis of the resulting geographic feature matrix $\mathbf{Z}_{L,F}$.

In Chapter 4, we propose a flexible approach to extract geographic feature signals based on geographic observations with arbitrary influence functions. We show that the model is able to extract robust distributions (see Section 2.3.4), handle user redundancy, and normalizes the background distribution of the observations.

A special instance of geographic feature extraction is concerned with the extraction of a single feature that represents an apriori known phenomenon. Such applications are presented by [Xu et al., 2012] and [Sakaki et al., 2010]. Their applications are specifically tuned to extract particular phenomena (trajectories) [Sakaki et al., 2010] or to consider data-specific side information (user locations) [Xu et al., 2012]. In the context of our framework, we treat phenomenon recovery just as a special instance of geographic feature extraction, or in other words, as a domain-specific realization of the geographic feature extraction task.

3.6.3 Geographic Feature Comparison

Given a geographic feature matrix $\mathbf{Z}_{L,F}$ with a probably large number of geographic feature candidates, an initial data mining task is to select features that have interesting properties (see Section 3.5) and to categorize features by similar semantics.

We use the term *geographic feature comparison* to describe all kinds of selection, filtering, and categorization tasks. This is motivated by the fact that these tasks underlies a pairwise comparison of the feature signals, either to the other signals in the matrix, or to a well-known reference signal (e.g., a uniform distribution or a background population). Moreover, thresholding by statistics like the mean value or the variance of a signal can be seen as a comparison of these statistics to instances of well-known distributions.

In [Chen and Roy, 2009; Hays and Efros, 2008; Zhang et al., 2012b], the authors extract clusters of features that have similar spatio-temporal distributions. By using K-means as the clustering technique, they actually compute the similarity between the signal vectors using the Euclidean distance. This corresponds to a categorization tasks that determines those feature as similar that occur at the same points in space and time.

In Chapter 5, we present a novel representation of the feature signals that allows to select and categorize features on the basis of their type. Finding features that are of similar type allows to build a more specific geographic feature matrix for subsequent mining tasks, or to compare user-generated data sources on the basis of their kinds of covered geographic information (as we will show later).

3.6.4 Latent Geographic Feature Extraction

Even after performing feature selection, filtering, and categorization tasks using appropriate feature comparison methods, the number of candidate features in a geographic feature matrix $\mathbf{Z}_{L,F}$ can still be large. Moreover, there likely exist several features describing the same phenomenon. A primary output of geographic feature mining is, however, a small number of highly interesting feature signals, corresponding to real-world processes and entitites. We call this analytics task *geographic phenomenon discovery.*

To extract a small and informative number of features, those candidate features with similar semantics should be combined into high-level features. Clustering of geographic features by their spatio-temporal signals can be seen as such a geographic phenomenon discovery task [Chen and Roy, 2009; Hays and Efros, 2008; Zhang et al., 2012b]. However, ordinary clustering will result in a binary association of feature candidates with the discovered high-level features. Moreover, the obtained clusters show highly redudant semantics (as we will show later).

In Chapter 6, we introduce a novel approach to extract informative and distinct geographic phenomena from user-generated data, called *latent geographic feature extraction.* The approach is based on dimensionality reduction of the geographic feature matrix. Latent geographic features allow to explore the semantics of geographic space as perceived by the users in the user-generated data source. Moreover, the latent geographic feature signals constitute high-level geographic features themselves, and can be extracted and persisted as geographic raster data for subsequent tasks. Using the terminology of our geographic feature mining framework, the latent geographic features can be used to define a low-dimensional geographic feature matrix $\mathbf{Z}_{L,Q}, n_Q < n_F$, for subsequent data mining and learning tasks, such as predictive models, geographic segmentation, or context-aware recommendation.

3.7 Summary

In this chapter, we introduced a conceptual data mining framework that allows to realize geographic data mining tasks and applications in a highly data-independent manner. For this, we introduced a process to represent the qualitative and geographic information in user-generated data records as geographic feature signals. The aim of the data mining process is then to discover and extract informative feature signals and feature signal combinations from the candidate signals.

We first reviewed a heterogeneous set of related works to identify general problems, as well as underlying models and concepts. Then, we introduced geographic observations, geographic features, the geographic feature matrix, and a knowledge discovery process, to formulate and address different kinds of problems identified before. The fundamental tasks of the framework are (1) geographic feature extraction, (2) geographic feature comparison, and (3) latent geographic feature extraction. Together, they allow to extract a small number of highly informative geographic feature signals, each representing a dimension to describe geographic space.

This chapter provides an exhaustive overview of the topics covered in the next chapters. There, we detail the introduced framework tasks and develop techniques to address their specific challenges.

Chapter 4

Geographic Feature Extraction: A Flexible and Robust Approach

4.1 Introduction

In this chapter we present a novel approach to extract geographic feature signals from user-generated data. Our approach is flexible by being based on geographic observations instead of particular user-generated data sources. The model is also aimed at extracting highly robust signals even in the presence of a strong background distribution and user bias.

The resulting geographic feature signals $z_f(c)$ are important for a variety of tasks and applications:

- To compare features (tags, terms, image elements) on the basis of their geographic semantics and to select features with place and event semantics [Chen and Roy, 2009; Cheng et al., 2010; Rattenbury et al., 2007a].
- To detect places, events, and other geographic phenomena within the spatio-temporal signals of a huge number of features [Chen and Roy, 2009; Hays and Efros, 2008; Leung and Newsam, 2010; Zhang et al., 2012b].
- To predict the location and posting time of records by their content [Cheng et al., 2010; Serdyukov et al., 2009; Wing and Baldridge, 2011].
- To summarize and describe areas based on representative features [Ahern et al., 2007; Crandall et al., 2009; Kennedy, 2008; Kennedy et al., 2007].

Geographic phenomenon recovery can be seen as a special case of geographic feature extraction [Sakaki et al., 2010; Xu et al., 2012]. Here, the spatio-temporal signal of a known phenomenon q should be recovered from user-generated data records, where the geo-referenced records act as binary observations (positive or negative). In the two mentioned papers the binary feature is extracted in a pre-processing step by training a classifier (using manually labeled tweets) or by using heuristics (keywords occurring in the tweets). Then, the signal $z_q(c)$ is estimated based on the positive and negative observations. In this chapter we handle the signal extraction of every feature f similar to the extraction of a phenomenon q. However, we generalize the problem such that not binary observations but arbitrary feature influences can be given as input data.

The remainder of this chapter is structured as follows. First, in Section 4.2 we detail the problem statement and the contributions. In Section 4.3, we present existing approaches to extract geographic feature signals from user-generated data. Then, in Section 4.4 we introduce our novel probabilistic model based on a Bayesian network. In Section 4.5, we discuss the similarity of the proposed models under certain assumptions. Finally, in Section 4.6 we present a detailed evaluation of the parameter influence and compare the models using ground-truth signals.

4.2 Problem Statement and Contributions

The task of geographic feature extraction can be defined as follows:

Given a user-generated data source, extract one or more geographic features $f \in F$ and their respective geographic feature signals $z_f(c)$ based on the qualitative and geographic information in the records.

In the previous chapter we introduced a general representation of the input data in the form of geographic observations $O = \{o_1, \ldots, o_n\}$ with $o = (E_F, E_C, E_U)$, where E_F, E_C, and E_U are positive real-valued influence functions over the feature, spatio-temporal, and user domain, respectively (see Section 3.3.1). In our framework we define geographic feature extraction as follows:

Given a set of geographic observations O, extract a geographic feature matrix $\mathbf{Z}_{L,F} \in \mathbb{R}_+^{n_L \times n_F}$ over the features F on a lattice L. A geographic feature signal $z_{f_i}(c), c \in L$ is represented by column i in the matrix $\mathbf{Z}_{L,F}$.

Since the input data is uncertain and has a high level of noise, the extraction of meaningful and robust signal is non-trivial. In order for a signal to be meaningful, we assume that it adheres to the following properties:

(1) It should be robust to small variations in the input data (see Section 2.3.4).

(2) It should be a piecewise smooth function in the order of the scale level of interest (e.g., a country-level phenomenon should not have a lot of peaks within the country scale). The scale level must be given by the user (see Section 2.3.3).

(3) It should represent the influence (importance) of the feature in space and time. This is the most application-specific property since different notions of importance might be of interest for different applications. Here, we focus on signals that can have arbitrary shaped distributions, that are independent of a possible user bias, and that are not governed by a background distribution.

We propose a novel approach to extract geographic features signals from geographic observations based on a Bayesian network model. It allows to model observation with arbitrary positive influence functions for features, geographic context, and users. Other works mostly assume the input to be given as records with a bag-of-word feature representation, point locations, and single associated users [Ahern et al., 2007; Chen and Roy, 2009; Cheng et al., 2010; Crandall et al., 2009; Rattenbury et al., 2007a]. Since geographic observations include this particular type of measurement as a special case, our model is able to process this input representation as well. Beside this apparent flexibility of our model, the approach allows to:

(1) Extract more robust signals than the existing binomial model. This is an important property, since results just occurring by chance need to be separated from true signals occurring in the noisy and uncertain data.

(2) Extract signals that are not governed by a background distribution in the data. This is needed since all signals will follow the background distribution, which limits their usage to judge about the feature specific importance in space and time.

(3) Extract signals that are not affected by a possible user bias. This is important since in some cases (e.g., user photo series), a single user will govern the signal of some features at some points in space and time. We propose a parametrization that allows to vary the impact of the user redundancy continuously from pure feature-based signals to pure user-based signals.

(4) Since the above parameters have a negative effect on the signal strength, we propose a parametrization that allows to extract stronger signals by optimizing the confidence in positive/negative observations.

4.3 Existing Approaches

We now review existing approaches to extract arbitrary shaped geographic feature signals. Other approaches make assumptions on the distributions in describing the signals by (1) several bumps or dense regions (clustering) [Ahern et al., 2007; Crandall et al., 2009; Kennedy, 2008; Kennedy et al., 2007], (2) a single bump (unimodal distribution) [Backstrom et al., 2008], or (3) a trajectory (HMM) [Sakaki et al., 2010]. An arbitrarily shaped signal $z_f(c)$ might be further processed to determine its spatio-temporal type (see Chapter 5). Then, the extraction of $z_f(c)$ using the following models can be seen as a pre-processing step to extract robust signals at a particular scale level, similar to the idea presented in [Chen and Roy, 2009].

4.3.1 Input Data

For the following approaches, the input can be described as follows. The data set is given as a number of records

$$R = \{r_1, \ldots, r_n\}. \tag{4.1}$$

Each record is described as tuple

$$r = (r_F, r_c, r_u), \tag{4.2}$$

where r_F is a bag-of-words vector of categorical features $f \in F$, $r_c \in D_C$ is a spatio-temporal point coordinate given as timestamp and a spatial coordinate (mostly an associated GPS measurement), and $r_u \in U$ is a single user. We use r_{f_i} do denote the number of times feature f_i occurs in the record and $f \in r_F$ to represent a feature in the record. Note that this representation is similar to a geographic observation where r_f, r_c, and r_u are particular types of influence functions E_F, E_C, and E_U, respectively.

4.3.2 Count Models

Most of the existing works extract the signal just on the basis of the counts of categorical features in space and time [Ahern et al., 2007; Backstrom et al., 2008; Chen and Roy, 2009; Crandall et al., 2009; Kennedy, 2008; Kennedy et al., 2007; Leung and Newsam, 2010; Zhang et al., 2012b]. This happens either explicitly (by counting the number of records or features in the cells of a spatio-temporal lattice), or implicitly (by finding the spatial cluster centers on the basis of the density or record or features). Even in [Sakaki et al., 2010] the hidden Markov model uses the density of positive labeled records in space and time.

We focus on count models defined on a spatio-temporal lattice (discrete spatio-temporal space) instead of density estimates in continuous spatio-temporal space. As discussed in Section 2.3, there is no fundamental difference between the two techniques when comparable bandwidths are used.

The most simple approach to extract an arbitrarily shaped signal based on categorical features is counting of features in the lattice cells $L = \{c_1, \ldots, c_n\}$. The record-based count over the cells $c_1, \ldots, c_n$ is defined as

$$x_i^{(j)} = \text{count}_r(f_j, c_i), \tag{4.3}$$

with

$$\text{count}_r(f_j, c_i) = \sum_{r \in R} \mathbf{1}\{r_c \in c_i\} \cdot \mathbf{1}\{f_j \in r_F\}. \tag{4.4}$$

Here, the records represent binary observations of a feature f in space and time. Other counting schemes can be used, for example, based on the frequency of a feature in the records. This feature-based count is defined as

$$x_i^{(j)} = \sum_{r \in R} \mathbf{1}\{r_c \in c_i\} \cdot r_{f_j}, \tag{4.5}$$

where r_{f_j} denotes the number of times feature f_j occurs in record r. In this case the number of feature occurrences in a record can be seen as a positive and additive influence of a feature in a record. For records with a small number of features (such as words in tweets) , this counting scheme will be similar to the record-based counts. However, such a counting scheme will lead to different results for longer text documents. We see the feature-based counts as a more fine-granular type of observation, which should be used if possible (such as for long texts).

For both counting schemes, the aggregated number of features in a cell is used as the geographic feature signal

$$z_{f_j}(c_i) = x_i^{(j)}. \tag{4.6}$$

Note that the signal follows an existing background distribution and both are sensitive to a user bias.

User Redundancy

User redundancy can be tackled by using a counting scheme based on the number of distinct users using a feature in a spatio-temporal cell

$$x_i^{(j)} = \mathrm{count}_u(f_j, c_i), \tag{4.7}$$

where

$$\mathrm{count}_u(f_j, c_i) = |\{r_u | r \in R \wedge r_c \in c_i \wedge f_j \in r_F\}|. \tag{4.8}$$

This limits the influence of users having a large number of records at a particular point in space and time. Note that with this user-based counting scheme (1) the number of features in a record cannot be taken into account, such as in (4.5), and (2) the user influence of a record must be a binary relationship (associated or not).

As we stated before, it is reasonable to exploit the frequency of features in records, and, as discussed in Section 3.3.6, to consider more complex relationships between users and records. In our proposed model we allow to handle the user redundancy given arbitrary feature influence functions (corresponding to different counting schemes).

Robustness

The extracted count-based signal $z_f(c)$ is highly robust since it is based on the records themselves. If the records change slightly the signal will change slightly, since there is a direct relationship between the number of records and the signal intensity (see Section 2.3.4).

Transformations

The extracted signals can be further processed to obtain more meaningful results. For example, since the signal intensity will clearly follow a background distribution, this influence can be limited by down-weighting high intensity values.

We propose three transformations that are also used in Chapter 6. All of them reduce the influence of high signal values:

- The square-rooted signal

$$z'_f(c) = \sqrt{z_f(c)}. \tag{4.9}$$

- The logged signal
$$z'_f(c) = \log(1 + z_f(c)). \tag{4.10}$$

- The binarized signal
$$z'_f(c) = \mathbf{1}\{z_f(c) > 0\}. \tag{4.11}$$

A binarized signal takes a single observation as the only evidence needed to turn a cell into a positive signal or not, making the result totally independent from an assumed background distribution.

Runtime Complexity

We now discuss the runtime complexity for the count model in a lattice with n_C cells, a set of n_R records, and n_U users. This discussion (as well as the discussion for the space complexity in the next section) will also be valid for the binomial model and the Poisson model explained later, since both are based on the counts of features or users in the lattice.

To extract the record- or feature-based counts of a single feature in a spatio-temporal lattice with n_C cells and n_R records we need to process the records once and sum up the record or feature counts. We can determine the cell c of a spatio-temporal point r_c in $O(1)$ time in a lattice defined on a regular grid. For the feature-based counts in (4.4) and (4.5) the runtime complexity will hence be $O(n_R)$.

Given an irregular grid (such as the US states) the runtime complexity to find the corresponding cell will be $O(\log n_C)$ if an appropriate spatial index structure is used. Then, the runtime complexity to build the spatio-temporal lattice will be $O(n_R \log n_C)$, and $O(n_R n_C)$ if no index structure is used.

When counting the number of users we need to update a set of users in each cell to determine if the user was already inserted or not. Let n_U be the total number of users. Each cell would need a data structure to check for duplicate users. We assume a standard binary search tree to realize those set operations. We can reduce the problem to an insert of every record in a single binary search tree, to test if a user has already been inserted or not. Given that the duplicate check will be performed when inserting the records (in time $O(\log O_R)$), the total runtime complexity will be $O(n_R \log n_R)$ in a regular lattice and $O(n_R(\log n_R + \log n_C))$ in an irregular lattice with index support.

One can use an approximation of the user counts by using the hashing-based Flajolet-Martin algorithm [Leskovec et al., 2014, p. 124], which determines the number of distinct items in a set in $O(1)$ amortized time. The average error of the estimate is

$$err(v) = 0.77351 \log_2(v), \tag{4.12}$$

where v is the size of the set [Martin, 1985]. By this, the total runtime for the regular lattice case will also be $O(n_R)$ and for the irregular case with index support $O(n_R \log n_C)$, on the cost of a small counting error (4.12).

Space Complexity

A lattice with n_C cells will need an array of that size to hold the counts, resulting in $O(n_C)$ space. However, there will be a large number of cells having no count such that a sparse representation can be used. There, only cells having a count greater than zero are recorded. Given the sparsity factor

$$\kappa_C = \frac{\delta_C}{n_C}, \tag{4.13}$$

with

$$\delta_C = |\{c \in C \wedge z_f(c) > 0\}|. \tag{4.14}$$

The space complexity will then be $O(n_C \kappa_C) = O(\delta_C)$. Note that smoothing of the signal, e.g., by convolution with a kernel, will increase κ_C and hence the space complexity. In the worst case, such as smoothing by a kernel with infinite support (e.g., a non-bounded Gaussian), κ_C will be 1 and $\delta_C = n_C$. In such cases, a lazy smoothing strategy can be used that applies the kernel just when the smoothed signal is needed (such as before visualization or before further processing). One can hold the complete data in a sparse matrix $\mathbf{Z}'_{L,F}$ and apply a pre-processing routine around the signal each time it is requested

$$z_f(c) = \theta(z'_f(c)). \tag{4.15}$$

In order to manage the signals more efficiently also compressed matrix representations can be employed, on the cost of an additional runtime overhead. Such techniques will not be discussed in this thesis, however.

In case of determining user counts, a set of distinct users needs to be recorded for each cell (e.g., in a binary search tree). Given the average number of users in a cell is $\delta_{cell,U}$, and the sparsity being $\kappa_{cell,U} = \delta_{cell,U}/n_U$. Then, the space complexity will be $O(n_C n_U \kappa_C \kappa_{cell,U})$ to compute the result, and $O(n_C \kappa_C)$ to hold the final result.

4.3.3 Binomial Model

Using counts is a common and robust way to represent the influence of a feature in space and time. It is, however, strongly influenced by an underlying background distribution. In the following we introduce a geographic feature extraction technique that normalizes the signal according to an unknown background distribution.

As input we assume binary count data (feature is present or not). The number of positive events can then be modeled by a binomial distribution. Let $x_i^{(j)} = \text{count}_r(f_j, c_i)$ and $m_i = \text{count}_r(c_i)$, then

$$x_i^{(j)} \sim Binomial(x_i^{(j)}; p_i^{(j)}, m_i), \tag{4.16}$$

with $p_i^{(j)}$ being the probability that a positive event will happen in cell c_i. The MLE estimate of $p_i^{(j)}$ is

$$\hat{p}_i^{(j)} = \frac{x_i^{(j)}}{m_i}. \tag{4.17}$$

We represent the signal as the probability that the feature will be used in a cell.

$$z_{f_j}(c_i) = \hat{p}_i^{(j)}. \tag{4.18}$$

This is an appropriate model to normalize an unknown background distribution. In this case the total number of records is assumed to represent the overall population.

Given a geographic feature matrix of count signals $\mathbf{Z}_{L,F}$. Note that in such a matrix each signal can be described by the distribution $\mathrm{p}(f, c)$. The binomial model is derived from the count-based matrix by

$$\mathrm{p}(f|c) = \frac{\mathrm{p}(f,c)}{\sum_{f \in F} \mathrm{p}(f,c)} = \frac{\frac{\mathrm{count}_r(f,c)}{\mathrm{count}_r(\cdot)}}{\frac{\sum_{f \in F} \mathrm{count}_r(f,c)}{\mathrm{count}_r(\cdot)}} = \frac{\mathrm{count}_r(f,c)}{\sum_{c \in L} \mathrm{count}_r(f,c)}, \tag{4.19}$$

since for the record- and feature-based count functions

$$\mathrm{count}_r(c) = \sum_{f \in F} \mathrm{count}_r(f,c). \tag{4.20}$$

This establishes the relationship between the record- and feature-based count models and the binomial model.

User Redundancy

User redundancy can be tackled in the same way as for the count model by counting the users instead of the features. Obtaining $\mathrm{p}(f|c)$ from the count-based geographic feature matrix $\mathbf{Z}_{L,T}$ such as in (4.19) needs, however, a separate computation, since

$$\mathrm{count}_u(c) \neq \sum_{f \in F} \mathrm{count}_u(f,c). \tag{4.21}$$

This is because users in a cell might support several features and will be counted repeatedly. In this case the binomial signal can be obtained by using $\mathrm{p}(c|f)$ as extracted from a count-based matrix $\mathbf{Z}_{L,F}$ and computing $\mathrm{p}(f)$ separately by the number of distinct users using feature f as in (4.7). Then, the binomial signal is extracted by

$$\mathrm{p}(f|c) = \frac{\mathrm{p}(c|f)\,\mathrm{p}(f)}{\sum_{f' \in F} \mathrm{p}(c|f')\,\mathrm{p}(f')}. \tag{4.22}$$

Robustness

The binomial signal is not robust to small changes of the input data. For this note that

$$z_{f_j}(c_i) = \frac{x_i^{(j)}}{m_i} \tag{4.23}$$

directly depends on the total number of records in a cell m_i. If m_i is small, the fraction will jump if $x_i^{(j)}$ differs by a small quantity. More precisely, the standard error of the estimate $\hat{p}_i^{(j)}$ is

$$err\left(\hat{p}_i^{(j)}\right) = \sqrt{\frac{\hat{p}_i^{(j)}\left(1-\hat{p}_i^{(j)}\right)}{m_i}} \tag{4.24}$$

and depends inversely on the number of total counts in the cell m_i. Hence, the error will be high in cells with a small total number of records (a small background population), and the error will be the highest for signal values around 0.5.

Runtime and Space Complexity

Since the binomial model is directly based on the counts of positive and negative observations the runtime and space complexity are the same as for the count-based models (see Section 4.3.2)

4.3.4 Linked Poisson Model

In the binomial model the total count of records or features in a cell has been used an indicator of the background population. One can also assume an arbitrary positive signal indicating the strength of a background distribution. In [Xu et al., 2012] the authors propose a linked Poisson distribution to model the positive number of tweets (records having a feature f indicating a positive observation) given an arbitrary background distribution.

We first introduce the standard Poisson model. There, the positive counts are modeled as

$$x_i^{(j)} \sim Poisson(x_i^{(j)}; \lambda_i^{(j)}), \tag{4.25}$$

where $\lambda_i^{(j)}$ is the intensity parameter indicating the average number of counts in a cell. Given just a single count value for each cell, the MLE estimate is simply the number of counts itself (similar to the count model)

$$\hat{\lambda}_i = x_i^{(j)}. \tag{4.26}$$

In the following we omit the supscripts of $\lambda_i^{(j)}$ and $x_i^{(j)}$, and just use λ_i and x_i for representing the intensity and the counts of a feature f, respectively. Given an arbitrary background signal in the cells $b_1, \ldots, b_n, b_i \in \mathbb{R}_+$ (not necessarily the total number of counts). The background is treated as an intensity, similar as λ. One can introduce a link function between the positive counts and the background intensity. This function might represent an arbitrary relationship but is usually given as

$$\gamma(x_i, b_i) = x_i \cdot b_i. \tag{4.27}$$

The link function is introduced in the Poisson model as

$$x_i \sim Poisson(\gamma(x_i, b_i); \lambda_i). \tag{4.28}$$

The MLE estimate of λ_i is then

$$\hat{\lambda}_i = \frac{x_i}{b_i}. \tag{4.29}$$

Note that if b_i is the total number of counts (sum of positive and negative records) then this is just the same signal as for the binomial model.

The Poisson model is an important tool in statistical epidemiology since it can be used to introduce other covariate information (other indicators) by introducing them in the link function [Lawson, 2001]. Also, other link functions have been introduced in this context. However, the above multiplicative relationship is a common choice and also used in [Xu et al., 2012].

The intensity is finally used as the geographic feature signal

$$z_{f_j}(c_i) = \hat{\lambda}_i^{(j)}. \tag{4.30}$$

Since for our problem we are not given an external background distribution we assume the total number of records as an indicator for it, as in [Xu et al., 2012]. The authors still used the Poisson model to realize explicit smoothing based on a neighborhood matrix **W**. By this, they are able to smooth the signal given their assumed coarse US-state lattice. This regularization-based method is, however, very expensive. Since our focus is on the extraction of smooth signals given a possibly high resolution regular lattice, we smooth the signal using efficient kernel convolution methods, which will be detailed later.

Also, the authors extend the link function to include the Twitter-specific location names in the user profiles as covariate information. Since this technique cannot be generalized to other data sources easily, we do not consider this extension in our comparison.

Since we do not make use of these two extensions, the Poisson model is reduced to the binomial model. In this case the user redundancy handling as well as the runtime and space complexity for the Poisson model are the same as for the binomial model. Hence, we use the count and the binomial model as comparative approaches.

Both, the binomial and the Poisson model are only well defined on count data. In our Bayesian network model we generalize the above models to be used with arbitrary influence functions on the features, geographic context, and user.

4.4 Bayesian Network Model

In the following we introduce a flexible and robust approach to extract geographic feature signals from user-generated data. The expected input is a set of geographic observations, with each observations having arbitrary positive influence functions as defined in Section 3.3.1. This makes the model far more flexible that the count or binomial model. The resulting geographic feature signals can be represented in a geographic feature matrix (see Section 3.4.1). This makes the approach an important sub-task in the geographic feature mining framework.

In addition to its flexibility to take various kinds of features, geographic information, and user information provided by a data sources into account, our proposed model has the following benefits:

- It allows to extract more robust signals than the binomial model by allowing to parametrize an appropriate prior signal.

- Our model allows to vary the impact of redundant user observations on the extracted signal continuously, making it an instance of the record count and the user count models.

- The model allows to mix arbitrary geographic context information on the basis of the influence functions. By this, we are able to exploit different kinds of information in the records in a unifying way.

- Since the signal prior has a negative effect on the signal strength, we propose a parametrization that allows to extract stronger signals by optimizing the confidence in positive/negative observations.

4.4.1 Bayesian Networks

We use the language of Bayesian networks to define our model. A Bayesian network BN consists of a graph $\mathcal{G}$ whose nodes $\mathcal{X} = \{X_1, \dots, X_n\}$ are random variables and where each edge corresponds to the direct influence of one node on another node. Each node is represented by a conditional probability distribution (CPD), describing the probability of a node value given the values of the parent nodes. The graph $\mathcal{G}$ can be viewed as a skeleton for representing the joint distribution over all random variables. The joint distribution is defined via the chain rule for Bayesian networks

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i | pa(X_i)), \tag{4.31}$$

where $pa(X_i)$ denotes the parents of node X_i in the graph $\mathcal{G}$. Here, we assume that all variables are known such that we deal with a fully observed Bayesian network. Moreover, w.l.o.g, we assume that all variables are discrete. Note that a Bayesian network is also called a directed graphical model such as described in Section 2.4.2. We use $val(X) = \{x_1, \dots, x_k\}$ to denote the k discrete states of variable X. For binary random variables, we use $val(X) = \{x^0, x^1\}$, with x^0 denoting the false state and x^1 denoting the true state.

A probability distribution is denoted $P(X)$. The probability of X for value x is $P(X = x) = P(x)$. We now use the upper-case letter P to denote a probability distribution, since this is the standard notation used in Bayesian network literature [Kollar and Friedman, 2009].

A probability distribution $P(X, Y)$ can be restricted by setting variables to known states such as $P(X, Y = y) = P(X, y)$. Now the probability distribution $P(X, y)$ is only a distribution on X since the state y is inserted as evidence in the network. In a situation where the structure of the network and the probabilities of the CPDs are known inference consists of:

(1) Inserting evidence into the model by setting variables to known states.

(2) Marginalizing (integrating) out variables, e.g., $P(X) = \sum_{y \in Y} P(X, Y)$, and applying Bayes rule.

For example, from $P(R, Q, C)$ we derive the conditional probability $P(q|c)$ by marginalization of R and applying Bayes' rule

$$P(q|c) = \frac{\sum\limits_{r \in R} P(r, q, c)}{\sum\limits_{r' \in R} \sum\limits_{q' \in Q} P(r', q', c)}. \tag{4.32}$$

For a detailed introduction to Bayesian networks and graphical models we refer the reader to [Kollar and Friedman, 2009].

4.4.2 Probabilistic Signal Extraction

We start with assuming a given signal $z_q(c)$ of a geographic phenomenon q. The signal represents the phenomenon in the value domain $[0, 1]$, with $z_q(c) = 1$ if the signal occurs at c and $z_q(c) = 0$ otherwise. This behavior is reasonable for different scenarios:

- The signal represents the confidence in presences or absence. Given a signal between 0 and 1 means there is some uncertainty in knowledge about the presence of the phenomenon.
- The signal represents how much of the spatio-temporal window has been covered. Let the spatio-temporal space be discrete with $c \in L$. A process might cover the whole cell ($z_q(c) = 1$) or it might not cover the cell at all ($z_q(c) = 0$). A value between 0 and 1 indicates a partial coverage of the cell.
- The signal represent an intensity of an underlying process. For example, let the phenomenon represent the crime rate. A value of 1 then represents the maximal crime rate while a value of 0 represent no crime at all.

All of the interpretations above are reasonable to understand the semantics of a signal, and the interpretations might overlap. It is up to the analyst to interpret the extracted signal in an appropriate way, depending on the input data (geographic observations) and the type of qualitative features. We now introduce a probabilistic model to extract such a signal from arbitrary geographic observations.

The variable C represent the geographic context space. The distribution $P(C)$ is the background distribution with $P(c)$ being the probability that measurements occur at cell c. A binary random variable $Q = \{q^0, q^1\}$ denotes presence or absence of a phenomenon q. We make the necessary assumption that the background distribution and the phenomenon signal are independent. Then, the joint probability that a phenomenon q has been observed at c is

$$P(q^1, c) = z_q(c)P(c). \tag{4.33}$$

The intuition behind this statement is that $z_q(c)$ represents the mass of phenomenon q found at c, and $P(c)$ is the probability that a measurement of this phenomenon is performed. We can then interpret the phenomenon signal $z_q(c)$ as the context-conditional probability to find the phenomenon

$$P(q^1|c) = \frac{P(q^1, c)}{P(c)} = z_q(c). \tag{4.34}$$

We need the binary variable Q to distinguish between positive and negative observations in the Bayesian network. Given we have a feature f that represents a positive observation q^1, the above probability equals $\mathrm{p}(f|c)$, which we used in the previous chapter to represent a geographic feature signal.

Let $P(Q, C, \mathcal{X})$ be a probability distribution with variables Q and C (as described above) and also a number of other variables $\mathcal{X}$. The inference task in any such distribution to extract the geographic feature signal for phenomenon q is

$$z_q(c) = P(q^1|c) = \frac{\sum\limits_{\mathcal{X}} P(q^1, c, \mathcal{X})}{\sum\limits_{\mathcal{X} \cup Q} P(Q, c, \mathcal{X})}. \tag{4.35}$$

4.4.3 Random Variables

We now define the random variables that will occur in our model. The random variables represent the input data as follows:

- R is a discrete random variable representing the measurements. Each record r is an individual measurement and provides information about qualitative features, their spatio-temporal context, and about what users created the record. The value domain of R is the set of records in our data set, i.e., $val(R) = \{r_1, \dots, r_{n_R}\}$. In the case of geographic observations, we have $R = O$.

- Q is a binary random variable describing if a phenomenon has been observed (q^1), or not (q^0). This variable represents a feature f by two states (present or absent). Note that here Q is not used to denote a set of phenomena but the observation state of a single phenomenon.

- C is a random variable representing the geographic context. W.l.o.g we assume a discrete space given by a spatio-temporal lattice L. The space consists of spatio-temporal cells $val(C) = \{c_1, \dots, c_{n_L}\}$. The probability distribution $P(C)$ can be seen as the background distribution.

- U is a discrete random variable representing the users. The value domain is the set of individual users, i.e., $val(U) = \{u_1, \dots, u_{n_U}\}$.

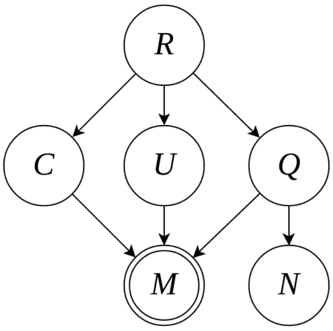


Figure 4.1: Bayesian network structure of the measurement domain.

Apart from the domain variables, we introduce two auxiliary variables to model dependencies and prior assumptions:

- M is a binary constraint variable. It is introduced to model the dependence between context, users, and phenomena. Only the m^1 state of the variable will be of interest, and this state will be given as evidence in our Bayesian network.
- N is a binary virtual evidence variable. It is used to represent the overall confidence we have in our semantic measurements. Only the n^1 state of the variable will be of interest, and this state will be given too as evidence in our Bayesian network.

As stated above, we keep $M = m^1$ and $N = n^1$ fixed such that we only consider the joint distribution $P(R, C, U, Q, m^1, n^1)$. Given the distribution the inference task is:

$$z_q(c) = P(q^1|c, m^1, n^1) = \sum_u \sum_r \frac{P(r, c, u, q^1, m^1, n^1)}{\sum_{u'} \sum_{r'} \sum_q P(r', c, u', q, m^1, n^1)}. \tag{4.36}$$

4.4.4 Network Structure

The Bayesian network structure to define the joint probability over our domain variables is shown in Figure 4.1. In this structure the geographic context C, the users U, and the phenomenon Q are conditionally independent given the records R. This can be written as follows

$$P(C, U, Q|R) = P(C|R)P(U|R)P(Q|R). \tag{4.37}$$

This independence assumption is a simplification. As stated in Section 4.4.2 the independence assumption between the context C and Q is reasonable since we do not expect that the total mass of measurements in geographic space C to depend on the kinds of phenomenon Q occurring there (and vice versa). As we discussed earlier there are, however, reasons for not making such an independence assumption between geographic context, the phenomenon and the users, since some users might exist that are heavily biased to observe particular phenomena at particular points in space and time. Because of this, dependency between C, U, and Q is introduced by the auxiliary constraint variable

M. We will use a specific functional form of $P(M|C,U,Q)$ that allows to continuously vary the influence of redundant observations on the signal. Also, a virtual evidence variable N is introduced to parametrize the evidence of the phenomenon distribution $P(N|Q)$. By this, we are able to control the influence of positive and negative observations to optimize the resulting signal. Finally, $P(R)$ is the probability that a record is a measurement at all. This is assumed to be

$$P(r) = \frac{1}{n_R}. \tag{4.38}$$

This just reflects the prior probability that a record in the data set is considered a measurement. Since we do not distinguish between important and non-important measurements, this is a constant probability for all records.

The joint distribution is defined as follows

$$P(R,C,U,Q,M,N) = P(C|R)P(U|R)P(Q|R)P(N|Q)P(M|C,U,R)P(R). \tag{4.39}$$

Since we use M and N to introduce a parametrization on Q and on the dependency between C, U, and Q, we will always activate these CPDs by inserting evidence in the network. Thus, we are dealing with the following distribution

$$P(R,C,U,Q,m^1,n^1) = P(C|R)P(U|R)P(Q|R)P(n^1|Q)P(m^1|C,U,R)P(R). \tag{4.40}$$

Equations (4.39) and (4.40) provide a skeleton to compute the joint distribution once we are given the factors (CPDs). Based on the joint distribution the signal can then be extracted using (4.36).

In the following we detail for $P(C|R)$, $P(U|R)$, and $P(Q|R)$ how they can be extracted from the input data, and define the functional form of $P(N|Q)$ and $P(M|C,U,R)$ to introduce model parameters.

Geographic Record Influence

We call the CPD $P(C|R)$ the geographic record influence. For a given a record r, the spatio-temporal distribution $P(C|r)$ represents its influence over the cells $c \in L$. The distribution is similar to a normalized spatio-temporal influence variable $\dot{z}_q(c)$ (see Section 2.2.1). Given a geographic observation o, the distribution $P(C|r)$ is just its spatio-temporal influence E_C.

In the case of a given set of user-generated data records (e.g., photos or tweets), we extract the geographic influence using a spatio-temporal variable as detailed in Section 2.2.1. If the records have an associated GPS coordinate we can use a piecewise constant step function (2.10) to represent the influence in space and time. If the records are associated with a polygon or bounding box we can use (2.11) to represent $P(C|r)$. Each record can have its own specific geographic record influence. This allows to use different kinds of geographic information for each record (user location, GPS coordinate, bounding box).

We can model spatial dependency directly within this CPD. For a single associated point coordinate (e.g., GPS coordinate) the resulting Gaussian influence function (2.12) can be used, where a given standard deviation σ specifies the strength of the spatio-temporal interaction. Such interaction assumptions are crucial to model uncertainty. For example, a GPS coordinate will clearly not specify an infinitely small point on Earth with total certainty, but can be described as a bump with highest certainty (the center) at this point. The Gaussian representation allows to model exactly this behavior.

In some cases an indicator of the scale level of a given spatial point is given explicitly, e.g., by a type attribute. An example are spatial coordinates provided as metadata in Wikipedia documents. Additional to the point coordinate itself, this metadata attribute contains a scale level indicator (street, city, country, continent). This scale level can be used to set the standard deviation in (2.12) such that, e.g., countries and cities are represented as bumps with corresponding shapes.

Phenomenon Record Influence

We call $P(Q|R)$ the phenomenon record influence. The distribution that a record r is a measurement of phenomenon q is then represented as $P(Q|r)$. The CPD is a distribution of the two binary states q^1 and q^0, with $P(q^1|r)$ denoting the probability that the record is a positive observation, and $P(q^0|r) = 1 - P(q^1|r)$ being the probability that the record is a negative observation. We need this separation to make the positive and negative cases dependent on a parametrization introduced by $P(N|Q)$ later.

In the case of a given categorical feature we can set $P(q^1|r) = 1$ if a feature f occurs in the record r. We can also realize the feature-based counting scheme by using the feature frequency within the record as $P(q^1|r)$. Note that $P(q^0|r)$ does not need to be defined as it follows from the positive state.

In [Sakaki et al., 2010] the authors train a classifier to label a record as a positive observation. We can use the confidence output of the classifier as $P(q^1|r)$ to distinguish between confident and non-confident feature influences. In [Xu et al., 2012] the authors use a heuristic on the basis of occurring keywords in a record. We can set $P(q^1|r) = 1$ if record r contains a keyword. Furthermore, we can use arbitrary rules to define our CPD, e.g., if keyword k_a occurs we set $P(q^1|r) = p_a$, if keywords k_b and k_c occur, we set $P(q^1|r) = p_{bc}$, and $P(q^1|r) = 0$ otherwise.

Given a geographic observation o, the qualitative feature signal E_F represents $P(q^1|r)$ (see Section 3.3.4).

In the following experiments we use a simple rule-based CPD on the basis of the textual content of tweets with $P(q^1|r) = 1$ if a keyword string occurs in the tweet, and $P(q^1|r) = 0$ otherwise. By this we are able to make a fair comparison between our results and the results obtained from the count model and the binomial models.

Phenomenon Confidence

In the experiments we show that the overall confidence we have in positive and negative observations can be used to optimize the signal. For this, we introducing the auxiliary

Q	$P(Q\|r)$
q^0	**0.5**
q^1	**0.5**

N	Q	$P(N\|Q)$
n^0	q^0	0.1
n^1	q^0	**0.9**
n^0	q^1	0.3
n^1	q^1	**0.7**

N	Q	$P(N,Q\|r) = P(N\|Q)P(Q\|r)$
n^0	q^0	$0.1 \cdot 0.5 = 0.05$
n^1	q^0	$\mathbf{0.9 \cdot 0.5 = 0.45}$
n^0	q^1	$0.3 \cdot 0.5 = 0.15$
n^1	q^1	$\mathbf{0.7 \cdot 0.5 = 0.35}$

Table 4.1: Probability tables for $P(N,Q|r) = P(N|Q)P(Q|r)$.

variable N and define $P(N|Q)$ as follows:

$$P(n^1|Q) = \begin{cases} \beta_0 & \text{if } Q = q^0 \\ \beta_1 & \text{otherwise} \end{cases} \tag{4.41}$$

with $\beta_0, \beta_1 \in [0,1]$ and $P(n^0|Q) = 1 - P(n^1|Q)$. The parameter β_0 represents the overall confidence we have in the negative observations and β_1 the confidence we have in the positive observations. A parameter value of $\beta_0 = 0.9$ means that we are 90% sure that a negative observation is correct, a value of $\beta_1 = 0.7$ means that we are 70% sure that a positive observation is correct. Both parameters can be set independently. The impact of the parameters is shown in Table 4.1 for $\beta_0 = P(n^1|q^0) = 0.9$ and $\beta_1 = P(n^1|q^1) = 0.7$. In this example the record is equally likely a positive or negative observation. However, after applying the confidence parameters we treat the record as being more likely a negative observation, since now $P(n^1, q^0|r) > P(n^1, q^1|r)$. Note that only the n^1 state is of interest and is given as evidence in the network.

User Record Influence

$P(U|R)$ specifies the probability that a user has an influence on a record. By defining the user influence by a separate factor within the model we can set it independently of the geographic record influence and the phenomenon record influence. Note that this is a huge difference to the restrictions in the other models. There, user counts cannot be used with feature-based counts. In our model we are able to use arbitrary $P(Q|R)$ CPDs.

Since we integrate over the users in (4.36) the $P(U|R)$ CPD will have no effect on the resulting signal so far. We use the variable M in the next section to introduce dependence between users U, geographic context C, and phenomenon Q.

For the experiments, we assume that each record has a single associated user r_u and we set $P(r_u|r) = 1$ (resulting in a zero probability for all other users). This again allows to compare the different models appropriately.

User Redundancy

To realize dependence between U, C, and Q we introduce the auxiliary variable M in the model by $P(M|Q,C,U)$. The aim of variable M is to penalize observations that show a high redundancy with respect to a joint user, context, and phenomenon occurrence.

We define a relation $J(u,c,q,r)$ between user, geographic context, phenomenon, and records. A co-occurrence of the variables (a measurement) is assumed to have a positive impact on the signal. However, subsequent measurements of the same co-occurrence are expected to provide less evidence than the first occurrence. Since the co-occurrence depends on the CPDs introduced above, we define our relation on their basis

$$J(u,c,q,r) = \begin{cases} 1 & \text{if } P(u|r) > 0 \wedge P(c|r) > 0 \wedge P(q^1|r) > P(q^0|r) \\ 0 & \text{otherwise} \end{cases} \tag{4.42}$$

Then, we define the number of co-occurrences a user u contributed about q at context c by

$$m(u,c,q) = \sum_{r \in R} J(u,c,q,r). \tag{4.43}$$

Now, the CPD $P(m^1|U,C,Q)$ is defined as a function of $m(u,c,q)$, penalizing large co-occurrences. It is defined as

$$P(m^1|u,c,q) = \begin{cases} \frac{1}{m(u,c,q)^\epsilon} & \text{if } m(u,c,q) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.44}$$

with $P(m^0|u,c,q) = 1 - P(m^1u|,c,q)$. The parameter ϵ is used to control the strength of penalization. For $\epsilon = 0$ the influence is not penalized at all and for $\epsilon = 1$ the influence is inverse proportional to $m(u,c,q)$.

4.4.5 Inference

Given the CPDs $P(C|R)$, $P(Q|R)$, $P(U|R)$ and the parameters β_0, β_1, and ϵ the signal

$$z_q(c) = P(q^1|c,n^1,m^1) \tag{4.45}$$

at geographic context c can be computed by summing over all records and users as shown in (4.36). The resulting runtime of this brute force approach is $O(n_R n_U)$. To obtain the signal over the complete geographic context the runtime will be $O(n_R n_U n_C)$. Note that this is much higher than $O(n_R \log n_R)$ (user-based count or binomial model) or (n_R) (record- or feature-based count or binomial model). In the following we introduce an efficient algorithm to compute the signal and discuss its runtime and space complexity.

Algorithm and Runtime Complexity

The above brute force attempt assumes that all records might have an influence on all users U and the complete context space C. However, those influences are likely to be bounded. We use $\delta_{rec,U}$ to denote the average number of users having $P(U|R) > 0$ and $\delta_{rec,C}$ to denote the average number of cells having $P(C|R) > 0$. First, we rewrite the general inference equation

$$\begin{aligned} P(q^1|c,m^1,n^1) &= \sum_u \sum_r \frac{P(r,c,u,q^1,m^1,n^1)}{\sum_{u'} \sum_{r'} \sum_q P(r',c,u',q,m^1,n^1)} \\ &= \frac{P(c,q^1,m^1,n^1)}{P(c,q^0,m^1,n^1) + P(c,q^1,m^1,n^1)} \end{aligned} \tag{4.46}$$

with

$$\begin{aligned} P(c,q,m^1,n^1) &= \sum_r \sum_u P(r,c,u,q,m^1,n^1) \\ &= \sum_r \sum_u P(q|r)P(c|r)P(u|r)P(m^1|u,c,q)P(n^1|q)P(r) \end{aligned}. \tag{4.47}$$

Once we are able to compute (4.47) efficiently for q^0 and q^1, we can obtain the signal by (4.46). Efficient calculations are possible by exploiting sparsity and by using intermediate results.

For this we reorder (4.47) and drop $P(r)$ because it is a constant and will cancel out in (4.46)

$$\begin{aligned} \tilde{P}(c,q,m^1,n^1) &= P(n^1|q) \sum_u P(m^1|u,c,q) \sum_r P(c|r)P(u|r)P(q|r) \\ &= P(n^1|q) \sum_u \frac{1}{m(u,c,q)^\epsilon} \sum_r P(c|r)P(u|r)P(q|r) \\ &= P(n^1|q) \sum_u \frac{w(u,c,q)}{m(u,c,q)^\epsilon}. \end{aligned} \tag{4.48}$$

To achieve this we inserted (4.44) and used

$$w(u,c,q) = \sum_r P(c|r)P(u|r)P(q|r). \tag{4.49}$$

The functions $w(u,c,q)$ and $m(u,c,q)$ can be computed by a single run. Given that a record has influence only over a small number of cells and users, both can be processed by traversing over all records once with a runtime complexity of $O(n_R \delta_{rec,U} \delta_{rec,C})$. Since $\delta_{rec,U}$ and $\delta_{rec,C}$ are assumed to be small we almost achieve $O(n_R)$. Given $w(u,c,q)$ and $m(u,c,q)$ we need to traverse over all users

$$\tilde{P}(c,q,m^1,n^1) = P(n^1|q) \sum_u \frac{w(u,c,q)}{n(u,c,q)^\epsilon} = P(n^1|q) W(c,q,\epsilon), \tag{4.50}$$

where

$$W(c,q,\epsilon) = \sum_u \frac{w(u,c,q)}{n(u,c,q)^\epsilon}. \tag{4.51}$$

The function $W(c,q,\epsilon)$ (4.51) is the user aggregated and redundancy penalized influence over all records and all users. This operation needs $O(n_U n_C)$ time using a brute force approach. However, the spatio-temporal signal values of a user are at least as sparse as the final signal, which we denoted as $\kappa_C = \delta_C / n_C$ in Section 4.3.2. Hence, the runtime can assumed to be $O(n_U n_C \kappa_C) = O(n_U \delta_C)$. Once we have computed (4.51) for q^0 and q^1 we can derive

$$\tilde{P}(c,q^0,m^1,n^1) = P(n^1|q^0) W(c,q^0,\epsilon) = \beta_0 W(c,q^0,\epsilon) \tag{4.52}$$

and

$$\tilde{P}(c, q^1, m^1, n^1) = P(n^1|q^1)W(c, q^1, \epsilon) = \beta_1 W(c, q^1, \epsilon). \tag{4.53}$$

From these two statements we can compute the final signal by

$$\begin{aligned} z_q(c) &= \frac{P(c, q^1, m^1, n^1)}{P(c, q^0, m^1, n^1) + P(c, q^1, m^1, n^1)} \\ &= \frac{\tilde{P}(c, q^1, m^1, n^1)}{\tilde{P}(c, q^0, m^1, n^1) + \tilde{P}(c, q^1, m^1, n^1)} \\ &= \frac{\beta_1 W(c, q^1, \epsilon)}{\beta_0 W(c, q^0, \epsilon) + \beta_1 W(c, q^1, \epsilon)}. \end{aligned} \tag{4.54}$$

Hence, once we have computed $W(c, q, \epsilon)$ the signal can be determined in $O(1)$ time. In summary it will need $O(n_U \delta_C + n_R \delta_{rec,U} \delta_{rec,C})$ time to compute the signal over the complete context space, which is much faster than the brute force attempt $O(n_U n_C n_R)$. Algorithm 4.1 details the computation of $W(c, q)$ given a parameter ϵ. One can see that the computations factor into two sub-routines to compute $w(u, c, q)$ and $m(u, c, q)$, given the CPDs, and to compute $W(u, c, \epsilon)$, given ϵ. By this, we can reuse $w(u, c, q)$ and $m(u, c, q)$ as long as the CPDs do not change.

Space Complexity

The approach first needs to compute $w(u, c, q)$ and $m(u, c, q)$ resulting in a space complexity of $O(n_U n_C)$ (since $n_Q = 2$ is a constant) to hold the counts and influences for each user and each geographic context cell. Again we exploit the sparsity of the signal and the user distribution to reduce the space complexity. Let κ_C be the sparsity factor of the signal in context space and $\kappa_{cell,U}$ be the sparsity factor on the number of users contributing to a cell. The resulting space complexity to compute $w(u, c, q)$ and $m(u, c, q)$ is then $O(n_U n_C \kappa_C \kappa_{cell,U})$. Afterwards, $w(u, c, q)$ and $m(u, c, q)$ are aggregated into $W(c, q)$, which needs $O(n_C \kappa_C)$ space. The total space complexity is hence

$$O(n_U n_C \kappa_C \kappa_{cell,U} + n_C \kappa_C) = O(n_C \kappa_C (n_U \kappa_{cell,U} + 1)) = O(n_C n_U \kappa_C \kappa_{cell,U}). \tag{4.55}$$

which is the same as for the binomial model.

4.4.6 Signal Prior

Our model allows to extract robust results because we are able to model uncertainty in the CPDs. One standard approach is to assume a prior distribution on the CPDs. For this, we assume that each record has a small probability γ_{phen} to be a positive phenomenon observation $\tilde{P}(q^1|r) = P(q^1|r) + \gamma_{phen}$. Also, we we assume that each cell has a small probability γ_{ctx} to contain the phenomenon $\tilde{P}(c|r) = P(c|r) + \gamma_{ctx}$. Both CPDs can easily be normalized by

$$P(c|r) = \frac{\tilde{P}(c|r)}{\sum_c' \tilde{P}(c'|r)} \tag{4.56}$$

Algorithm 4.1 Algorithm to compute $W(c, q)$.

Input: Parameter ϵ, records R, users U, discrete context space C, with each record $r \in R$ having the CPDs $P(C|r)$, $P(Q|r)$, and $P(U|r)$

Assumptions:

$$|\{c \in C \land P(c|r) > 0\}| \simeq \delta_{rec,C}$$

$$|\{u \in U \land P(u|r) > 0\}| \simeq \delta_{rec,U}$$

$$|\{c \in C \land w(u, c, q) > 0\}| \simeq \delta_C$$

Routine:

```
(1) Create sparse matrices $w(u, c, q), m(u, c, q) \in \mathbb{R}^{n_U \times n_C \times 2}$

    for each $r \in R$:

        for each $u \in P(U|r) > 0, c \in P(C|r) > 0, q \in \{q^0, q^1\}$:
            if $P(u|r) > 0 \land P(c|r) > 0 \land P(q^1|r) > P(q^0|r)$:
                $m(u, c, q)$ += $1$
            $w(u, c, q)$ += $P(c|r)P(u|r)P(q|r)$

(2) Create sparse matrix $W(c, q) \in \mathbb{R}^{n_C \times 2}$

    for each $u \in U, q \in \{q^0, q^1\}$:

        for each $c \in w(u, \cdot, q) > 0$:
            $W(c, q)$ += $w(u, c, q)/m(u, c, q)^\epsilon$
```

Output: $W(c, q)$

and

$$P(q|r) = \frac{\tilde{P}(q|r)}{\sum_{q' \in Q} \tilde{P}(q'|r)}. \tag{4.57}$$

This allows to assume a prior signal distribution, which is valid even if no influences have been measured at all. Modeling the distribution in such a way will, however, induce a huge runtime overhead, since the CPDs will no longer be bounded (see Section 4.4.5). Therefore, we introduce a signal prior $\gamma = (\gamma_p, \gamma_w)$ directly on $z_q(c)$:

$$z_q(c; \gamma) = \frac{\beta_1 W(c, q^1, \epsilon) + \gamma_p \gamma_w}{\beta_0 W(c, q^0, \epsilon) + \beta_1 W(c, q^1, \epsilon) + \gamma_w}. \tag{4.58}$$

The prior adds a probability $\gamma_p \in [0, 1]$ to each cell (the prior probability of a positive observation). $\gamma_w \in \mathbb{R}_+$ is the strength of the prior affecting how fast the background distribution will be superimposed by the record influences. We will discuss appropriate settings for the prior in the experiments.

4.4.7 Spatial Dependency

Until now the spatial dependency has been considered by choosing an appropriate bandwidth for the spatio-temporal lattice. The signal can, however, further be smoothened to account for spatial interaction between the cells in the lattice. This resembles the smoothing scheme proposed in Section 2.3.2 to estimate signals on the basis of histograms that are more similar to continuous KDE estimates.

The scheme can be realized by using smoother geographic record influence functions $P(C|R)$. However, by this we will increase the sparsity factor $\kappa_{rec,C}$, thus resulting in a higher runtime. To allow for efficient modeling of the interaction between neighboring cells we use a kernel directly on the extracted signal, a well known technique in digital image processing [Gonzalez and Woods, 2007]. For this, we use a Gaussian kernel $K_{\sigma'}$ to convolute the signal

$$z_q(c;\gamma,\sigma') = z_q(c;\gamma) * K_{\sigma'}(c). \tag{4.59}$$

We define the kernel as a $\sigma' \times \sigma'$ matrix that covers a 2-dimensional symmetric Normal distribution with standard deviation 3.0 (this means the outer cells have almost zero probability mass).

4.5 Model Comparison

Given our model, different instances can be defined by choosing appropriate CPDs and parameters. In this section we show that using simple CPDs and an appropriate parametrization leads to the binomial model based on record or user counts. We will use θ to refer to a particular set of CPDs and parameters and denote $z_q(c;\theta)$ as a particular instance of our model.

Binomial Feature Count Model

As described in Section 4.3.3 the binomial model estimates the signal by

$$z_q(c) = \frac{\text{count}_r(q,c)}{\text{count}_r(c)}, \tag{4.60}$$

where we use q to denote the corresponding indicator feature.

Using our model we choose $P(c|r)$ to be 1 if the record falls into context c, $P(q|r)$ being 1 if the record is a positive phenomenon observation, and $P(u|r) = 1$ if the user u is associated with r. We set $\epsilon = 0$, $\beta_0 = 1$, $\beta_1 = 1$, and $\gamma_w = 0$.

Since then $w(u,c,q) = \text{count}_r(u,w,q)$ is just the number of records of user u measuring q at c and $m(u,c,q)$ vanishes because of $\epsilon = 0$, we have

$$z_q(c;\theta^{rec}_{binom}) = \frac{\sum_u w(u,c,q^1)}{\sum_q \sum_u w(u,c,q)} = \frac{\sum_u \text{count}_r(u,c,q)}{\sum_q \sum_u \text{count}_r(u,c,q)} = \frac{\text{count}_r(c,q)}{\text{count}_r(c)}. \tag{4.61}$$

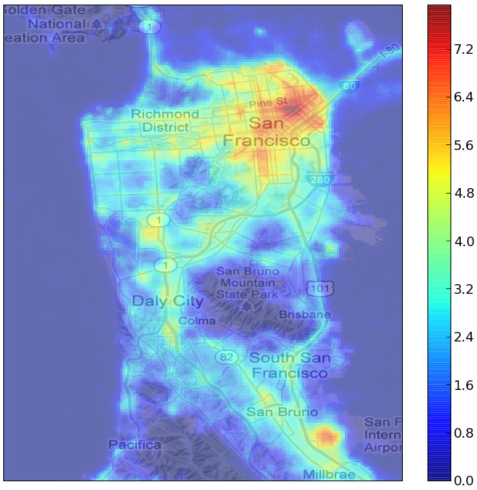


Figure 4.2: Logged counts of twitter user contributions in San Francisco.

Binomial User Count Model

The binomial model based on user counts is

$$z_q(c) = \frac{\text{count}_u(q, c)}{\text{count}_u(c)}. \tag{4.62}$$

By using the same CPDs and parameters as above but choosing $\epsilon = 1$, we get

$$z_q(c; \theta^{user}_{binom}) = \frac{\sum_u \frac{w(u,c,q^1)}{m(u,c,q^1)^\epsilon}}{\sum_q \sum_u \frac{w(u,c,q)}{m(u,c,q)^\epsilon}} = \frac{\text{count}_u(c, q)}{\text{count}_u(c)}, \tag{4.63}$$

since now $w(u, c, q) = m(u, c, q)$.

4.6 Experiments

We now present experiments conducted with different instances of our model. First, we discuss the impact of the parameters to optimize the signal and show how different kinds of geographic information can be mixed to obtain better results. Then, we compare the count, the binomial model, and our model by measuring the similarity of their extracted signals to a ground-truth phenomenon signals. For all models we use a keyword-based phenomenon record influence $P(Q|R)$ and an indicator-based user record influence $P(U|R)$ as proposed in Section 4.4.4.

4.6.1 Data and Setup

As input data we use tweets crawled from the public Twitter stream API from 12/03/28 to 13/04/11 using a bounding box covering San Francisco (-122.5552, 37.5938, -122.3478, 37.8410). The resulting data set consists of $10M$ records of which around $2.7M$ have a point coordinate and almost all records having an associated bounding box. The point coordinate reflects the GPS user location where the tweet was submitted and the bounding box represents a coarser tweet location (point of interest, district, city, state). To compare the count, the binomial, and our model we only use tweets having a point coordinate, resulting in $|R| = 2.7M$ records and $|U| = 142,631$ users. For the other experiments, we use all tweets.

We use a fine grained spatial lattice with a resolution of 180 x 151 cells. The spatial areas have an approximate extent of 153 meters. For our experiments, we did not use a temporal context space due to presentation considerations. Hence, we only discuss the spatial distribution of phenomena. The parameter discussion and the evaluation results can however assumed to be equally valid if a temporal domain is taken into account. The resulting geographic context space has size of $n_C = 27,180$.

The experiments have been performed using a single 2.0GHz CPU on a machine with 48GB RAM. The records have been completely loaded in memory (using about 7GB RAM) to generate the intermediate results w and m. Creating these matrices for a given set of CPDs needs around 3 minutes. Calculating the signal $P(Q|C)$ given the parameters β, γ, σ and ϵ needs time in the order of seconds. The inference is implemented in Python without particular optimization strategies except sparse matrix algorithms using the SciPy library [1].

4.6.2 Characteristic Keywords

In the following we extract geographic feature signals based on characteristic keywords. A characteristic keyword is assumed to act as a feature that represents a known phenomenon. For example, we expect the word 'bridge' to be representative for the bridges in San Francisco, or the word 'beach' to be representative for the beach areas. This is of course a strong simplification, since clearly not all tweets containing these words have been submitted with these phenomena in mind, nor have they been tweeted at those particular locations. However, we still assume that there is a higher chance that these words have a connection to the mentioned major phenomena in San Francisco. The aim of evaluating the signals using these keywords is hence, to find good signals even in the presence of a high noise level and uncertainty.

Other works, such as [Xu et al., 2012], use the same heuristics to extract phenomena, like car accidents in the United States. For real-world phenomenon recovery applications more sophisticated phenomenon record influences $P(Q|R)$ need to be developed or learned from human annotations.

In the following we assume the characteristic keywords to be noise indicators, and assume that models that perform well on such heuristics will also extract meaningful signals for other noisy and uncertain features.

[1] http://www.scipy.org

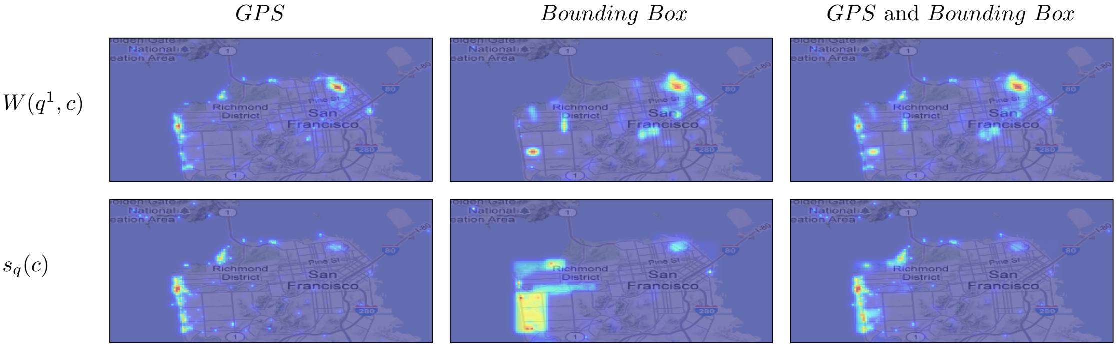


Figure 4.3: Aggregated influence and extracted signal using GPS, bounding box, and combined information.

4.6.3 Merging Spatio-temporal Information

First, we demonstrate the flexibility of our model to combine different kinds of spatio-temporal information. For this, we consider records having a GPS coordinate r_{gps} and/or a bounding box r_{bbox} and mix the spatial information as follows:

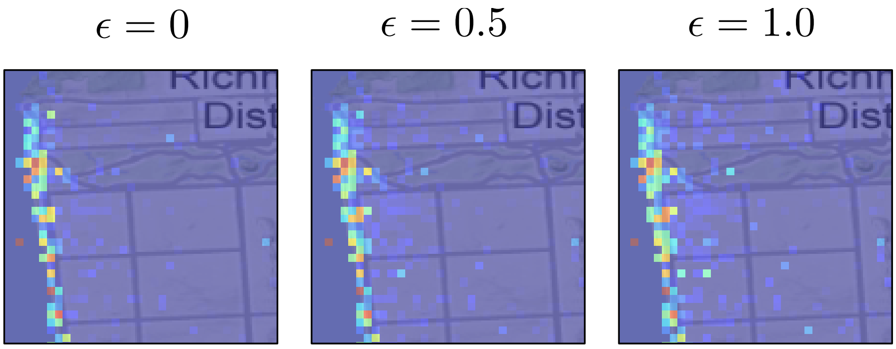
$P(c|r_{gps}) = 1$ if the GPS coordinate falls into context cell c, and 0 otherwise. $P(c|r_{bbox})$ follows a 2D Gaussian over the bounding box area such as shown in (2.11). We define a CPD that uses either the GPS, the bounding box, or both kinds of information to compute the geographic record influence

$$P(c|r;\pi) = P(c|r_{gps}, r_{bbox};\pi) = \begin{cases} P(c|r_{gps}) & \text{only GPS} \\ P(c|r_{bbox}) & \text{only bbox} \\ \pi P(c|r_{gps}) + (1-\pi)P(c|r_{bbox}) & \text{both} \end{cases} \tag{4.64}$$

where the mixing coefficient π controls the weight of the GPS and the bounding box information. Figure 4.3 shows the results using $\pi = 0.5$. The top row shows the aggregated influence $W(q^1, c)$ using either only the GPS, the bounding box, or the combined CPD. One can clearly see the differences of the GPS (left), bounding box (center), and mixed (right) aggregated influences. The bottom row shows the corresponding signals extracted using a $P(Q|R)$ CPD with indicator string 'beach' and parameters $\gamma_w = 0$, $\sigma = 3.0$, $\epsilon = 1$, $\beta_0 = 1, \beta_1 = 1$. One can see a stronger signal at the beach areas when using the mixed context influence while preserving the high resolution information from the GPS coordinates. Hence, signal extraction can clearly benefit from using different spatio-temporal information sources, which can be easily achieved by defining an appropriate mixing CPD.

4.6.4 User Redundancy

We now discuss the influence of the user redundancy parameter ϵ using the Twitter GPS records. For this, we activate the constraint variable by inserting evidence m^1 in the Bayesian network. Figure 4.4 shows the resulting 'beach' signal using $\gamma_w = 0$,



$\epsilon = 0$ $\epsilon = 0.5$ $\epsilon = 1.0$

Figure 4.4: Influence of varying redundancy penalization.

$\beta_0 = 1, \beta_1 = 1$ for varying ϵ. There is a slight increase in the signal strength and emergence of some noise cells when increasing ϵ. The increasing signal strength comes from a lower aggregate intensity $W(q^0, c)$, compared to $W(q^1, c)$. This is intuitive, since a user will likely contribute other records except those for phenomenon q. However, setting $\epsilon = 1.0$ leads to the emergence of noise at cells where almost only user u contributed records. This shows that values between 0 and 1 can be beneficial to increase the signal strength while suppressing the influence of noisy cells.

Overall, for our data set of Twitter records the influence of ϵ is not strong. Especially, the signal prior γ will superimpose its influence. However, we expect the influence of ϵ being much larger in data sets with a huge amount of duplicate records, e.g., in photo collections.

4.6.5 Model Parametrization

We discuss the influence of the signal prior (γ_p, γ_p), the phenomenon confidence (β_0, β_1) and spatio-temporal smoothing σ jointly, as they show a strong interaction.

Generally, (γ_p, γ_w) limits the influence of cells where only a small amount of positive phenomenon influence has been observed, and hence reduces the aggregated influence $W(q^1, c)$. A heuristic setting of γ_p is the probability that positive observations about q occur in the data set (e.g., the number of records containing the indicator feature). Using $P(Q|R)$ with the indicator string 'bridge' results in a fraction of 0.0012 tweets containing the string. Using this as γ_p means the assumed probability to see a tweet containing 'bridge' over the whole context space is expected to be that fraction. Now, having a cell with only one record containing the word 'bridge' will result in $P(q^1|c) = 1$ when using the binomial model. Using γ_p and setting the pseudo count γ_w to a reasonable number will lead to a large decrease of the probability at that context cell.

In Figure 4.5 the influence of different γ_w values is shown. A huge reduction in signal intensity at cells around the bridge can be observed when increasing γ_w. For high values, almost all cells not overlapping with the Bay Bridge and the Golden Gate Bridge have been decreased to small values.

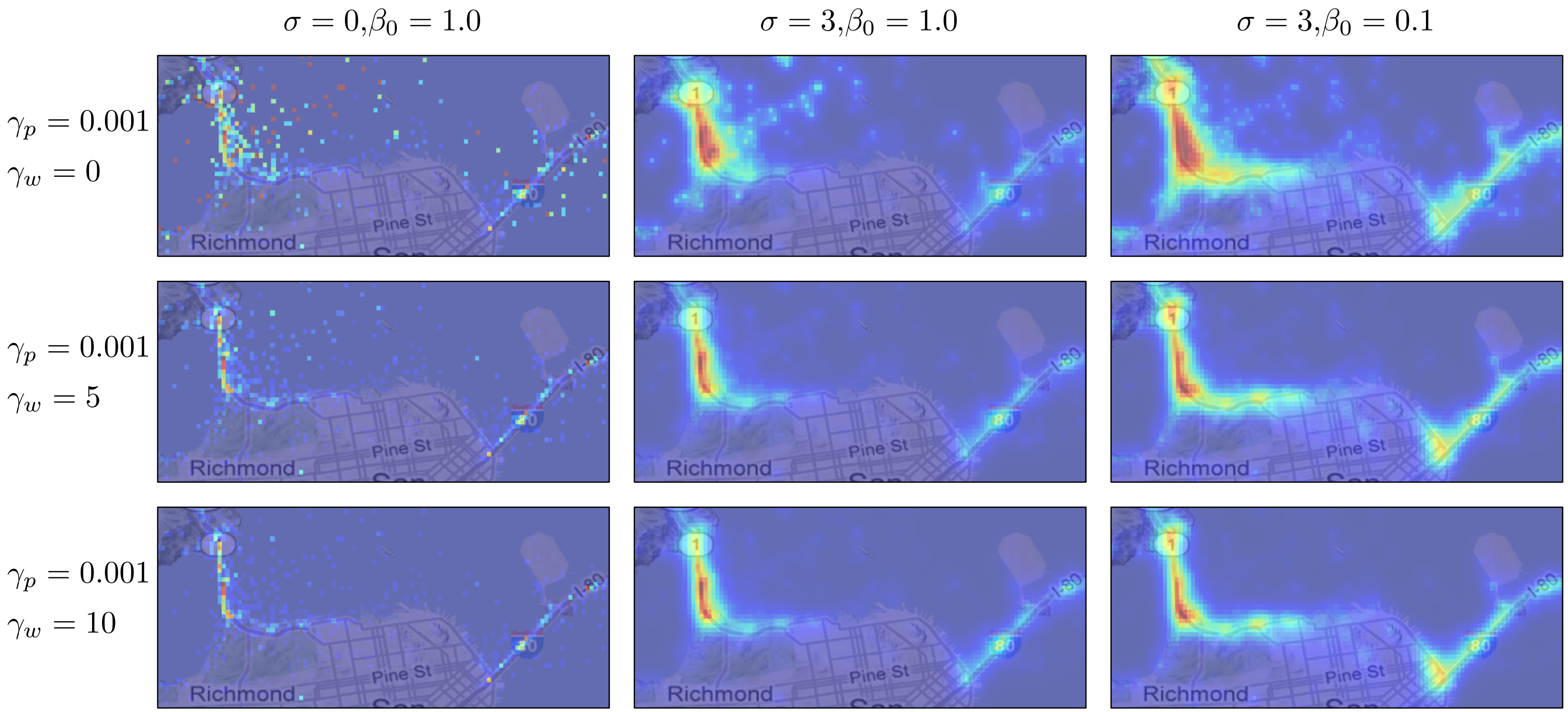


Figure 4.5: Parameter dependency between signal prior γ, phenomenon confidence β_0, and spatial smoothing σ.

The effect of additional spatio-temporal smoothing is shown in the center column. When applying a kernel with $\sigma = 3$ the noisy cells stay present when no γ prior is used (top center). Hence, using a spatio-temporal kernel alone will not give adequate results. However, the smoothed signal with γ prior shows a smooth distribution with most of the signal mass distributed over the bridges.

As a side effect of using a γ prior, the overall signal looses intensity also in those cells having a reasonable number of records. In Figure 4.5 one can see that the Bay bridge (right) becomes a small line when applying a strong γ prior. To increase the signal at cells having a high positive influence $W(q^1, c)$, the confidence parameter β_0 can be used. By looking at (4.36) we see that a reduction of β_0 leads to a decreasing influence of negative observations $W(q^0, c)$. Given a γ prior, then only the total positive influence and not the fraction related to the negative influences becomes significant. In Figure 4.5 we see that decreasing β_0 will increase the signal at those locations having a larger number of positive observations. Combining this with a strong γ_w parameter results in strong signals around the phenomenon cells while noise cells stay removed.

We use the described parameter dependency to derive a heuristic optimization strategy for our model as follows:

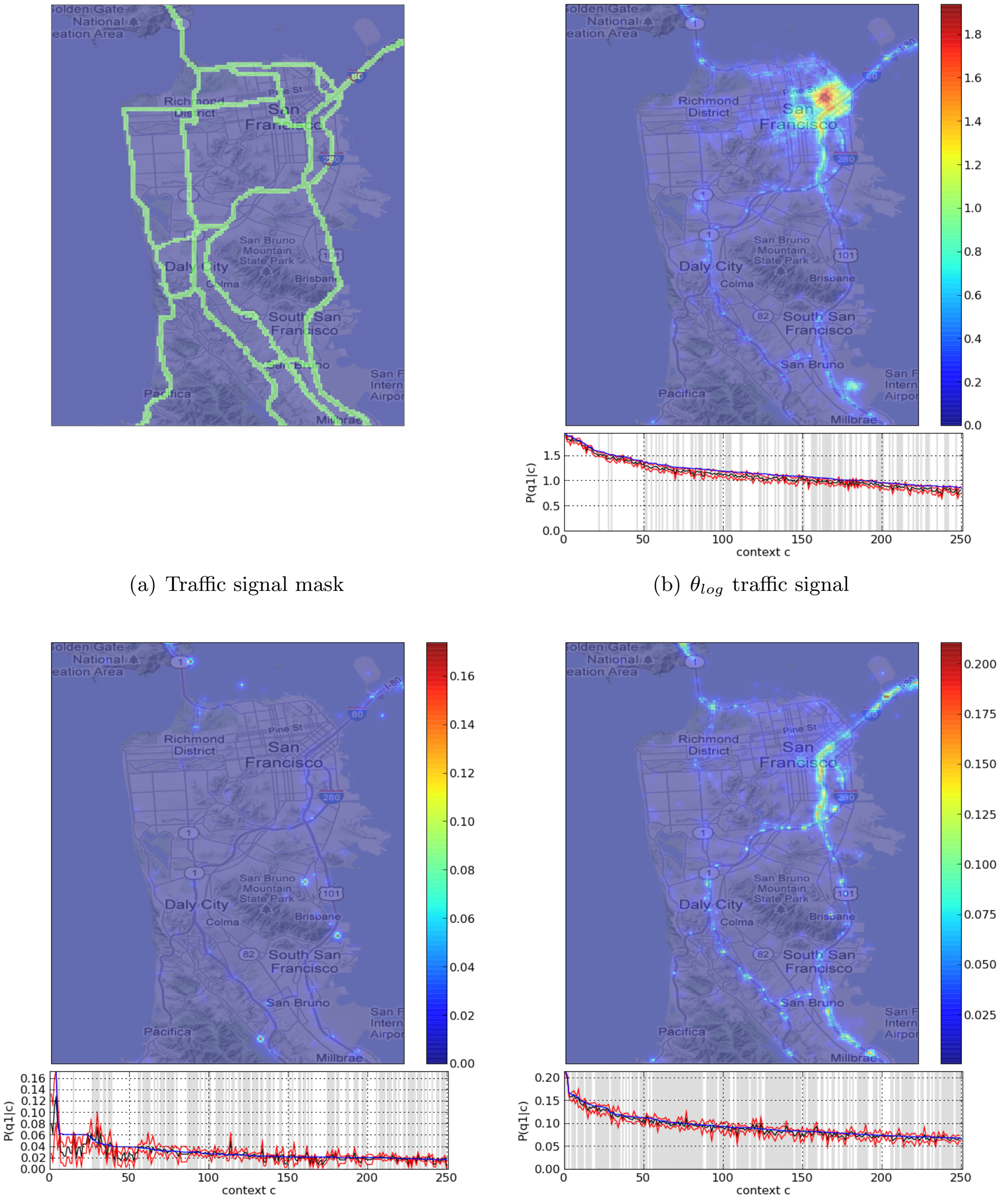
Heuristic Optimization. Given a set of CPDs, we start with $\beta_0 = 1$, $\beta_1 = 1$, $\gamma_w = 0$ and γ_p being set in the order of the overall probability that the phenomenon will occur as an observation in the dataset. Then, we increase γ_w until the signal will not change heavily (we observe a parameter of $\gamma_w = 10$ to give good results). Then, spatio-temporal smoothing is applied such that the signal becomes smooth but small-scale characteristics are still present. Then, the phenomenon confidence β_0 is decreased until the characteristics of the aggregated influence $W(q^1, c)$ will become present.

4.6.6 Ground-Truth Evaluation

We now evaluate the count, binomial, and Bayesian network model against ground-truth signals to judge about their quality and robustness. Our aim is to evaluate the intrinsic ability of the models to describe heterogeneous and noisy measurements derived from user-generated data. For this, we use a fixed spatio-temporal smoothing factor $\sigma = 3.0$ for all models and just use user counts by setting $\epsilon = 1$. To be able to compare the models we use indicator $P(Q|R)$ CPDs using the keywords 'beach' and 'traffic' and only use the $2.7M$ tweets having a GPS coordinate.

The two ground-truth signals are depicted in Figure 4.6 and 4.7. The 'beach' signal $z_{beach}(c)$ has a value of 1 at all locations in San Francisco where major beaches can be found. The 'traffic' signal $z_{traffic}(c)$ has a value of 1 along the highways and major roads in San Francisco. The binary signals represent presence or absence of the phenomenon.

We compare three different models. θ_{log} is a count model where the signal just consists of the logged user counts of the positive observations. To make it a probabilistic signal $z_q(c; \theta_{log}) = P(q|c)$ we normalize it to the unit interval. θ_{binom} is the binomial user count model. This model reflects the intuitive approach to extract a signal from binary



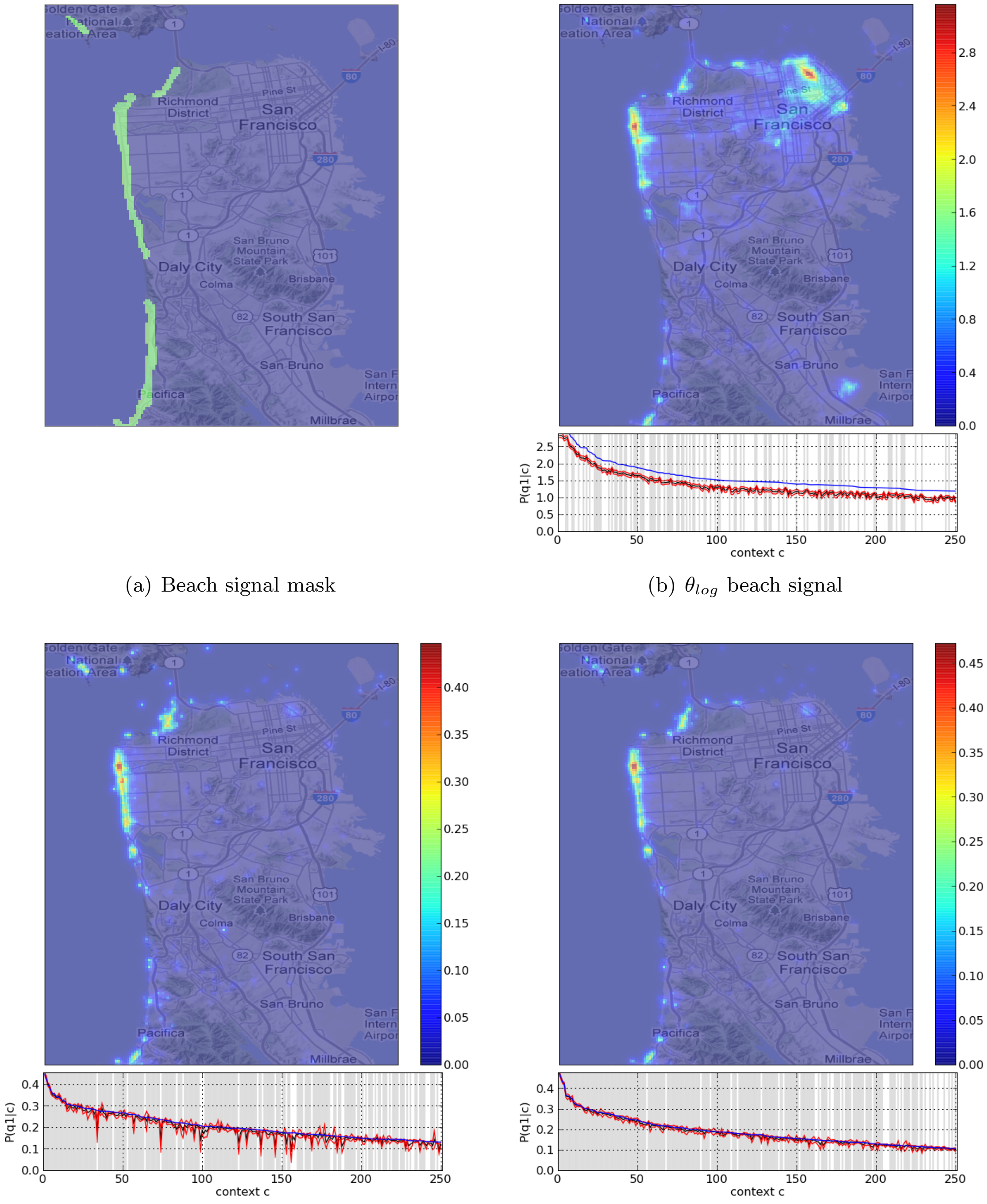
(a) Traffic signal mask

(b) θ_{log} traffic signal

(c) θ_{binom} traffic signal

(d) θ_{opt} traffic signal

Figure 4.6: Signal extraction results for the 'traffic' signal using different models.



(a) Beach signal mask

(b) θ_{log} beach signal

(c) θ_{binom} beach signal

(d) θ_{opt} beach signal

Figure 4.7: Signal extraction results for the 'beach' signal using different models.

Traffic (2303 cells)					
Model	Y_{10}	Y_{100}	Y_{1000}	Y_{tot}	cossim
θ_{log}	0.000	0.270	0.443	0.415	0.398
θ_{binom}	0.100	0.400	0.459	0.411	0.312
θ_{opt}	**0.700**	**0.830**	**0.615**	**0.478**	**0.527**

Beach (573 cells)					
Model	Y_{10}	Y_{100}	Y_{250}	Y_{tot}	cossim
θ_{log}	0.400	0.500	0.388	0.323	0.388
θ_{binom}	**1.000**	0.910	0.808	0.574	0.651
θ_{opt}	**1.000**	**0.940**	**0.828**	**0.595**	**0.660**

Table 4.2: Overlap $Y_k(q, \theta)$ and cosine similarity of extracted signals using different model instances.

positive and negative phenomenon observations. The model corresponds to the approach proposed in [Xu et al., 2012]. θ_{opt} is the model tuned by the heuristic optimization strategy proposed in Section 4.6.5.

Figure 4.6 shows the resulting signals for the 'traffic' phenomenon. The map shows the signal intensity distribution, and the underlying graph shows the top 250 signal intensities in decreasing order. The graph shows the signal in blue color. Furthermore, the mean (black) and the 95% confidence interval of the signal (red) calculated on the basis of a 10-fold bootstrapping. The mean and the confidence interval allow to judge about the robustness of the signal in terms of changing the input data. The graph is colored with a gray background if the corresponding signal cell matches the signal mask. Hence, a graph with a pure gray background means that the top 250 signal values fully explain the corresponding signal mask.

To quantitatively judge about how precisely a signal describes the phenomenon the amount of overlap of the top-k signal values with the phenomenon mask is used. Let $C_{\theta,k}$ be the k cells with the highest signal value $z_q(c; \theta)$. The overlap is computed by

$$Y(q, \theta) = \frac{\sum_{c \in C_{\theta,k}} \mathbf{1}\{z_q(c) > 0 \land z_q(c; \theta) > 0\}}{k}. \tag{4.65}$$

We use this measure instead of standard precision and recall because the signal might not have an absolute 0 value, and hence the signal would describe the mask completely. However, what we are really interested in is how well the best signal values describe our true signal. In Table 4.2 we show the resulting precision values for varying k (depending on the size of the signal mask) and for the complete signal mask (Y_{tot}). A value of 1 for Y_{tot} means that the top signal values fully cover the signal mask. To measure the similarity between the extracted signal and the true signal in terms of their distribution similarity, we use the cosine similarity between the model signal $z_q(c; \theta)$, normalized from 0 to 1, and the signal mask $z_q(c)$.

For the ’traffic’ signal, we see that the θ_{log} model is clearly dominated by the total population counts. The major roads are visible, however, the top signal values are located in the downtown area and the airport (apart from the highways). The signal is robust as it is not dominated by cells having a small number of observations, as can be seen in the graph.

The ’traffic’ signal of the binomial model θ_{binom} shows a slight increase in the overlap with the signal mask. However, the signal is dominated by some noisy peaks apart from the highways. In Table 4.2 one can see that the overall precision Y_{tot} for the binomial model is not higher than for the log model. Also, the signal graph shows a heavy variation of the mean signal and large confidence bounds, indicating that the signal is not robust to changing data. One can also see that especially those signal cells that have a low sample mean are those cells not overlapping with the mask.

The optimized model shows a much better signal of the traffic mask in terms of quality and robustness. By using our heuristic optimization strategy, we end up with $\gamma_p = 0.002$, $\gamma_w = 5$, and $\beta_0 = 0.01$. As can be seen in the graph and by the precision values, the top-k signal values overlap with the traffic mask nicely. Moreover, the signal is robust in terms of changing the input data. Table 4.2 shows that the tuned model outperforms the other models for each measure.

The ’beach’ signal results are similar, with the logged count model being dominated by the background population. The binomial model shows better results. However, by looking at the signal graph, one can see that those cells not overlapping the beach mask have a low sample mean, and hence being non-robust estimates. By tuning the model we end with $\gamma_p = 0.01$, $\gamma_w = 4$, $\beta_0 = 0.1$. The optimized model looks similar to the binomial model, however, with the noisy cells removed. In Table 4.2 the precision measures and the cosine similarity for the optimized model is overall better than for the binomial model. However, the differences are not that intense than for the ’traffic’ signal.

Overall, the ’traffic’ signal is much harder to extract since there are fewer positive observations, and the signal is spread in regions with high and low background population. In these cases, the optimized Bayesian network model shows much better results than the binomial model. For signals with a larger number of observations that are not overlapping regions with high background population, the quality of the binomial model improved. However, even then we can increase the quality of the signal by using our parametrized model.

4.7 Summary

In this chapter, we introduced a novel approach to extract geographic feature signals from different types of user-generated data sources. Since the extraction of geographic signals provides the basis to represent the qualitative and geographic information of user-generated data in a geographic feature matrix, this technique is fundamental to our framework. The main advantages of our proposed approach is its robustness to different kinds of noise, such as uncertainty of the feature semantics, spatio-temporal noise, and existing background distributions. Moreover, the approach allows to exploit a variety of qualitative and geographic information in the records, resulting in stronger and more meaningful signals.

We achieve this robustness and flexibility by addressing the singal extraction problem by an appropriately structured Bayesian network model. This allows to represent qualitative, geographic, and user information in the records independently from each other using robust and expressive conditional probability distributions. Also, by this probabilistic formulation we are able to use prior distributions to tackle the problem of sparsity in the spatio-temporal distributions.

By extracting signals from large real-world social media data sets, we showed that our approach extracts more meaningful signals for well-known geographic entities than standard techniques used in related work. The signals are not only better distributed along the entity locations, they are also more robust to variations in the input data, which proves the robustness of our approach.

Chapter 5

Geographic Feature Comparison: Categorization and Selection by Interaction Characteristics

5.1 Introduction

By using unstructured data sources to explore and extract geographic phenomena we almost always need to cope with a huge number of geographic features and their corresponding geographic feature signals. For example, when using geo-referenced text messages (such as tweets) to extract and explore geographic phenomena we are faced with a large set of candidate features, such as the set of all frequent terms or hashtags. Consequently, an important pre-processing step involves the selection of promising features among the thousands of feature candidates. In the previous chapter we proposed methods to extract robust spatio-temporal feature signals from the raw data. Now, we focus on techniques to select and categorize the extracted geographic features by characteristics of their spatio-temporal signals. Such selection and categorization tasks are important to explore the semantics of large sets of feature candidates, to build concise geographic feature spaces, and to compare data sets to each other.

We use the term *geographic feature comparison* to refer to the general task of comparing two geographic feature signals against each other. This might be a comparison between two feature candidates (to asses if they are similar) or between a candidate feature and a distribution with well-known characteristics (e.g., a uniform distribution or a parametric unimodal distribution). Feature comparison can then be used to (1) select a number of features that exhibit certain characteristics (e.g., being uniformly distributed, dominant in a certain area, or similar to a reference feature or a background distribution) and (2) to categorize features by clustering them into groups with similar distribution characteristics.

Pairwise comparisons of features are realized by a distance or similarity function and by an appropriate representation of the feature signal. A frequently used approach compares features on the basis of their raw vectorized signals (later called their intensity representation) by using the Euclidean distance or the cosine similarity [Zhang et al., 2012b]. Such a comparison determines those features as similar that occur at the same points in space and time. Thus, those features are considered similar that describe the same phenomenon (e.g., a particular city, region, or event). Such direct comparisons have been used to extract and summarize events or places by clustering, such as shown in [Chen and Roy, 2009; Hays and Efros, 2008; Zhang et al., 2012b].

In this chapter we introduce novel methods that allow to compare features on the basis of their *spatio-temporal type* instead of the concrete phenomenon they represent. As defined in Section 3.5.1, the spatio-temporal distribution type (or spatio-temporal type for short) denotes the kind (type) of geographic phenomenon a feature represents. Such types include landmarks and events at certain scale levels, place and regions types (phenomena occurring in several disjoint areas), recurring events, or trajectories. For this, we introduce the idea to represent geographic feature signals by *interaction characteristics*. The term interaction stems from spatial point pattern analysis, where it describes how points interact with each other. For example, points might tend to avoid each other, leading to a regular structure, or they might tend to occur in close proximity, leading to clustered structure. Using statistics of the point interaction allows to describe how a signal is distributed, e.g., to distinguish between signals that have a single peak or several peaks.

From an application point of view, we will show that a comparison on the basis of the spatio-temporal type allows to categorize features and to summarize data sources concisely. This allows for feature selection and filtering decisions to build more concise geographic feature matrices for subsequent data mining task, for example, a geographic feature matrix with highly predicative landmark features. In the following we introduce a set of feature representations that exhibit the spatio-temporal type of the signal and evaluate their qualitative performance to distinguish between different signal types.

The remainder of this chapter is structured as follows. First, in Section 5.2 we detail the problem statement and the contributions. Then, in Section 5.3 we explain the difference between intensity and interaction characteristics of spatio-temporal signals, which builds the theoretical foundation to define feature representations exhibiting spatio-temporal type semantics. In Section 5.4, we define three different feature representations and evaluate their performance to distinguish between signals of different spatio-temporal types. Finally, in Section 5.5 we present different data mining tasks that make use of the novel comparison method using real-world data sources.

5.2 Problem Statement and Contributions

Within our framework, the usual input to geographic feature comparison are geographic feature signals. A *geographic feature signal* is a positive-valued discrete or continuous spatio-temporal variable (see Section 2.2.1). Here, however, we also describe features by sets of points. The set of points corresponding to a feature is called its *point pattern*. The points represent observations of the feature in space and time (such as by GPS coordinates and timestamps). A geographic feature signal can be extracted from such point patterns by using a geographic feature extraction method proposed in the previous chapter. Defining the methods also on point patterns establishes the link to the underlying foundations of interaction characteristics and illustrates the general applicability of the techniques.

The aim of *geographic feature comparison* is to compare geographic feature signals against each other or against a well-known type of distribution. We think of a comparison as a distance or similarity function that indicates if two feature signals are similar or dissimilar to each other. A pairwise comparison of the raw signals allows to realize the following tasks:

- *Feature Selection and Filtering*: By comparing the features against a well known type of distribution (e.g. a uniform distribution or a reference feature) we can select those features that correspond to known characteristics. This is important from an exploratory point of view to asses how many of the candidate features describe particular patterns, such as a covered area or a reference phenomenon. These comparisons can be realized by computing the Euclidean or cosine distance of the feature signals directly, since two signals will be similar if they occur at the same locations in space and time.

- *Feature Categorization*: By clustering features into groups in which their signals are pairwise similar, we can summarize and compare data sets on the basis of their spatial semantics. By using a distance measure on the signals directly, the clusters will be comprised of features that describe the same phenomenon. Such a clustering approach has for example been used to find and describe places and events [Chen and Roy, 2009; Hays and Efros, 2008; Zhang et al., 2012b].

In this chapter we extend the feature comparison idea and focus on the problem of comparing features by their *spatio-temporal distribution type*. The spatio-temporal distribution type of a signal or a point pattern (in the following only called *spatio-temporal type*) is a category of the underlying geographic phenomenon. Such basic categories comprise places, events, trends, or trajectory, among others. For a detailed definition of different kind of spatio-temporal distribution types see Section 3.5.1. In the following we restrict ourselves to spatial data, i.e., spatial point patterns and spatial signals. As in the previous chapter, the presented techniques are equally valid for spatio-temporal signals if a joint spatio-temporal distance function exist, such as shown in (2.2).

We now give a more specific categorization of spatial types, which we use to describe particular patterns in the remainder of this chapter. Point patterns and signals exposing the following types are shown in Figure 5.1.

- *(Small/large) Landmark Feature*: A landmark feature describes a unique place on Earth, like a city or a place name. In this case, the point pattern shows a single cluster or the signal shows a single peak. Different kinds of landmark features can be distinguished by their size. A city-level landmark occurs in a small region while a country-level landmark occurs in a larger region. Therefore we can distinguish between landmark features on different scales, here generally referred to as small and large landmark features.

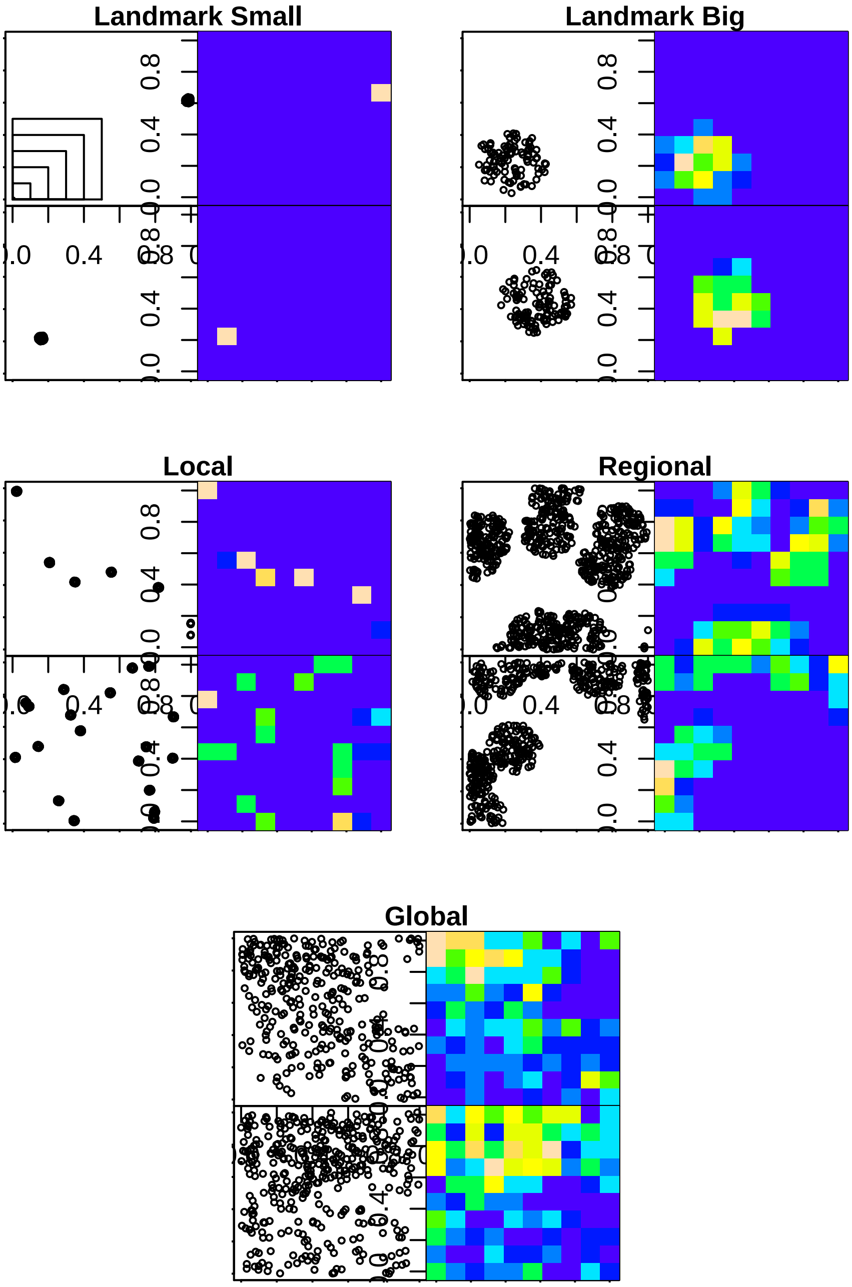


Figure 5.1: Five types of spatial feature distributions in 2D space. Each of the five blocks shows two different instances of a spatial feature type in the top and the bottom row, respectively. In each row a point pattern is shown in the left column and the corresponding feature signal (intensity distribution) in the right column.

- *Local Feature*: A local feature occurs within a small area but at several places on Earth. Hence, the point pattern is clustered at several locations or the signal shows several peaks. For example, a feature occurring predominantly in cities will show a clustered distribution, with each cluster covering an area at city-level scale. Typical examples of such geographic features are place categories, or moods and actions of people.
- *Regional Feature*: A regional feature occurs in a wider area than a local feature and might also occur at several places on Earth. Examples of such features are attributes describing a physical phenomenon ('forest', 'mountains', etc.) or regional behavior (language, actions, etc.). The pattern has several large clusters or the signal shows several wide peaks.
- *Global Feature*: A global feature is distributed in the whole area of interest and does not show strong location-specific variation. For example, the spatial distribution of records containing a stop-word (a word occurring in almost all of the records) will not describe a specific geographic phenomenon, but only the distribution of the observations in general. This does not mean that the points and the signal are distributed uniformly in the area. Such global features will show a number of small and/or large clusters according to the background distribution of the data, if the signal is not normalized accordingly.

We will introduce methods that allow to asses how many of the candidate features are of a particular spatial type, for example features exposing a landmark type or a global type. Extracting features of a particular type is important to build geographic feature spaces tailored for specific data mining tasks. E.g., a geographic feature space with landmark features will be highly predictive and descriptive for locations. Those landmark features will be promising candidates to extract place labels or to predict record locations. On the other side, local and regional features are promising candidates to segment space into different semantic categories (e.g., into rural and urban areas, or into areas with affinity and non-affinity for particular behavior).

By being able to cluster the features into groups that adhere to the same spatio-temporal type, data sets can also be summarized and compared to each other easily. From an exploratory point of view this is an important task to understand the geographic semantics of a data set, e.g., what kind of phenomena are described by the features. Moreover, this allows to compare data sets independently of their area of interest. This will allow to merge features of the same type from different data sets, to build a global data sets for a particular type of phenomenon. We summarize the general problem statement as follows:

> Given a set of spatio-temporal signals (or point patterns) of features, how can we compare the features, such that those features exhibiting the same spatio-temporal distribution type are similar to each other?

We propose a vector-based representation of the feature signals based on their *interaction characteristics*. The term interaction derives from spatial point pattern analysis, where it describes how points interact with each other. For example, points might tend to avoid each other, leading to a regular structure, or they might tend to occur in close proximity, leading to clustered structure. In the following we introduce the theoretical underpinnings of point interaction to derive different vector-based feature representations that can be used in distance and similarity functions. Then, we show that a representation based on the K-function together with the Canberra distance [Landauer et al., 1967] has the best qualitative performance in distinguishing between different types of signals. Finally, we show how this setting can be used to categorize features in real-world data sets and to summarize data sources based on their covered geographic feature types.

5.3 Distribution Characteristics

In the following we detail the theoretical foundations of different characteristics of geographic feature signals. We make use of the terminology and methods of spatial point patterns to develop a feature representation that allows to compare signals on the basis of their spatio-temporal type.

5.3.1 Point Pattern Analysis

In spatial statistics the field dealing with patterns and descriptions of spatial point sets is called *point pattern analysis* [Baddeley, 2010]. Spatial point pattern analysis extends to the spatio-temporal domain D_{ST} if an appropriate spatio-temporal distance between the points is defined, such as in (2.2). A spatial point pattern is a set of points in the area of interest D_S,

$$S = \{s_1, \ldots, s_n\} \subseteq D_S. \tag{5.1}$$

We assume a given distance function d between points

$$d(s_j, s_j) \in \mathbb{R}_+, s_i, s_j \in D_S. \tag{5.2}$$

A pattern S is assumed to be a sample of a random point process Y. We can think about a point process as a model that allows to generate patterns that have the same characteristics.

The number of points of a point pattern S located inside a region $A \subset D_S$ is denoted $N(S \cap A)$. If the points are distributed randomly, then for any region A the expected number of points is proportional to its area $|A|$,

$$\mathbb{E}[N(S \cap A)] = \lambda\ |A|. \tag{5.3}$$

The constant λ is called the *intensity* of the point pattern with the same meaning as in the Poisson distribution shown in (4.25). For the complete area of interest D_S, the intensity is then

$$\lambda = \frac{n_S}{|D_S|}. \tag{5.4}$$

The intensity λ is one characteristic of a point process Y, namely the first moment of the point process. See [Baddeley, 2010; Ripley, 1981; van Lieshout, 2010] for a formal introduction to point pattern analysis.

5.3.2 Intensity Characteristics

A point process that has a random distribution is, however, a rare case. Such a point processes will almost only occur in carefully designed experiments. For real-world data, the intensity (the expected number of points) will likely vary in the area of interest D_S. This variation is modeled using an intensity function instead of a scalar intensity parameter. The function returns the intensity at every point in the area of interest

$$\lambda(s) \in \mathbb{R}_+, s \in D_S. \tag{5.5}$$

The intensity function is the continuous version of the intensity distribution over cells in a spatial lattice, as proposed in (2.2). The intensity signal is also identical to a geographic feature signal as defined in (3.52). This builds the connection between a point pattern and a geographic feature signal. The intensity function can be estimated by density estimation. Known approaches are histograms and Kernel Density Estimation (KDE), as introduced in Section 2.3.2.

Here, we focus on the usage of the intensity function as a characteristics of a point pattern. If we compare two point patterns directly on the basis of their intensity function, they will be similar if the points occur at the same locations. Since, in our context, the intensity function describes where a feature occurs on Earth, they will be similar if they represent the same phenomenon. We say that the intensity function represents the *intensity characteristics* of a point pattern or of a signal.

Determining the similarity between features is achieved by using distance or similarity functions on the intensity representations, i.e., vectors representing the intensity function (5.5). Frequently used distance functions are the Euclidean distance, the Cosine similarity, and the mean squared error (MSE). We will not detail the comparison of features based on their intensity characteristics in more detail, since this idea is successfully used in a variety of related works [Chen and Roy, 2009; Hays and Efros, 2008; Zhang et al., 2012b] and is considered a standard approach to determine features that describe the same phenomena.

5.3.3 Interaction Characteristics

Different from the intensity function, the patterns can also be compared in terms of how the points interact. If the points occur randomly in the whole area, each point behaves independently from the others, and the pattern is called independent (or random). A randomly distributed pattern is a pattern generated from a Poisson process [van Lieshout, 2010]. If the points are likely to occur in close spatial proximity, the pattern is called *clustered*. If the points of a pattern tend to avoid each other, the pattern is called *regular*. Note that there is not one single type of clustered or regular interaction.

For example, points might tend to occur in several small clusters, or in a single large cluster.

Given a function that describes this interaction, we say that it represent the *interaction characteristics* of a point pattern. We extend this concept to arbitrary geographic feature signals in the next section.

Figure 5.1 show examples of point patterns with different interaction characteristics. The landmark patterns have a single cluster, while the local and regional patterns show several clusters. The global pattern has a varying intensity over the area of interest but is not clearly clustered.

A common method to describe the interaction between points are distance statistics. For this, the frequencies of distances between points are determined to describe a location distance distribution that is invariant to the particular point locations. A well-known distance statistic is the K-Function. It is based on the pairwise distances of the points in the pattern [Ripley, 1977] and is defined as

$$K(r) = \frac{1}{\lambda}\,\mathbb{E}[N(S \cap b(s,r)) - 1]. \tag{5.6}$$

Here, $b(s,r)$ is the disc around point s with radius r. The subtraction of 1 comes in because the center point itself is not counted. Hence, $\lambda \cdot K(r)$ is the expected number of points within a distance r for any point in the pattern.

The K-function describes a cumulative distribution function. If the radius r is larger than the complete area of interest, then $K(r) = 1$, since the expected number of points lying around any point in the pattern includes all other points.

If we compare two point patterns on the basis of their K-function, they will be similar if the points occur with the same expected distances from each another. Given that the points represent observations of a phenomenon, we expect that the patterns or signals will be similar if they describe the same type of geographic phenomenon. We state this as a hypothesis and evaluate how well the K-function describes the types of geographic features in the next sections.

5.4 Interaction-based Feature Comparison

In this section we present three approaches to describe a set of points or geographic feature signals by their interaction characteristics. The result of each approach is a numerical vector or quantity that describes how the points or the signal are distributed, called its *interaction characteristics representation.*

We present an approach based on the Gaussian covariance matrix, an approach based on the cumulative distribution of the signal's intensity values, and an approach using the K-function. Then, we will evaluate the qualitative performance of these approaches in combination with different distance functions.

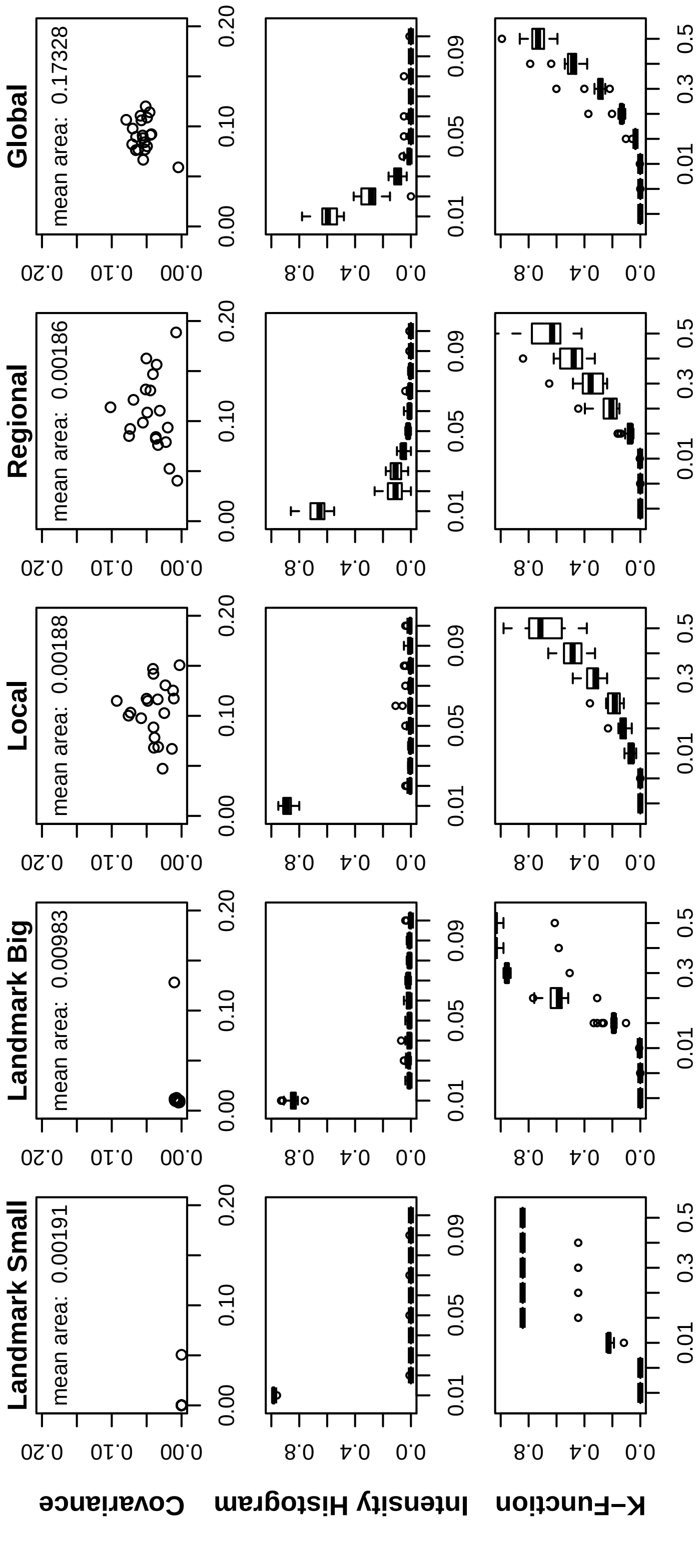


Figure 5.2: The three proposed interaction characteristics for the five types of point patterns (as shown in Figure 5.1).

5.4.1 Gaussian Covariance

The first approach describes the points of the signal by a single Gaussian distribution. There, we can use the parameters of the Gaussian to describe its spatio-temporal type.

Given a point pattern, we need to fit a Gaussian distribution on the set of spatial points. We already showed how to fit a k-dimensional multivariate Gaussian on a set of points in Section 2.3.1.

We now discuss the fitting procedure for discrete spatio-temporal signals. Given a discrete geographic feature signal $z_f(c)$. We can use the same estimation approach as for a point pattern by interpreting the signal as a weighted set of points. For this, we assume that the center points of the lattice on which the discrete signal is defined are

$$C = \{c_1^*, \ldots, c_m^*\}. \tag{5.7}$$

The weights are given by the normalized signal $\dot{z}_f(c)$. To estimate the mean of a Gaussian we use

$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} \dot{z}_f(c) \cdot c_i^*. \tag{5.8}$$

To estimate the covariance matrix we use

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{m} \sum_{i=1}^{m} \dot{z}_f(c_i)(c_i^* - \hat{\boldsymbol{\mu}})(c_i^* - \hat{\boldsymbol{\mu}})^\top. \tag{5.9}$$

Given the fitted parameters of the Gaussian, we describe the interaction characteristics by the 2×2 covariance matrix $\hat{\boldsymbol{\Sigma}}$, since it represents the shape of the distribution in 2-dimensional space, invariant to its location (mean).

Two kinds of representations are considered:

(1) We determine the Eigenvalues $e_f = (e_{f,1}, e_{f,2}) \in \mathbb{R}^2$ of the covariance matrix $\hat{\boldsymbol{\Sigma}}$. This is achieved by a singular value decomposition of the covariance matrix. This is a standard procedure to determine the Eigenvalues and has negligible (constant) runtime on a 2×2 matrix.

The first Eigenvalue $e_{f,1}$ describes the magnitude of the vector showing in the direction with the largest variance. The second Eigenvalue $e_{f,2}$ describes the magnitude of the vector with second-largest variance. Hence, the two quantities are a rotation-invariant description of the shape of the Gaussian.

(2) We extract the 95 percent confidence area of the Gaussian and compute its size, denoted $a_f = area_{0.95}(\Sigma_f)$. For this, we use the Eigenvalues e_f and compute the 95 percent confidence area by

$$a_f = 2\pi e_{f,1} e_{f,2}. \tag{5.10}$$

Figure 5.2 shows scatter plots between the Eigenvalues $e_{f,1}$ and $e_{f,2}$, and the area a_f for different pattern types. The scatter plots for each type are based on 20 patterns that are similar to the patterns shown in Figure 5.3. The data comes form a synthetic point pattern dataset, which is detailed in Section 5.4.5.

One can see that the magnitude of the Eigenvalues and the size of the area depends on the total spread of the distribution. A landmark feature will be characterized by a small area a_f and small eigenvalues e_f of the covariance matrix. A local, regional, and global feature will be characterized by larger areas and Eigenvalues, compared to the landmark distributions. This over-simplifying approach will not perform well in distinguishing local, regional, and global features. However, it will successfully identify landmarks and separate them from the others feature types.

5.4.2 Intensity Distribution

The second representation of interaction characteristics is based on the distribution defined over the signal values $z_f(c) = \lambda(c), c \in L$. Given a point pattern we assume that it is first transformed into a discrete spatio-temporal signal, as detailed in the previous chapter. The normalized signal values $\dot{z}_f(c)$ are used to extract a cumulative distribution function over the intensity domain

$$P(\dot{z}_f(c) < x) = \frac{1}{n_L} \sum_{i=1}^{n_L} \mathbf{1}\{\dot{z}_f(c) < x\}, x \in [0, 1]. \tag{5.11}$$

We represent this distribution by an equi-interval histogram evaluated over k bins

$$I_f = \left(\frac{1}{n_L} \sum_{c \in L} \mathbf{1} \left\{ \frac{i}{k} < \dot{z}_f(c) \leq \frac{i+1}{k} \right\} \right)_{i=0}^{k-1}. \tag{5.12}$$

This vector characterizes how often the signal values occur in the cells. In Figure 5.2 the intensity histogram of different signal types is shown in a boxplot. Each plot shows the histograms of 20 patterns with similar characteristics (landmark, local, regional, global). The histogram shows the intensity values on the x-axis and the probabilities on the y-axis. The landmark and even the local patterns show that most of the cells have a very low signal value. The regional and global patterns, however, have a higher number of cells with larger intensities, indicated by the increasing left tail. The plots show a significant difference between the vectors in the first three and the second last columns, indicating that the landmarks and local features will be clearly separated from the regional and global features.

5.4.3 K-Function

Finally, we employ the K-function to extract a representation of interaction characteristics. For this, we need an estimate of the K-Function $\hat{K}(r)$, which can be calculated for a given radius r by

$$\hat{K}(r) = \frac{1}{\lambda n_S} \sum_{s_i \in S} \sum_{s_j \in S, j \neq i} \mathbf{1}\{d(s_i - s_j) \leq r\}. \tag{5.13}$$

The normalizing constant $1/(\lambda n_S)$ can be dropped, since for a relative comparison of two patterns the absolute value of the function has no impact. Moreover, one can consider the induced error by counting the point itself as small. Therefore, the following function is used to describe the interaction characteristics of a pattern

$$K'(r) = \sum_{s_i \in S} \sum_{s_j \in S} \mathbf{1}\{d(s_i - s_j) \leq r\}. \tag{5.14}$$

The above estimate is based on a point pattern. We now introduce an alternative view on the K-function that allows to compute the interaction characteristics for arbitrary signals $z_f(s)$. For this, w.l.o.g., we assume that the signal $z_f(s)$ is defined on the basis of the intensity function $\lambda(s)$, and hence represents the average number of points in space.

Since the signal represents the average number of points at a location s, we can integrate over the disc $b(s, r)$ instead of counting the number of points within a given radius to obtain the covered number of points

$$J(s, r) = \sum_{s_j \in S} \mathbf{1}\{d(s - s_j) \leq r\} = \int_{b(s,r)} z_f(s_j) ds_j. \tag{5.15}$$

Similarly, instead of traversing over each point to compute the expected value, we can integrate over the area of interest and use the value $z_f(s)$ as the multiple of $J(r)$. Hence, $K'(r)$ is computed as

$$K'(r) = \int_{D_S} z_f(s_i) J(s_i, r) ds_i = \int_{D_S} z_f(s_i) \int_{b(s_i,r)} z_f(s_j) ds_j ds_i. \tag{5.16}$$

Given that $z_f(s)$ is a discrete spatial signal defined over the cells $L = \{c_1, \ldots, c_n\}$ and the distance between two cells is defined by the distance between their center points $d(c_i^*, c_j^*)$, $K'(r)$ is computed by

$$K'(r) = \sum_{c_i \in L} z_f(c_i) \sum_{c_j \in L, d(c_i^*, c_j^*) \leq r} z_f(c_j). \tag{5.17}$$

The K-function is then represented as a vector over k radii

$$K_f = \left(K'(r_i)\right)_{i=1}^{k}. \tag{5.18}$$

The boxplots for these vectors are shown in Figure 5.2. They shows significant differences between features having a landmark characteristic (first two columns) and the local, regional, and global distributed features. Also, the two landmark types show significant differences.

5.4.4 Performance Considerations

The number of geographic features signals in an unstructured geographic information source can be large. Hence, performance in the extraction of interaction characteristics is a crucial aspect. In the following we discuss the runtime complexity of the above approaches. We assume point patterns with n_R points, discrete geographic feature signals with n_C cells, and $n_C \ll n_R$. We assume that the discrete signal is sparse with a sparsity factor of $\kappa_C = \delta_C / n_C$, where δ_C denotes the number of non-zero cells in the lattice.

We discuss runtime complexities for point patterns and discrete spatial signals. In Section 4.3.2 and Section 4.4.5 the runtime complexities to transform discrete signals from a set of geographic observations (point patterns) have been discussed, ranging from $O(n_R)$ for a simple count-based approach to $O(n_U \delta_C + n_R \delta_{rec,U} \delta_{rec,C})$ to extract optimized signals for varying feature weights and user influences. To denote the runtime of the geographic feature extraction (GFE) method we use $O(\text{GFE})$.

Gaussian Covariance and Intensity Distribution

For a set of points, the Gaussian covariance matrix can be estimated in $O(n_R)$ time, and for a given discrete signal, the runtime decreases to $O(n_C \kappa_C)$ (see Section 5.4.1).

The intensity histogram approach needs to transform the points into a discrete signal. Then, the intensity histogram can be obtained in $O(n_C \kappa_C)$ time, resulting in a total of $O(\text{GFE} + n_C \kappa_C)$ time for a point pattern.

Both, the Gaussian covariance matrix and the intensity histogram approach are efficient and will allow to process a huge number of features easily.

K-Function

In order for the K_f vector approach to be efficient, one needs to look at the sparsity factors and use approximate solutions. In the following, we present the runtime complexities for the case that the input is provided as a point pattern, for the case that the input is a discrete signal defined on a regular lattice, and for the case that the point pattern is approximated by a signal.

The vector represents the function at k different radii. The runtime is, however, only affected by the largest chosen radius r_{max}, since the values for the smaller radii can be computed while traversing over the points or cells needed to process r_{max}. We denote the average number of points or cells within radius r_{max} as $\delta_{rad,R} = n_R \kappa_{rad,R}$ and $\delta_{rad,C} = n_C \kappa_{rad,C}$, respectively.

Point Pattern Input. First, we consider the runtime of point pattern input. Using a brute force approach we need to traverse over all points and process all other points to determine those instances within r. This will need $O(n_R^2)$ time. To determine the points within r_{max} more efficiently, a spatial index structure can be built in a pre-processing step in $O(n_R \log n_R)$ time. Then, accessing the $\delta_{rad,R}$ points around a given point will need in the order of $O(\log n_R)$ time (the matching points need however still to be processed).

Using the index structure the total runtime reduces to

$$\begin{aligned}
&O(n_R \log n_R + n_R(\log n_R + \delta_{rad,R})) \\
=&O(n_R \log n_R + n_R \log n_R + n_R \delta_{rad,R}) \\
=&O(n_R(\log n_R + \log n_R + \delta_{rad,R})) \\
=&O(n_R(\log n_R + \delta_{rad,R})).
\end{aligned} \tag{5.19}$$

For this technique to be more efficient than the brute force approach the average number of points $\delta_{rad,R}$ must be less than

$$\delta_{rad,R} < n_R - \log n_R, \tag{5.20}$$

which is a very reasonable assumption for appropriate choices of r_{max}.

Discrete Signal Input. Given a discrete spatio-temporal signal on a regular grid. The number of non-zero cells within radius r_{max} around a cell in the lattice is $\delta_{rad,C} = n_C \kappa_{rad,C}$.

Within a regular grid we can determine the indexes of a cell in $O(1)$ time. Given a cell c, the rectangle around the cell including the circle with radius r_{max} can hence be indexed directly. However, we still need to traverse the set of non-zero cells in that block of cells. Given that we have a sparse matrix representation, we can access the columns having non-zero cells directly and then traverse the non-zero rows (or vice versa). We assume that the cost of accessing the cells in the block that are outside of the circle with radius r_{max} is negligible, resulting in a runtime of $O(n_C \kappa_{rad,C})$ to access the surrounding cells.

We need to traverse over the sparse signal, which needs $O(n_C \kappa_C)$ time, and collect the values for each of those non-zero cells. This will need time of

$$\begin{aligned}
&O(n_C \kappa_C n_C \kappa_{rad,C}) \\
=&O(n_C^2 \kappa_C \kappa_{rad,C}).
\end{aligned} \tag{5.21}$$

Hence, the runtime is heavily affected by the factors κ_C and $\kappa_{rad,C}$. If $\kappa_C \leq 1/\sqrt{n_C}$ and $\kappa_{rad,C} \leq 1/\sqrt{n_C}$ (which are reasonable assumptions) the runtime will be less than linear in the order of cells in the lattice and is considered fast.

In this case the technique is also a reasonable choice to compute an approximated K_f vector for point pattern input, resulting in a runtime of $O(\text{GFE} + n_C^2 \kappa_C \kappa_{rad,C})$, which reduces to $O(n_R + n_C)$ if a fast count-based feature extraction approach is used and the above sparsity assumptions hold.

5.4.5 Evaluation

We evaluate the above interaction characteristics representations by computing the pairwise similarities of features from a synthetic data set of point patterns. Different distance functions are used to determine the similarity between the feature representations. Since we know the types of the patterns in advance, a high quality feature representation and distance function combination should result in an ideally segmented distance matrix.

Synthetic Dataset

The synthetic data set was generated by sampling from a Matern cluster processes using the R *spatstat* package[1] [Baddeley, 2005]. The Matern cluster process is a parametric point process to describe point patterns with different point interactions. The parameters are given by an average number of cluster centers, a radius around cluster centers, and an average number of points inside a cluster. Details are provided in [Isham, 2010]. We choose 5 parameter settings, representing the distribution types described in Section 5.2.

For each of these types, 20 patterns (instances) are generated, resulting in a total of 100 patterns. The discretized signals of these patterns represent a set of 100 geographic feature signals $z_{f_1}(c), \ldots, z_{f_{100}}(c)$. Figure 5.1 shows the types and two patterns for each of the them. Each plot shows the point pattern on the left and its intensity distribution (geographic feature signal) on the right.

Distance Matrix

We compute the distance matrices for a number of combinations between interaction characteristics representations and distance functions. The following distance functions have been used:

- *Euclidean distance*: This is just the metric between two vectors $a, b \in \mathbb{R}^k$ in Euclidean space, defined as

$$d_{\text{Eucl}}(a, b) = ||a - b||_2 = \sqrt{\sum_{i=1}^{k} (a_i - b_i)^2}. \tag{5.22}$$

  The Euclidean distance is a common first choice for numeric vectors. However, it treats each vector element independently, resulting in large distances even if two neighboring vector elements are similar.

- *Maximum distance*: This distance is defined as

$$d_{\max}(a, b) = ||a - b||_\infty = \max\{||a_1 - b_1||, \ldots, ||a_1 - b_1||\}. \tag{5.23}$$

  Here, only the vector element with the largest difference contributes to the distance.

- *Earth mover's distance (EMD)*: This distance is frequently used to compare histograms in image processing [Rubner et al., 2000]. We represent the input histogram by the vectors $a, b \in \mathbb{R}^k_+$. The EMD computes the amount of work needed to transform a given histogram a into a histogram b. The amount of work is determined by a flow matrix $\mathbf{F} \in \mathbb{R}^{k \times k}_+$. The total amount of work is defined as

$$\text{work}(a, b, \mathbf{F}) = \sum_{i=1}^{k} \sum_{j=1}^{k} d_{ij} f_{ij}, \tag{5.24}$$

[1] http://www.spatstat.org

subject to the conditions

$$f_{ij} \geq 0, i \in [1;k], j \in [1;k], \tag{5.25}$$

$$\sum_{j=1}^{k} f_{ij} \leq a_i, i \in [1;k], \tag{5.26}$$

$$\sum_{i=1}^{k} f_{ij} \leq b_i, j \in [1;k], \tag{5.27}$$

$$\sum_{i=1}^{k} \sum_{j=1}^{k} f_{ij} = \min\left(\sum_{i=1}^{k} a_i, \sum_{j=1}^{k} b_j\right). \tag{5.28}$$

In (5.24) d_{ij} is the distance of histogram bins (here the absolute distance of the index in the vector). The work is finally normalized, resulting in the EMD distance defined as

$$d_{\text{EMD}}(a,b) = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} d_{ij} f_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} f_{ij}}. \tag{5.29}$$

This distance is expensive to compute, since the minimization problem above needs to be solved and thus resulting in a sup-linear runtime in the order of the number of vector elements. The distance is, however, a promising choice for histogram-like vectors, since it takes distances and values between all elements in the vector into account.

- *Canberra distance*: This distance between two vectors $a, b \in \mathbb{R}^k$ is defined as

$$d_{\text{Canberra}}(a,b) = \sum_{i=1}^{k} \frac{|a_i - b_i|}{|a_i| + |b_i|}. \tag{5.30}$$

The Canberra distance takes the absolute value of the elements in the vector into account. The distance between vector elements with high values will be generally smaller (since the denominator will be large). Hence, differences of small values have more impact.

The aim of the evaluation is to best separate five groups of point pattern types. For this, we plot the patterns in a distance matrix as shown in Figure 5.3. The patterns are ordered by groups, with the groups themselves ordered by landmark small, landmark big, local, regional, global from left to right and bottom to top. For a perfect result, the patterns within a group are perfectly similar (small distance) and intra-group patterns are totally dissimilar (large distance), as shown in the pairwise separated case.

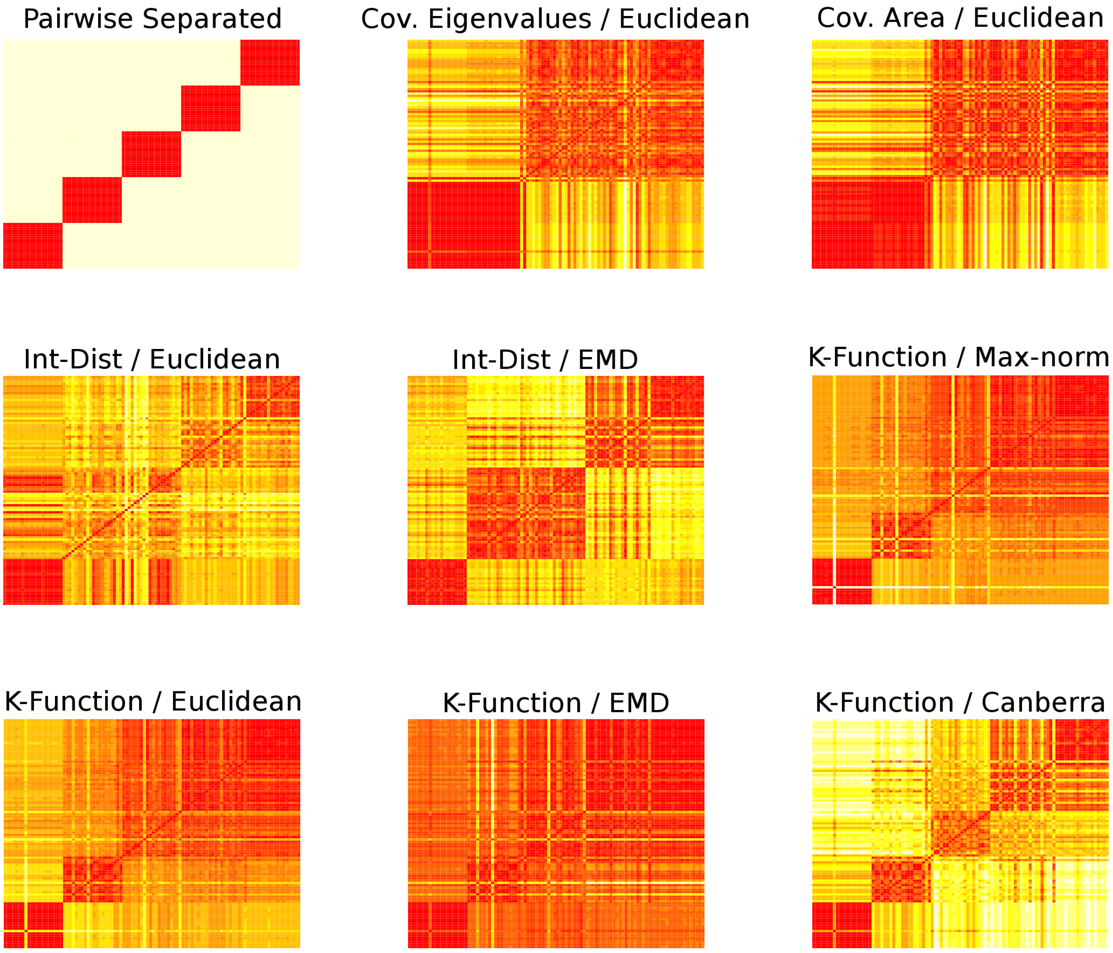


Figure 5.3: Distance matrices using covariance, intensity histogram, and K-function characteristics in combination with different distance functions.

Results

We evaluated the interaction characteristics of the above three feature representations using different distance functions. The resulting distance matrices are given in Figure 5.3. The upper-left distance matrix shows the ideal pairwise separation of the feature point patterns in the synthetic dataset.

The covariance/Euclidean matrix shows the distances using the Euclidean distance with the Eigenvalues e_f as input. The matrix shows that the landmarks and the local features are nicely separated from the regional and global features. However, there is a lack of separation between the local, regional, and global features as well as between the landmark features. The same can be observed for the covariance area using the absolute norm (Manhattan distance).

The intensity histogram/Euclidean matrix (Int-Hist/Euclidean) shows the distances using Euclidean distance with I_f vectors as input. The matrix shows a nice separation between the small landmark and the global features. There is also a remarkable separation between the groups of landmarks and local features, and the regional and global features. The latter separation becomes even more clear by the matrix showing distances using the EMD with I_f vectors (Int-Hist/Earth Mover).

The remaining matrices show the results for the K_f vectors. The usage of the maximum and the Euclidean distance function shows similar results. Even if the small and large landmark are pairwise separated, and together are separated from the local, regional, and global features, the overall performance is poor, compared to the intensity histogram with EMD. The same is true for the EMD of the K_f vectors. The best results are obtained with the Canberra distance. This can be explained as differences between small probabilities in the CDF have more impact in distinguishing between K_f vectors.

In our evaluation the Canberra distance together with the K_f vectors performed best. The performance of the EMD using the I_k vectors in also remarkable. Given the evaluation results we use the K_f/Canberra setting for the clustering task in the following Section.

5.5 Feature Exploration

We now present different exploration tasks based on interaction characteristics using real-world data sets. These data sources provide a large number of features (words, tags, keys) and each feature is represented by a point pattern. We will use the count-based signals derived from the point patterns, $z_f(c)$, to compare the features using the K_f vector representation and the Canberra distance.

Three kinds of exploration tasks are proposed that allow to select and filter features in the data sources. First, the features are clustered by their respective K_f vectors and the Canberra distance. We show that the resulting clusters provide a meaningful categorization of the features into spatio-temporal types.

Then, we propose a data source summarization tasks that allows to describe and compare different data sets quickly by boxplot visualizations of the K_f vectors in the

Source	*BBox*	# records (n_R)	# features (n_F)
Flickr	USA	4,518,322	27,552
	GER	528,281	4,536
Twitter	USA	247,743	2,047
	GER	46,111	480
OSM-Keys	USA	3,113,495	319
	GER	2,400,857	507
OSM-POIs	USA	1,230,709	362
	GER	1,325,678	483

$$\begin{aligned} \text{BBox GER} &= ([47.159, 55.103], \quad [5.888, 15.029]) \\ \text{BBox USA} &= ([25.244, 50.064], \quad [-124.892, -52.558]) \end{aligned}$$

Table 5.1: Data source statistics.

clusters. This technique allows for a comparison of different data sources (inter-source comparison) with respect to their feature types.

Finally, we present scatter plots between the feature's covariance area a_f (interaction characteristics) and the feature frequency. This allows for an intuitive and insightful understanding of how the feature type characteristics are related to feature frequency, and how these relationship differs between sources (inter-source comparison) and between areas (inter-area comparison).

5.5.1 Data Sources

We use Flickr, Twitter, and OpenStreetMap (OSM) data to extract different sets of geographic features. For each data source, we use data from the US and Germany. The Flickr features are tags of the geo-referenced photos. The Twitter features are hashtags of geo-referenced tweets. Both, the tags and the hashtags are normalized to lower case characters. From the OSM data we extract two kinds of features. The OSM-Keys data source consists of the attribute-keys of the OSM node records [2]. The keys can be seen as the attribute names. The users are allowed to specify arbitrary attribute names, hence, an exploration task to understand the spatio-temporal semantics of those keys is a meaningful analysis. The OSM-POI data source consists of the concatenated key-value pairs of selected attributes describing points-of-interests (POIs). For each data source, only those features are considered that occur at least 20 times. See Table 5.1 for detailed statistics of the data sources.

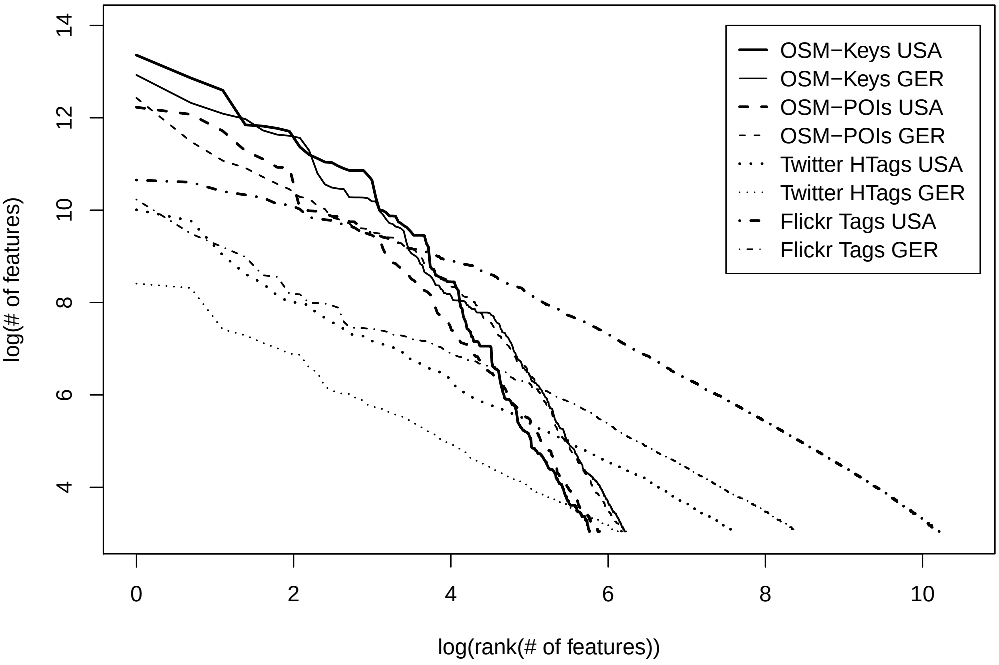


Figure 5.4: Log-log rank plot of feature frequencies.

5.5.2 Feature Frequency

For a better comparison of the features in the next section, we group features into different frequency groups. In each data source, the frequency of the features in the records can be described by a long-tail distribution. Figure 5.4 shows the log-log plot of the feature frequencies against their frequency rank. All the data sources show an almost linear log-log relationship. This means that the features (tags, hashtags, keys, attributes) follow a power law distribution, with a small number of very frequent features (head-group), a medium number of medium frequency (body-group), and a large number of very infrequent features (tail-group). For each data source, the features are first categorized into these groups by partitioning them into three equi-sized bins along the logged frequencies.

5.5.3 Feature Categorization

The aim of the feature categorization task is to determine and understand different classes of feature types in the data. For this, we cluster the features on the basis of their K_f vectors. The resulting clusters are assumed to contain features that have similar spatio-temporal types (e.g., landmarks, local, regional, global features). The clusters

[2] Each OSM record has an arbitrary number of associated attributes, with each attribute being represented as a key-value pair.

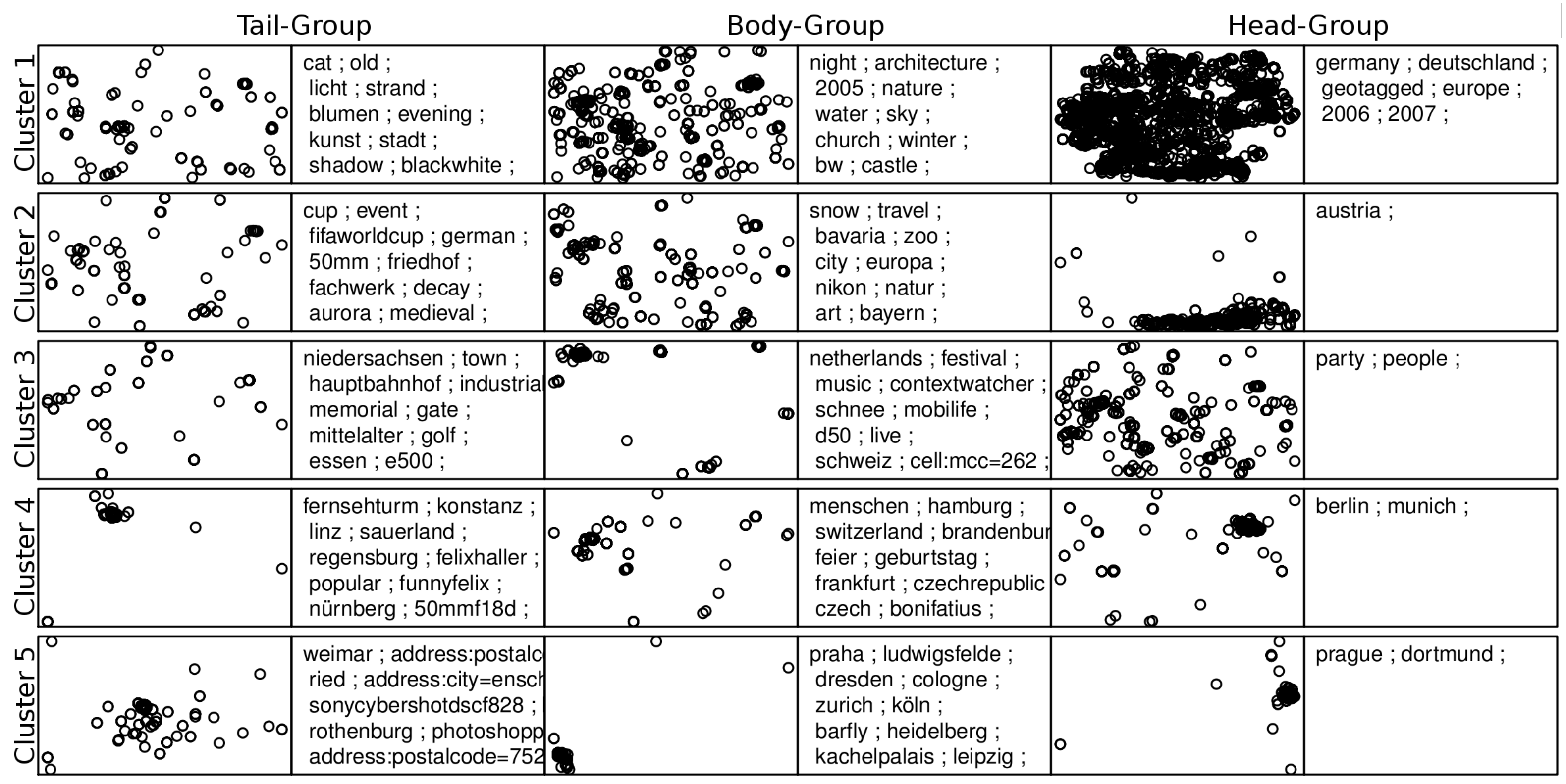


Figure 5.5: Canberra distance/K_f-vector clustering result for the Flickr-GER data set. Each row represents a cluster of features and the columns show the feature subset intersecting the respective frequency group. Each column has a left part, showing features associated with the respective cluster and frequency group combination, and a right part, showing the point pattern of a representative feature in the list.

can then be used to define a geographic feature matrix of geographic signals covering a particular spatio-temporal type. For example, for a predictive spatio-temporal model, we will need features that allow to distinguish between points on Earth (landmark features). To segment regions of coherent geographic semantics, however, we will need local or regional features that represents place types.

Clustering

The features are clustered using a fixed number of 5 clusters. This rather small number is motivated by our goal to find clusters that describe significantly different interaction characteristics. It also follows our assumption of having four different interaction characteristics classes (landmark, local, regional) plus one cluster for some unexpected characteristics.

As input data we use the Flickr GER data source where the features are tags of geo-referenced photos. To cluster the features we use the K-medoid clustering approach together with the K_f vector representation and the Canberra distance. For the K_f vector extraction we chose a maximum radius of 4.0 degrees (corresponding to $\sim$ 444 km) to capture interaction patterns up to a region-level scale.

Results

The resulting clusters are presented in an *interaction-frequency matrix* in Figure 5.5. The interaction-frequency matrix displays the clusters in the rows and the frequency groups in the columns. Each cell contains the most frequent features of the corresponding cluster intersecting the corresponding frequency group. In the matrix, also the point pattern of a representative feature is displayed for each cluster and frequency group.

The interaction-frequency matrix shows the most global features being of cluster 1 (row 1) in the head-group. The shape of Germany is clearly visible as expected for a global feature in the area of Germany (note that the plot is squeezed along latitude). The point patterns of the body- and the tail-group of the same cluster can be seen as thinned patterns, still showing the global characteristic of the point distribution. Note that this means that the features in the global clusters have a coherent spatio-temporal semantics independent of their total frequency.

Some dense clusters of points are visible in those thinned versions, representing major cities with a large number of feature instances. Hence, the global cluster still contains clustered (non random) point patterns, which is an expected behavior in the presence of a strong background distribution. The features in the first row can be interpreted as the background distribution and have little predictive value on the spatial domain (e.g. 'germany', 'europe', '2007').

The features of clusters 2 and 3 show a more local and regional semantics. Here, cluster 2 shows a more regional behavior in the tail- and the body-group than cluster 3. The head-group of cluster 3 clearly shows a landmark ('austria'). However, its size leads to not assigning it to the landmark clusters 4 or 5. The features in cluster 2 and 3 mostly represent types of places ('event','fachwerk','medieval', 'snow','travel',town','party'). These features a valuable instances to segment space into regions with coherent semantics or to assign them as attributes to points in space.

Clusters 4 and 5 clearly show a landmark semantics, where the points in the body- and head-group of cluster 4 are more widely spread than in cluster 5. These feature are highly predictive to identify a particular point or region in space. Overall, we interpret the result as showing global features in cluster 1, regional features in cluster 2, local features in cluster 3 and landmark features in clusters 4 and 5.

5.5.4 Data Source Summarization

We now turn to the task of summarizing data sources on the basis of the interaction characteristics. This allows to describe and compare sources as a whole. The tasks can be used to identify promising data sources and to merge features of similar spatio-temporal semantics from different data sources into a single geographic feature matrix.

5.5.5 Boxplot Visualization

For the inter-source comparison task, we clustered the sources by interaction characteristics using the same parameters as in the feature categorization task. Each cluster can be described by a boxplot of the K_f vectors of its features. We plotted the clusters of all data sources in the German area of interest as rows in Figure 5.6. The rows correspond to the datasets (Flickr, Twitter, OSM-Tags, OSM-POIs) and the columns to the clusters. The clusters are roughly ordered by their interaction characteristic (from global to landmark). The small matrix on the bottom-left of the figure is a rough indication about which clusters show similar characteristics. In the following we denote Flickr-X as being cluster X of the Flickr source (corresponding to row 1 column X).

Comparing Flickr and Twitter, both have a global and a landmark cluster (cluster 1 and 5, respectively). Also, Flickr-4 and Twitter-2 show similar characteristics. The geographic semantics of the latter cluster is a larger landmark cluster, e.g., containing predictive features of large cities ('berlin', 'hamburg') or regions ('czech', 'switzerland'). Interestingly, also OSM-POIs-1 is similar to these characteristics, containing features like 'natural_tree' and 'sport_skiing'. Hence, these features exhibit a landmark characteristics on a large-scale.

The OSM-Keys and OSM-POIs sources do not have plain landmark clusters. This is an expected result because the keys and POI-attributes are assumed to be general types used over the whole area of interest. The strongest landmark characteristic of the OSM sources is shown by OSM-Tags-5 with features like 'kms:zip' (a OSM attribute holding a zip code), which clearly has a landmark semantics.

Clusters 3 and 4 of the Twitter source have no corresponding partner in the Flickr data source, thus indicating significant differences in these data sources. Overall, the Flickr and Twitter data sources show interaction characteristics different from the more coordinated OSM sources.

Figure 5.7 shows summaries for the US datasets. One can see similar clusters as for the Germany data, with the Flickr and Twitter sources having strong landmark and

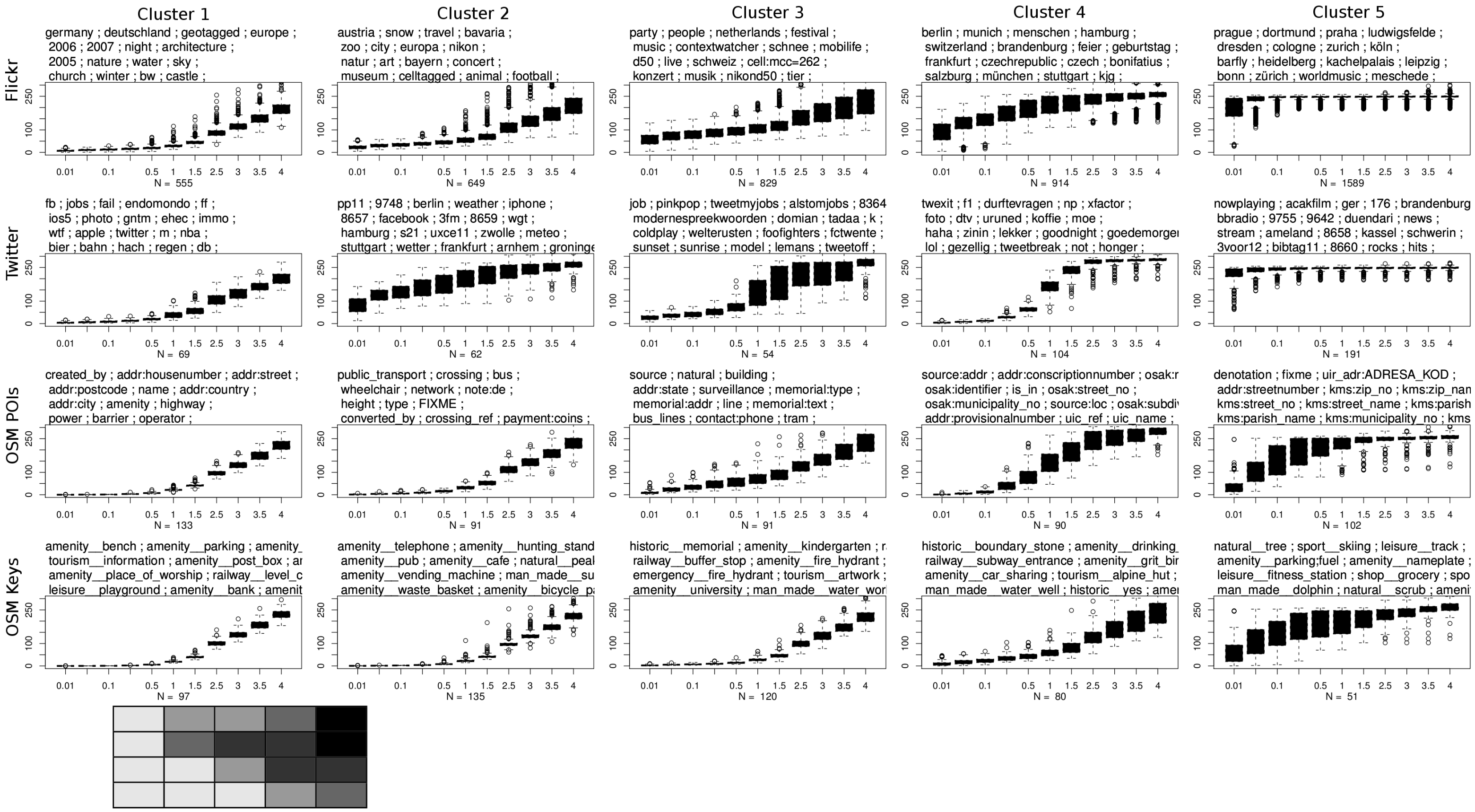


Figure 5.6: Inter-source cluster summary matrix for the Germany data sets. The four rows represent the data sets and the five columns the feature clusters. Each cell shows the features of the respective cluster on the top, and a K_f-vector boxplot of these features below. The small matrix on the lower-left indicates which boxplots are similar to each other by encoding the cells with identical color.

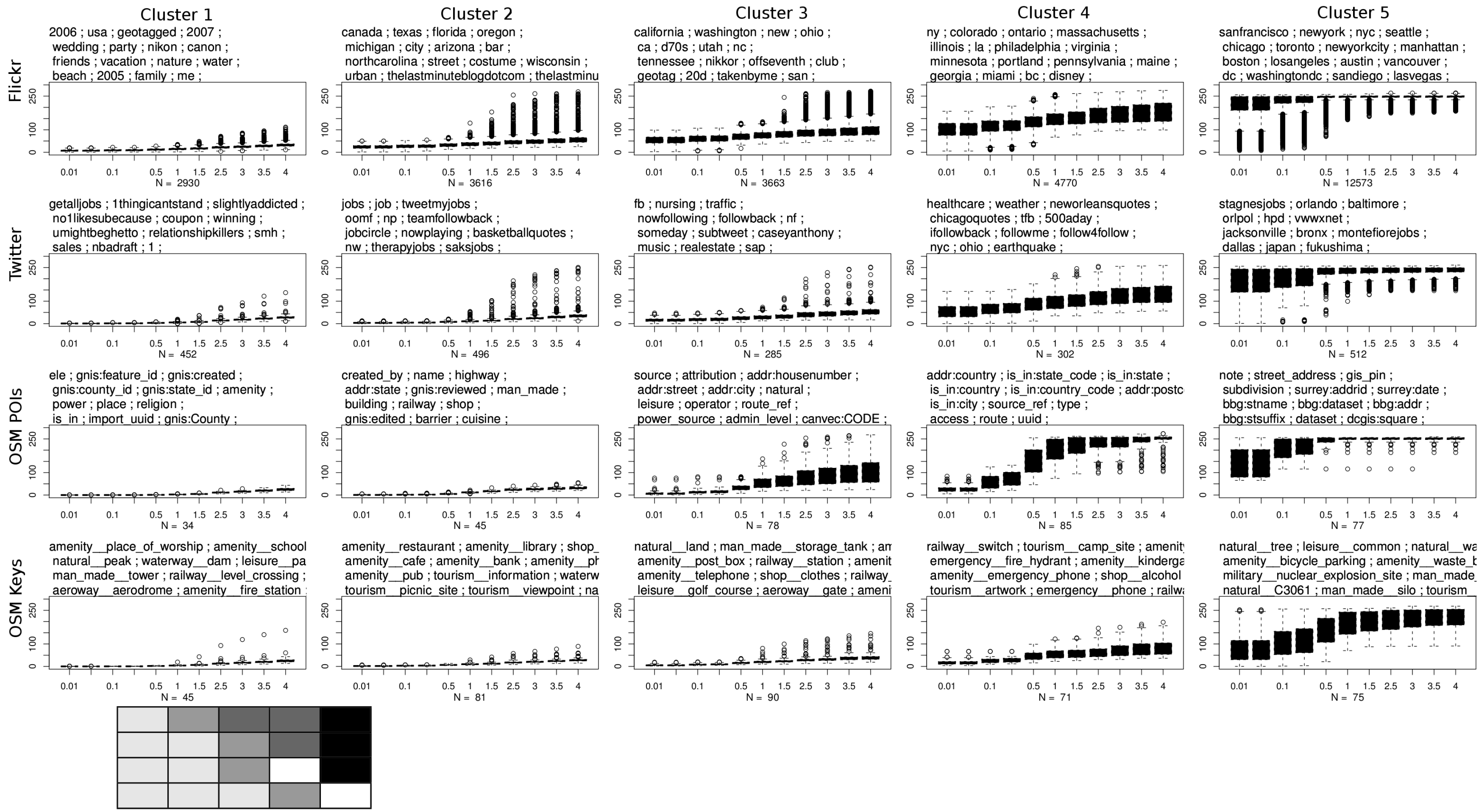


Figure 5.7: Inter-source cluster summary matrix for the US data sets. The four rows represent the data sets and the five columns the feature clusters. Each cell shows the features of the respective cluster on the top, and a K_f-vector boxplot of these features below. The small matrix on the lower-left indicates which boxplots are similar to each other by encoding the cells with identical color.

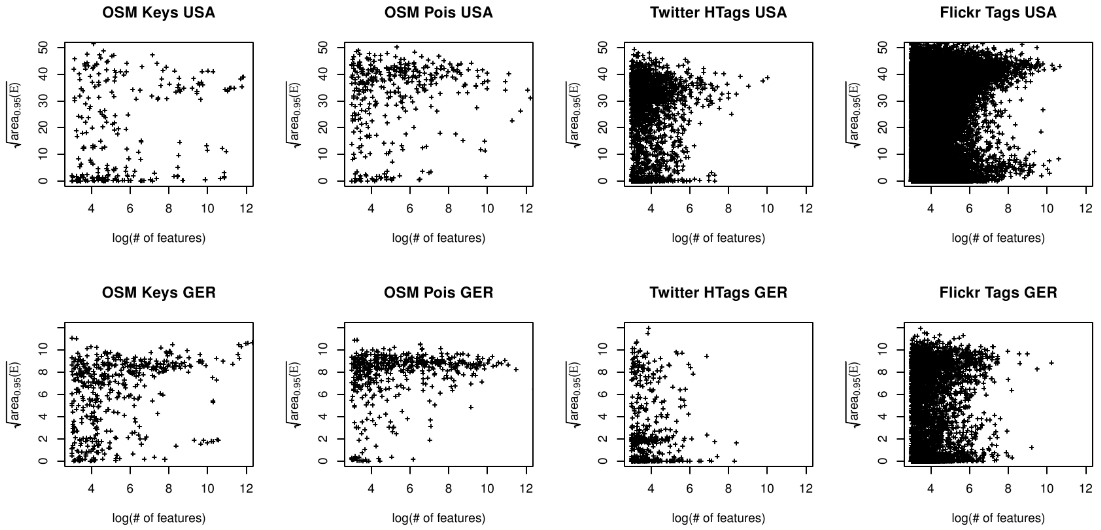


Figure 5.8: Area-frequency scatter plots. Each plot represents the frequency on the x-axis and the area on the y-axis. The plots in the top row show the results for the Germany data sets and the bottom row shows the results for the US data sets.

regional features. The Flickr, OSM-Keys, and OSM-POI clusters are very similar to the respective data sources in Germany. Only the Twitter-5 cluster has a less dominant landmark semantics than its German counterpart. This can be explained by a higher number of high frequency landmark features (hashtags) that have a larger scale level than in Germany. This is something we expect for the US, with a large number of landmark features in the main metropolitan areas.

5.5.6 Area-Frequency Scatter Plot

Finally, we describe a simple data visualization by using the covariance property a_f (the 95% confidence area of a point pattern of the Gaussian covariance approach). Note that this interaction characteristics is more limited, only allowing for a separation of landmark and non-landmark features by a single quantity. However, this allows us to relate the feature frequency to the landmark semantics easily in a scatter plot.

Figure 5.8 shows the scatter plots of a_f against the feature frequency for each feature in a data source. Eight plots are arranged in a matrix with the data sources in the columns and the regions in the rows. Note that a cluster of points in a scatter plot represents a dominant area-frequency subset in the data.

Each plot shows a smaller number of medium-area features compared to small-area and large-area features. This indicates that we have a larger number of landmark and global features in the data, while features covering a medium-scale area are less prevalent. One can see that the OSM-Key USA data source has fewer global features than its

German counterpart. This is an indicator that the OSM data in the USA is much less standardized, resulting in a higher number of keys with landmark semantics. On the other hand, the Twitter data source in the USA has more global features than its German counterpart, indicating that hashtags in Germany are much less coordinated than in the US. Hence, for a geographic feature matrix covering both, the US and Germany, the features of the Twitter and the OSM-Key sources need to be carefully selected, since they have different spatio-temporal semantics in these areas.

In contrast to the Twitter and OSM-Key sources, the OSM-POI and the Flickr sources show a more similar behavior in the US and Germany. Hence, these data sources are comparable regarding the contained spatio-temporal types. In summary, the scatter plot summarization allows to compare data sources or areas of interest easily in order to guide feature merging and data source selection decisions.

5.6 Summary

In this chapter we introduced novel methods and applications to compare spatio-temporal signals and point patterns on the basis of their spatio-temporal type. For this, we used the concept of interaction characteristics to derive vector-based representations of distribution properties. These representations can then be compared by distance and similarity functions. We showed that a representation derived from the K-function, in combination with the Canberra distance, performed best in separating different types of point patterns. However, also the intensity histogram together with the Earth-Mover distance showed remarkable results.

Given these novel techniques, we demonstrated how features in real-world data sets can be described, categorized and filtered by their type, making this feature comparison method a promising exploratory tool to pre-process and analyze large sets of candidate features. Finally, we introduced boxplot summarizations of the K-function vectors of clustered features to extract a compact visualization of the types in a data set. This allows for a comparison of features from different sources or from different areas of interest to guide feature selection and merging decisions.

In this chapter we did not deal with comparison methods that are based on the raw signals. As detailed before, such methods are valuable to filter and categorize features that refer to the same phenomenon instead of the same phenomenon type. Other works have successfully employed these methods to extract and summarize places and events. From an exploratory point of view, these methods focus on an alternative dimension to analyze and pre-process the set of feature candidates in identifying and merging redundant information. Accordingly, we see our novel method as an additional tool to analyze and explore the wealth of geographic information contained in unstructured geographic information sources.

Chapter 6

Latent Geographic Feature Extraction: Phenomenon Discovery using Dimensionality Reduction

6.1 Introduction

From the perspective of a data analyst, the primary objective is to extract a small but informative number of patterns from the data in a highly automated fashion. In this chapter, we deal with the task of extracting a small but informative number of distinct *geographic phenomena* from user-generated data. We call this task *geographic phenomenon discovery.* A geographic phenomenon is generally defined as any social or physical process or entity that can be identified in space and time. As discussed in Section 3.5.1 and detailed in Chapter 5, types of such phenomena include places (landmarks), events, trends, and trajectories of objects.

The motivation to do this is well-justified, since the records of user-generated data sources exhibit the opinions, interests, and impressions of people, together with geographic information about where and when they occur. We refer the reader to Chapter 3 for a detailed discussion on how user-generated data can be used as a source of geographic observations, and to [Kennedy et al., 2007] for an application-oriented introduction to place and event extraction from social media.

A number of works address specifically the extraction of geographic phenomena with place and event semantics [Ahern et al., 2007; Cheng et al., 2010; Crandall et al., 2009; Deng and Lemmens, 2009; Kennedy, 2008; Rattenbury et al., 2007a; Rattenbury and Naaman, 2009; Yin and Cao, 2011]. In Section 3.2, we reviewed these models and categorized them in clustering-based and uni-modal distribution-based approaches. Different from these approaches, the following work will focus on the extraction of arbitrarily-shaped spatio-temporal distributions from user-generated data. This includes complex spatio-temporal distributions of geographic phenomena such as coastlines, mountain regions, urban regions, and social behavior, among others. To extract such spatio-temporal patterns, the place and event extraction methods, which focus on bump-finding in the distributions, will not work adequately. Therefore, our work is more related to recent approaches of spatio-temporal topic modeling [Hong et al., 2012; Mei et al., 2006; Yin et al., 2011] and the extraction of tags with similar spatio-temporal semantics [Zhang et al., 2012b].

We propose a novel and fundamental approach to obtain geographic phenomena from user-generated data, called *latent geographic feature extraction*. This approach works directly on the geographic feature matrix, as introduced in Section 3.4.1, and represents a sub-task in the geographic feature mining framework. In this approach, a small number of *latent geographic features* is assumed to be hidden in a huge number of candidate signals. The aim of latent geographic feature extraction is to extract these latent features from the geographic feature matrix. Using the idea of a generative model, the latent geographic features can be seen as underlying factors that generate the candidate features. For example, several geographic features of a textual record source will refer to the same phenomenon, such as the terms 'beach', 'boat', 'sea', 'sand', 'water'. The aim of latent geographic feature extraction is to discover a phenomenon that explains the joint occurrence of these terms in geographic space, namely, a latent 'coast' feature. Of course, extracting phenomena that are already known is not so much of interest from an exploratory point of view. However, discovering joint occurrences of social, market, or health related terms will be highly informative to describe such processes and entities from a geographic perspective, and to utilize this information in domain-specific applications.

In our evaluation, we will often use well-known processes and entities to judge about the quality of the phenomenon discovery results. This is a necessary proxy evaluation for developing concepts and techniques that allow to find novel and unexpected phenomena on the basis user-generated data in general. Will we present some unexpected and meaningful phenomena found for the Los Angeles area using Flickr data in a dedicated exploration section of this chapter.

The introduced approach is based on *dimensionality reduction* of the geographic feature matrix. Since dimensionality reduction is closely connected to *latent variable models* [Hastie et al., 2009, p. 678], this justifies the naming of our approach. In the following, we compare several types of dimensionality reduction approaches that make use of different statistical properties of the data. In an exhaustive evaluation, we show that techniques assuming sparse and statistically independent feature combinations result in the most informative latent features, and that our proposed latent geographic feature extraction approach is able to discover more informative phenomena than existing document-centric approaches for data sets exhibiting very high noise levels. Moreover, we show that the normalization strength of the candidate feature signals can be used as a parameter to find different types of phenomena. Preliminary results of these evaluations have also been published in [Sengstock and Gertz, 2012b] and [Sengstock and Gertz, 2012a].

As for the output of existing place and event extraction approaches, latent geographic features are a valuable analytic output. They allow to explore the semantics of geographic space as perceived by the users and reflected in the data. Moreover, the latent geographic feature signals constitute high-level geographic features themselves, and can be extracted and persisted as geographic raster data for subsequent tasks. Using the terminology of our geographic feature mining framework, the latent geographic features can be used to define a low-dimensional geographic feature matrix for subsequent data mining and learning tasks, such as predictive models [Gallagher, 2010; Shekar et al., 2002], geographic segmentation [Leung and Newsam, 2010], or context-aware recommendation [Adomavicius and Tuzhilin, 2011].

The remainder of this chapter is structured as follows. First, in Section 6.2, we detail the problem statement and the contributions. Then, we review related topic modeling approaches in Section 6.3. In Section 6.4, we introduce the concept of dimensionality reduction on the geographic feature matrix and discuss different dimensionality reduction techniques. In Section 6.5, we evaluate their applicability for phenomenon discovery in several exploration tasks. Finally, in Section 6.6 we introduce how the records can be associated with the extracted phenomena, in support of categorization, browsing, and search tasks.

6.2 Problem Statement and Contributions

A *geographic phenomenon signal* is represented as a spatio-temporal variable

$$z_q(c) \in \mathbb{R}_+, c \in D_C. \tag{6.1}$$

The *geographic feature candidates* are given as a set

$$F = \{f_1, \dots, f_p\}, \tag{6.2}$$

with each feature $f \in F$ having a *geographic feature signal*

$$z_f(c) \in \mathbb{R}_+, c \in D_C. \tag{6.3}$$

We expect that the feature signals are defined on a spatio-temporal lattice $L = \{c_1, \dots, c_n\}$, and we describe the discrete signal $z_f(c)$ as a vector over a discrete number of cells

$$\mathbf{z}_f = (z_f(c_1), \dots, z_f(c_n)) \in \mathbb{R}^n. \tag{6.4}$$

A *geographic feature matrix* $\mathbf{Z}_{L,F} \in \mathbb{R}_+^{p \times n}$ is a matrix in which each row corresponds to an n-dimensional geographic feature signal.

To represent columns and rows in matrices we use the following notation: Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we use $col_i(\mathbf{A}) \in \mathbb{R}^n$ to refer to the vector in the ith column, and $row_j(\mathbf{A}) \in \mathbb{R}^m$ to refer to the vector in the jth row.

A geographic phenomenon can be described by a combination η of the feature signals $z_{f_1}(c), \dots, z_{f_p}(c)$ denoted

$$z_q(c) = \eta_{\boldsymbol{\alpha}_\mathbf{q}}(c). \tag{6.5}$$

The variable $\boldsymbol{\alpha}_q \in \mathbb{R}^p$ denotes a vector of feature weights for the features $f \in F$. The proposed dimensionality reduction methods assume that the feature combinations are linear, leading to the following feature combination function

$$\eta_{\boldsymbol{\alpha}_\mathbf{q}}(c) = \alpha_0 + \alpha_1 z_{f_1} + \cdots + \alpha_p z_{f_p}. \tag{6.6}$$

We generally omit the intercept α_0 and assume $\boldsymbol{\alpha}$ to be a p-dimensional vector of weights. The corresponding signal vector of $z_q(c)$ is denoted $\mathbf{z}_q \in \mathbb{R}^n$. The general idea of latent geographic feature extraction does not depend on linearity of the feature combination, as long as a feature weight vector $\boldsymbol{\alpha}_q$ can be derived from the combination.

For the sake of convenience, we assume that the resulting geographic phenomenon signal $\mathbf{z}_q$ and the feature weights $\boldsymbol{\alpha}_q$ can have both, positive and negative values. To interpret the resulting signal vector $\mathbf{z}_q$ as a positive intensity and the elements in $\boldsymbol{\alpha}_q$ as importance weights, we flip the sign of $\boldsymbol{\alpha}_q$ if the largest absolute value in the vector is negative. In this case, we also flip the sign of the signal vector $\mathbf{z}_q$. Then, we only make use of the positive signal values and feature weights to describe the intensity of the geographic phenomenon and the importance of feature weights for that phenomenon. We will see, however, that the most promising techniques will produce signals and weights that are solely positive or negative, and can hence always be flipped to a strictly positive value domain.

The features with the highest weight in $\boldsymbol{\alpha}_q$ are used as a summary of the phenomenon q. Since we employ textual features in this work, we can use them as a textual description given as an ordered list of terms. The description summarizes what a phenomenon is about and the signal describes where and when it occurs.

Thus, each phenomenon q is represented by a spatio-temporal signal vector $\mathbf{z}_q$ and a feature weight vector $\boldsymbol{\alpha}_q$. Geographic phenomenon discovery needs to extract a small number $k \ll p$ of geographic phenomena

$$Q = \{q_1, \ldots, q_k\}, \tag{6.7}$$

with each phenomenon $q \in Q$ being described by a spatio-temporal distribution $\mathbf{z}_q$ and a feature weight vector $\boldsymbol{\alpha}_q$,

$$q = (\mathbf{z}_q, \boldsymbol{\alpha}_q). \tag{6.8}$$

An extracted geographic phenomenon q is said to be *informative* or *meaningful* if it describes a real-world social or physical phenomenon. Since the technique of geographic phenomenon discovery is inherently unsupervised, we evaluate informativeness by presenting phenomenon descriptions and signals and discuss if the combination is a reasonable result. For this, (1) the signal should clearly describe high and low intensity areas, (2) the description should describe a meaningful geographic process or entity, and (3) the signal-description combination should be coherent. Besides informativeness, the extracted signals should be distinct from each other. A result in which several extracted phenomena represent the same real-world process or entity is hence considered a bad result. The problem statement can now be defined as follows:

> Given a set of p geographic features candidates F, extract k geographic phenomena Q, with $k \ll p$, where the phenomena $q \in Q$ are highly informative and represent distinct geographic processes or entities.

The contributions are summarized as follows:

- We show that dimensionality reduction of the geographic feature matrix allows to discover informative and distinct geographic phenomena from user-generated data. In an exhaustive evaluation we show that this approach is able to discover more informative phenomena than spatio-temporal topic models and clustering approaches. For this, we introduce a set of carefully selection qualitative criteria, to judge about the informativeness of the results.

- We show that the choice of the dimensionality reduction technique has a strong impact on the quality of the results. We find that techniques assuming sparse and statistically independent feature combinations result in the most informative phenomena.

- We show that the strength of normalization of the signals in the geographic feature matrix can be used as a parameter to find different types of geographic phenomena, such as results having landmark, regional, or global semantics.

- We demonstrate how extracted latent geographic features can be associated with the records in the input data source. This allows to address problems in the field of information organization and retrieval, such as categorization, browsing, and search tasks.

6.3 Document-centric Approaches

First, we describe approaches that are aimed to extract textual topics from document collections. The primary aim of these models is to describe each document by a small number of topics that allow for efficient processing of large collections while preserving the essential statistical relationships that are useful for classification, novelty detection, summarization, and similarity computations [Blei et al., 2003]. Given a set of geo-referenced documents, the spatio-temporal information can be taken into account to describe how the topics are distributed in space and time.

We introduce a *topic-pivot* approach that uses latent Dirichlet allocation (LDA) to extract a small number of topics in a pre-processing step, and then computes the spatio-temporal topic distribution using the geographic information of the records. The second introduced approach is an extension of probabilistic latent semantic analysis (PLSA) that takes the spatial location of documents explicitly into account, called latent geographic topic analysis (LGTA).

6.3.1 Geo-referenced Documents

The document-centric approaches work directly on the corpus of geo-referenced documents, with each document being represented as a bag of words. As in Section 4.3.1, we describe this type of data source as

$$R = \{r_1, \ldots, r_m\}, \tag{6.9}$$

where each record r (document) is described as a tuple

$$r = (r_F, r_c[, r_u]). \tag{6.10}$$

We use r_{f_i} do denote the number of times feature f_i occurs in the record, $r_c \in D_C$ to denote the GPS coordinate and r_u to denote the user. The user information is not considered in the introduced approaches and is therefore ignored in the remainder of this section.

6.3.2 Topic-pivot Approaches

Different approaches exist to extract topics from a document corpus. A non-probabilistic approach is latent semantic analysis (LSA) [Deerwester et al., 1990]. LSA extracts the topics by a singular value decomposition of the $m \times p$ record-feature matrix. This decomposition has a close connection to principal component analysis (PCA), which will be discussed later. In [Hofmann, 1999], a probabilistic version of LSA, called probabilistic latent semantic analysis (PLSA) has been introduced. Different from the SVD approach, this is based on a Bayesian model (a mixture model of multinomial distributions) and has a sound statistical interpretation. Recently, PLSA was improved by a generative topic model called latent Dirichlet allocation (LDA) [Blei et al., 2003]. In the following we shortly review LSA, since it is closely connected to PCA and LDA, which is a de-facto standard in topic modeling and has been used in a variety of geographic data mining applications, such as in [Adams and McKenzie, 2012] and [Chae et al., 2012].

Latent Semantic Analysis (LSA)

We shortly detail LSA, since its usage of the singular value decomposition (SVD) is also needed later for PCA. LSA is also one of the first topic modeling approaches and is still widely used in text mining and information retrieval [Manning et al., 2008].

LSA extracts the topics by a singular value decomposition of the $p \times m$ feature-record matrix (also called term-document matrix)

$$\mathbf{T} \in \mathbb{R}_+^{p \times m}, \tag{6.11}$$

where the matrix elements t_{ij} denote the number of times feature f_i occurs in record r_j. The singular value decomposition of this matrix is

$$\mathbf{T} = \mathbf{UDV}^\top, \tag{6.12}$$

where the columns of $\mathbf{U} \in \mathbb{R}^{p \times p}$ are the left singular vectors, $\mathbf{D} \in \mathbb{R}^{p \times p}$ is a diagonal matrix with eigenvalues of both, $\mathbf{TT}^\top$ and $\mathbf{T}^\top\mathbf{T}$, ordered by $d_1 \geq \cdots \geq d_m$ and being called the singular values, and the rows of $\mathbf{V}^\top \in \mathbb{R}^{p \times m}$ are the right singular vectors [Manning et al., 2008, p. 374]. The singular value decomposition allows to represent the matrix $\mathbf{M}$ by a rank-k approximation $\mathbf{T}_k$. For this, the smallest $m - k$ eigenvalues in $\mathbf{D}$ are set to zero [Manning et al., 2008, p. 376]. By removing the last $m - k$ singular values from $\mathbf{D}$ and correspondingly removing the last $m - k$ columns from $\mathbf{U}$ and rows from $\mathbf{V}^\top$, one can represent the original $p \times m$ term-document matrix by a $p \times k$ approximation.

For the sake of topic modeling, the weights of the p features in the k topics can be found in the columns of $\mathbf{UD}$, and the topic weights of a record can be found in the columns of $\mathbf{V}^\top$. Note that the resulting weights can be positive and negative, which prevents interpreting them as probabilities.

Runtime Complexity. Different algorithms exist to decompose a matrix by SVD. In general, the runtime of the algorithms are difficult to express, since they depend on how fast the iterative algorithms converge, and on the admissible numerical error of the decomposition. In [Trefethen and Bau, 1997] the authors give the runtime of a two-step bidiagonalization/QR factorization approach for a $p \times m$ matrix as

$$O(pm^2 + m^2) = O(m^2(p + 1)) = O(pm^2). \tag{6.13}$$

The runtime is hence quadratic in the number of records m, which is problematic for data sets with a huge number of records.

Latent Dirichlet Allocation (LDA)

LDA is a probabilistic topic model and a successor of PLSA. It is based on a generative mixture model. The generative process of the data set can be described by the following schema:

```
For each document $r \in R$
    Sample a topic distribution for the document $\theta_r \sim Dirichlet(\alpha)$
    For each word in $r_F$
        Sample a topic $q \sim Multinomial(\theta_r)$
        Sample a word $f \sim Multinomial(\beta_q)$
```

The parameters represent the record-conditional topic probability

$$\mathrm{p}(q|r) = \theta_r \tag{6.14}$$

and the topic-conditional feature probability

$$\mathrm{p}(f|q) = \beta_q. \tag{6.15}$$

These two distributions allow to describe each record as a mixture of topics, and each topic as a distribution over features. We will show in the next section how these probabilities can be used to compute a spatio-temporal distribution of a topic.

Runtime Complexity. The authors in [Blei et al., 2003] propose a variational inference algorithm to estimate the parameters θ_r and β_q. Different algorithms exist, however, to estimate the parameters, including sampling-based approaches [Xiao and Stibor, 2010], variational EM-algorithms [Nallapati et al., 2007], and online learning approaches [Hoffman et al., 2010]. The variational approaches are considered to be fast and are, different from the sampling approaches, deterministic. The runtime of the variational algorithm proposed in [Blei et al., 2003] is reported by the authors in [Nallapati et al., 2007] to be

$$O(l(mh^2k + pk)), \tag{6.16}$$

where l is the number of iterations needed to converge to a solution, m is the number of records (documents), h is the average number of features in a record, k is the number of topics, and p is the number of features. Their inference approach is not quadratic in the number of records such as LSA. However, it strongly depends on the number of iterations needed until convergence (which heavily depends on the admissible error). One can, however, assume that $l \ll m$ for large corpora, making the runtime complexity of LDA smaller than LSA.

Spatio-temporal Topic Distribution

The two topic models described above discover a number of topics $q_1, \ldots, q_k$ in the data. In the probabilistic LDA model, the resulting topics are described by the distributions $\mathrm{p}(f|q)$ and $\mathrm{p}(q|r)$.

In [Adams and McKenzie, 2012] and [Chae et al., 2012] the authors' aim is to extract geographic phenomena from geo-referenced data. In their work, the authors first extract a number of k topics using LDA in a pre-processing step. Using our notation of geographic phenomena, the distribution $\mathrm{p}(f|q)$ represents the feature weight vector

$$\boldsymbol{\alpha}_q = (\mathrm{p}(f|q))_{f \in F}. \tag{6.17}$$

The spatio-temporal distribution of the phenomena is extracted from $\mathrm{p}(q|r)$ by aggregating the geographic information of the records. For this, a weighted histogram approach is used

$$z_q(c) = \sum_{r \in R} \mathbf{1}\{r_c \in c\}\, \mathrm{p}(q|r). \tag{6.18}$$

We call this the *topic-pivot approach*, since the topics are first extracted and then, the spatio-temporal distribution is determined on the basis of the topics. Note that in the LSA approach we can represent the feature weight vector by the columns of $\mathbf{UD}$,

$$\boldsymbol{\alpha}_q = col_i(\mathbf{UD}). \tag{6.19}$$

The topic weights of a record, represented in the columns of $\mathbf{V}^\top$ are not strictly positive. Hence, there is no well-defined technique to extract the spatial distribution $z_q(c)$ for topic q. We can, however, use the transformation proposed in Section 6.2, to extract a positive value range for the feature weights.

6.3.3 Geographic Topic Models

The topic-pivot approach extracts topics without taking the geographic information of the records into account. Extensions to topic models have been proposed that make use of additional attributes in the data, however. Such attributes include temporal information, such as the document creation time [Blei and Lafferty, 2006; Hong et al., 2011; Wang et al., 2008], and geographic information, such as associated GPS coordinates [Hong et al., 2012; Yin et al., 2011] (see Section 3.2.7 for more examples of spatio-temporal topic models). Even if these topic models are primarily meant to describe document topics, they can be used for geographic phenomenon discovery, if the models allow to extract spatio-temporal distributions of the topics.

Latent Geographic Topic Analysis (LGTA)

For the comparative evaluation, we use the model proposed in [Yin et al., 2011], since it represents a straightforward extension of PLSA for geo-referenced documents. Moreover, it is based on simple geo-referenced records, different from models that take additional domain-specific attributes like language, links, or social networks into account.

To realize a spatio-temporal topic model, the authors introduce the concept of a region. Each region $l \in L$ is represented by a Gaussian distribution in spatial space D_C

$$Gaussian(c; \mu_l, \Sigma_l). \tag{6.20}$$

The textual topics are now extracted on the basis of the features occurring in the regions instead of the documents. By this, the model is actually a combination of a Gaussian mixture model and PLSA. The generative process to create the geo-referenced data source can be described as

```
For each document $r \in R$
    Sample a region from region importances $l \sim Discrete(\gamma)$
    Sample a document location from the region $r_c \sim Gaussian(\mu_l, \Sigma_l)$
    For each word $f \in r_F$:
        Sample a topic from the region topic importances $q \sim Multinomial(\phi_l)$
        Sample a word from the topic importances $f \sim Multinomial(\theta_q)$
```

The following probabilities are describe by the model

$$\mathrm{p}(f|q) = \theta_q. \tag{6.21}$$

This is the importance of the features $f \in F$ for a topic q and is denoted in our notation as

$$\boldsymbol{\alpha}_q = (\mathrm{p}(f|q))_{f \in F}. \tag{6.22}$$

The distribution of a topic in geographic space is described by the regions $l \in L$. The probability at any point $c \in D_C$ for a given region l is given by

$$\mathrm{p}(c|l) = Gaussian(c; \mu_l, \Sigma_l) \tag{6.23}$$

and the importance of a topic for a region is given as

$$\mathrm{p}(q|l) = \phi_l. \tag{6.24}$$

Note that the regions represent a Gaussian mixture model with the region importances

$$\mathrm{p}(l) = \gamma. \tag{6.25}$$

The distribution of a topic in space is thus given as

$$\mathrm{p}(c|q) = \sum_{l \in L} \mathrm{p}(c|l)\, \mathrm{p}(q|l)\, \mathrm{p}(l), \tag{6.26}$$

which we describe in our notation as

$$z_q(c) = \mathrm{p}(c|q). \tag{6.27}$$

Runtime Complexity. The authors use a nested EM approach to estimate the parameters of the model. The runtime complexity of this approach is stated as

$$O(l_1(kmp + mh + md + l_2kmp)), \tag{6.28}$$

where l_1 and l_2 are the number of iterations needed for the outer and the inner EM loop to converge, respectively, m is the number of records, p is the number of features, h is the average number of features per record, and k is the number of topics. Again, the runtime complexity heavily depends on the number of iterations l_1 and l_2, but is linear in the number of records m. Hence, if a weak convergence criterion is used this approach can be seen as efficient for large data sets.

6.4 Dimensionality Reduction

The previous approaches use the documents directly. We now propose techniques to extract geographic phenomena from a geographic feature matrix. We call these methods *latent geographic feature models*. These models have the following advantages:

- The signals of the candidate features can be extracted using sophisticated feature signal extraction models as shown in Chapter 4. This includes the handling of user redundancy, influence of the background distribution, robustness, and appropriate scale level selection.

- Features can be filtered and selected using appropriate geographic feature comparison tasks in a pre-processing step.

- The candidate signals can be appropriately normalized in a pre-processing step. We show that appropriate normalization allows to extract latent geographic features of different spatio-temporal type.

The above benefits are a result of defining geographic phenomenon discovery as a subtask in the geographic feature mining framework (see Chapter 3). There, instead of building complex models whose primary aim is to model documents, terms, or other corpus information, we focus on the spatio-temporal signals and their spatio-temporal semantics, and provide intuitive and domain-specific techniques to extract geographic knowledge.

In the following, we first introduce the main objectives of dimensionality reduction in vector-space data representations. Then, we review the semantics of a geographic feature matrix. Finally, we detail three selected dimensionality reduction techniques, and discuss their semantics in the context of geographic phenomenon discovery.

6.4.1 Background

Dimensionality reduction tries to project data points of a high-dimensional data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, with $n = p$ or even $n < p$, onto a low-dimensional subspace $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times k}$, with $k \ll p$. One way to accomplish such a task is to *select* features from the data matrix. Such a selection can be based on heuristics or on *feature selection* algorithms in a supervised learning context [Guyon and Elisseefi, 2006]. In the latter case, those features are selected that increase the performance of a predictive model. Our aim is, however, to extract features in an unsupervised fashion. Such techniques are known as *unsupervised feature selection* [Guyon and Elisseefi, 2006] or *feature extraction* approaches [Bishop, 2006, p. 2].

Latent feature extraction is a popular technique in exploratory data analysis. There, the aim is to study unobserved (latent) characteristics of the data. Our input data is made of spatio-temporal cells, with the dimensions being geographic feature candidate signals. The extracted latent features thus represent a subspace of the feature candidate dimensions.

We discuss the application of principal component analysis (PCA), independent component analysis (ICA), and sparse PCA (SPCA) as dimensionality reduction techniques. PCA and ICA are among the most popular techniques (see [Bishop, 2006, p. 561] and [Hastie et al., 2009, p. 558]). The application of ICA as an explorative feature extraction technique is also known as projection pursuit [Hastie et al., 2009, p. 557]. We then discuss SPCA as a recent technique proven useful for high-dimensional data [Hastie et al., 2009; Peter and Van de Geer, 2011; Zou et al., 2006].

6.4.2 Geographic Feature Matrix

The input to vector-space based dimensionality reduction techniques is an $n \times p$ matrix with the n rows representing the observations, and the p columns representing the attributes. In our task, the input is a geographic features matrix, where the n rows represent cells in a discrete spatio-temporal domain (also called data points or locations in the following), and the p columns represent the feature candidates

$$\mathbf{Z}_{L,F} \in \mathbb{R}_{+}^{n \times p}. \tag{6.29}$$

Using dimensionality reduction, the input matrix is transformed into a low-dimensional representation

$$\mathbf{Z}_{L,Q} \in \mathbb{R}^{n \times k}, \tag{6.30}$$

with $k \ll p$. We assume that the essential statistical structure of the matrix $\mathbf{Z}_{L,F}$ is preserved in the low-dimensional matrix $\mathbf{Z}_{L,Q}$, such that the resulting latent features exhibit informative and distinct information about geographic phenomena. The columns of the low-dimensional matrix represent the latent geographic feature signals $\mathbf{z}_{q_1}, \ldots, \mathbf{z}_{q_k}$. In addition, the techniques return a feature weight vector for each latent dimension $\boldsymbol{\alpha}_{q_1}, \ldots, \boldsymbol{\alpha}_{q_k}$, which allows to describe the latent feature on the basis of the candidate features.

For our input data, the feature signals are treated as a high-dimensional measurement over the spatio-temporal cells in a lattice. Since these candidate features are assumed to be noisy, exhibiting redundant information, or no essential geographic semantics at all (e.g., by just following the background distribution), the aim of dimensionality reduction is to find a new set of dimensions that describe the geographic space in a more informative way. For this, each dimension should have a distinct semantics, noise should be reduced, and the background distribution be removed (if not done so already in the input matrix). The different dimensionality reduction techniques proposed in the following allow to extract such low-dimensional representations by assuming different statistical properties of the resulting latent features.

6.4.3 Spatio-temporal Tag Clustering

First, we present a simple solution based on ordinary clustering of the signals, which can be seen as a naive dimensionality reduction technique. In [Zhang et al., 2012b], the authors propose a method to find clusters of tags in geo-referenced social media with similar spatio-temporal semantics. For this, the authors create a matrix similar to the geographic feature matrix, with the cells of discretized spatio-temporal space represented in the rows, and the tags in the columns. The resulting clusters can hence be interpreted as latent geographic features in our sense.

To find clusters of similar tags, the columns of the matrix are clustered using K-means. Since the input are the spatio-temporal distributions of the tags $\mathbf{z}_{f_1}, \ldots, \mathbf{z}_{f_p}$, the resulting k clusters $q_1, \ldots, q_k$ are describes by non-overlapping sets of features

$$\{F_{q_1}, \ldots, F_{q_k}\}, F_{q_i} \subseteq F, F_{q_i} \cap F_{q_j} = \emptyset \forall i \in [1; k]. \tag{6.31}$$

The cluster centroid represents the average spatio-temporal signal over the cluster features

$$\mathbf{z}_q = \frac{1}{|F_q|} \sum_{f \in F_q} \mathbf{z}_f. \tag{6.32}$$

By using K-means, there is no notion of a feature weight vector for a cluster and the authors in [Zhang et al., 2012b] do not consider them at all. The Euclidean distance between the feature distribution and the cluster centroid might be used as an inverse indicator of the weight of a feature to a cluster. However, then a Gaussian mixture model (GMM) would be a better choice for this task.

K-Means and GMM are closely connected to PCA. However, their statistical objective in the context of dimensionality reduction is not well understood. See [Ding, 2004] for a discussion on the similarities between PCA and K-Means, and [Bishop, 2006, p. 443] for details on the relationship between K-Means and GMM.

The major benefit of the K-means approach is the runtime efficiency for large datasets. The runtime complexity is $O(lknp)$, where l is the number of iterations to converge (usually in the order of 10), k is the number of latent features, n is the number of spatio-temporal cells, and p is the number of candidate features. Since we already compiled the usually large number of m records in the geographic feature matrix, the K-means approach is very fast since it is linear in the number of cells n, and the number of candidate features p. This runtime benefit is, however, not anymore valid if GMM is used for the task. Because of this, we only choose K-means as an alternative to the following dimensionality reduction approaches, since the runtime is the primary advantage.

6.4.4 PCA

PCA seeks linear combinations of the original variables (features) that maximize the variance over the data points (cells). Those combinations describe new extracted variables and are called principal components (PCs). The first k components, ordered by their amount of variance, are used as dimensions to approximately describe the data in a reduced k-dimensional space.

The PCs of a geographic feature matrix $\mathbf{Z}$ represent the latent geographic feature signals. In accordance to the PCA terminology, we now also use the term components to refer to the latent geographic features. The components can be found by singular value decomposition (SVD). For this, the matrix $\mathbf{Z}$ is first centered such that all columns have zero mean. The matrix is then factorized as

$$\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^\top. \tag{6.33}$$

As described in Section 6.3.2 the resulting matrices $\mathbf{U}$, $\mathbf{D}$, and $\mathbf{V}^T$ represent the left singular vectors, the singular values, and the right singular vectors respectively. By only choosing the k largest singular values, and removing the $n-k$ columns and the $n-k$ rows from $\mathbf{U}$ and $\mathbf{V}^\top$, respectively, the principal components can be found in the columns of $\mathbf{UD}$, and the feature weights in the rows of $\mathbf{V}^\top$. The feature weights are also called the loadings of a principal component in PCA terminology.

By using our notation for latent geographic features, the latent geographic features signals (components) are given as the columns of **UD**

$$\mathbf{z}_{q_i} = row_i(\mathbf{UD}) \tag{6.34}$$

and the feature weights (loadings) are given as

$$\boldsymbol{\alpha}_{q_i} = col_i(\mathbf{V}^\top). \tag{6.35}$$

The PCs have the property to be mutually uncorrelated and are ordered by their amount of variance they contribute to describe data points. The first property means that the latent geographic features have the least mutual correlation of all possible linear combinations of the candidate feature signals. This is a desired property, since it leads to highly distinctive latent geographic features. The second property induces an order on the latent geographic features. The first latent feature describes most of the variance in the spatio-temporal distributions of the candidate features. Hence, this variation will be found in most of the features. In our case, this will most likely be the background distribution, since it occurs in every signal. Hence, we can use PCA to extract the background distribution with the first latent geographic features, and look at the other $k-1$ latent features as describing distinct geographic semantics.

PCA was successfully used for dimensionality reduction for other kind of data. For example, to extract latent dimensions in microarray data [Zou et al., 2006] or in images showing faces [Turk and Pentland, 1991]. For all these kinds of data, PCA extracts a small number of dimensions that sufficiently explain the data records. These dimensions are also called prototypes [Hastie et al., 2009, p. 459], since they represent prototypical dimensions of the input data. For the microarray data, these prototypes are also called Eigen-genes [Zou et al., 2006], and for the face-image data Eigen-faces [Turk and Pentland, 1991].

We can use the same terminology and call the low number of latent geographic features the *geographic Eigen-features*, in the sense of being a prototypical dimension to describe geographic space.

Runtime Complexity. PCA is based on the SVD of the geographic feature matrix. As stated in Section 6.3.2, the runtime to decompose an $n \times p$ matrix is

$$O(np^2). \tag{6.36}$$

Hence, PCA is linear in the number of spatio-temporal cells in the lattice and quadratic in the number of feature candidates. Thus, a very large set of candidate features (in the order of thousands) should first be reduced using domain-specific feature selection strategies (such as proposed in Chapter 5) to efficiently compute the PCs.

6.4.5 ICA

Independent component analysis (ICA) is a technique to separate a multivariate signal into source components that are mutually statistical independent, rather than statistically uncorrelated [Hastie et al., 2009, p. 557]. The source components correspond to the principal components in PCA, and hence to the latent geographic feature signals. The statistical independence between the latent features, enforced by ICA, is an even more strict assumption between distributions than non-correlation. This property of ICA has been successfully exploited to distinguish individual speakers in audio signals, to reduce noise in images, for latent factor discovery in financial data, and to extract basis face representations (similar to Eigen-faces) for face recognition [Bartlett et al., 2002; Hyvärinen and Oja, 2000]. In all those applications a representation of the original signal by statistical independent source components captures essential structure in the data by a small number of dimensions.

Given a $n \times p$ data matrix $\mathbf{Z}$, ICA is described as the factorization

$$\mathbf{Z}^\top = \mathbf{AS}, \tag{6.37}$$

where $\mathbf{A} \in \mathbb{R}^{p \times p}$ is the mixing matrix, and $\mathbf{S} \in \mathbb{R}^{p \times n}$ is the source matrix. The rows of $\mathbf{S}$ represent the source components (similar to the principal components in PCA), and the columns of $\mathbf{A}$ the mixing coefficients (similar to the loadings in PCA).

An ICA algorithm needs to compute both matrices, $\mathbf{A}$ and $\mathbf{S}$. This is achieved by computing a un-mixing matrix $\mathbf{W} \in \mathbb{R}^{p \times p}$ with the property

$$\mathbf{S} = \mathbf{WZ}^\top \tag{6.38}$$

and

$$\mathbf{A} = \mathbf{W}^{-1}. \tag{6.39}$$

The FastICA algorithm to compute the un-mixing matrix as proposed in [Hyvärinen and Oja, 2000] is shown in Algorithm 6.1.

Applying ICA to a geographic feature matrix finds a set of statistically independent latent geographic features. The mapping is given as

$$\mathbf{z}_{q_i} = row_i(\mathbf{S}) \tag{6.40}$$

and

$$\boldsymbol{\alpha}_{q_i} := col_i(\mathbf{A}). \tag{6.41}$$

ICA has the property that the source components are mutually independent which corresponds to both, non-Gaussian nature and low entropy of the distributions $z_q(c), q \in Q$. ICA finds the same number of source components as there are input features. To extract a smaller number of source components the dimensionality of the input is reduced to k dimensions using PCA in a pre-processing step. Then, ICA is performed on this low-dimensional representation.

In the context of a geographic features matrix, independence between latent geographic features can be stated as follow: Two latent geographic features $z_{q_i}(c)$ and $z_{q_j}(c)$ are independent if, for all cells $c_1, \ldots, c_n$, we are not able to predict the signal $z_{q_i}(c)$ on the basis of the signal $z_{q_j}(c)$, and vice versa. Hence, the latent features should contain no redundant information about each other.

Algorithm 6.1 FastICA Algorithm.

Input:

- Transposed geographic feature matrix (input matrix) $\mathbf{X} = \mathbf{Z}^\top \in \mathbb{R}^{p \times n}$ where each column is an p-dimensional sample.
- Functions $g(\mathbf{f}), g'(\mathbf{f})$. The function $g(\mathbf{f})$ applies the transformation $\tanh(x)$ to each element of the vector $\mathbf{f}$. The function $g'(\mathbf{f})$ applies the transformation $1 - \tanh^2(x)$ to each element of the vector $\mathbf{f}$. See [Hyvärinen and Oja, 2000] for details about these choices and alternative functions.

Routine:

```
for $i \in [1; p]$
    Create vector $\mathbf{w}_i \in \mathbb{R}^{p \times 1}$ with random values
    $\mathbf{w}_i = \mathbf{w}_i / \sqrt{\sum_{j=1}^{p} w_{ij}^2}$
    while $\mathbf{w}_i$ changes
        $\mathbf{t} = \mathbf{w}_i^\top \mathbf{X}$
        $\mathbf{w}_i = \frac{1}{n}\mathbf{X} g(\mathbf{t})^\top - (\frac{1}{n}\sum_{j=1}^{n} g'(\mathbf{t})_j)\mathbf{w}_i$
        $\mathbf{w}_i = \sum_{j=1}^{i} \mathbf{w}_i^\top \mathbf{w}_j \mathbf{w}_j$
        $\mathbf{w}_i = \mathbf{w}_i / \sqrt{\sum_{j=1}^{p} w_{ij}^2}$
```

Output: $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_p)$

Other than PCA, the components of ICA have no natural order. If the number of extracted latent geographic features is small, such that an analyst can easily traverse through the complete result, ICA will work as an exploratory tool. However, if the number of latent features is large, the analyst should be guided by a score of the meaningfulness of the latent features. In such a setting, the extracted latent geographic features can be ordered according to a given criterion. For example, by the similarity to the background distribution (see Section 5.2), by the aggregated dominance of the features in the input data (see Section 3.5.2), or by the representativeness of the signal in a spatio-temporal window (see Section 3.5.3).

Runtime Complexity. In [Hyvärinen and Oja, 2000], the authors propose a fix-point iteration scheme to compute the independent components called FastICA. The authors do not specify the runtime complexity of the algorithm explicitly, however, we state the following runtime complexity on the basis of Algorithm 6.1

$$O(pl(np + np + p^2 + p)) = O(pl(2np + p^2 + p)) = O(p^2 l(n + p)), \tag{6.42}$$

where l is the number of iterations needed for convergence of the $\mathbf{w}_i$s, n is the number of cells in the lattice, and p is the number of features. The number of iteration to converge is usually very small (in the order of 10), as stated by the authors and observed by our experiments using the *scikit* implementation[1].

To extract the independent components, the input matrix must first be reduced by PCA. Hence, the total runtime of ICA is

$$O(np^2 + k^2 l(n+k)), \tag{6.43}$$

where k is the number of components to be found. Note that $k \ll p$. Hence, the runtime of the FastICA routine will become much faster resulting in a small runtime overhead compared to PCA.

6.4.6 Sparse PCA

In PCA and ICA, the resulting latent geographic features are assumed to be combinations of all candidate features. Hence, a latent geographic feature will have a non-zero feature weight for almost all of the features, with many of the feature candidates having weights close to zero. From an exploratory point of view it would be more meaningful to represent a latent geographic feature by only a few high-informative feature candidates.

From a statistical point of view, such a sparsity assumption has many advantages. High-dimensional statistics suffers from the problem that not enough observations exist to describe the dependencies between a large number of variables. A recently introduced concept is the assumption of sparsity of the underlying latent dimensions. Hence, a latent dimension is only made of a few variables with weights different from zero. Given this assumption, less data is needed to estimate the model [Peter and Van de Geer, 2011].

In SPCA, the PCs are described by a maximum number of v variables, with $v \ll p$. SPCA was introduced in [Zou et al., 2006]. The authors propose a regularized linear regression approach to estimate the principal components and the loadings, called the *elastic net*. This approach expects a vector of sparsity parameters

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k), \tag{6.44}$$

where λ_i denotes the number of features that should have a non-zero loading for the principal component i. Interestingly, these parameters induce a trade-off between sparsity and the amount of variance that is captured by the components [Zou et al., 2006]. Hence, the analyst can decide to let the principal components capturing most of the variance (large λ_i) or to have a more meaningful interpretation according to the feature candidate weights, but slightly redundant latent geographic feature results.

The mapping of sparse PCs to latent geographic features is the same as in PCA. The sparsity assumption results in latent geographic features $(\mathbf{z}_q, \boldsymbol{\alpha}_q)$ having only v or less features with a weight different from zero in $\boldsymbol{\alpha}_q$.

[1] https://github.com/scikit-learn

Algorithm 6.2 Sparse PCA Algorithm.

Input:

- A centered geographic feature matrix $\mathbf{Z} \in \mathbb{R}^{n \times p}$.
- A vector of sparsity numbers $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$ for each principal component.
- A global penalty λ being set to a small value.

Routine:

```
Compute the principle component loadings $\mathbf{A} = (a_1, \ldots, a_k) = \mathbf{V}^\top$ using the SVD $\mathbf{Z} = \mathbf{UDV}^\top$

Until convergence of $\mathbf{A}$

    for $j \in [1; k]$
        $\beta_j = \underset{\beta}{\operatorname{argmin}} (a_j - \beta)^\top \mathbf{Z}^\top \mathbf{Z} (a_j - \beta) + \lambda ||\beta||^2 + \lambda_j ||\beta||_1$
    Set $\mathbf{B} = (\beta_1, \ldots, \beta_k)$
    Compute SVD of $\mathbf{Z}^\top \mathbf{Z} \mathbf{B} = \mathbf{UDV}^\top$
    Update $\mathbf{A} = \mathbf{UV}^T$
```

Runtime Complexity. The algorithm to estimate the sparse principal components is shown in Algorithm 6.2. The runtime of SPCA using elastic net regularization as reported in [Zou et al., 2006] is

$$O(lk(pvn + v^3)), \tag{6.45}$$

where l is the number of iterations needed until convergence, k is the number of latent features, p is the number of candidate features, v is the number of non-zero coefficients, and n is the number of spatio-temporal cells. As for the previous approaches, the algorithm heavily depends on the number of iterations needed until convergence. However, as the authors state and as observed in our experiments, the number of iterations is very small. Note that for a high sparsity (small v) the algorithm is efficient even for a huge number of features p.

6.5 Comparative Experiments

In this section, we present a comparison of the different geographic phenomenon discovery approaches introduced in the previous section. For this, we first describe the user-generated data sources used for evaluation and detail the comparison criteria. Then, we conduct a qualitative evaluation of all approaches with respect to the informativeness and the distinctiveness of the extracted phenomena. Afterwards, we focus on a comparison between the proposed dimensionality reduction approaches and then, quantitatively

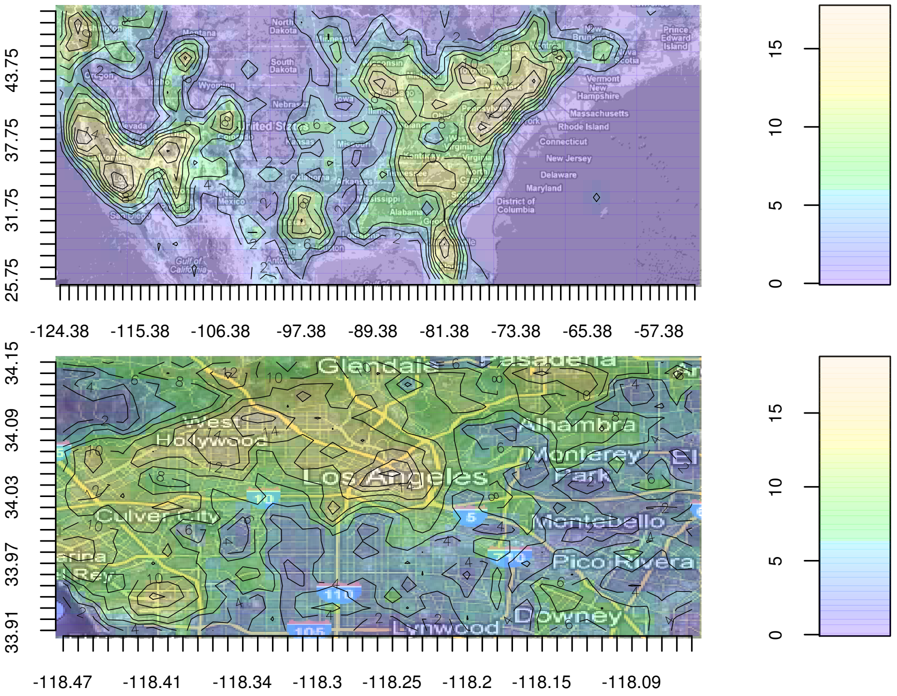


Figure 6.1: Total number of distinct users of the US and LA Flickr data set in log-scale aggregated using a geo-rectangular grid with $\delta_S = 1.0$ (111 km) and $\delta_S = 0.01$ (1.11 km), respectively.

evaluate the impact of normalization on the resulting phenomena types. Finally, we use the most promising technique (SPCA) for an exploratory analysis of the Los Angeles (LA) area. The results show that SPCA-based phenomenon discovery is a valuable tool for data analysis and geographic knowledge discovery.

In our experiments we only focus on the extraction of spatial phenomena. Hence, we use only a single infinite interval to describe the temporal dimension, resulting in purely spatial signals. The reasons for this restriction are as follows: (1) The number of spatio-temporal bins would be much larger in a spatio-temporal setting. To extract robust spatio-temporal feature signals we would need even more data than used for the subsequent comparison (in the order of millions for the US area). At the time of conducting the experiments, we did not have access to such a huge data set. (2) Spatio-temporal phenomena are much harder to describe and interpret. Hence, for evaluating the qualitative performance, a focus on spatial phenomena allows for a more meaningful comparison.

	Flickr US	Flickr LA
# rec m	5,976,689	245,312
# features p	30,323 (695 filtered)	58,141 (3235 filtered)
# users	3,081	5,796
Area D_S	US bounds (-124.87,25.25,-52.61,50.06)	LA city bounds (-124.87,25.25,-52.61,50.06)
Interval D_T	[Jan-2008, Jan-2011)	[Jan-2010, Jan-2012)
Grid	$\delta_S = 1.0$ (111 km) $72 \times 24 = 1728$ cells	$\delta_S = 0.01$ (1.11 km) $43 \times 25 = 1075$ cells

Table 6.1: Flickr data set statistics.

6.5.1 Data and Setup

As input data we use geo-referenced documents retrieved by the Flickr API[2]. For this, we collected photos from a query over the US area within the time interval Jan-2008 to Jan-2011, and from a query over the Los Angeles (LA) area for the time interval Jan-2010 to Jan-2012. The details of the data sets are given in Table 6.1.

We create a geographic feature matrix $\mathbf{Z}_{L,F}$ using the count-based user model proposed in Section 4.3.2. The spatial lattice is based on a geo-rectangular grid. We use a cell-width $\delta_{US} = 1.0$ for the US-area and $\delta_{LA} = 0.01$ for the LA-area, resulting in a lattice of size $n_{US} = 1728$ and $n_{LA} = 1075$. As features, the tags associated with the geo-referenced photos have been used. Those textual tags are mostly self-describing, allowing for a meaningful interpretation of the phenomenon semantics. From both data sets we removed those features (tags) that occur in less than 5 cells and whose user contribution over all cells is less than 10. This results in $|F_{US}| = 695$ and $|F_{LA}| = 3235$ features. Figure 6.1 shows the spatial distribution of user counts for both data sets.

We do not use a background-normalized input (such as proposed in Chapter 4), to (1) test the intrinsic ability of the models to cope with a background-polluted distribution, and (2) to allow for a fair comparison to the document-centric approaches (which are not able to use a priorly normalized signal). Note that this pre-processing ability is a great advantage of the latent geographic feature extraction approaches and relies on the input being a pre-processed geographic feature matrix, instead of the documents themselves.

6.5.2 Comparison Criteria

As mentioned in Section 6.2, geographic phenomenon discovery should result in a small number of informative and distinct geographic phenomena. Since a phenomenon is represented by a spatial signal and a description (extracted from the features having the highest weight), the evaluation should take both of them into account. In the following we clarify the criteria to judge about good and bad results:

[2] http://www.flickr.com/services/api

- *Discriminative spatial signal*: The extracted spatial signal of a phenomenon should clearly describe high and low intensity areas. A good signal will clearly assign a high value to those regions where the phenomenon occurs, and zero to the other parts. Moreover, high intensity areas should be contiguous, without gaps or noisy peaks.

- *Meaningfulness of description*: The description, extracted from the features of highest weight, should represent a meaningful geographic process or entity. We distinguish between features having landmark semantics and regional semantics (see Chapter 5). Consequently, the phenomena can be more landmark-ish or regional. Landmark semantics are represented by placename features, such as city or country names ('newyork', 'california'). Regional semantics are represented by attribute features such as 'nature', 'water', 'mountain', etc. Furthermore, global features are supposed to have no geographic meaning in the area of interest at all. In Figures 6.2 to 6.8 we grayed-out the global features, since they are supposed to have no value to describe the semantics of a phenomenon. A phenomenon having a huge number of global features with high weight is considered a bad result. The following tags are supposed to have global semantics within the US and/or LA area:

  > Global area features: 'us', 'usa', 'america', 'unitedstates', 'northamerica' (in the US area), additionaly, 'la', 'los', 'angeles', 'losangeles', 'california' (in the LA area).
  >
  > Years: '2008', '2009', '2010, '2011'.
  >
  > Camera names and photo properties: 'nikon', 'canon', 'iphone', 'olympus', 'fuji', 'nikkor', bw', 'color', 'photo', 'image', 'hdr', 'd90', 'd700', 'd5000', 'geotagged'.
  >
  > Service names: 'instagram', 'instagramapp', 'iphoneography', 'hipstamatic'.

- *Signal-description coherence*: The extracted signal and the description should belong to each other. If a description contains placenames and the signal has no intensity at these locations, this is considered a bad result.

- *Phenomenon redundancy*: Within the set of extracted phenomena, their semantics (represented by the description-signal combination) should be distinct from each other. A result with many redundant phenomena is considered a bad result.

6.5.3 Parameter Selection

All of the reviewed and newly introduced approaches for geographic phenomenon discovery need a parameter k to specify the number of phenomena to be extracted. This important parameter has a direct impact on the quality of the results. Choosing a too large k will lead to a huge number of non-informative and redundant phenomena, choosing a too small k results in a few phenomena that reflect the landmarks with the most user contributions (e.g., New York, Los Angeles), along with a phenomenon reflecting the background distribution.

Basically, choosing k can be done by evaluating the results over a range of choices. Given an appropriate quantitative measure, this selection process can be automated by an exhaustive search over a fine-grained parameter space. However, for the problem of geographic phenomenon discovery, we have not been able to develop a suitable quantitative measure. Instead, we presented a number of qualitative criteria to judge about the informativeness of the results in the previous section. For the following experiments, we run the phenomenon discovery tasks using a small number of promising parameter choices,

$$k \in (10, 20, 50, 100). \tag{6.46}$$

We selected the parameter that results in the most informative results using ICA and SPCA. Then, we compared the results by using the same parameter value for all approaches. We also tested other parameter choices for the LGTA and the topic-pivot LDA approach. However, we have not been able to extract more informative results than by using the parameter as detected above.

In general, finding a promising quantitative measure to evaluate the quality of geographic phenomenon discovery is a highly interesting direction of future research. We discuss this problem in the conclusions in Section 7.2.

6.5.4 Comparison of all Approaches

We use a user-count based geographic feature matrix $\mathbf{Z}$ as input for the dimensionality reduction and clustering approaches. For the document-centric approaches (LDA, LGTA) we use the geo-referenced tag sets directly. Thereby, we remove those features (tags) from the sets that have been removed from the geographic feature matrix. We found a parameter value of $k = 20$ to result in the most informative phenomena for the two data sets.

We use the identifier <approach-n> to refer the the nth extracted phenomenon of a particular approach. The corresponding phenomena are given in the Figures 6.2 to 6.8.

PCA. By using PCA, the extracted geographic phenomena are ordered by the amount of variance they describe in the geographic feature space. Because of this, the phenomenon PCA-1 follows the background distribution of the data, since locations are primarily distinguished by the intensity of signal. The description indicates that 'newyork', 'california', 'sanfransico', and 'city' are dominant in the records and closely describe the total variation of the spatial signal.

In the other extracted phenomena we find landmarks, e.g., PCA-2 (California), PCA-6 (Florida), PCA-7 (Canada), or PCA-18 (Arizona). However, some phenomena have a mixed landmark semantics, e.g., PCA-3 (Chicago, LA, Washington) or PCA-4 (Chicago, San Franciso). Moreover, the landmark features occur redundantly in several phenomena, e.g., 'sanfrancisco' in PCA-1, PCA-2, and PCA-4.

The spatial signals are not very discriminative. They show several positive and negative peaks, which prevents segmenting the signal into high and low intensity areas. This is based on the fact that PCA extracts signals with strong positive and negative values.

ICA. Other than for PCA, the ICA results show highly discriminative spatial signals. Each signal has clear high and low intensity areas. All phenomena describe small to medium sized landmarks, e.g., ICA-1 (Pennsylvania), ICA-2 (Seattle), ICA-3 (Florida), ICA-4 (Oregon), etc. The descriptions show a mixture of landmark features (place names) and regional features ('landscape', 'nature', 'beach', etc.) that are characteristics for the area. This shows that ICA combines feature candidates with different distributions, other than just picking landmark features. Finally, the resulting phenomena are highly distinct from each other, and the signal-description combinations are coherent.

SPCA. Similar to the PCA result, the phenomenon SPCA-1 can be seen as describing the background distribution. The remaining SPCA results mainly represent landmark phenomena, such as SPCA-2 (California), SPCA-3 (Chicago), SPCA-4 (San Franciso), etc. The landmark phenomena are highly distinct from each other.

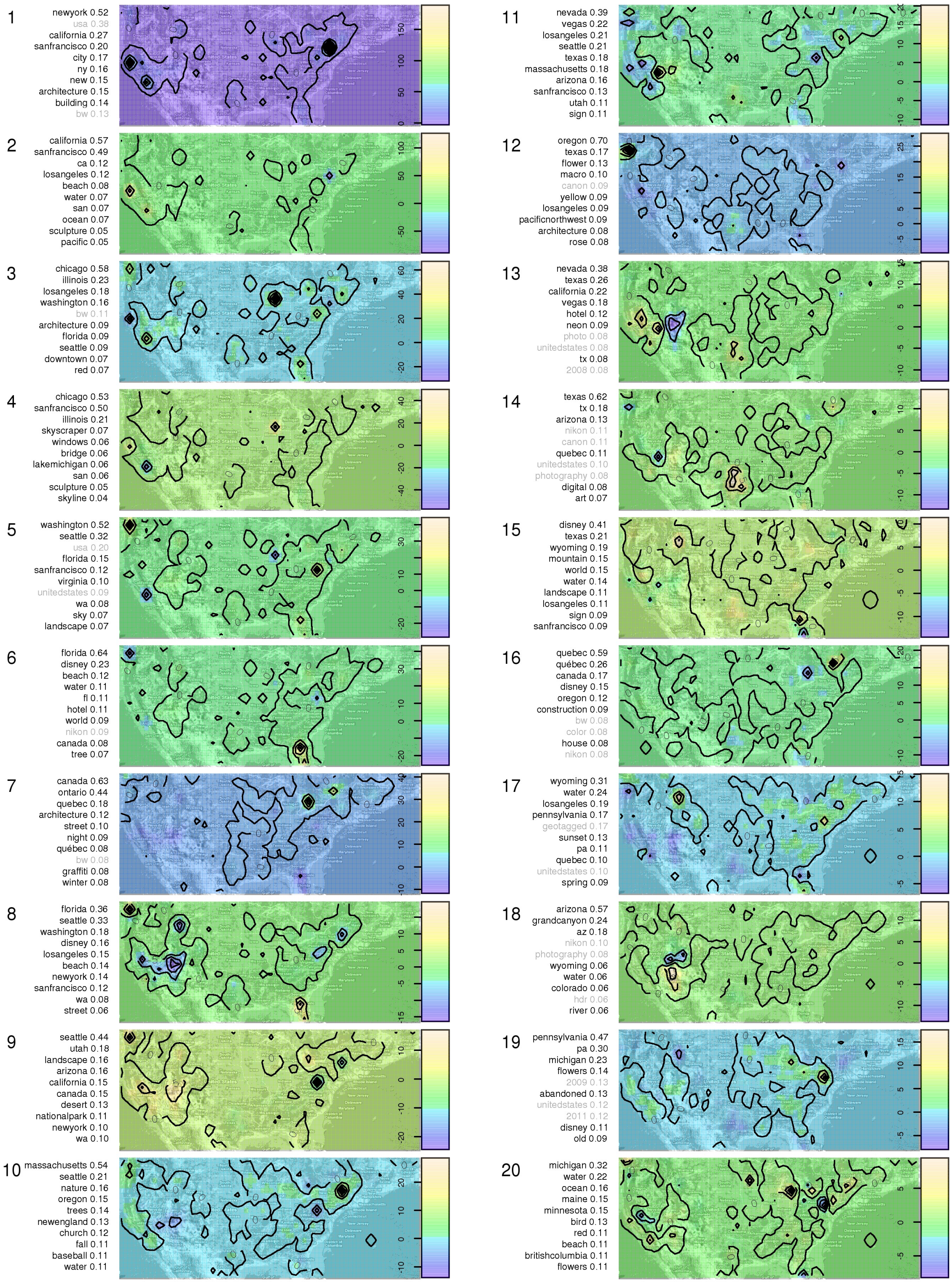
Since this approach uses a sparsity constraint, the resulting descriptions only have a small number of features with a weight different from zero. SPCA-16 (Quebec landmark feature) even only has a single weighted feature. Compared to PCA and ICA, this results in more compact descriptions that are easier to interpret. Also, compared to PCA, this results in much more meaningful descriptions in general. Different from the ICA result, there exists a phenomenon SPCA-17 ('water','landscape','nature') with pure regional characteristics. Such regional phenomena are important if space should be segmented into categories. Only a single phenomenon (SPCA-5) shows a mixed landmark semantics. The spatial signals are highly discriminative, as they are in the ICA result.

K-Means. The K-Means result shows highly discriminative signals. Mostly, only a single peak is shown with the remaining area being zero. The semantics of the phenomena are of the landmark and regional type. A major difference to the previous results is the number of 6 phenomena having only a single feature, and 3 having only 2 features with a weight different from zero. These phenomena with highly sparse descriptions are landmarks with the feature being the respective placename features. However, these landmarks occur redundantly, such as in K-Means-9 and K-Means-13 (New York), K-Means-3 and K-Means-10 (Chicago/Illinois). Moreover, the signals of K-Means-1, K-Means-2, and K-Means-14 are highly similar but the descriptions are very different. K-Means-12 shows a mixed landmark semantics (Washington and Seattle). Despite the fact that K-Means extracts highly discriminative signals and sparse descriptions, the result is poor because of the small number of distinct extracted phenomena and the ambiguity of descriptions for similar spatial signals.

LDA. The spatial signal of the LDA result are not discriminative at all. Every phenomenon signal is described by an almost random number of peaks over the whole area of interest. Also, the descriptions have no intuitive interpretation as geographic phenomena. This can be expected, since the extraction of feature combinations is performed without taking the spatial proximity of the records into account (topic-pivot approach).

LGTA. For LGTA, we show results using two parameterizations. In LGTA-A, the record-feature counts are used directly without taking the frequencies of the features in the corpus into account ($\lambda = 0$). In LGTA-B, the feature counts are down-weighted by the total feature frequency ($\lambda = 0.5$). This limits the influence of features that occur often in the records. The results of both LGTA approaches show discriminative spatial signals. The extracted LGTA-A phenomena mainly show landmark signals. However, the corresponding descriptions often do not relate to them. For example, LGTA-A-4 shows a peak at Chicago, however, the description has high weight for the features 'flowers', 'nature', 'water', 'lake'. Similarly, LGTA-A-3 shows a strong peak at California, however, the features with high weight are 'florida' and 'pennsylvania'. The LGTA-B results are similar. The phenomena show less weight for frequently used place names (such as 'newyork', 'california', etc.). However, the resulting phenomena still have a mixed semantics and non-coherent signal-description combinations. We find this result surprising, since the generative model is explicitly meant to model spatial topic distributions. However, for the given Flickr data set we have not been able to extract meaningful phenomena.

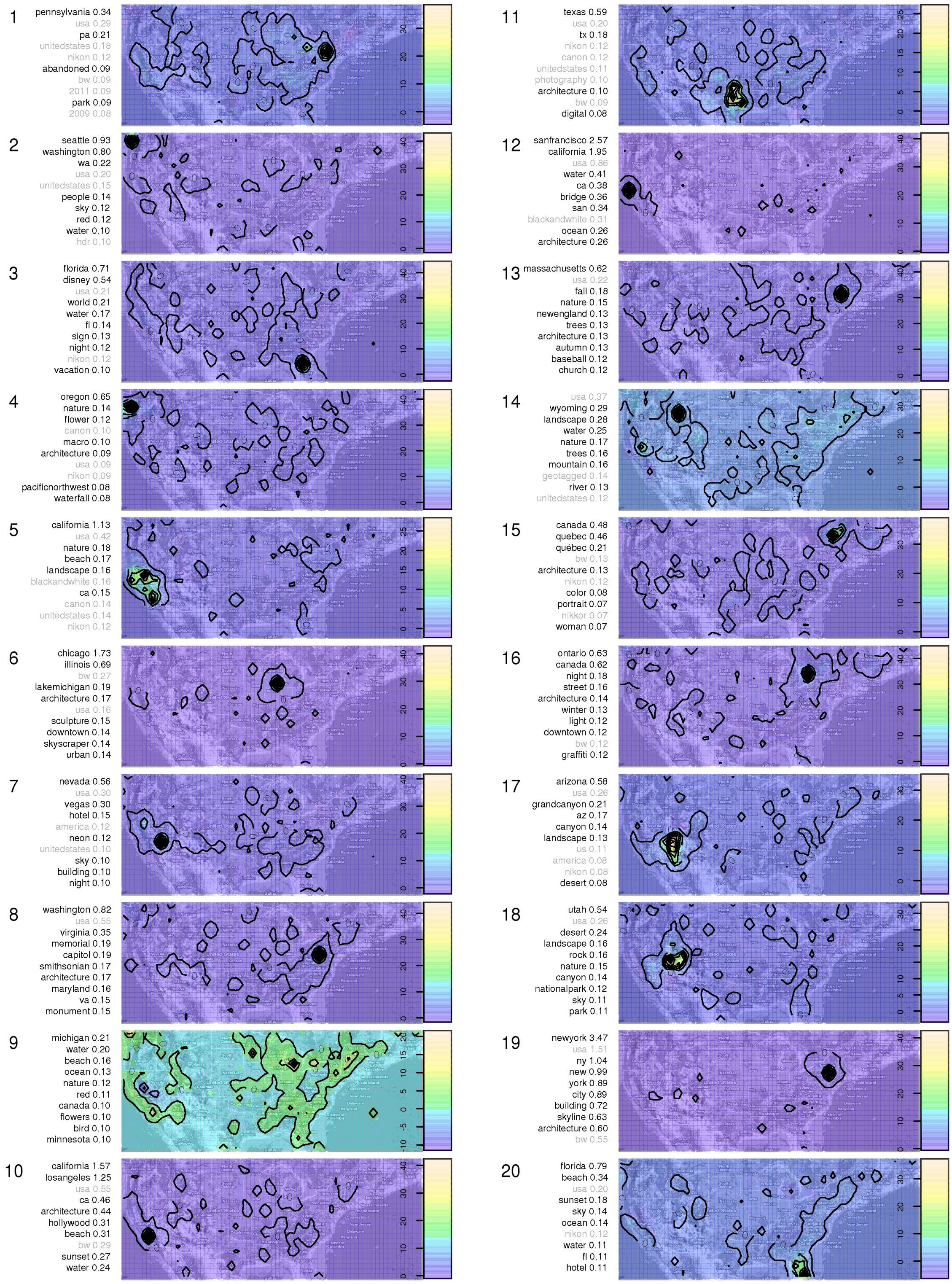
Evaluation. Given the results from the different approaches, we find ICA and SPCA to perform best. The resulting phenomena have meaningful semantics, the spatial signals are highly discriminative, and the signal-description combinations are coherent. Interestingly, PCA fails to extract good phenomena, despite being very similar to SPCA and ICA. Hence, the sparsity constraint as well as the independence assumption between signals turn out to be promising modeling concepts. K-Means performed well in extracting single features with landmark semantics. However, the resulting phenomena are not distinct from each other. Moreover, K-Means extracts phenomena with similar spatial signals but different descriptions, which shows that this approach fails in suitably combining the feature candidates. For our used Flickr data set, LDA fails completely in extracting informative phenomena and LGTA lacks in signal-description coherence and distinctiveness.



(a) Phenomena 1-10

(b) Phenomena 11-20

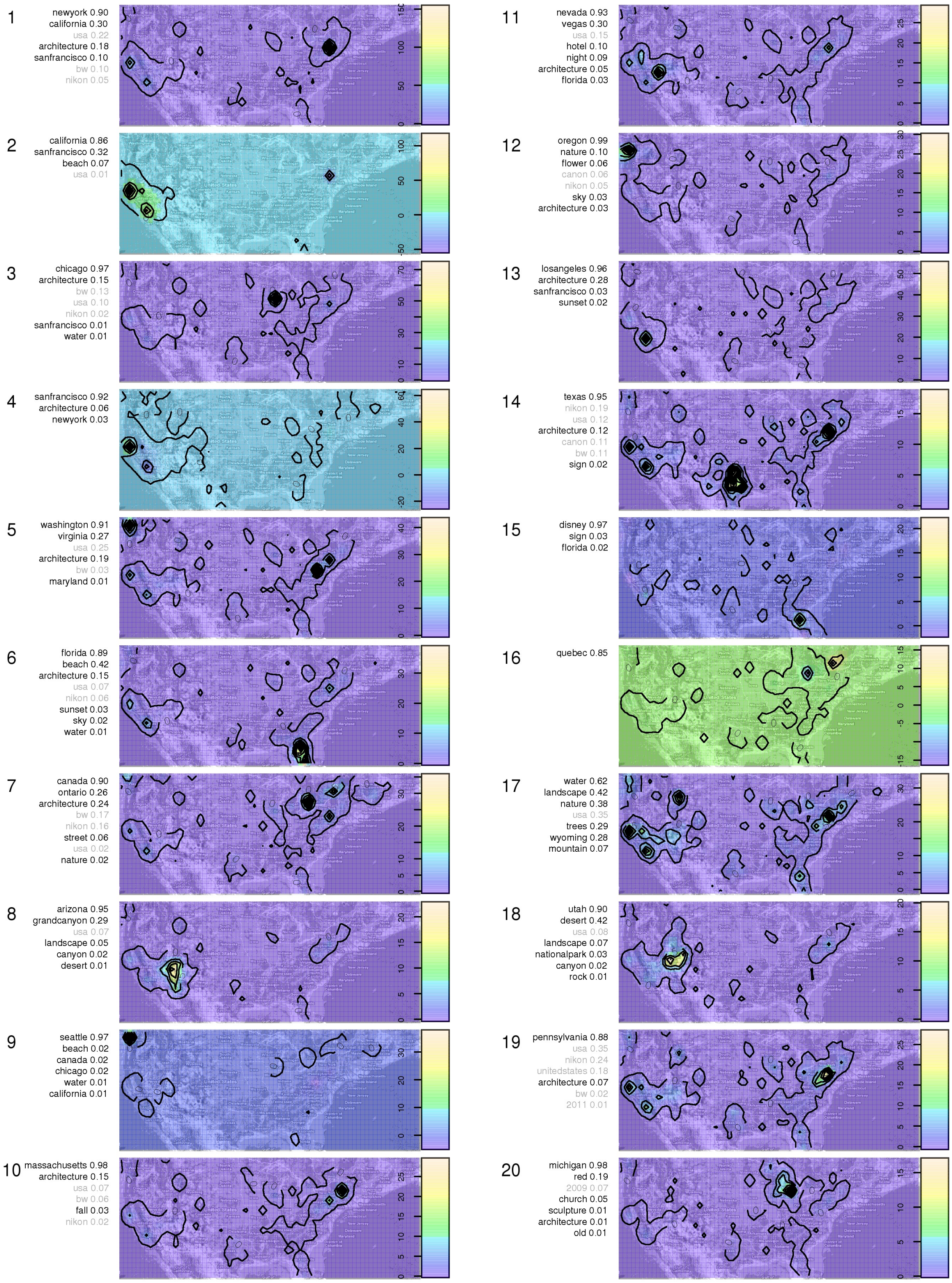
Figure 6.2: All extracted PCA phenomena from US Flickr data set ($k = 20$).



(a) Phenomena 1-10

(b) Phenomena 11-20

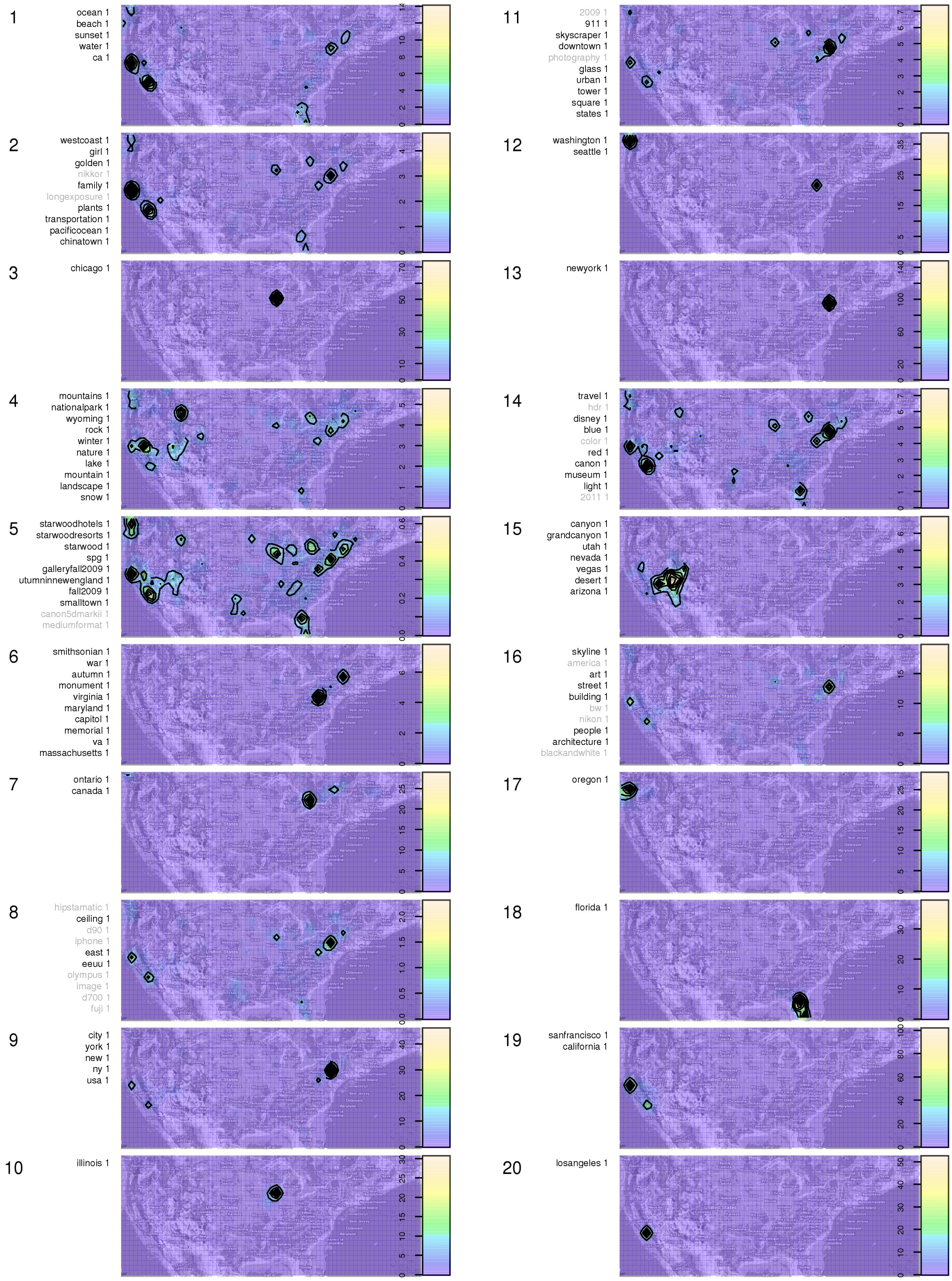
Figure 6.3: All extracted ICA phenomena from US Flickr data set ($k = 20$).



(a) Phenomena 1-10

(b) Phenomena 11-20

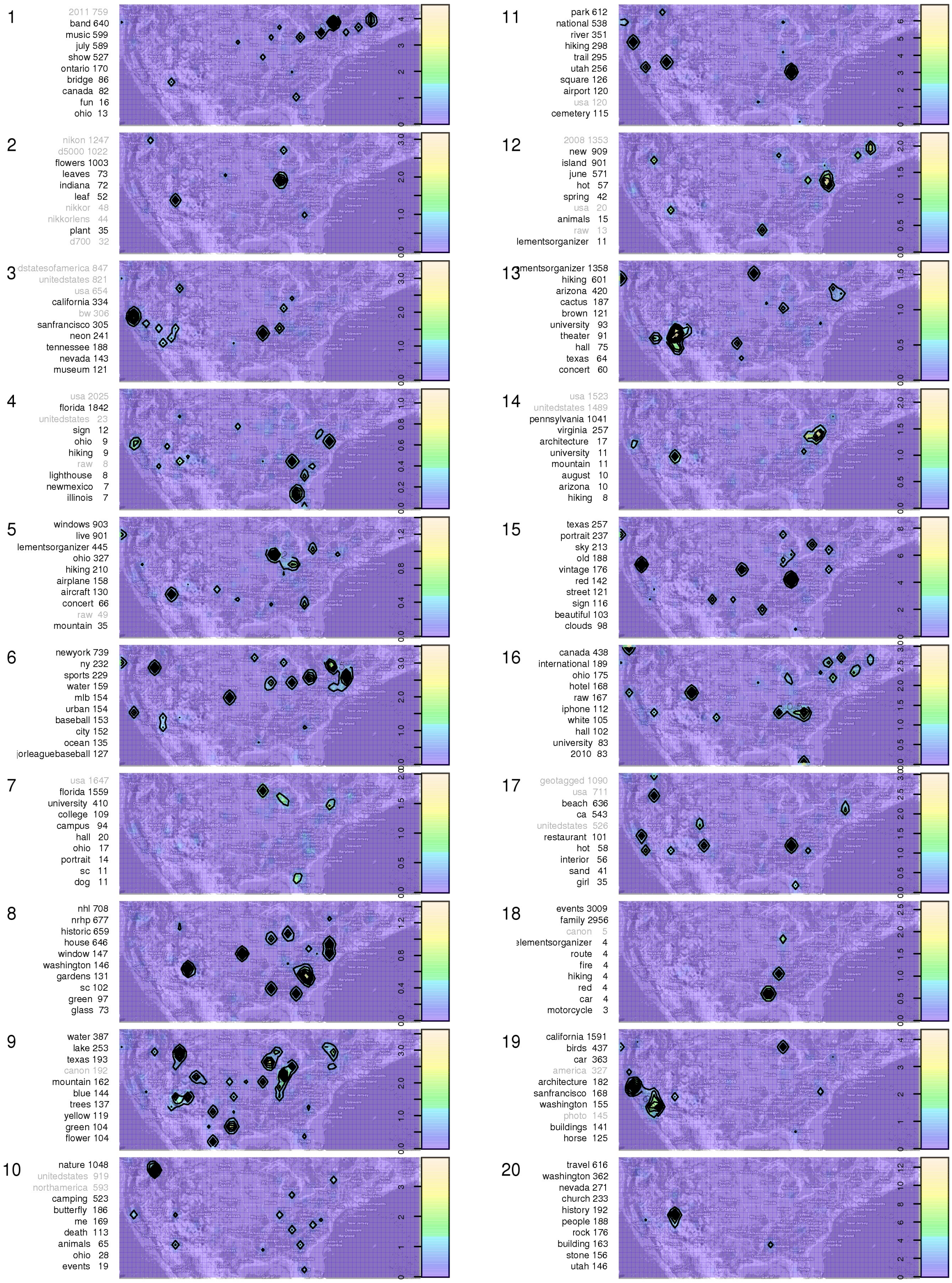
Figure 6.4: All extracted SPCA phenomena from US Flickr data set ($k = 20$).



(a) Phenomena 1-10

(b) Phenomena 11-20

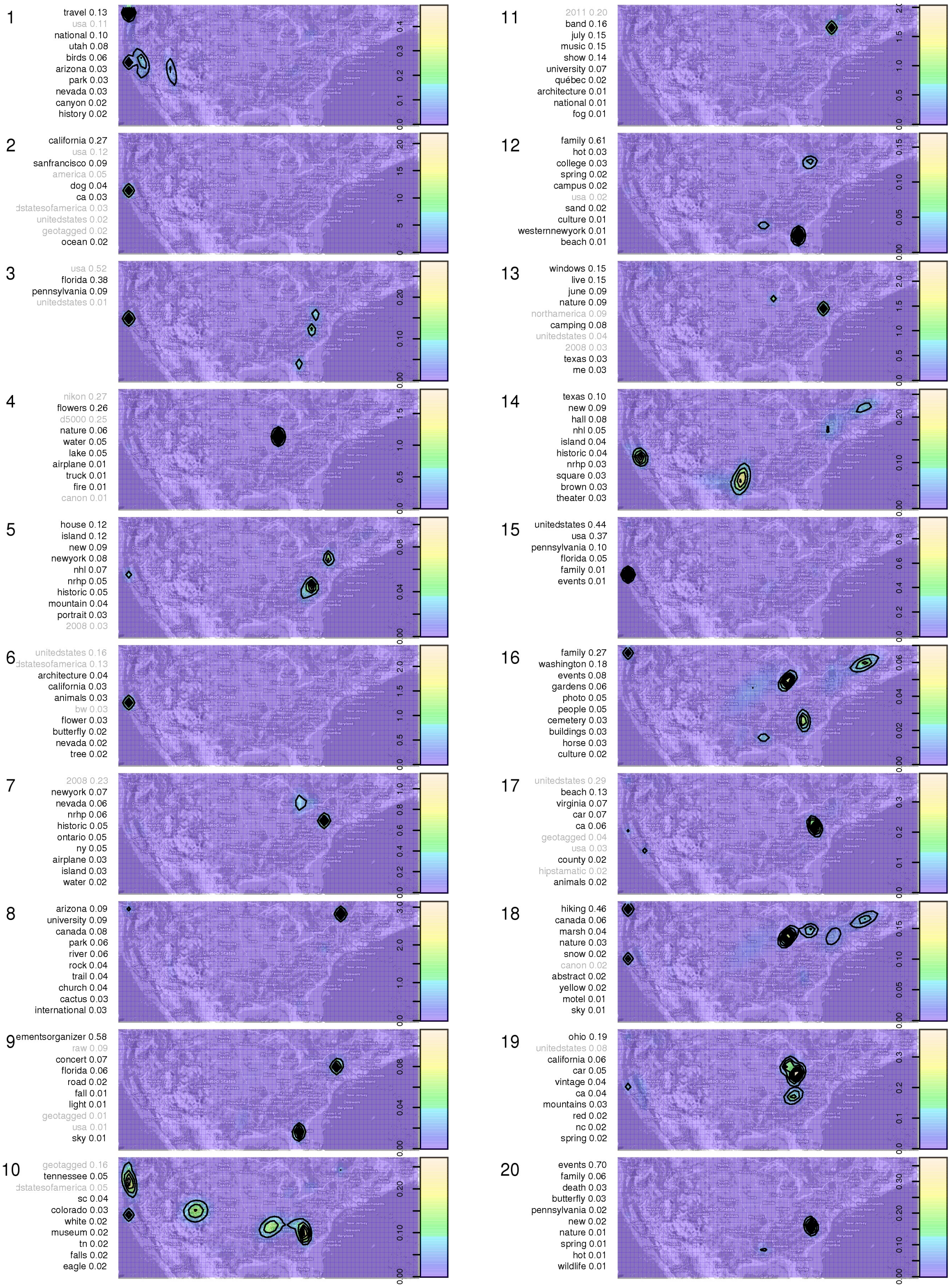
Figure 6.5: All extracted K-Means phenomena from US Flickr data set ($k = 20$).



(a) Phenomena 1-10

(b) Phenomena 11-20

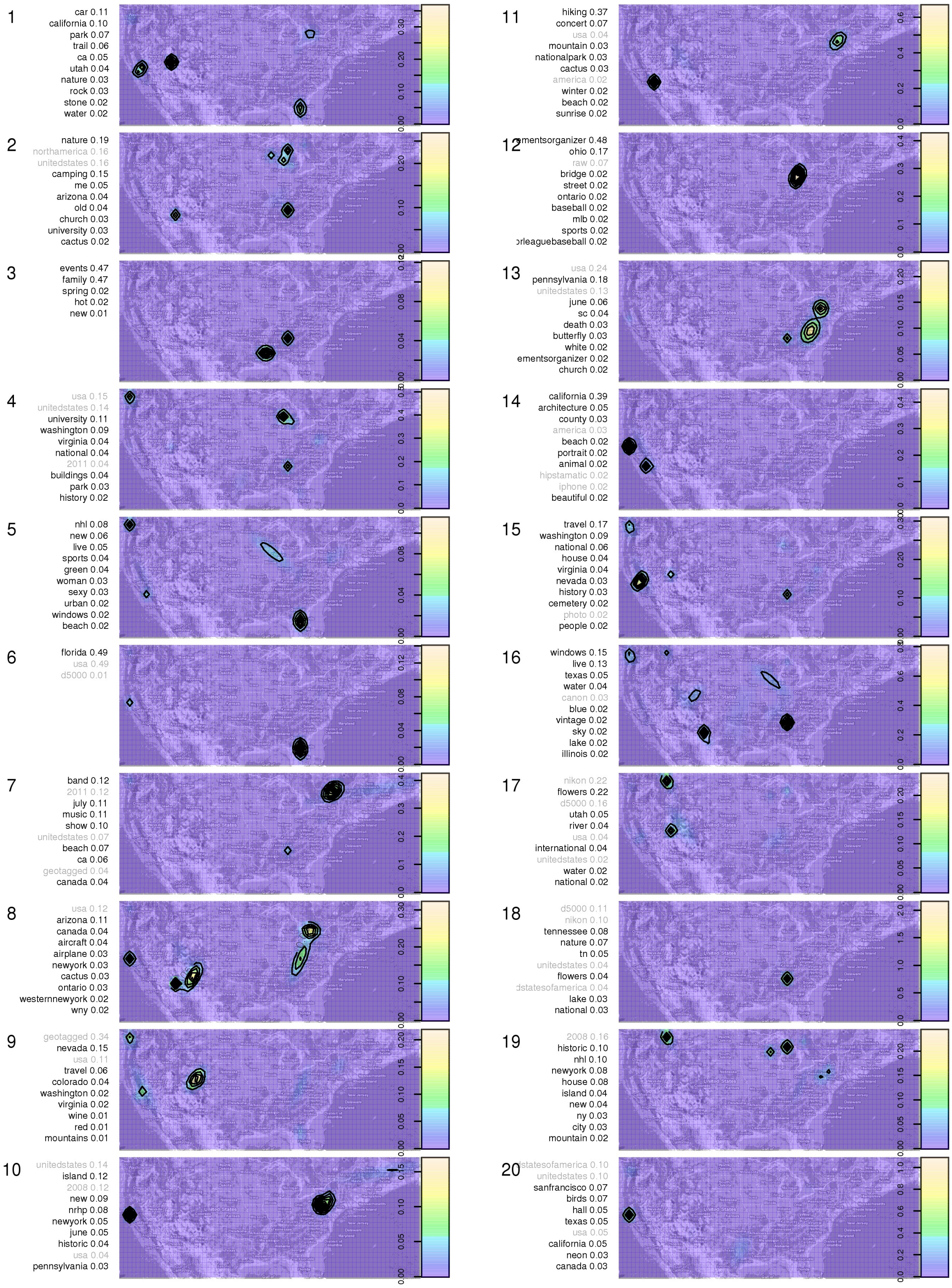
Figure 6.6: All extracted LDA phenomena from US Flickr data set ($k = 20$).



(a) Phenomena 1-10

(b) Phenomena 11-20

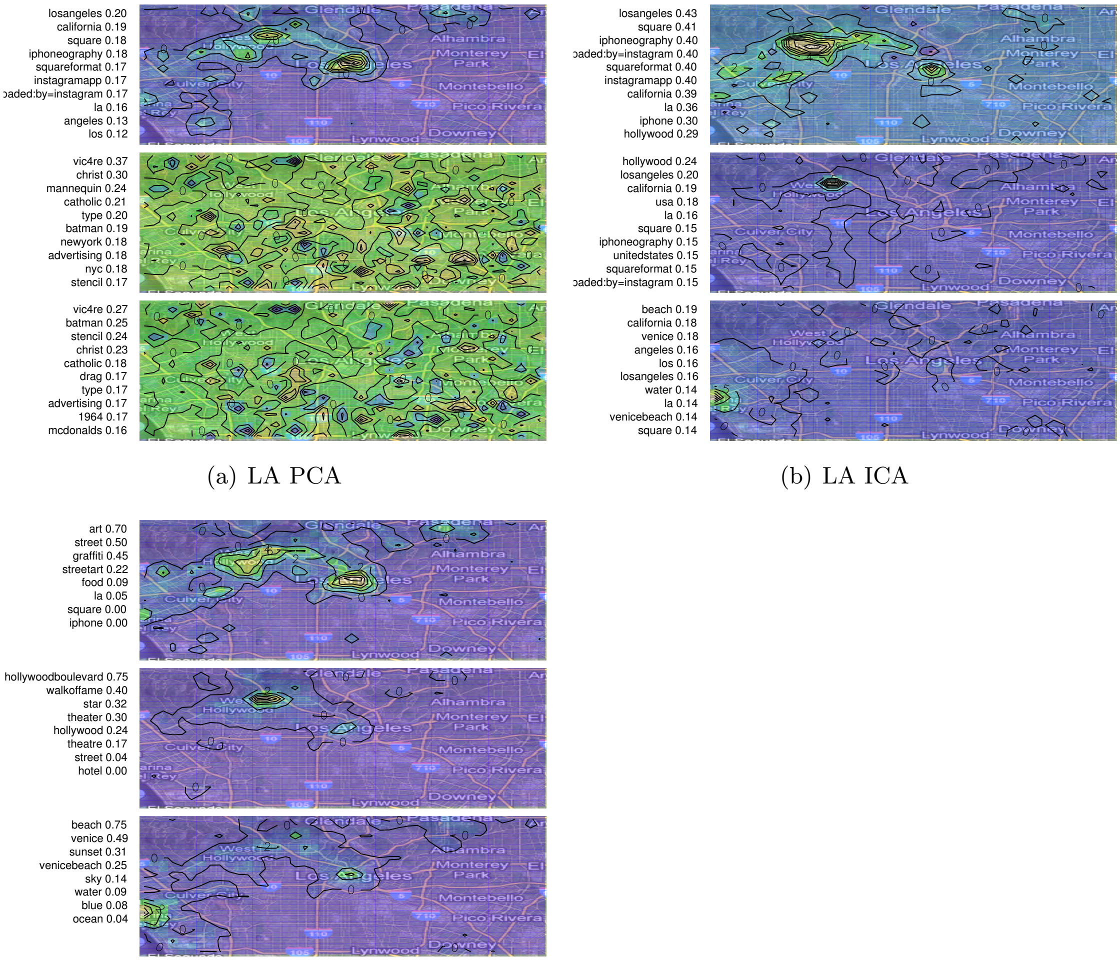
Figure 6.7: All extracted LGTA phenomena from US Flickr data set $(k = 20, \lambda = 0)$.



(a) Phenomena 1-10

(b) Phenomena 11-20

Figure 6.8: All extracted LGTA phenomena from US Flickr data set ($k = 20, \lambda = 0.5$).



(a) LA PCA

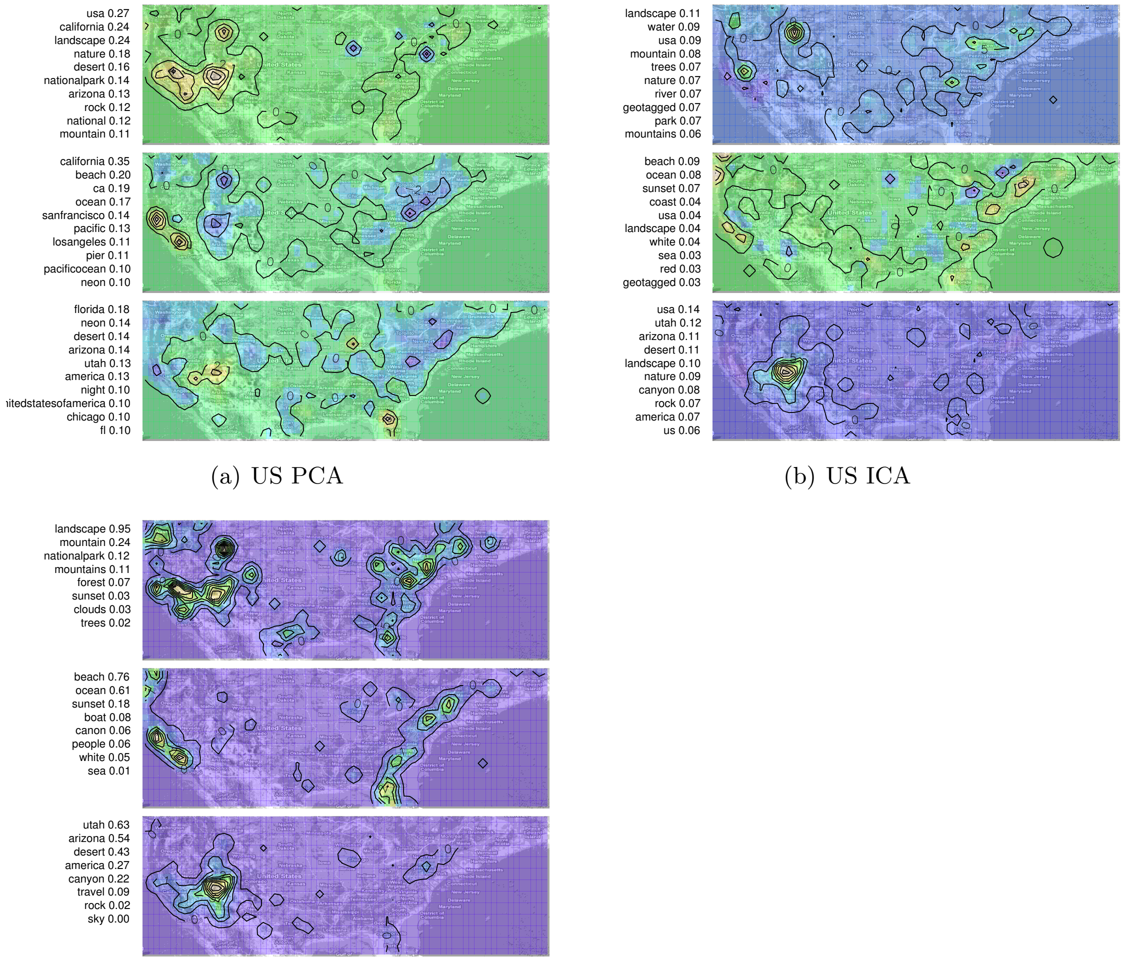
(b) LA ICA

(c) LA SPCA

Figure 6.9: Three selected phenomena extracted from the LA Flickr data set using PCA, ICA, and SPCA ($k = 20$). See Section 6.5.5 for a discussion about the unique characteristics.

6.5.5 Dimensionality Reduction Comparison

We now compare the three dimensionality reduction approaches PCA, ICA, and SPCA to evaluate the impact of the different statistical assumptions. For this, we select three phenomena from the SPCA results using $k = 20$ and find the most similar counterparts in the PCA and ICA results on the basis of their spatial signals. Then, we compare the descriptions to each other. We now use a log-normalized input signal. Figure 6.9 and Figure 6.10 show the three selected phenomena for each technique and for both data sets, respectively.



(a) US PCA

(b) US ICA

(c) US SPCA

Figure 6.10: Three selected phenomena extracted from the US Flickr data set using PCA, ICA, and SPCA ($k = 20$). See Section 6.5.5 for a discussion about the unique characteristics.

In the LA data set, the selected SPCA-phenomena candidates can be described as having *city*, *hollywood*, and *(venice) beach* semantics. For PCA, we have to choose the first three extracted phenomena ordered by their variance, because only the first phenomenon has a meaningful signal. This can be explained by the fact that the geographic feature matrix $\mathbf{Z}$ is under-specified ($p > n$). Although it is possible to extract principal components, it is not well-defined for such cases. This results in noisy and non-informative phenomena except for the first one, which just follows the background distribution.

ICA extracts more informative phenomena and the selected candidates from the SPCA results are found. We see that the city phenomenon is described by features having global semantics ('losangeles','square'). Consequently, it is similar to the background distribution (see Figure 6.1 for baseline in log-scale). The hollywood phenomenon has a signal around the Hollywood area. It is only described by a single representative feature ('hollywood'), with the rest of the features having global semantics. The beach phenomenon has a signal around Venice beach (lower-left of the map) and is described by the features 'beach', 'venice', 'water' and 'venicebeach'. Less representative global features are 'california' and 'losangeles'.

Differently, SPCA extracts highly informative phenomena with distinct characteristics. The features are less polluted by global features like 'losangeles', 'california', or 'instagramapp'. The city phenomenon features are highly informative for an inner city environment ('art', 'street', 'graffiti', 'food'). The hollywood phenomenon features are highly specific for this area ('hollywoodboulevard', 'walkoffame', 'star'). Furthermore, SPCA also extracts other phenomena close to the Hollywood area (universal studio, griffith park; see Figure 6.13), which we will present in Section 6.5.7. The beach phenomenon is highly informative for the coastal environment found at Venice beach ('beach', 'sunset', 'sky', 'blue', 'ocean') and is not polluted by global features at all.

From the US data set we select the SPCA-phenomena *nature*, *coast*, and *desert region* for comparison. Here, the geographic feature matrix $\mathbf{Z}_{L,F}$ has a smaller number of features, with $n > p$, resulting in more meaningful results for PCA. PCA extracts a nature phenomenon with dominant signal on the west coast. It is strongly influenced by global features ('usa', 'california'). ICA extracts the nature phenomenon with a signal better distributed over the US, with high intensities in the mountain regions. Still, the phenomenon shows high weights for non-informative global features ('usa','geotagged'). SPCA extracts the phenomenon with a signal distributed evenly in the mountain areas. All features are highly informative for the regions ('landscape', 'mountain', 'nationalpark', among other). The same observations hold for the extracted coast phenomenon and the desert region phenomenon.

Evaluation. For the input data sets and the selected phenomena we see that SPCA results are highly informative regarding their signal distribution and their feature weights. ICA performs better than PCA, with PCA being useless if the geographic feature matrix is not of full rank ($p > n$). This shows that modeling the feature weights as sparse vectors plays a key role in extracting informative phenomena.

6.5.6 Normalization

Normalization of the geographic feature matrix has a strong impact on the characteristics of the resulting phenomena. Without normalization, the discovered phenomena are dominated by features having intense peaks. Those features are found at locations with high user contributions (large cities, populated places) and mostly describe landmarks (city-names, place-names). The extracted phenomena then have landmark characteristics with features being informative for this area.

Logging reduces the impact of the peaks, resulting in less impact of these landmark features. The resulting phenomena are hence less dominated by landmark-ish features.

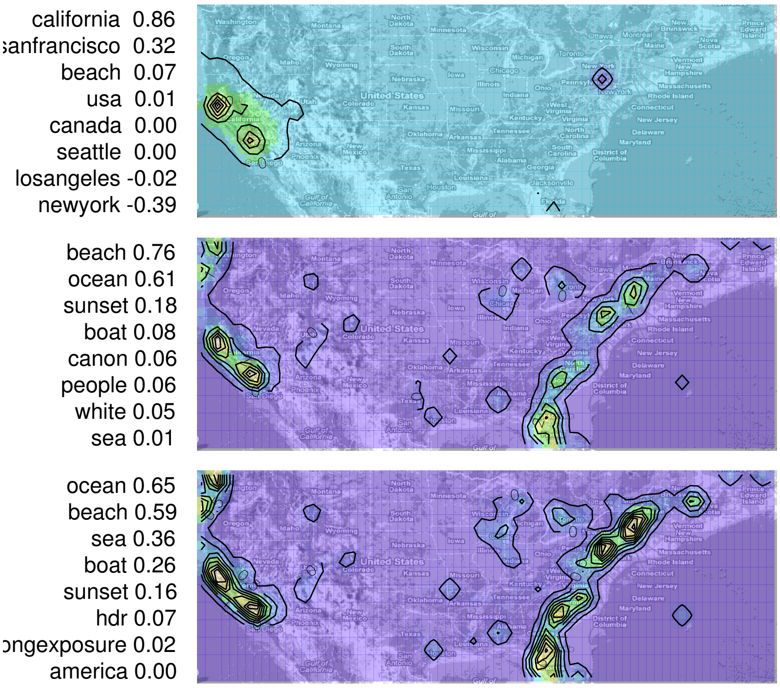


Figure 6.11: Selected SPCA phenomena ($k = 20$) with highest weight of the feature 'beach' extracted using non-normalized input (top), logged input (center), and binarized input (bottom). The 'beach' phenomenon for the non-normalized input describes a landmark ('California'), while the 'beach' phenomenon for the binarized input describes a regional feature (coastal regions in the US).

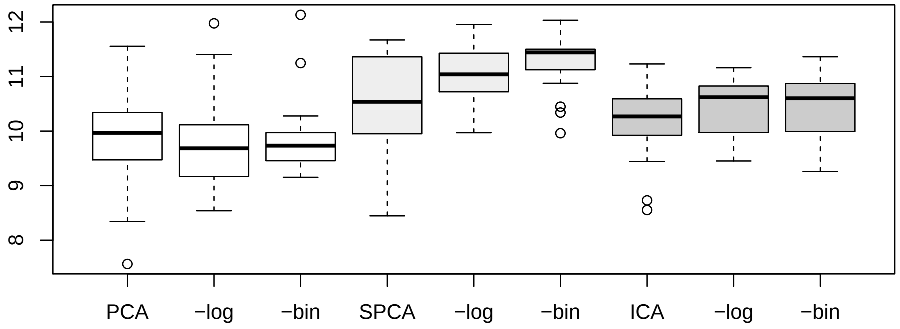


Figure 6.12: Boxplot of entropies of phenomena signals using different approaches and normalizations ($k = 20$). A small entropy indicates a landmark signal, large entropy indicates a global signal. The SPCA signals show the strongest effect on an increasing normalization level of the input matrix.

Binarization diminishes the impact of peaks, resulting in extracted phenomena with high weights for features that are widely distributed in the area of interest (corresponding to regional semantics).

Figure 6.11 shows the extracted phenomena (SPCA) with highest weight for the feature 'beach' in the US data set using non-normalized, log-normalized, and binarized input, respectively. The non-normalized phenomenon shows a signal in California, which is the area with highest user contributions for the beach feature. However, the most dominant features are the landmark features 'california' and 'sanfrancisco'. The phenomenon can hence be described as a California landmark phenomenon, with 'beach' being a representative feature. Using logged input, the signal is spread on both coasts, still with a higher intensity on the West Coast. The features are not dominated by landmark features anymore but represent coastal features. Using binarization, the signal is evenly distributed on both coasts. Moreover, informative coastal features like 'ocean', 'beach', 'sea', and 'boat' have higher weights (compared to the logged input).

Evaluation. In an explorative task the user might want to discover more landmarkish, regional, or global phenomena. We use the strength of normalization as a parameter of geographic phenomenon extraction. Figure 6.12 shows the impact of normalization to the entropies of the signal distributions of $k = 20$ extracted phenomena in a box plot. Recall that a signal with low entropy corresponds to a landmark feature, while a signal with high entropy corresponds to a global feature (see Section 2.2.5). We see that ICA and SPCA respond to the log and bin normalization with higher entropies. Thereby SPCA responds much better than ICA. No response can be seen for PCA. Hence, SPCA is the primary candidate for explorative settings, according to (1) the informativeness of the extracted phenomena, and (2) the response to geographic feature matrix normalization.

6.5.7 Exploratory Analysis

We finally present some interesting discovered phenomena for the LA data set using the best performing technique, SPCA. Figure 6.13 and Figure 6.14 show five selected phenomena using log and bin normalization, respectively. The log normalized phenomena correspond to interesting landmark phenomena in the LA area. We discovered three interesting phenomena for Hollywood (*universal studios*, *griffith park*, *hollywood boulevard*), a *little tokyo* phenomena around the central station, and an *airport* phenomena. Other phenomena not presented here are: *venice beach*, *downtown*, *passadena*, *lacma*, among others. Note that, although those phenomena are discovered in an unsupervised fashion from a noisy geographic information source, they are highly informative to explore the LA city area.

Using bin normalization, the extracted phenomena show more regional characteristics. They can be understood as attributes of geographic space. The phenomena can be labeled as: *inner city*, *going out*, *tourist related*, *walking*, and *nature*. Interestingly, the tourist phenomenon has a high signal around the presented LA-landmark phenom-

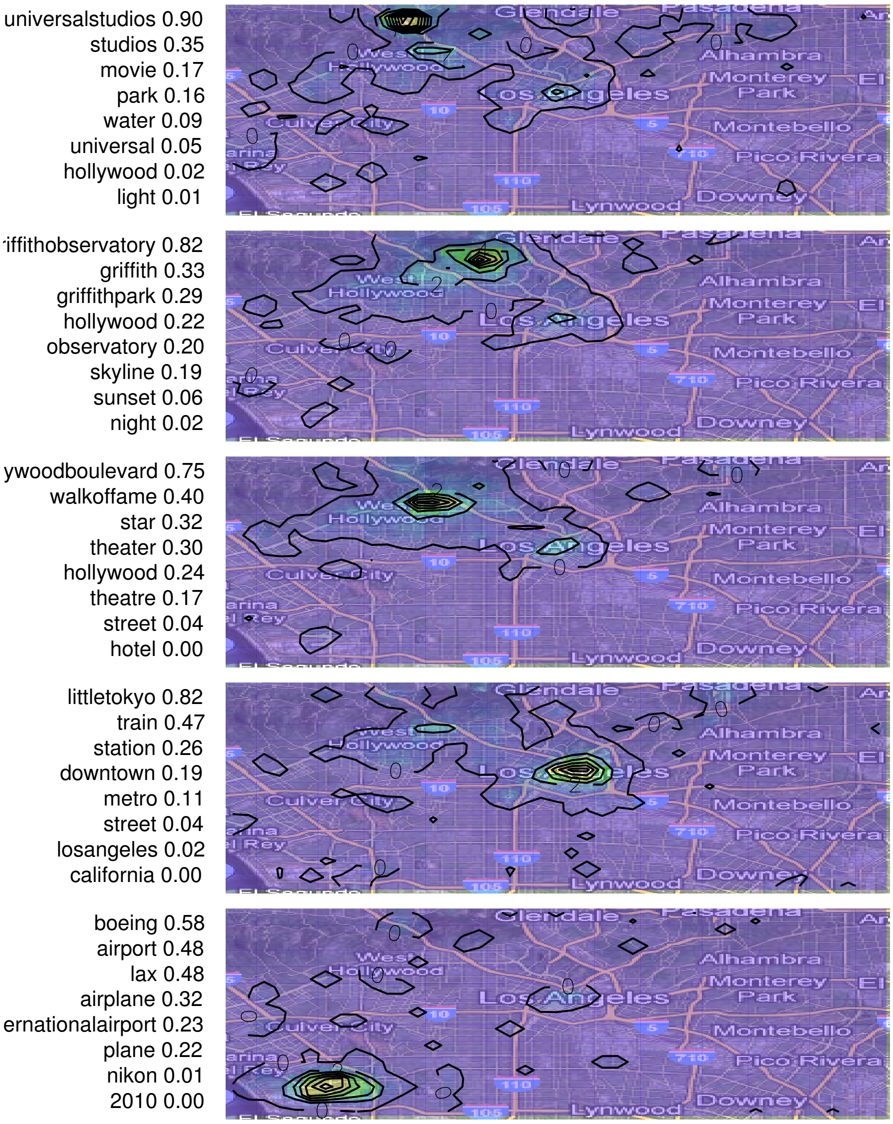


Figure 6.13: Exploratory phenomenon discovery result showing five selected SPCA phenomena from LA data set using logged input and $k = 20$.

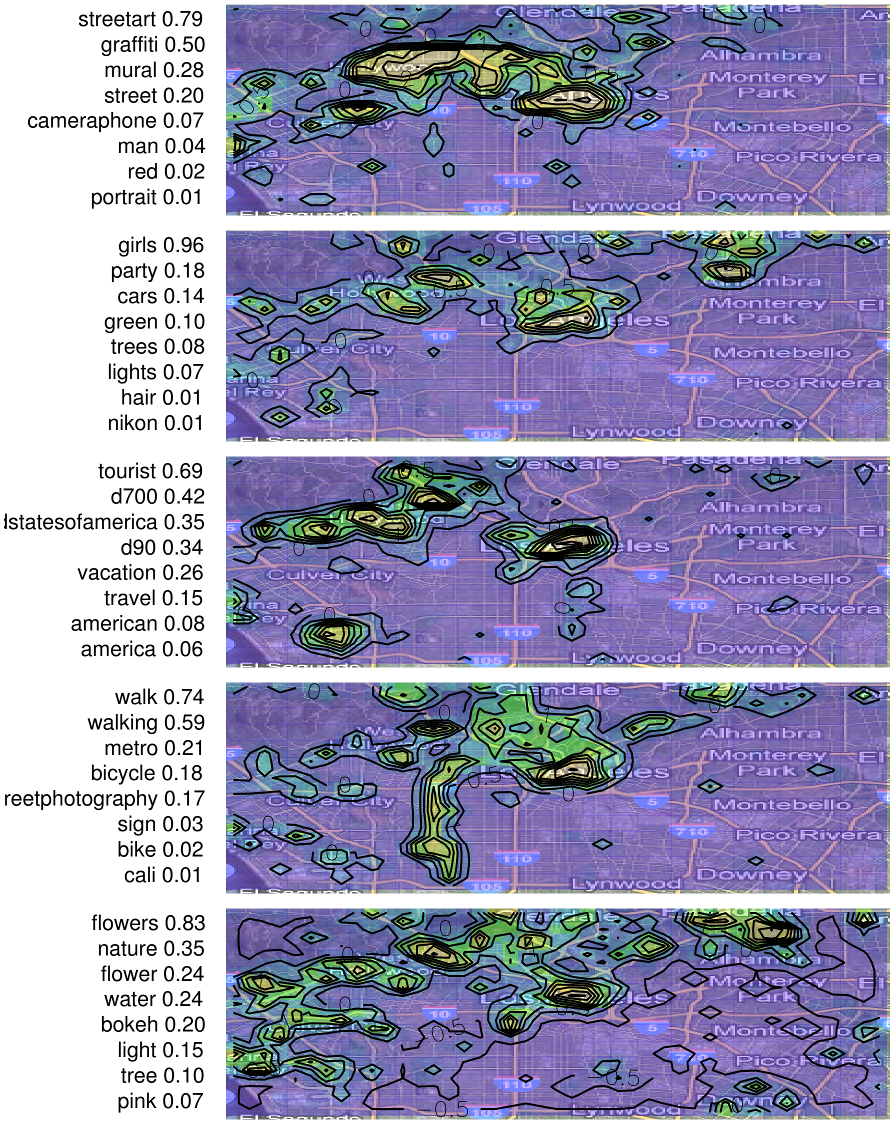


Figure 6.14: Exploratory phenomenon discovery result showing five selected SPCA phenomena from LA data set using binarized input and $k = 20$.

ena. This clearly makes sense, as those landmark phenomena are of high interest to LA tourists.

We see that an unsupervised geographic phenomenon discovery task can be used to discover semantics of geographic space, using a source of noisy user-generated geographic observations as input. The discovered signals are discriminative to segment the area into regions of different semantics and describe geographic processes and entities that are important to the area of interest.

6.6 Record-Phenomenon Relationship

In this chapter, we presented an approach to discover a small number of geographic phenomena from user-generated data by using dimensionality reduction on a previously extracted geographic feature matrix $\mathbf{Z}_{L,F}$. We showed that phenomena discovered using this technique are more informative than results obtained through comparable document-centric approaches. However, different from the topic-pivot LDA model (see Section 6.3.3) and the LGTA model (see Section 6.3.2), our approach is not modeling the relationship between the records (documents) and the resulting phenomena explicitly. This stems from the feature influence aggregation step when building the geographic feature matrix $\mathbf{Z}_{L,F}$ (see Chapter 4). There, we loose the direct link between the features and the records. Consequently, we also loose the relationship between the records and the phenomena, since they are extracted by reducing the dimensions of the geographic feature matrix.

The record-phenomenon relationship is, however, a useful property for a variety of tasks in information organization and retrieval, for example, to support browsing of record collections by geographic semantics, and to search for records of similar geographic semantics by representing the records in latent (geographic) feature space. We refer the reader to [Sizov, 2010] for a list of applications that exploit record-phenomenon relationships, and to [Hofmann, 1999] for search in latent feature space.

In this section, we present a simple proof-of-concept technique to establish links between records and latent geographic features that have been extracted by using our dimensionality reduction approach. By this, we show that our approach can be used to address problems in information organization and retrieval, despite the fact that connections between records and phenomena are not modeled explicitly. Parts of this section have already been published in [Sengstock and Gertz, 2012a].

6.6.1 Record Similarity Measures

To establish links between the records $r \in R$ and the extracted phenomena $q \in Q$ we employ similarity measures between a record and a phenomenon. In the following, we introduce a simple measure based on the inner product of vector-valued record representation and the phenomena.

When extracting the geographic feature signals from the input records, we first identified a set of record features. We now use these features to represent each record by a bag-of-feature vector

$$\mathbf{f}_r \in \mathbb{R}_+^p. \tag{6.47}$$

This vector aggregates the record feature influences as specified in Section 3.3.4. Moreover, we assume that each record has an associated discrete spatio-temporal signal, as defined in Section 3.3.5. This spatio-temporal distribution is denoted as

$$\mathbf{z}_r \in \mathbb{R}_+^n. \tag{6.48}$$

For records being geo-referenced by a single GPS coordinate, this distribution has a single cell with a value of one, and all other cells being zero.

As introduced in Section 6.2, the extracted phenomena $Q = \{q_1, \ldots, q_k\}$ are represented by two quantities, the feature weight vectors

$$\{\boldsymbol{\alpha}_{q_1}, \ldots, \boldsymbol{\alpha}_{q_k}\}, \boldsymbol{\alpha}_q \in \mathbb{R}^p, \tag{6.49}$$

and the spatio-temporal distributions

$$\{\mathbf{z}_{q_1}, \ldots, \mathbf{z}_{q_k}\}, \mathbf{z}_q \in \mathbb{R}^n. \tag{6.50}$$

We can now associate the records with the phenomena by comparing either the records' feature vector $\mathbf{f}_r$ to the phenomenons' feature weight vector $\boldsymbol{\alpha}_q$, or by comparing the records' spatio-temporal distribution $\mathbf{z}_r$ to the phenomenons' distribution $\mathbf{z}_q$. For this, we define two similarity functions, a feature similarity function and a geographic similarity function. Both are based on the inner product of the respective record vector and the phenomenon vector. The inner product is the un-normalized version of the cosine similarity, which is the de-facto standard for similarity computations between bag-of-word representations [Feldman and Sanger, 2007]. For this proof-of-concept approach, we used the simpler inner product technique because of efficiency considerations. However, the results are comparable to the cosine similarity results.

The feature similarity is defined as

$$s_F(r_i, q_j) = \mathbf{f}_r^\top \boldsymbol{\alpha}_q. \tag{6.51}$$

The geographic similarity is defined as

$$s_G(r, q) = \mathbf{z}_r^\top \mathbf{z}_q. \tag{6.52}$$

Both similarities return high values if the record and the phenomenon vector are similar to each other.

In the following, we use the two similarity measures to associate the records with the phenomena, by using either the records' feature information (the bag-of-feature representation) or the records' geographic information (the spatio-temporal distribution).

6.6.2 Data and Setup

We conducted experiments on 111, 166 Wikipedia abstracts occurring within the US area ($[-124, -54] \times [26, 50]$). The data was downloaded from *DBpedia*[3] and the features have been extracted by splitting the abstracts by word boundary characters (spaces and sentence delimiters). Only terms occurring at least 20 times in the collection have been kept and no stopwords have been removed, resulting in $|F| = 17,266$ document features (terms). The spatio-temporal lattice is represented by a regular grid with stepwidth 1.0 degree over the US bounds, resulting in a geographic feature matrix $\mathbf{Z}_{L,F}$ with $|L| = 1728$ rows (72×24 grid). We normalized $\mathbf{Z}_{L,F}$ to use only the binary occurrence information of a feature (binarization) to extract phenomena with regional to global semantics.

We use ICA to extract latent geographic features, as this dimensionality reduction technique showed promising results in the previous experiments. For the choice of the number of phenomena to be extracted, we found that $k = 10$ gave the most informative results for this data set (using the criteria proposed in the previous section).

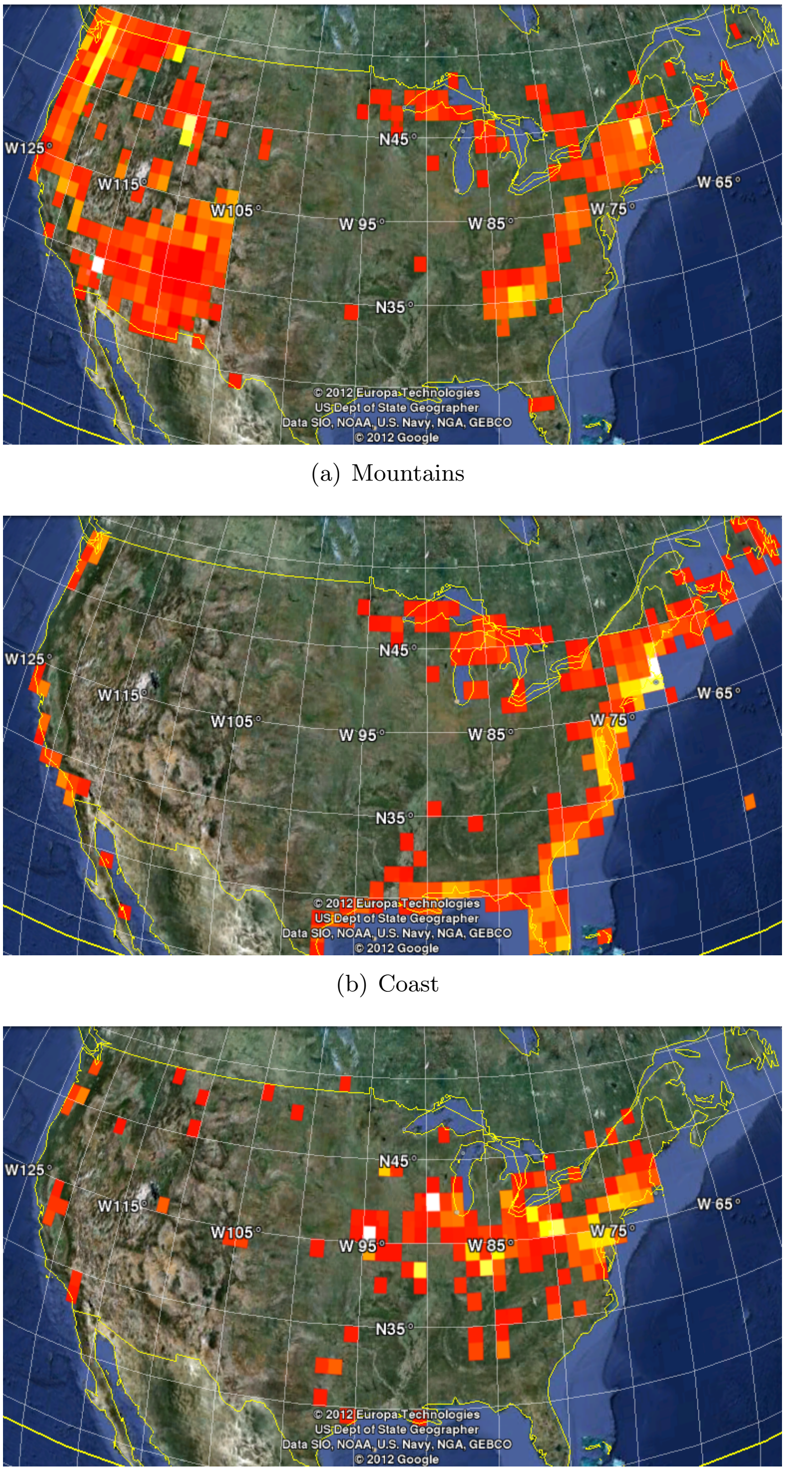
6.6.3 Results

By using the terms and the geographic information of Wikipedia abstracts for latent geographic feature extraction, one obtains a small number of informative geographic phenomena. Figure 6.15 shows the spatial signals of three selected phenomena, and Table 6.2 shows the corresponding top-10 features with highest weight (description). Phenomenon q_1 clearly represents the mountains in the US, q_2 the coastal regions, and q_3 can be interpreted as being related to historic-industrial places (with high signal intensity in the east of the US). Other informative, non-presented phenomena are 'major cities', the 'Canadian border', 'California', as well as a phenomena reflecting the background distribution.

We compute the similarity of the abstracts to the three selected phenomena using the feature similarity s_F (6.51) and the geographic similarity s_G (6.52). Table 6.3 shows the top-8 Wikipedia abstracts for each phenomenon, as determined by the feature similarity. The corresponding geographic similarities are shown in the right three columns of the table.

Feature Similarity. As shown in Table 6.3, the top-ranked Wikipedia abstracts for each of the three phenomena belong clearly to the respective phenomena semantics. E.g., 'Humback Mountain Cs.', 'Schofied Pass Nevada', and 'Schofied Pass Nevada' are clearly associated with the mountain phenomenon. As well are 'Bay Island Bermuda', 'North Dumpling Light', and 'Long Beach Light' clearly associated with the coast phenomenon. The document-centric approaches model the association between a record and a phenomenon by a weight distribution, and select the top-weighted phenomenon or phenomena as the respective labels. Here, we can use the same approach, by choosing the phenomenon with the highest similarity score. By this, we are able to label

[3] http://dbpedia.org


W125°
W115°
W105°
W 95°
W 85°
W 75°
W 65°
N45°
N35°
© 2012 Europa Technologies
US Dept of State Geographer
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
© 2012 Google

(a) Mountains

(b) Coast

(c) Historic

Figure 6.15: Geographic phenomenon signals of three selected phenomena extracted from the geo-referenced Wikipedia abstracts using ICA and $k = 10$. The signals show very distinct distributions from each other, representing the mountain areas, the coast areas, and the historic-industrial regions in the US.

q_1: Mountains	q_2: Coast	q_3: Historic
mountains mountain summit peak wilderness hiking range forest flows highest	bay coast islands peninsula beach island port coastal boat ocean	steel cementery trains 1900 joseph mills society pennsylvania tracks cost

Table 6.2: Geographic phenomenon descriptions of three selected phenomena extracted from the geo-referenced Wikipedia abstracts using ICA and $k = 10$. The table shows the top-10 weighted features of the three selected phenomena q_1, q_2, and q_3. By the semantics of the terms, the phenomena have been labeled 'Mountains', 'Coast', and 'Historic', respectively.

Wikipedia Articles	$s_F(r, q_1)$	$s_F(r, q_2)$	$s_F(r, q_3)$	$s_G(r, q_1)$	$s_G(r, q_2)$	$s_G(r, q_3)$
Humback Mountain Cs.	**0.16**	0.03	0.07	**5.78**	-0.16	-0.85
Schofield Pass Nevada	**0.15**	0.03	0.05	-0.06	**-0.04**	-0.26
Schoflied Pass Wyoming	**0.15**	0.02	0.06	0.15	-0.42	**0.21**
Red Mountain Cascades	**0.14**	0.04	0.07	**5.78**	-0.16	-0.85
Conjeos Peak	**0.13**	0.03	0.05	**0.58**	-0.71	-0.08
Red Mountain Rossland	**0.12**	0.05	0.06	**1.42**	-0.30	0.00
Willow Creek Pass Col.	**0.12**	0.04	0.05	**2.01**	-0.32	0.06
Stampeda Pass	**0.12**	0.05	0.05	**5.78**	-0.16	-0.85
Bay Island, Bermuda	0.03	**0.12**	0.04	0.03	**3.17**	-0.97
North Dumpling Light	0.05	**0.11**	0.06	1.47	**6.31**	4.85
Long Beach Light	0.06	**0.11**	0.05	-1.71	**3.72**	0.24
Mapeque Bay, Prince Edw.	0.03	**0.11**	0.06	-1.47	**0.89**	-0.53
Cornelius Island	0.03	**0.11**	0.04	-1.08	**7.83**	4.15
Monomoy National W. R.	0.05	**0.11**	0.04	-0.46	**8.34**	0.29
Nosuch Bay, Bermuda	0.03	**0.11**	0.03	0.03	**3.17**	-0.97
Bedwell Bay, British Colum.	0.05	**0.11**	0.07	**1.98**	-1.52	-0.35
Mount Vernon Cementry	0.04	0.04	**0.09**	-0.53	-2.02	**8.65**
Boulevard Heights, St. L.	0.00	0.05	**0.09**	-2.17	-2.24	**8.32**
Acheson Tunnel	0.06	0.05	**0.09**	-0.09	-0.56	**7.67**
Washington Trust Build.	0.04	0.07	**0.09**	-0.09	-0.56	**7.67**
St.Thomas Syro-M. Church	0.00	0.07	**0.09**	-0.12	1.40	**10.17**
Theatre Passe Muraille	0.03	0.05	**0.09**	**-0.01**	-8.97	-3.00
Crystal Mall British Colum.	0.03	0.06	**0.09**	**1.77**	-0.61	-0.49
Reynolda Gardens	0.02	0.07	**0.09**	**1.12**	-0.14	1.07

Table 6.3: Top-8 Wikipedia abstracts ordered by the feature similarity $s_F(r, q)$ to the three phenomena q_1, q_2, and q_3. Each row-group of eight abstracts corresponds to one phenomenon. The feature similarly for each phenomenon is shown in the left three columns, the geographic similarity for each phenomenon is shown in the right three columns.

the records by their most similar phenomena to support information organization and retrieval tasks.

Geographic Similarity. By using the geographic similarity function, those records will have a high similarity value that occur at locations where a phenomenon has a strong signal. Of course, such an association is ambiguous if several phenomena have a strong influence at the records' location. In such a case, one needs to employ the feature similarity function to decide which phenomenon is most representative for the record. Consequently, the geographic similarity is less valuable to organize the records by their geographic semantics. However, this technique can be employed if no obvious phenomenon of a record can be determined (if all feature similarity scores are small). Then, the association will be established by choosing the spatio-temporal similarity score.

Table 6.3 shows some abstracts that have a high feature similarity but a low geographic similarity for a specific phenomenon. Two of such examples for the 'mountain' phenomenon are 'Schofield Pass Nevada' and 'Schofield Pass Wyoming'. Those abstracts are clearly linked to the 'mountain' phenomenon, however, they occur in regions with a low signal of that phenomenon. These differences give interesting opportunities to improve and evaluate the resulting phenomenon discovery approaches in future work. Moreover, such differences give rise to explore the record collection by patterns of different feature- and/or geographic preferences.

6.7 Summary

In this chapter, we studied the problem of exploratory geographic phenomenon discovery from user-generated data. We reviewed document-centric approaches that extract geographic phenomena by finding document topics and associated spatio-temporal distributions in the input collection, and introduced our novel approach of latent geographic feature extraction that extracts phenomena by dimensionality reduction of a previously generated geographic feature matrix. We compared the approaches by evaluating the informativeness of the resulting phenomena using a set of carefully defined qualitative criteria. By this, (1) we showed that the choice of the dimensionality reduction technique has significant impact on the quality of the results. In particular, the statistical independence property of ICA and the sparsity property of SPCA result in more compact descriptions, more distinct phenomena, and more coherent signal-description combinations compared to K-Means or PCA. Moreover, (2) we showed that the document-centric approaches fail to find highly informative phenomena in the considered data sets. This is particularly interesting for the LGTA model, which was specifically designed to find document topics with geographic semantics. Since this model was originally evaluated on a document collection of pre-selected records covering a high-level topic (e.g., 'cars', 'electronics'), this may be due to the higher noise level in our test data sets. We also showed that (3) the strength of normalization of the geographic feature matrix can be used to control the type of the extracted geographic phenomena, such as phenomena having a landmark, regional, or global semantics. Finally, (4) we presented a simple technique to associate the records with discovered phenomena that have been extracted using our latent geographic feature extraction approach. By this, we showed that the geographic feature mining framework can be used to address problems in the information organization and retrieval domain.

Chapter 7

Conclusions

7.1 Summary

In this thesis, we developed a conceptual data mining framework to extract highly informative dimensions of geographic space from user-generated data. By revealing commonalities in the models and challenges of existing works, we introduced a process to represent the qualitative and geographic information in user-generated data records as geographic feature signals. The aim of the data mining process is then to discover and extract informative feature signals and feature signal combinations from the candidate signals. The fundamental tasks defined and studied in this thesis are (1) geographic feature extraction, (2) geographic feature comparison, and (3) latent geographic feature extraction.

In Chapter 4, a probabilistic model for geographic feature extraction based on a Bayesian network has been proposed. This approach allows to extract robust feature signals and leverage different kinds of qualitative and geographic information from the records.

In Chapter 5, a novel technique to categorize and select geographic feature signals on the basis of their spatio-temporal type has been introduced. This helps to select informative sub-sets from the feature candidates gathered in the geographic feature extraction task. The technique is based on a representation of the signals by their interaction characteristics. This representation can then be used by distance and similarity functions to cluster and summarize features, or to filter them by comparison to well-known distributions.

In Chapter 6, the task of latent geographic feature extraction has been covered. This tasks allows to extract a small number of informative feature combinations from the candidate set. We introduced a model that extracts such latent features using dimensionality reduction and showed the superior informativeness of the results compared to document-centric (and hence data-centric) approaches.

In summary, this thesis develops a set of fundamental data mining tasks to utilize user-generated data as a source of geographic knowledge. These tasks can be seen as providing a layer between the user-generated data and particular domain-specific applications, and by this, help to make the field of geographic knowledge discovery more prolific for application-driven research in other domains.

7.2 Discussion

Our work is the first comprehensive study of a common framework and of fundamental tasks to utilize user-generated data as a general sensor for geographic phenomena. Of course, this gives raise to a lot of questions about the model and design decisions, as well as about the particular challenges addressed in this work.

- *Generality*: Even if the ideas have been developed by revealing the specific techniques and models in a huge number of existing data- and application specific works, a generalization will never allow to realize and support all kinds of applications. For the sake of fruitful research in application-driven domains, we assume, however, a set of clearly defined concepts and problem statements as indispensable. This research can be seen as exploring the problem from the side opposite to the application-driven approaches. Bringing them together by interdisciplinary research is a valuable goal for the future.

  On the other side, the generality exposed in this work allows to find techniques in related fields that address a number of the common challenges. These are spatial statistics, computer vision, and information retrieval, as frequently referred to in this work.

- *Candidate Features*: The most fundamental decision to geographic feature mining is the set of candidate features. In this work, we mostly used textual features, such as tags or terms tokenized from documents, and showed that they are a powerful source to describe geographic space. We experienced, however, that the data sources exhibit very different semantics between and within each other. For example, Twitter can be seen as a medium for a variety of purposes [Gill, 2005; Kwak et al., 2010], and transporting content of varying quality [Agichtein et al., 2008]. The purpose of Twitter, or micro-blogs in general, includes communication between users, providing a forum for user opinions, allowing for automated status update of applications, and advertisement from companies and public services, among others. To extract informative dimensions of geographic space from these heterogeneous semantics, the selection of appropriate data sources and subsets of these sources constitutes an important perquisite. The task of geographic feature comparison, as proposed in this thesis, can be seen as a core component in the framework to help users select appropriate feature candidates for the subsequent applications.

  Furthermore, in developing the different concepts and evaluating the tasks, we increasingly got the impression that abstract image features can provide a general yet discriminative source to describe the semantics of geographic space. This includes colors and textures (discriminating weather conditions, urban/natural areas, night/day) and common sub-structures (discriminating objects or people). Moreover, by identifying objects and people in images and linking them to external knowledge sources, the images will be a highly objective source of real-world observations. Exploiting techniques from computer vision will hence be a promising direction of research in the future.

- *Intrinsic and Extrinsic Evaluation*: Extracting informative dimensions to describe geographic space in an application-independent manner makes it hard to separate good from bad signals. Intrinsic quality measures, that is, measures that judge about the quality of the tasks by the resulting data only and not taking the user feedback into account, are nevertheless an important tool to evaluate the tasks in a fast and comparable manner. In this work, we introduced properties of geographic feature signals, such as dominance and representativeness. Moreover, we defined the type semantics of spatio-temporal signals and provided qualitative criteria to judge about the informativeness of extracted features. We believe that such basic measures are an important step towards a general framework and see this as a prolific direction of research for geographic knowledge discovery and geographic information science.

  However, extrinsic evaluations, that is, evaluations relying on user feedback from real-world applications, will finally judge about the qualitative performance of underlying techniques and models. With a variety of different applications emerging in this data-rich environment in the near future, the quality of common concepts and techniques can be better understood to improve common abstractions and intrinsic measures. Using the common models in existing and emerging applications and evaluating the user feedback appropriately is hence an important necessity from a practical point of view.

7.3 Future Work

In addition to the general thoughts about prolific future research directions given in the discussion above, a number of concrete ideas for future work emerged during this thesis:

- *Spatio-temporal regularization*: We addressed the problem of spatio-temporal data sparsity by using appropriate smoothing bandwidths in non-parametric models. This approach is fast and allows explicitly to take the scale of the resulting signals into account. However, smoothing all points in space and time identically can lead to inaccurate results (e.g., by smoothing a signal into a region without any user contributions). An alternative approach is to penalize spatio-temporal distributions that are non-smooth while learning statistical models. This approach allows to use sophisticated penalty functions that reflect the local characteristics of the data. One such approach that extends non-negative matrix factorization (NMF) by a regularization factor has been introduced in [Cai et al., 2011]. NMF is not yet well understood from a statistical point of view [Gaussier and Maupertuis, 2005].

However, it shows promising results as a dimensionality reduction technique for multivariate data [Lee and Seung, 2001]. We started with initial experiments using a spatio-temporal regularization factor based on the idea proposed in [Cai et al., 2011]. We expect this adapted dimensionality-reduction approach to improve the informativeness of the latent geographic feature extraction results by exploring novel penalty functions of spatio-temporal signals to tackle spatio-temporal data sparsity.

- *Markov Random Fields*: Other than smoothing and spatio-temporal regularization, Markov random fields (MRF) model the dependency between neighboring points in space and time explicitly. MRFs are much harder to fit than the models proposed in this thesis, and using them for massive data is an active field of research [Hastie et al., 2009, p. 643]. Nevertheless, we see a great potential to improve (latent) geographic feature extraction using these models and plan to utilize them in the future.

- *Online algorithms*: As we stated throughout this work, we want to use user-generated data as a steady source of geographic phenomenon signals. Hence, the actuality of the resulting dimensions and the utilization of novel data are essential aspects. One approach to address this problem is to split the data into temporal intervals and to perform geographic feature mining on these data subsets. By using a sliding window approach [Leskovec et al., 2014, p. 116] the resulting signals will be available in a reasonable small temporal resolution and change smoothly over time. However, by such a simple approach we loose the connection between the extracted signals in the different intervals. Hence, we have no explicit information if the signal in the new time interval is a changed version of a former signal, a new one, or a combination. This is particularly important for the latent geographic feature extraction task, since the discovered signals are already combinations of features. One way to address this problem is to use the extracted signals of the current time interval as input to the following interval. By using PCA, ICA, and SPCA an initial set of latent features can be used as a starting point to converge to a new solution. This property can be used to capture the dependencies between the results in the time intervals. We plan to extend latent geographic feature extraction on this basis in the future.

Bibliography

Adams, B. and McKenzie, G. (2012). Frankenplace: An Application for Similarity-Based Place Search. In *Proceedings of the Sixth International Conference on Weblogs and Social Media, ICWSM, 2012.*

Adomavicius, G. and Tuzhilin, A. (2011). Context-aware recommender systems. In *Recommender Systems Handbook*, pages 217–253. Springer.

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining, WSDM 2008*, pages 183–194.

Ahern, S., Naaman, M., Nair, R., and Yang, J. H.-I. (2007). World Explorer: Visualizing Aggregate Data from Unstructures Text in Geo-Referenced Collections. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2007*, pages 1–10.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic Detection and Tracking Pilot Study: Final Report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.*

Anagnostopoulos, A., Kumar, R., and Mahdian, M. (2008). Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2008*, pages 7–15.

Backstrom, L., Kleinberg, J., Kumar, R., and Novak, J. (2008). Spatial variation in search engine queries. In *Proceedings of the 17th International Conference on World Wide Web, WWW 2008*, pages 357–366.

Backstrom, L., Sun, E., and Marlow, C. (2010). Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010*, pages 61–70.

Baddeley, A. (2005). Spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42.

Baddeley, A. (2010). Modeling Strategies. In *Handbook of Spatial Statistics*, pages 339–369. CRC Press.

Bartlett, M. S., Movellan, J. R., and Sejnowski, T. J. (2002). Face Recognition by Independent Component Analysis. *IEEE Transactions on Neural Networks*, 13(6):1450–1464.

Becker, H. and Gravano, L. (2010). Event Identification in Social Media. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 291–300.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning, ICML 2006*, pages 113–120.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

Cai, D., He, X., and Han, J. (2011). Graph Regularized Non-negative Matrix Factorization for Data Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1548–1560.

Cao, L., Yu, J., Luo, J., and Huang, T. S. (2009). Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In *Proceedings of the seventeen ACM international conference on Multimedia, MM 2009*, pages 125–134.

Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D. S., and Ertl, T. (2012). Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology, VAST 2012*, pages 143–152.

Chen, L. and Roy, A. (2009). Event detection from flickr data through wavelet-based spatial analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009*, pages 523–532.

Cheng, Z., Caverlee, J., and Lee, K. (2010). You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM 2010*, pages 759–768.

Cheng, Z., Caverlee, J., Lee, K., and Sui, D. Z. (2011). Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, ICWSM 2011*, pages 81–88.

Comaniciu, D. and Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619.

Crandall, D., Backstrom, L., Huttenlocher, D., and Kleinberg, J. (2009). Mapping the Worlds Photos. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009*, pages 761–770.

Crandall, D. and Snavely, N. (2012). Modeling people and places with internet photo collections. *Communications of the ACM*, 55(6):52.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-temporal Data.* John Wiley & Sons.

Davies, E. R. (2012). *Machine Vision: Theory, Algorithms, Practicalities.* Elsevier.

Davis, H. B. (2006). HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, to Read. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, HYPERTEXT 2006*, pages 31–40.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Deng, D. P. and Lemmens, R. (2009). Conceptualization of Place via Spatial Clustering and Co-occurrence Analysis. In *Proceedings of the 2009 International Workshop on Location Based Social Networks, LBSN 2009*, pages 49–56.

Ding, C. (2004). K-means Clustering via Principal Component Analysis. In *Proceedings of the Twenty-first International Conference on Machine Learning, ICML 2004*, pages 29–39.

Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., and Tomkins, A. (2007). Visualizing tags over time. *ACM Transactions on the Web*, 1(2):1559–1131.

Earle, P., Gu, M., Buckmaster, R., Ostrum, C., Horvath, S., and Vaughan, A. (2010). OMG Earthquake! Can Twitter Improve Earthquake Response? *Seismological Research Letters*, 81(2):246–251.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010*, pages 1277–1287.

Emily, M., Kleban, J., Jiejun, X., and Manjunath, B. S. (2009). Not all tags are created equal: learning flickr tag semantics for global annotation. In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, ICME 2009*, pages 1452–1455.

Feldman, R. and Sanger, J. (2007). *The Text Mining Handbook.* Cambrige University Press.

Gallagher, A. (2010). The Wisdom of Social Multimedia: Using Flickr For Prediction and Forecast. In *Proceedings of the International Conference on Multimedia, MM 2010*, pages 1235–1244.

Gaussier, E. and Maupertuis, D. (2005). Relation between PLSA and NMF and Implications. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005*, pages 206–207.

Gill, K. E. (2005). Blogging, RSS and the Information Landscape: A Look At Online News. In *Proceedings of the WWW 2005 workshop on the weblogging ecosystem.*

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

Gonzalez, R. C. and Woods, R. E. (2007). *Digital Image Processing.* Pearson.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4):211–221.

Gorski, K. M., Hivon, E., Banday, a. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., and Bartelmann, M. (2005). HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. *The Astrophysical Journal*, 622(2):759–771.

Guralnik, V. and Srivastava, J. (1999). Event detection from time series data. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 1999*, pages 33–42.

Guyon, I. and Elisseefi, A. (2006). An Introduction to Feature Extraction. In *Feature Extraction: Foundations and Applications*, pages 1–25. Springer.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining - Concepts and Techniques.* Morgan Kaufmann.

Hao, Q., Cai, R., Wang, C., Xiao, R., Yang, J.-m., Pang, Y., and Zhang, L. (2010). Equip Tourists with Knowledge Mined from Travelogues. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010*, pages 401–410.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

Hays, J. H. and Efros, A. A. (2008). Im2gps: estimating geographic information from a single image. In *Proceedings of the IEEE Conference of Computer Vision and Pattern Recognition, CVPR 2008.*

Hoffman, M. D., Blei, D. M., and Bach, F. (2010). Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems, NIPS 2010*, pages 856–864.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR 1999*, pages 50–57.

Hollenstein, L. and Purves, R. (2010). Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*, 1(1):21–48.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A., and Tsioutsiouliklis, K. (2012). Discovering Geographical Topics In the Twitter Stream. In *Proceedings of the 21st International Conference on World Wide Web, WWW 2012*, pages 769–778.

Hong, L., Yin, D., Guo, J., and Davison, B. D. (2011). Tracking Trends: Incorporating Term Volume into Temporal Topic Models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2011*, pages 484–492.

Hyvärinen, a. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430.

Isham, V. (2010). Spatial Point Process Models. In *Handbook of Spatial Statistics*, pages 283–298. CRC Press.

Jaffe, A. (2006). Generating Summaries and Visualization for Large Collections of Geo-Referenced Photographs. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, MIR 2006*, pages 89–98.

Johnson, H. A., Wagner, M. M., Hogan, W. R., Chapman, W., Olszewski, R. T., Dowling, J., and Barnas, G. (2004). Analysis of Web access logs for surveillance of influenza. *Studies in health technology and informatics*, 107(2):1202–1206.

Kennedy, L. (2008). Generating Diverse and Representative Image Search Results for Landmarks. In *Proceedings of the 17th International Conference on World Wide Web, WWW 2008*, pages 297–306.

Kennedy, L., Naaman, M., Ahern, S., Nair, R., and Rattenbury, T. (2007). How Flickr Helps us Make Sense of the World: Context and Content in Community-Contributed Media Collections. In *Proceedings of the 15th International Conference on Multimedia, MULTIMEDIA 2007*, pages 631–640.

Kisilevich, S., Mansmann, F., and Keim, D. (2010). P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application, COMGEO 2010*, pages 381–384.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002*, pages 91–101, New York, New York, USA. ACM Press.

Kollar, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

Koperski, K. and Han, J. (1995). Discovery of Spatial Association Rules in Geographic Information Databases. In *Proceedings of the 4th International Symposium on Advances in Spatial Databases, SSD 1995*, pages 47–66.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010*, pages 591–600.

Landauer, T., Foltz, P., and Laham, D. (1967). Mixed-Data Classification Programs I - Agglomerative Systems. *Australian Computer Journal*, 1(1):15–20.

Lawson, A. B. (2001). *Statistical Methods in Spatial Epidemiology.* John Wiley & Sons.

Le Gall, D. (1991). MPEG: A Video Compression Standard for Multimedia Applications. *Communications of ACM*, 34(4):46–58.

Lee, D. D. and Seung, H. S. (2001). Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems, NIPS 2001*, pages 556–562.

Lehmann, J., Gonçalves, B., and Cattuto, C. (2012). Dynamical Classes of Collective Attention in Twitter. In *Proceedings of the 21st International Conference on World Wide Web, WWW 2012*, pages 251–260.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). Mining of Massive Datasets. *Stanford Technical Report.*

Leung, D. and Newsam, S. (2010). Proximate Sensing: Inferring What-Is-Where From Georeferenced Photo Collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010*, pages 2955–2962.

Liang, C.-K., Hsieh, Y.-T., Chuang, T.-J., Wang, Y., Weng, M.-F., and Chuang, Y.-Y. (2010). Learning Landmarks by Exploiting Social Media. In *Proceedings of the 16th International Conference on Advances in Multimedia Modeling, MMM 2010*, pages 207–217.

Lieberman, M. D. (2010). Geotagging: Using Proximity , Sibling , and Prominence Clues to Understand Comma Groups. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, pages 1–8.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Ma, J. and Perkins, S. (2003). Online novelty detection on temporal sequences. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2003*, pages 613–618.

Manning, C. D., Raghavan, P., and Schuetze, H. (2008). *Introduction to Information Retrieval.* Cambrige University Press.

Martin, G. N. (1985). Probabilistic Counting Algorithms for Data Base Applications. *Journal of Computerand System Sciences*, 31(2):182–209.

Mei, Q. (2006). A Mixture Model for Contextual Text Mining. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2006*, pages 649–655.

Mei, Q., Cai, D., Zhang, D., and Zhai, C. (2008). Topic modeling with network regularization. In *Proceedings of the 17th International Conference on World Wide Web, WWW 2008*, pages 101–110.

Mei, Q., Liu, C., Su, H., and Zhai, C. (2006). A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th International Conference on World Wide Web, WWW 2006*, pages 533–542.

Miller, H. J. and Han, J. (2009). Geographic Data Mining and Knowledge Discovery: An Overview. In *Geographic Data Mining and Knowledge Discovery*, pages 1–26. Taylor & Francis Group.

Naaman, M., Boase, J., and Lai, C.-H. (2010). Is it Really About Me? Message Content in Social Awareness Streams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW 2010*, pages 189–192.

Nallapati, R., Cohen, W., and Lafferty, J. (2007). Parallelized Variational EM for Latent Dirichlet Allocation: An Experimental Evaluation of Speed and Scalability. In *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops, ICDMW 2007*, pages 349–354.

O'Hare, N. and Murdock, V. (2012). Modeling locations with social media. *Information Retrieval*, 16(1):30–62.

Paul, M. J. and Dredze, M. (2011). You Are What You Tweet: Analyzing Twitter for Public Health. In *Proceedings of Fifth International AAAI Conference on Weblogs and Social Media, ICWSM 2011*, pages 265–272.

Peter, B. and Van de Geer, S. (2011). *Statistics for High-Dimensional Data.* Springer.

Radinsky, K., Teevan, J., Bocharov, A., and Horvitz, E. (2012). Modeling and Predicting Behavioral Dynamics on the Web. In *Proceedings of the 21st International Conference on World Wide Web, WWW 2012*, pages 599–608.

Rajaraman, K. and Tan, A.-h. (2001). Topic Detection, Tracking and Trend Analysis Using Self-organizing Neural Networks. In *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2001*, pages 102–107.

Rattenbury, T., Good, N., and Naaman, M. (2007a). 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, SIGIR 2007*, pages 103–110.

Rattenbury, T., Good, N., and Naaman, M. (2007b). Towards Extracting Flickr Tag Semantics. In *Proceedings of the 16th international conference on World Wide Web, WWW 2007*, pages 1287–1288.

Rattenbury, T. and Naaman, M. (2009). Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web*, 3(1):1–30.

Ripley, B. D. (1977). Modelling Spatial Patterns. *Journal of the Royal Statistical Society - Series B*, 39(2):172–212.

Ripley, B. D. (1981). *Spatial Statistics*. John Wiley & Sons.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Rue, H. and Held, L. (2010). Discrete Spatial Variation. In *Handbook of Spatial Statistics*, pages 171–200. Taylor & Francis Group.

Sahr, K., White, D., and Kimerling, A. J. (2003). Geodesic Discrete Global Grid Systems. *Cartography and Geographic Information Science*, 30(2):121–134.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010*, pages 851–860.

Sankaranarayanan, J., Teitler, B. E., Samet, H., Lieberman, M. D., and Sperling, J. (2009). TwitterStand: News in Tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2009*, pages 42–51.

Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S. (2007). Collaborative Filtering Recommender Systems. In *The Adaptive Web*, pages 291–324. Springer.

Sengstock, C. and Gertz, M. (2011a). CONQUER : A System for Efficient Context-aware Query Suggestions. In *Proceedings of the 20th International Conference Companion on World Wide Web, WWW 2011 Companion*, pages 265–268.

Sengstock, C. and Gertz, M. (2011b). Exploration and Comparison of Geographic Information Sources using Distance Statistics. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2011*, pages 329–338.

Sengstock, C. and Gertz, M. (2012a). Latent Contextual Indexing of Annotated Documents. In *Proceedings of the 21st International Conference Companion on World Wide Web, WWW 2012 Companion*, pages 593–594.

Sengstock, C. and Gertz, M. (2012b). Latent Geographic Feature Extraction from Social Media. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems, GIS 2012*, pages 149–158.

Sengstock, C., Gertz, M., Abdelhaq, H., and Flatow, F. (2013a). Reliable Spatio-temporal Signal Extraction and Exploration from Human Activity Records. In *Proceedings of the 13th International Conference on Advances in Spatial and Temporal Databases, SSTD 2013*, pages 484–489.

Sengstock, C., Gertz, M., and Canh, T. V. (2012). Spatial Interestingness Measures for Co-location Pattern Mining. In *Proceedings of the 12th IEEE International Conference on Data Mining Workshops, ICDMW 2012*, pages 821–826.

Sengstock, C., Gertz, M., Flatow, F., and Abdelhaq, H. (2013b). A probablistic model for spatio-temporal signal extraction from social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2013*, pages 264–273.

Serdyukov, P., Murdock, V., and van Zwol, R. (2009). Placing flickr photos on a map. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009*, pages 484–491.

Shekar, S. and Huang, Y. (2001). Discovering Spatial Co-location Patterns: A Summary of Results. In *Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Databases, SSTD 2001*, pages 236–256.

Shekar, S., Schrater, P. R., Vatsavai, R. R., Wu, W., and Chawla, S. (2002). Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. *IEEE Transactions on Multimedia*, 4(2):174–188.

Shekhar, S. and Chawla, S. (2003). *Spatial Databases: A Tour.* Pearson Education.

Shepitsen, A., Gemmell, J., Mobasher, B., and Burke, R. (2008). Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys 2008*, pages 259–266.

Sheth, A. (2009). Citizen Sensing, Social Signals, and Enriching Human Experience. *IEEE Internet Computing*, 13(4):87–92.

Singh, V. K. (2010). Social Pixels: Genesis and Evaluation. In *Proceedings of the International Conference on Multimedia, MM 2010*, pages 481–490.

Sizov, S. (2010). GeoFolk: Latent Spatial Semantics in Web 2.0 Social Media. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM 2010*, pages 281–290.

Strötgen, J. and Gertz, M. (2013). Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

Tan, P.-N., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining.* Pearson.

Toyama, K., Logan, R., and Roseway, A. (2003). Geographic location tags on digital images. In *Proceedings of the eleventh ACM international conference on Multimedia, MULTIMEDIA 2003*, pages 156–166.

Trefethen, L. N. and Bau, D. (1997). *Numercial Linear Algebra.* SIAM.

Turk, M. A. and Pentland, A. P. (1991). Face Recognition Using Eigenfaces. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 1991*, pages 586–591.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of artificial intelligence research*, 37(1):141–188.

van Lieshout, M.-C. (2010). Spatial Point Process Theory. In *Handbook of Spatial Statistics*, pages 263–282. CRC Press.

Viola, P. and Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001*, pages 511–518.

Wang, C., Blei, D., and Heckerman, D. (2008). Continuous Time Dynamic Topic Models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, UAI 2008*, pages 1–8.

Wang, C., Wang, J., Xie, X., and Ma, W.-Y. (2007). Mining geographic knowledge using location aware topic model. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, GIR 2007*, pages 65–70.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference.* Springer.

Weeds, J., Weir, D., and Mccarthy, D. (2004). Characterising Measures of Lexical Distributional Similarity. In *Proceedings of the 20th international conference on Computational Linguistics, COLING 2004*, pages 1–7.

Weiss, S. M., Indurkhya, N., Zhang, T., and Damerau, F. J. (2005). *Text Mining - Predictive Methods for Analyzing Unstructured Information.* Springer.

Wing, B. P. and Baldridge, J. (2011). Simple Supervised Document Geolocation with Geodesic Grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT 2011*, pages 955–964.

Xiao, H. and Stibor, T. (2010). Efficient Collapsed Gibbs Sampling For Latent Dirichlet. In *Proceedings of the 2nd Asian Conference on Machine Learning, ACML 2010*, pages 63–78.

Xu, J.-M., Bhargava, A., Nowak, R., and Zhu, X. (2012). Socioscope: Spatio-Temporal Signal Recovery from Social Media. In *Proceedings of the 2012 European Conference on Machine Learning and Knowledge Discovery in Databases, ECML-PKDD 2012*, pages 644–659.

Ye, M., Yin, P., Lee, W.-C., and Lee, D.-L. (2011). Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011*, pages 325–334.

Yin, Z. and Cao, L. (2011). Diversified Trajectory Pattern Ranking in Geo-Tagged Social Media. In *Proceedings of the 2011 SIAM International Conference on Data Mining, SDM 2011*, pages 980–991.

Yin, Z., Cao, L., Han, J., Zhai, C., and Huang, T. (2011). Geographical Topic Discovery and Comparison. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011*, pages 247–256.

Yoo, J. S., Shekhar, S., and Celik, M. (2005). A Join-Less Approach for Co-Location Pattern Mining: A Summary of Results. In *Proceedings of the Fifth IEEE International Conference on Data Mining, ICDM 2005*, pages 813–816.

Yuan, J., Zheng, Y., and Xie, X. (2012). Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining, KDD 2012*, pages 186–194.

Zhang, H., Korayem, M., Crandall, D. J., and Lebuhn, G. (2012a). Mining Photo-sharing Websites to Study Ecological Phenomena. In *Proceedings of the 21st International Conference on World Wide Web, WWW 2012*, pages 749–758.

Zhang, H., Korayem, M., You, E., and Crandall, D. J. (2012b). Beyond Co-occurrence: Discovering and Visualizing Tag Relationships from Geo-spatial and Temporal Similarities. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM 2012*, pages 33–42.

Zheng, Y., Zhang, L., Ma, Z., Xie, X., and Ma, W.-Y. (2011). Recommending Friends and Locations based on Individual Location History. *ACM Transactions on the Web*, 5(1):1–44.

Zheng, Y., Zhang, L., and Xie, X. (2009). Mining Interesting Locations and Travel Sequences from GPS Trajectories. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009*, pages 791–800.

Zhu, F., Yan, X., Han, J., and Yu, P. S. (2006). Mining Frequent Approximate Sequential Patterns. In *Next Generation of Data Mining*. CRC Press.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.