Latent Geographic Feature Extraction from Social Media

Christian Sengstock
Institute of Computer Science
Heidelberg University, Germany
sengstock@informatik.uni-heidelberg.de

Michael Gertz
Institute of Computer Science
Heidelberg University, Germany
gertz@informatik.uni-heidelberg.de

ABSTRACT
In this work we present a framework for the unsupervised extraction of latent geographic features from georeferenced social media. A geographic feature represents a semantic dimension of a location and can be seen as a sensor that measures a signal of geographic semantics. Our goal is to extract a small number of informative geographic features from social media, to describe and explore geographic space, and for subsequent spatial analysis, e.g., in market research. We propose a framework that, first, transforms the unstructured and noisy geographic information in social media into a high-dimensional multivariate signal of geographic semantics. Then, we use dimensionality reduction to extract latent geographic features. We conduct experiments using two large-scale Flickr data sets covering the LA area and the US. We show that dimensionality reduction techniques extracting sparse latent features find dimensions with higher informational value. In addition, we show that prior normalization can be used as a parameter in the exploration process to extract features representing different geographic characteristics, that is, landmarks, regional phenomena, or global phenomena.

Categories and Subject Descriptors
H.2.8 [Database applications]: Data Mining
; I.5.3 [Clustering]: Algorithms, Similarity measures

Keywords
Feature extraction, social media, geographic feature

1. INTRODUCTION
The Web is a huge and steadily growing source of unstructured geographic information, such as provided by georeferenced documents in social media. Several approaches exploit such sources for knowledge discovery, information retrieval, and computer vision [10, 19]. Tasks include the extraction of tag semantics [13, 15], geographical document topic discovery [6, 16], finding representative landmark images [8], and spatial learning [3, 17].

In this work we treat georeferenced social media, like photos or microblogs, as a noisy source of geographic semantics. For example, the terms restaurant or eating are expected to be found more often in documents contributed at locations where restaurants occur or where people eat. The amount of occurrences of those document features (terms) around a location can then be seen as a signal to describe a location, e.g., whether a location is a good choice for dinner or not. Such a signal is a geographic feature because it represents a semantic dimension of geographic space\(^1\). In social media we have a large number of possible geographic features (e.g., all tags/terms occurring in the collection). We can use each document feature as an individual sensor to measure a signal of (possibly abstract, unknown, or meaningless) geographic semantics, and we can treat the signals jointly as a multivariate measurement. The motivation of our approach is that such a noisy multivariate sensor derived from georeferenced social media can be used to describe and explore geographic space in a meaningful way.

To accomplish this, we study the unsupervised extraction of latent geographic features from georeferenced social media using dimensionality reduction. A latent geographic feature is a semantic dimension of geographic space that cannot directly be observed in a collection, but is exhibited by the spatial correlation of document features. Particularly, it is not possible to derive the signal from individual document feature counts. However, we assume that informative latent geographic features exist in the data, which can be seen as basic building blocks of geographic semantics. For example, a combination of the features beach, sand, and ocean might establish a latent geographic feature that can be labeled as the phenomenon coast. The signals of the individual features can then be seen as being generated from the underlying more informative coast feature, even if we are not able to measure this non-observed (latent) signal directly.

Our objective is to extract a small number of such latent geographic features that can be used as highly informative dimensions to describe and explore geographic space instead of looking at the large number of possibly meaningless and redundant individual feature signals. By that, we turn the noisy and unstructured geographic information in social media into an informative source of geographic information. Such a task is of high value for geographic knowledge discov-

\(^1\)The use of the term feature stems from its usage in the information retrieval literature.
ery and exploratory data analysis. In addition, the resulting latent geographic features can be used as spatial variables for subsequent learning and recommendation tasks, e.g., in market research or recommender systems.

The proposed framework for the extraction of latent geographic features is illustrated in Figure 1. For this, the unstructured geographic information in social media is (1) transformed into a noisy, high-dimensional multivariate signal of geographic semantics. Then, (2) the signal is sampled over geographic space, and (3) dimensionality reduction is used to extract latent geographic features. We show that dimensionality reduction techniques assuming sparsity of the extracted features perform best. Moreover, we show that prior normalization can be used as a parameter in the exploration process to extract features representing different geographic characteristics (landmarks, regional, global).

To validate our framework, we conduct several experiments using two large-scale Flickr data sets covering the LA area and the US. The results show that geographic feature extraction allows for rich exploration of geographic space using noisy, unstructured geographic information, such as georeferenced social media, as input.

The contributions of this paper can be summarized as follows: (a) We define the tasks of exploratory geographic feature discovery and latent geographic feature extraction from georeferenced social media. (b) We introduce a general framework to extract latent geographic features. (c) We discuss characteristics of georeferenced social media and their impact on latent geographic feature extraction. (d) We study different techniques to extract latent geographic features and compare their performance by an experimental evaluation using two large-scale Flickr data sets.

The remainder of this paper is structured as follows. In the following section, we discuss related work. In Section 3, we give definitions and detail the problem statement. General characteristics of georeferenced social media are described in Section 4. Different techniques to extract latent geographic features are presented in Section 5. Several experiments addressing the performance of our feature extraction techniques and the impact of normalization are discussed in Section 6.

2. RELATED WORK

We address related work from a number of research areas including event/landmark detection, spatial learning, and geographical topic discovery.

Analyzing tags and visual features from social media with respect to their spatio-temporal distributions has been studied extensively. In [2, 12, 14] approaches to discover event and landmark tags from the spatio-temporal distribution of Flickr tags are proposed. The motivation of these methods is to facilitate browsing and extraction of representative landmark tags. Other than in our work, the authors look at the individual distributions of tags, without taking correlations between different tags into account. Also, the semantics are restricted to landmarks, i.e., no distributed geographic phenomena are investigated.

Georeferenced social media has also been used in supervised learning to predict real-world phenomena on the basis of feature distributions. In [3] a general approach for temporal prediction is proposed. In [17] snow coverage is predicted on the basis of feature occurrences in Flickr. Both works show that georeferenced social media is valuable input to learn models about real-world phenomena. In this work we are interested in unsupervised learning of geographic features. However, we believe that our approach allows to extract informative dimensions that are also valuable input to learning tasks. We shortly describe differences of dimensionality reduction in a supervised and unsupervised learning context in Section 5.

In [18] a clustering of tags based on their spatio-temporal distribution was proposed to find geographically similar tags. The resulting tag-clusters can be interpreted as latent geographic features, as introduced in this work. However, by using clustering (Kmeans) instead of dimensionality reduction the tags are restricted to belong to a single feature and their contribution to a feature is not well defined. In the domain of text mining there has been recent interest in extending probabilistic topics models by additional attributes, like the geographic location of a document [6, 16]. Even if the resulting topics mainly represent document topics, they might be interpreted as geographic features in our sense. In [6, 16] they jointly model the occurrence of features in documents and the co-occurrence in geographic space. We focus on exploiting the spatial correlation among features to extract meaningful geographic signals and leave the study of document level correlations for geographic feature extraction for future work. The above mentioned approaches do not handle the exponential spatial distribution of documents as covered in our work using proper normalization.

A qualitative comparison of the different approaches in an unsupervised context is highly task specific. Our proposed framework and the used techniques have the advantage of a clear statistical justification with respect to spatial correlation, and/or independence between the extracted features. We plan to compare the different approaches mentioned above to our work based on a task-based evaluation in future work. In this work we focus on comparing different dimensionality reduction techniques and discuss the impact of normalization.

3. DEFINITIONS AND PROBLEM STATEMENT

In this section we introduce some basic definitions and formulate the problems of exploratory geographic feature discovery and latent geographic feature extraction.
3.1 Georeferenced Social Media

A georeferenced social media collection is a set of georeferenced documents $D = \{d_1, \ldots, d_m\}$. A document $d = (x, u, l, t)$ has a document-feature count vector $d.x$, a user id $u$, a geographic location $d.l$, and a timestamp $d.t$. Each component of the vector $x \in \mathbb{R}^p$ represents a feature from a global set of document features $F = \{f_i\}_{i=1}^n$. The count of a feature $f_i$ is $d.x_i$. We also use $d.x_i$ if the index is unknown. The location (georeference) of a document $d \in D$ is given by a point $l \in W \subseteq \mathbb{R}^2$, where $W$ denotes the geographic area of interest. In this paper, we use a sample of documents gathered for a time interval $T$ and an area of interest $W$. The collection $D$ provides us with an unstructured source of noisy geographic information.

3.2 Geographic Features

We define a geographic feature $f$ as a dimension representing the characteristic of a geographic location $l \in W$. This view is motivated by the use of the term feature in the information retrieval and machine learning literature. There, feature is mostly synonymous to attribute or input variable [4]. For instance, temperature, the amount of restaurants, or the number of mountains around a geographic location $l$ are geographic features based on the above definition. A geographic feature is synonymous to a spatial variable, as commonly used in statistical literature.

We denote geographic features as $f_1, \ldots, f_k$. Using the description of a geographic feature above, $f_i$ defines a function over locations in geographic space:

$$\phi_i : W \rightarrow \mathbb{R}$$

$$l \mapsto \phi_i(l)$$  \hspace{1cm} (1)

The function $\phi_i$ is called geographic feature sensor, and $\phi_i(l)$ is called geographic feature signal, that is, the influence of a feature $f_i$ at location $l$. Given a set of geographic features $f_1, \ldots, f_p$, the sensors $\phi_1, \ldots, \phi_p$ form a multivariate sensor:

$$\Phi(l) := (\phi_1(l), \ldots, \phi_p(l))^T$$  \hspace{1cm} (2)

In a multivariate sensor, we refer to the individual feature sensors as components. Using a feature sensor we are able to measure a multivariate feature signal at every location $l \in W$.

A spatial sampling scheme $L = (l_1, \ldots, l_n)$ is a set of locations at which the signal is measured. The most simple scheme is a regular grid over $W$ with a step width $w$. We use a sampling scheme to create a location sampling matrix:

$$X := (\Phi(l_1), \ldots, \Phi(l_n))^T \begin{pmatrix} \phi_1(l_1) & \cdots & \phi_p(l_1) \\ \vdots & \ddots & \vdots \\ \phi_1(l_n) & \cdots & \phi_p(l_n) \end{pmatrix}$$  \hspace{1cm} (3)

In $X$ each row $x_i$ corresponds to a location, and each column $X_j$ corresponds to a feature sensor. An entry of the matrix $x_{ij} := \phi_j(l_i)$ thus is a signal of feature $f_j$ measured at location $l_i$. We treat this matrix as a statistical sampling matrix with locations $x_1, \ldots, x_n$ being multivariate observations and columns $X_1, \ldots, X_p$ being random variables. From now on we omit the prefix geographic when clear from the context.

3.3 Signal Estimation

A social media collection $D$ provides the source to compute geographic features and to estimate the features’ signals. First, every document feature $f \in F$ is assumed to be a possible geographic feature. In this work we use the vocabulary of tag terms as document features $F$. Note, however, that other document features like words, labels (e.g., hashtags), urls, or classified entities could also be used. Each document feature (tag) $f_1, \ldots, f_p$ defines a feature sensor $\phi_1, \ldots, \phi_p$. The resulting set of sensors is described by the high-dimensional multivariate raw feature sensor $\Phi(l) \in \mathbb{R}^p$.

To estimate the sensor $\phi_i$ of a single feature $f_i$ we use the documents’ georeferences. Generally, a georeferenced document $d$ can be seen as noisy evidence that the document features $d.x_i > 0$ provide information about the location $d.l$. Given a large amount of documents, the evidence that features are informative for a location $l$ will increase if those features occur more often in documents nearby that location.

A single user can contribute an arbitrary number of documents to a collection $D$. In such a case, the user’s documents and the document features will often describe the same geographic phenomenon, e.g., the same city, district, place name, or place type. Our aim is to estimate an informative geographic feature signal reflecting a geographic phenomena. Users contributing a large number of documents about the same phenomena would introduce a user bias to the signal. Our assumption here is that a large number of different users that use the same feature provide more evidence than a single user using that feature in many documents. Therefore, we will use the number of different users using a feature $f_i$ around a location $l$ as basis for estimating $\phi_i$. This approach is also used in related work, e.g., [12, 17, 18] and is further discussed in Section 4.

Similar to approaches mentioned above, we use a regular rectangular grid to extract the signal of a feature at a location. The grid is defined as $C_w = (c_1, \ldots, c_n)$, with $w$ being the width/height of the cells $c_i$. For each cell, we count the number of different users that use a feature in their georeferenced documents. This can be equally formulated as a non-parametric density estimation with bandwidth parameter $w$: we try to find an estimator $h_f$ of the density of different users using a feature $f$ at a location $l \in W$. The signal is recovered by the product of the density with the feature signal $\phi_i$.

$$\phi_i(l) := |U_f| \cdot \hat{h}_f(l)$$  \hspace{1cm} (4)

Using the grid $C_w$ corresponds to a window counting (histogram) estimator with bandwidth $w$. Let $c_i \in C_w$ denote the cell that contains location $l$. Then the estimator is

$$\hat{h}_f(l) := |U_f|^{-1} \left| \{d.u \mid d \in D \wedge d.x_f > 0 \wedge d.l \in c_i \} \right|$$  \hspace{1cm} (5)

This estimator provides estimations that are non-smooth with respect to the grid resolution $w$. This is a reasonable input to our approach for the extraction of latent geographic features. However, other non-parametric estimators $h$ like Kernel-density estimation or Splines might be used to gather smooth estimates for $\phi_i$, if necessary.

Note that the bandwidth $w$ has a particular meaning in the geographic domain. Using a small bandwidth results in small scale variation of the feature signal while large band-widths will reflect variations on a larger scale. The bandwidth needs to be chosen according to the scale level of in-
terest. A feature signal reflecting the variation of a phenomenon on a global scale should have a larger bandwidth than a signal reflecting a city-scale phenomenon. For our experiments in Section 6, we choose \( w \) to reflect country scale and city scale phenomena according to the area covered by the data.

We use the grid \( C_n = (c_1, \ldots, c_n) \) also as the sampling scheme \( L = (l_1, \ldots, l_n) \) for the location sampling matrix \( \mathbf{X} \) (see Eq. 3). Each cell \( c_i \) corresponds to a location (cell) observation \( l_i \). By coupling the feature signal estimation and the location sampling step we set the density estimation bandwidth and the distance between location observations to the same value \( w \). An estimation bandwidth larger than the location distance will introduce spatial interaction between neighboring observations (smoothing), while a smaller bandwidth will introduce space that is not observed. In the latent feature extraction approaches we present below, we do not account for spatial interaction between location (cell) observations. Sampling the observations as independent (i.i.d) as possible (which is achieved by not smoothing) is a useful property from a statistical point of view and results in more informative features.

3.4 Latent Geographic Features

As discussed above, each location \( l \) can be described by a high-dimensional multivariate feature signal \( \Phi(l) \). Studying the univariate distributions \( \phi_1, \ldots, \phi_p \) allows to select features that are representative for a particular location (landmark extraction) [2, 9, 14]. Also, selected univariate sensors can be used as input for supervised learning tasks [17]. Here, instead of studying individual feature signals our idea is to extract informative dimensions from the given multivariate distribution. For this, we exploit the statistical structure contained in the location sampling matrix \( \mathbf{X} \) using dimensionality reduction techniques. Details on the process of extracting latent features are presented in Section 5. We first give a general definition of a latent geographic feature:

A latent geographic feature \( \tilde{f} \) is a signal that cannot be observed (estimated) directly in the collection. Generally \( \tilde{f} \) is a combination of the components \( f_1, \ldots, f_p \) of the raw feature sensor \( \Phi \). The component loadings (coefficients) of this combination reflect their importance for the extracted latent feature \( \tilde{f} \). We denote the \( p \) loadings of a latent feature \( \tilde{f} \) as the vector \( \mathbf{a} \). We then say a component has a high/low loading for a latent feature. We use the components \( \mathbf{a} \) ordered by their loadings to extract the most important components as a label for \( \tilde{f} \). For combinations where the loadings are positive and negative, we change the sign (of the loadings and the signal) such that the loading with the largest absolute value is positive.

A latent geographic feature sensor \( \tilde{f}_i \) is used to measure the influence of \( f_i \) at a location \( l \). A set of \( k \) extracted latent geographic features is called a latent geographic feature sensor, defined as:

\[
\tilde{\Phi}(l) := (\tilde{\phi}(l)_1, \ldots, \tilde{\phi}(l)_k)^T
\]  \hspace{1cm} (6)

3.5 Problem Statement

Given a georeferenced social media collection \( \mathcal{D} \). We assume that the collection is an interesting geographic information source for exploring geographic space. We are interested in extracting a small set of informative geographic features, which can be seen as basic building blocks of geographic semantics. We define this general task as:

Exploratory Geographic Feature Discovery: The exploratory process of discovering a small number of informative geographic features (spatial variables) from a noisy source of unstructured geographic information.

In our approach, we use a raw feature sensor \( \Phi \) of a collection \( \mathcal{D} \) to extract a low number of informative latent features. A raw sensor is expected to initially contain a large number of noisy signals of different geographic semantics. We describe the process to extract informative latent geographic features as:

Latent Geographic Feature Extraction: Given a georeferenced document collection \( \mathcal{D} \) and a raw sensor \( \Phi \in \mathbb{R}^p \) with a component for every document feature \( f \in \mathcal{F} \). Latent geographic feature extraction finds a number \( k \ll p \) of latent geographic features that are informative to describe geographic space \( \mathcal{W} \) of interest.

4. DATA CHARACTERISTICS AND NORMALIZATION

In this section we discuss characteristics of geographic features. Then, we develop statistical properties of georeferenced social media collections and discuss their implications for the process of discovering latent geographic features. Empirical results are obtained from the US Flickr data set introduced in Section 6.1. These results are also valid for the LA data set.

4.1 Geographic Feature Characteristics

We assume that feature signals have different geographic characteristics affecting the value of a feature to describe geographic space. The characteristics of a feature \( f \) are discussed based on the spatial distribution of its signal \( \phi_f(l) \), in particular based on the measured signal over a set of locations \( L \). Our objective is to find the following classes of geographic feature sensors in \( \Phi \):

- **Global**: A feature might be used in a substantial part of geographic space. Such a feature allows to identify a location and acts as a descriptive label. However, such a feature will not allow us to discriminate between large proportions of geographic space, except between the location where the label occurs and the rest. We call a feature exhibiting such a property a landmark feature.

- **Regional**: A feature might be used in a substantial part of geographic space. However, its usage might differ significantly from the baseline distribution \( B \). Such a feature is expected to contain relevant geographic information for large parts of the geographic space, allowing us to compare locations based on their semantics. We call such a geographic feature a regional feature.
It is important to note that this classification of geographic features based on the feature signal distribution depends on the chosen area of interest. A feature might be a global feature in a certain area (like the city name in the city area), but a landmark feature when looking at the whole country. Landmark and regional features might fit best our definition of informative geographic features. However, also a global feature might be a valuable signal if a description of the global density is of interest. Also, latent geographic features with global characteristics allow us to identify raw features (components) that can be thought of as non-informative geographic stop words.

To compare feature sensors on the basis of their geographic characteristics we use the entropy of the feature signals. Given a location sampling matrix $X$. Each column $X_1, \ldots, X_p$ represents a distribution of the feature signal over the $n$ locations. Let $p(X_j = l_i)$ be the density of feature $X_j$ at location $l_i$, i.e.,

$$p(X_j = l_i) := \frac{x_{ij}}{\sum_{k=1}^{n} x_{kj}}$$

The entropy of the feature signal $X_j$ is given as

$$H(X_j) := -\sum_{i=1}^{n} p(X_j = l_i) \log p(X_j = l_i)$$

The entropy is the average amount of information needed to encode the events that a feature occurs at locations $l_1, \ldots, l_n$. A high entropy means the feature occurs in large parts of geographic space, hence we need some information to encode such an event. A low entropy means the feature only occurs in a small part of the geographic space. Thus, we do not need much information to encode this event. A feature sensor with high entropy would correspond to global features, while feature sensors with low entropy correspond to landmark features. Figure 3 shows 5 features with highest, lowest, and around median entropy for our US Flickr data set.

We defined a global feature to be a feature whose spatial distribution is similar to the baseline distribution of a collection. A similarity measure in information theory to compare two distributions is the Kullback-Leibler (KL) divergence [5]. Figure 2(a) shows the scatter plot of Flickr features by their entropy and their KL-divergence to the baseline distribution. For our data set and the chosen location, sampling the entropy and the KL-divergence are highly correlated. Hence, in the following, we will discuss extracted latent geographic features with respect to their entropy. Depending on different normalization strategies and feature extraction approaches the resulting latent features might be either more global, regional, or landmark-ish.

### 4.2 Georeferenced Document Distribution

To motivate normalizations strategies of the raw feature sensor, we first discuss important statistical properties of the input data set. Given a collection of georeferenced documents $D$ and a grid $C_w = \{c_1, \ldots, c_n\}$ (analogous to Section 3.3, we use $l_1, \ldots, l_n$ to refer to the cells). For each location (cell) $l_i$, we calculate the following statistics based on the document georeferences, like in Eq. (5):

$$F(l) : \text{Total number of features at } l$$

$$F_d(l) : \text{Number of distinct features at } l$$

$$U(l) : \text{Number of distinct users at } l$$

$$D(l) : \text{Number of documents at } l$$

$$F(l, f) : \text{Count of feature } f \text{ at } l$$

$$U(l, f) : \text{Number of distinct users using } f \text{ at } l; \phi_f(l)$$

Figure 2(b) shows the distribution of the statistics in log-scale over locations $l \in L$ reversely ordered by $F(l)$. The plot only shows the results for 300 out of 1728 locations. Only 270 locations have a value greater than zero. This means that only 16% of the area $W$ contain a feature signal. This is clearly expected as georeferenced social media is likely not to occur at locations where no users are present. As a consequence, a latent feature extraction approach should be able to work with a small number of location observations.

The number of documents $D(l)$ and the total number of features $F(l)$ at a location show a similar distribution as do the number of users $U(l)$ and the number of distinct features $F_d(l)$. The difference in the distribution of the two groups shows the influence of the user bias, as stated in Section 3.3. However, all distributions show an exponential behavior (note the statistics are on log-scale). It is therefore important to account for the fact that some locations will have an exponentially larger number of user contributions when estimating feature signals.

Figure 2(c) shows the distribution of location statistics for the feature beach over $l \in L$ reversely ordered by $U(l, f)$. Only the first 60 locations out of 1728 are shown (51 locations have a signal). The feature number $F(l, f)$ and the number of distinct users $U(l, f)$ are given in log-scale. The difference between total feature counts and the number of contributing user (user bias) is clearly visible. However, still the distribution $U(l, f)$ has an exponential behavior, which

<table>
<thead>
<tr>
<th>Min</th>
<th>Median</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>massachusetts 0.78</td>
<td>sidewalk 2.69</td>
<td>nikon 4.49</td>
</tr>
<tr>
<td>seattle      0.72</td>
<td>waves 2.69</td>
<td>trees 4.49</td>
</tr>
<tr>
<td>losangeles   0.59</td>
<td>band 2.69</td>
<td>usa 4.61</td>
</tr>
<tr>
<td>sanfrancisco 0.47</td>
<td>smithsonian 2.69</td>
<td>landscape 4.67</td>
</tr>
<tr>
<td>chicago      0.40</td>
<td>olympus 2.69</td>
<td>nature 4.72</td>
</tr>
</tbody>
</table>

Figure 3: Geographic features (tags) from US Flickr data set grouped by maximal, median, and minimal entropy $H(X_i)$. Details of the plots are discussed in Section 4.
needs to be taken into account when defining a feature signal.

Summary: The social media data set from Flickr we studied (see Section 6) shows the following characteristics: (a) Only a small area of the total area of interest contains documents and hence a feature signal. For a given sampling scheme and feature signal estimation bandwidth $w$, this results in a low number of location observations. Hence, the location sampling matrix $X_{n \times p}$ possibly has a lower number of observations $n$ than features $p$. (b) The spatial distribution of documents, users, and user contributions to a feature shows exponential characteristics. Hence, a few locations will dominate the signal. To limit the influence of these locations, the signal needs to be normalized accordingly. (c) The few locations having a larger number of contributing users also have a higher number of distinct features, and hence a larger number of feature sensors with a signal. This fact makes the location observations heterogeneous. To reduce the influence of the number of signals to latent geographic feature extraction, we propose an approach that assumes the extracted latent geographic features to be sparse. A location with a large number of signals will then contribute to a variety of latent features, while locations with few signals might only contribute to a single latent feature. We discuss the sparsity assumption in Section 5.3.

Normalization: To account for the exponential characteristics of the feature signal distributions we normalize accordingly. Let $\phi_i$ be the signal normalized from the signal $\phi_i$. In Section 6, we discuss the effects of the following normalizations:

- **Logging:** $\phi_i'(l) := (\log(\phi_i(l)) + 1)^p_{p=1}$
- **Binarization:** $\phi_i'(l) := \{1(\phi_i(l) > 0)\}_{j=1}^p$

Logging follows from the discovered exponential characteristics of the distribution and transforms the signal in a linear domain. It sharply reduces the feature signal at locations having a very high value while letting it be higher than in locations with less signal. Binarization is a brute-force normalization that sets the signal to 1 if any signal was measured. Statistically the correlation between features then only depends on whether or not features occur together at locations.

5. LATENT GEOGRAPHIC FEATURE EXTRACTION

We now describe different approaches to extract informative latent geographic features from a raw feature sensor $\Phi$ and its corresponding location sampling matrix $X$. Generally, the techniques we use to extract latent geographic features belong to the class of dimensionality reduction techniques. Dimensionality reduction tries to project data points of a high-dimensional data matrix $X_{n \times p}$ (e.g., $n = p$ or even $n < p$) onto a low-dimensional subspace $\tilde{X}_{n \times k}$, with $k \gg p$. One way of accomplishing such a task is to select features (variables) from the data matrix. Such a selection can be based on heuristics or on feature selection algorithms in a supervised learning context. In the latter case, those features are selected that increase the performance of a predictive model. Our aim is to extract in an unsupervised fashion new features that exhibit informative geographic characteristics. Such techniques are known as feature extraction.

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 location observations with the dimensions being feature signals. We expect that the location sample matrix $X$ of a raw feature sensor contains an interesting statistical structure expressed in correlations of raw feature signals over the location observations.

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. We then discuss Sparse PCA as a recent technique proven useful for high-dimensional input [5, 20]. In the following, we briefly discuss the techniques with a particular focus on the interpretation of the resulting extracted features. For a detailed overview of dimensionality reduction, PCA, and ICA see [5]. SPCA was proposed in [20].

5.1 PCA

PCA seeks linear combinations of the original variables that maximize the variance of the data points. 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 data matrix $X_{n \times p}$ can be found by Singular Value Decomposition (SVD). Given a centered data matrix $X$, the SVD factorization is:

$$X_{n \times p} = U_{n \times p}D_{p \times p}V_{p \times p}^T$$

The $n$ data points described by PCs are the columns of $UD$. The variable loadings (coefficients) for the $i$th PC are in the $i$th column of $V$. With the data matrix being the location sampling matrix the mapping to latent geographic features is defined as

$$\Phi_i(l) := (UD)_i, \quad i = 1, \ldots, n$$

PCs have the property to be mutually uncorrelated and are ordered by their amount of variance they contribute to describe data points. Similar and dissimilar observations will be better described by the first $k$ dimensions of the data space. Representing an extracted geographic feature with a PC means that the features represent uncorrelated geographic phenomena.

In our setting many locations will have no or only very little signal while a few have a lot. A major distinction between locations is thus the amount of signal they have. The first PC will express this information. It will be less informative to describe geographic phenomena but will follow the density of the baseline distribution.

5.2 ICA

ICA is a technique to separate a multivariate signal into source components that are mutually statistical independent, rather than uncorrelated. The original signal can be

$^4$Application of ICA as an explorative feature extraction technique is also known as Projection Pursuit.

$^5$Note that here component means an extracted feature from the data matrix, not a component from a multivariate feature signal.

$^6$Statistical independence is an even more strict assumption between random variables than non-correlation.
recovered by mixing the source components appropriately. ICA has been applied successfully 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 for face recognition [1, 7]. In all those applications a representation of the original signal by statistical independent source components captures the essential structure of the data. Given a data matrix $X$, ICA is described as:

$$X_{n 	imes p} = S_{n 	imes p} A_{p 	imes p}$$

(12)

$S$ describes the source components in the columns. $A$ is the matrix of mixing coefficients, describing how the source components generate the observed input. Both, $S$ and $A$ are estimated by ICA\(^6\). ICA has the property that the source components (columns of $S$) $S_1, \ldots, S_p$ are mutually independent. A source component $S_i$ is a distribution over $n$ observations. Mutual independence then corresponds to both, non-Gaussian and low entropy of the distribution $S_i$. ICA finds the same number of source components as mixed components. To extract a smaller number of source components the dimensionality of the input is reduced to $k$ dimensions using PCA in a preprocessing step. Applying ICA to a location sampling matrix finds a set of statistically independent latent geographic features. The mapping is given as

$$\Phi(l_i) := s_i, \quad i = 1, \ldots, n$$

(13)

$$a_j := A_j, \quad j = 1, \ldots, p$$

(14)

In this context independence can be stated as follow: two latent geographic features $\phi_i$ and $\phi_j$ are independent if, for all locations $l_1, \ldots, l_n$, it is not possible to predict the signal $\phi_i(l_i)$ on the basis of the signal $\phi_j(l_i)$. This means that the extracted features should represent independent geographic phenomena.

5.3 Sparse PCA

In PCA and ICA the resulting components are assumed to be combinations of all variables. Hence, a latent geographic feature will have loadings for almost all of the raw features (with many raw features having loadings 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 raw features.

Moreover, 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 components. Hence, a component is only made of a few variables with loadings different from zero. Such assumptions correspond to restricting the statistical model explaining the data, allowing for more meaningful analysis even if we have few statistical observations [11].

In SPCA the PCs are described by a maximum number of $\alpha$ variables, with $\alpha \ll p$. SPCA was introduced in [20], showing how it can be solved using a linearized regression approach. The mapping of sparse PCs to latent geographic features is the same as in PCA. The sparsity assumption results in latent geographic features $\phi_i$ having only $\alpha$ or less components different from zero in $a_i$.

\(^6\)We use the FastICA algorithm to find the source and mixing matrix.

\(^7\)http://www.flickr.com/services/api

Figure 4: Distinct users $U(l)$ of the US and LA Flickr data set in log-scale. Please view in color.

6. EXPERIMENTS

In this section we present the performance results of our feature extraction techniques and their value as explorative geographic feature discovery tools. First, we describe the social media collections used as input for our experiments. Then, we qualitatively compare the different characteristics and the informational value of extracted latent geographic features for the presented techniques. Then, we study the impact of normalization to the characteristics of extracted latent geographic features and show that normalization can be used as a parameter in the explorative process. Finally, we present discovered latent geographic features from an LA data set.

6.1 Data

As input data we use georeferenced documents retrieved using the Flickr API\(^7\). We use a data set requested by a query over the US with timestamps from 2008 to 2011 and one query over the Los Angeles city area with timestamps from 2010 to 2012. The details of the data sets are given in the following table.

|       | $|F_{US}|$ | $|F_{LA}|$ |
|-------|----------|----------|
| $m$   | 5,976,689| 245,312  |
| $|F_{US}|$ | 30,323 (695 filtered) | 5,796 |
| $|U|$ | 3,081 | 5,796 |
| $W$ | US bounds $(-124.87, 25.25, -52, 50.06)$ | LA city bounds $(-111.87, 25.25, -52, 50.06)$ |
| $C_w$ | $w = 1.0 (111 \text{ km})$ | $w = 0.04 (1.11 \text{ km})$ |
|       | $72 \times 24 = 1728 \text{ cells}$ | $43 \times 25 = 1075 \text{ cells}$ |
|       | 585 cells with signal | 945 cells with signal |

From both data sets we removed those document features 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 4 shows the spatial distribution of users $U(l)$ for both data sets. For each data set, we created the location sample matrix $X$ using the corresponding grid $C_w$.

6.2 Comparison

We now compare the latent geographic features (LGFs) extracted by PCA, ICA, and SPCA. Generally, we select 3 LGFs from the SPCA results and find the most similar LGFs in the PCA and ICA results on the basis of their spatial signal distribution. For all techniques we use a log-normalized
input signal and $k = 20$ as the number of reduced dimensions. Figure 5 shows 3 LGFs for each technique and for both data sets. We chose similar LGFs between the different techniques as described above to compare their characteristics and informational value.

In the LA data set the selected LGFs in the SPCA results can be labeled as city feature (LA downtown related), hollywood feature, and (venice) beach feature. For PCA we have to choose the first 3 extracted features ordered by their variance because only the first latent feature has a meaningful signal. This can be explained by the fact that the location sample matrix $X_{n \times p}$ is fat ($p > n$), in which case the inverse of the covariance matrix is of no full rank. Although it is possible to extract PCs, it is not well defined for such cases. This results in noisy and non-informative LGFs except for the first latent feature, which follows the baseline distribution of documents.

ICA extracts more informative features and the mentioned LGFs are found. We see that the city feature is described by components having global feature characteristics ('losangels', 'square'), hence it is similar to the baseline distribution (see Figure 4 for baseline in log-scale). The Hollywood feature has a signal around the Hollywood area. It is only described by a single representative component ('hollywood'), with the rest of the components corresponding to global features. The beach feature has a signal around Venice beach (lower-left of the map) and is described by the components 'beach', 'venice', 'water' and 'venicebeach'. Less representative global components are 'california' and 'losangels'.

Differently, SPCA extracts highly informative LGFs with distinct characteristics. The features are less polluted by global components like 'losangeles', 'california', or 'instagramapp'. The city feature components are highly informative for an inner city environment ('art', 'street', 'grafitti', 'food'). The Hollywood feature components are highly specific for this area ('hollywoodboulevard', 'walkoffame', 'star'). Furthermore, SPCA also extracts other hollywood related features ('universal studio', 'griffith park'; see Figure 8(a)), which we will present in Section 6.4. The beach feature is highly informative for the coastal environment found at Venice beach ('beach', 'sunset', 'sky', 'blue', 'ocean') and is not polluted by global components at all.

From the US data set we select the features nature, coast, and desert region for comparison. Here, the input location matrix $X_{n \times p}$ is skinny ($n > p$), resulting in more meaningful results for PCA. PCA extracts a nature feature with dominant signal on the west coast. It is strongly influenced by global components ('usa', 'california'). ICA extracts the nature feature with a signal better distributed over the US, with high intensities in the mountain regions. Still, the feature is loaded by non-informative global components ('usa', 'geotagged'). SPCA extracts the feature with a signal distributed evenly in the mountain areas. All components are highly informative for the regions ('landscape', 'mountain', 'nationalpark', among other). The same observations hold for the extracted coast feature and the desert region feature.

Observations: For the input data sets and the selected LGFs we see that SPCA results are highly informative regarding their signal distribution and their component load-
ings. ICA performs better than PCA, with PCA being useless if the location sampling matrix is fat. This supports the fact that assuming sparsity of the extracted features plays a key role in extracting informative dimensions.

6.3 Normalization

Normalization has a strong impact on the characteristics of the resulting features. Without normalization the extracted features are dominated by components having a strong signal. Those components are found at locations with high user contributions (large cities, populated places) and describe landmarks (city-names, place-names). The extracted features then have landmark characteristics with components being informative for this area.

Logging reduces the impact of strong signals, resulting in less impact of such landmark components. The resulting feature components are less dominated by landmark-ish characteristics. Binarization almost diminishes the impact of strong landmark signals, resulting in extracted features with high loadings for components that are widely distributed.

Figure 6 shows the extracted latent geographic features (SPCA) with highest loading for the component 'beach' in the US data set using non-normalized, log-normalized, and binarized input. The non-normalized feature shows a signal in California, which is the area with highest user contributions for the beach component. However, the most dominant components are the landmark features 'california' and 'sanfrancisco'. The feature can hence be described as a California landmark feature, with 'beach' being a representative component. Using logged input the signal is spread on both coasts, still with a higher intensity on the west coast. The components are not dominated by landmark features anymore but represent coastal features. Using binarization the signal is evenly distributed on both coasts. Moreover, informative coastal components like 'ocean', 'beach', 'sea', and 'boat' have higher loadings (compared to the logged input).

Observations on Parameterization: In an explorative setting the user might want to discover more landmark-ish, regional, or global features. We use the strength of normalization as a parameter of latent geographic feature extraction. Figure 7 shows the impact of normalization to the entropies of the signal distributions of k = 20 extracted features 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. 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.

6.4 Exploration Results

We finally present some interesting discovered latent geographic features for the LA data set using the best performing technique, SPCA. Figure 8 shows 5 selected features using each, log and bin normalization. The log normalized feature correspond to interesting landmark features in the LA area. We discovered three interesting features for Hollywood (universal studios, griffith park, hollywood boulevard), a little tokyo feature around the central station, and an airport feature. Features not presented here are: venice beach, downtown, pasadena, lacma, among others. Note that although these features are discovered in an unsupervised fashion from a noisy geographic information source, they are highly informative to explore the LA city area.

![Figure 6: Selected SPCA features (k=20) with highest loading of the component 'beach' extracted using non-normalized input (top), logged input (center), and binarized input (bottom).](image)

![Figure 7: Entropies of latent geographic feature distributions.](image)

Using bin normalization the extracted features show more regional and global characteristics. They can be understood as attributes of geographic space. The feature can be labeled as: inner city, going out, tourist related, walking, and nature. Interestingly, the tourist feature has a high signal around the presented features using log normalization. This clearly makes sense as those landmark features are of high interest to LA tourists.

7. CONCLUSIONS AND ONGOING WORK

In this work we studied the task of exploratory geographic feature extraction from noisy unstructured geographic information sources. For this, we present a framework to extract latent geographic features using dimensionality reduction. Our experimental results show that (1) informative geographic features can be extracted by assuming sparsity of the latent feature signals and that (2) normalization allows to extract features of different geographic characteristics. We claim that geographic feature extraction has a variety of applications in market research and spatial analysis. For this, we plan to study feature extraction using other data sources and dimensionality reduction in a supervised learning context in our future work.

8. REFERENCES

Figure 8: Selected SPCA features from LA data set. Left column shows 5 landmark-ish features using logged input. Right column shows 5 regional features using binarized input.