On Quantifying Qualitative Geospatial Data: A Probabilistic Approach

Georgios Skoumas[*1], Dieter Pfoser[†2], and Anastasios Kyrillidis [‡3]

[1] Knowledge and Database Systems Laboratory, National Technical University of Athens, Greece

[2] Dept. of Geography and Geoinformation Science, George Mason University, USA

[3] Laboratory for Information and Inference Systems, Ecole Polytechnique Federale de Lausanne, Switzerland

November 21, 2013

Abstract

Living in the era of data deluge, we have witnessed a web content explosion, largely due to the massive availability of User-Generated Content (UGC). In this work, we specifically consider the problem of geospatial information extraction and representation, where one can exploit diverse sources of information (such as image and audio data, text data, etc), going beyond traditional volunteered geographic information. Our ambition is to include available narrative information in an effort to better explain geospatial relationships: with spatial reasoning being a basic form of human cognition, narratives expressing such experiences typically contain qualitative spatial data, i.e., spatial objects and spatial relationships.

To this end, we formulate a quantitative approach for the representation of qualitative spatial relations extracted from UGC in the form of texts. The proposed method quantifies such relations based on multiple text observations. Such observations provide distance and orientation features which are utilized by a greedy Expectation Maximization-based (EM) algorithm to infer a probability distribution over predefined spatial relationships; the latter represent the quantified relationships under user-defined probabilistic assumptions. We evaluate the applicability and quality of the proposed approach using real UGC data originating from an actual travel blog text corpus. To verify the result quality, we generate grid-based "maps" visualizing the spatial extent of the various relations.

1 Introduction

During the last decade, we have witnessed an explosion in the amount and variety of content available on the Web. Sophisticated analysis of such UGC has

[*] gskoumas@dblab.ece.ntua.gr

[†] dpfoser@gmu.edu

[‡] anastasios.kyrillidis@epfl.ch

become an important issue in many cutting edge research fields such as Geographical Information Science. In this work, our goal is to take advantage of such volunteered geographic information in geospatial data analysis. In particular, when applied to the geospatial domain, this translates to massively collecting and sharing knowledge in order to ultimately model and chart the world.

Traditionally, quantitative information in the form of spatial coordinates is the data used in virtually all geospatial applications. With spatial reasoning being a basic form of human cognition, qualitative spatial data in the form of spatial relationships (North, South, In, Close, Next, Far, etc.) is what people typically use in order to describe spatial scenarios. Such data makes a prime source for user-contributed geospatial content. Especially narratives expressing such experiences typically contain spatial knowledge. However, one of the drawbacks of this data is its lack of precision as qualitative relations are interpreted differently by the users (in contrast to coordinates).

As a motivational example we could consider the sentence; *"Big Ben is the nickname for the great bell of the clock at the north end of the Palace of Westminster"*. In this case, we want to quantify what people imply when they say *"North"* in terms of distance and direction. Having quantified *"North"* in this context and knowing the location of either *"Big Ben"* or *"Palace of Westminster"* will allow us to infer possible locations for the other. Eventually, by collecting more observations of this form, we will be able to refine the location and, thus, locate spatial objects that otherwise could not be geocoded. Figure 1 illustrates the underlying idea by relating our observation-based approach to the triangulation problem from surveying engineering, where an unknown location is determined by "observing" known locations.

To this end, we consider the following problem:

PROBLEM: *Given a set of spatial objects $\mathcal{K}$ whose positions in space are known, a set of spatial objects $\mathcal{U}$ whose positions in space are unknown, and a set of spatial relationships $\mathcal{R}$, find probabilistic estimates of the positions of objects of set $\mathcal{U}$ based on their spatial relationships $\mathcal{R}$ with objects of set $\mathcal{K}$.*

To achieve this, our approach follows a probabilistic path: the proposed method quantifies qualitative relations as probability measures based on crowdsourced multiple observations contained in texts. Each observation is roughly quantified using a spatial feature vector comprising distance and orientation. Then, a greedy Expectation Maximization-based (EM) method is used to train a probability distribution. The latter represents the quantified spatial relationships under a probabilistic framework, i.e., it provides a set of random variables (spatial feature vector) that have certain probability density functions (PDFs) associated with them, for a specific spatial relation.

In this work, we employ probabilistic models to represent spatial relationships. To the best of our knowledge, this is the first work that combines qualitative and quantitative spatial information for spatial probabilistic inference. The novelty of our approach lies in the process of mapping textual crowdsourced and uncertain location observations to their probable locations based on probabilistic spatial relationship models. The traditional machine learning techniques that we employ have never been used to achieve such a mapping. Moreover, this approach is one of the pivotal steps in developing automatic map-generation-from-text tools based on crowdsourced data.

The outline of the remainder of this work is as follows. Section 2 discusses

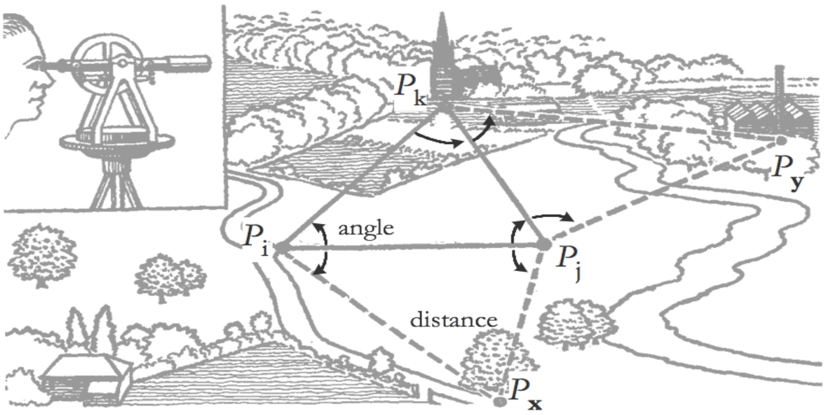


Figure 1: Intuitive problem formalization. P_i, P_j and P_k represent known locations that are used to compute several unknown locations (P_x and P_y) based on distance and angle observations.

related work. Section 3 discusses the specific qualitative data involved and introduces the spatial feature vectors used for quantification, while Section 4 introduces the tools necessary to derive quantification in the form of PDFs for the spatial relationships. Section 5 validates the proposed approach by means of "mapping" the obtained quantified relationships and using a similarity metric to assess the iterative quantification process. Finally, Section 6 presents conclusions and directions for future work.

2 Related Work

Work relevant to this paper includes qualitative modeling of spatial relations with application to spatial data management, and quantitative modeling of spatial knowledge.

Qualitative: The majority of works related to qualitative approaches for spatial information representation considers spatial relations. One popular spatial classification is constructed by topological relations (e.g., disjoint, overlap), direction relations (e.g., North, South), ordinal relations (e.g., inside, contain), and distance relations (e.g., far, near). The authors in [8, 9, 10, 17] present formal methods for qualitative representation of spatial relationships based on mathematical theories of order. Their applicability on spatial database systems and some key-role technical concepts are coherently discussed in [14, 24, 25]. Qualitative representation of spatial knowledge is discussed in [12, 20, 23]. The authors identify the common concepts of the qualitative representation and processing of spatial knowledge. They compare the representational properties of different systems and outline the computational tasks involved in relation-based spatial information processing.

Quantitative: Recent research on quantitative representation of spatial knowledge has been conducted in relation to situational awareness systems, robotics, and image processing. Modelling uncertain spatial information for situational awareness systems is discussed in [18] and [22]. The authors propose a bayesian probabilistic approach to model and represent uncertain event locations described by human reporters in the form of free text. They analyze several types of spatial queries of interest in situational awareness applications. Estimation of uncertain spatial relationships in robotics is addressed in [27]. The paper describes a representation of spatial information, called the stochastic map, and associated procedures for building it, reading information from it, and revising it incrementally as new information is obtained. The stochastic map contains the estimates of relationships among objects in the map, and their uncertainties, given all the available information. A probabilistic algorithm for the estimation of distributions over geographic locations is proposed in [15]. The authors use a data-driven scene matching approach in order to estimate geographic information based on images. In [28] the authors attempt to create a hierarchical probabilistic concept-oriented representation of space, based on objects. Their approach is based on learning from exemplars, clustering and the use of Bayesian network classifiers. Such a conceptualization and representation can enable robots to be more cognizant of their surroundings. Image similarity based on quantitative spatial relationship modeling is addressed in [31]. The authors propose a novel method for the representation of relative spatial relations between objects in images, applied to multimedia database applications. Finally, there has been some theoretic work on modeling spatial uncertainty using heuristics and fuzzy logic techniques. For example, in [32], a fuzzy decision tree algorithm is proposed to formalize spatial relations between linear objects.

3 Spatial Features from Qualitative Data

The main contribution of this work is to model qualitative spatial information in a quantitative and probabilistic way. Our main data source will be narratives and this section will survey our approach for extracting qualitative data from texts. Moreover, to be able to quantify qualitative spatial data, we need to have a respective means for representing it. Here, we present the spatial feature vector that models spatial relationships based on distance and orientation measures.

3.1 Dataset

Crowdsourced narratives are likely to contain spatial information. The more relevant the text is to "space", the more data it will contain. Our specific case considers travel blogs as a rich potential data source. This assumption is based on the intuition that people tend to describe their experiences in relation to their trips and places they have visited. This behavior results in "spatial" narratives.

To obtain such data, we used classical web crawling techniques as presented in [6] and we compiled a database[1] consisting of 120K user generated texts obtained from travel blogs[2].

[1] Available upon request

[2] TravelBlog, TravelJournal, TravelPod

3.2 Spatial Relations

Obtaining qualitative spatial data from text involves the detection (*i*) of spatial objects, i.e., Points-of-Interest (POIs) or toponyms and (*ii*) of spatial relationships between those POIs. The employed approach involves geoparsing, i.e., the detection of candidate phrases, and geocoding, i.e., linking the phrase/toponym to actual coordinate information. Using GATE's [4] text processing and semantic analysis components in combination with the algorithm presented in [6], we managed to extract 120k POIs from the text corpus. For the geocoding of the POIs, we rely on the open-source module GeoGoogle[3], a Java API utilizing the geocoding service that is part of the Google Maps API. This procedure associates (whenever possible) geographic coordinates with POIs that have been identified in the travel blog data. The final result of this stage is an index that contains the geographic coordinate information plus text information of all POIs, i.e., document, paragraph, sentence and word distance information.

The following excerpts are texts that contain relevant POIs and respective spatial relationship data.

- "...and then went out for tea at a lovely Italian restaurant in *Soho* **near** *Covent Garden*"
- "*Tate Modern* is a big modern art gallery, **on** *South Bank*, amazing building, has some great stuff in it as well."

These examples confirm our initial hypothesis for the existence of spatial knowledge in user generated content. As expected, we observe that POIs are more dense in urban places.

Having identified and geocoded the spatial objects, the next step is the localization of qualitative spatial relationships. This would ideally require efficient natural language processing (NLP) tools to automatically extract and map phrases to spatial relations linking POIs as contained in texts. Kernel methods for semantic relation extraction between entities in texts are developed in [3] and [34]. In [33] and [35], Support Vector Machine (SVM) approaches are used to extract spatial relations for spatial reasoning. In [19], the authors report on a novel task of spatial role labeling in text, based on machine learning methods to extract spatial roles and their relations. Finally, extraction of semantic relations from texts using dependency grammar patterns is addressed in [11] and [30]. Overall, while several of these techniques would be useful for spatial relationship extraction from texts, none either performed in a satisfying way or were available to us.

For the scope of this work, which lies on the area of probabilistic modelling of qualitative data but not on the extraction of qualitative spatial knowledge from text, we overcome this problem by using human annotation in combination with filtering of the input dataset. We restrict the data to be considered (*i*) to the geographic area of London and further reduce it by (*ii*) only considering sentences that include at least two POIs (which are needed to express a spatial relationship). This manageable dataset (sentences) is then annotated by humans to extract spatial relationships. Human annotation results into tuples of the form shown in Table 1. Here $\mathcal{P} = \{P_1, \ldots, P_m\}$ represents the set of spatial

[3] http://geo-google.sourceforge.net

objects participating in binary spatial relationships $\mathcal{R} = \{R_1, \ldots, R_n\}$ with $i, j \leq m$ and $k \leq n$.

$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$
P_1	R_1	P_2
P_3	R_1	P_4
P_3	R_1	P_5
$\vdots$	$\vdots$	$\vdots$
P_i	R_k	P_j

Table 1: Dataset denoting (geocoded) spatial objects and spatial relations.

3.3 Spatial Features

Statistical models are often used to represent observations in terms of random variables. These models can then be used for estimation, description, and prediction based on basic probability theory. In our approach, we model a spatial relation between two POIs $P_k \in \mathcal{K}$ and $P_u \in \mathcal{U}$ (k declares known and u declares unknown as described in Section 1) in terms of *distance* and *orientation*. We consider a labeled *spatial feature vector* as two random variables that model spatial relations in a probabilistic way. Assuming a projected (Cartesian) coordinate system, the distance is computed as the Euclidean metric between the two respective coordinates. The orientation is established as the counterclockwise rotation of the x-axis, centered at P_k, to the unknown point P_u.

Several instances of a spatial relation are used to create a dataset which will be used to train a probabilistic model for each spatial relation. Under a mathematical formalization, let us consider that for each instance of each relation we create a two-dimensional spatial feature vector $X = (X_d, X_o)^\intercal$ where X_d denotes the distance and X_o denotes the orientation between P_k and P_u. We end up with a set of two-dimensional feature vectors $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ for each spatial relation where the i-th vector of each set has the form $X_i = (X_{di}, X_{oi})^\intercal$. An example of the feature extraction procedure is illustrated in Figure 2, where four instances of spatial relation *Near* are used in order to create the respective set of spatial feature vectors $\mathcal{X}_{near} = \{[X_{d1}, X_{o1}]^\intercal, [X_{d2}, X_{o2}]^\intercal, [X_{d3}, X_{o3}]^\intercal, [X_{d4}, X_{o4}]^\intercal\}$. In this scenario, $\mathcal{K} = \{A, D, E, G\}$ and $\mathcal{U} = \{B, C, F, H\}$.

4 Spatial Relation Modeling

In this section we discuss the methods and algorithms we used to train probabilistic models that can efficiently represent spatial relationships based on our dataset. More specifically, we start by describing the approach we use to populate our dataset by using Kernel Density Estimation (KDE) which is a state-of-the-art method for the estimation of a multi-dimensional probability density function. We continue by analyzing the Gaussian Mixture Model (GMM), which is the probabilistic model we employ for the quantitative representation of spatial relations and we outline a greedy learning algorithm for parameter estimation of the GMM. Finally, we discuss Kullback-Leibler (KL) divergence,

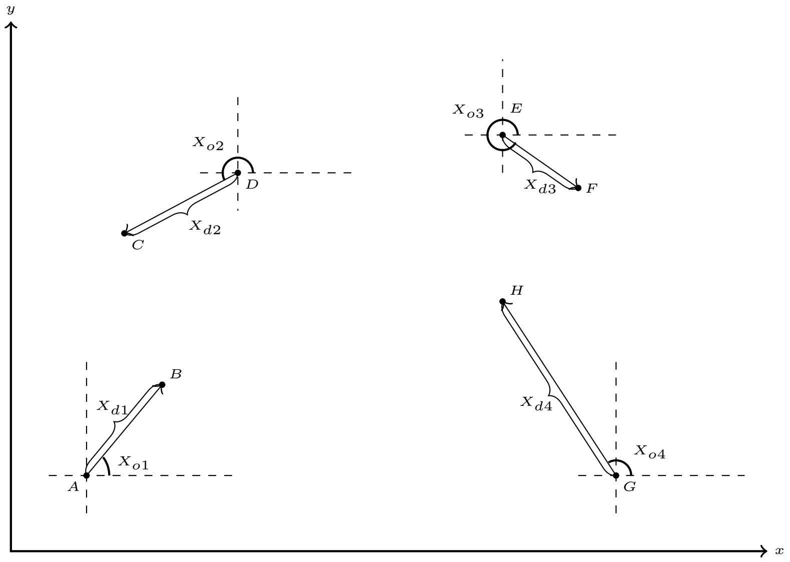


Figure 2: Distance and orientation feature extraction procedure. In this case B is near A, C is near D, F is near E and H is near G.

which is utilized to assess the similarity between GMMs, i.e., comparing GMM estimation stages of the same spatial relation, but also to compare different relations.

4.1 Populating a Spatial Feature Dataset

The collected data includes 120K texts from travel blogs; however, with a focus on a specific geographic area (London), the dataset does not include enough spatial relationship instances to train a two-dimensional probabilistic model.

To obtain more data we use KDE [2]. In our scenario, KDE techniques provide new density samples based on a small amount of ground-truth data. These estimates are then used in order to generate additional spatial feature vector data (semi-synthetic) to train probabilistic models (GMMs).

Relating KDE to our problem, let $X = (X_d, X_o)$ follow a two-dimensional true density f defined over $\mathbb{R}^2$. Let $\mathcal{X} = \{X_1,\ldots, X_n\}$ be an independent random sample set (initial spatial feature vector set in our case) drawn from f. The general form of the kernel density estimation function of f is:

$$\hat{f}_H(x; H) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i) \tag{1}$$

where $x = (x_1, x_2)^\intercal$ is a generic vector that depends on the Kernel used, e.g. Gaussian, Epanechnikov, Cosine etc., $X_i = (X_{di}, X_{oi})^\intercal$ with $1 \leq i \leq n$, $K_H(x) = |H|^{-\frac{1}{2}} K(H^{-\frac{1}{2}} x)$, and n denotes the number of instances of each spatial relation. In our case, $K_H(x)$ is a Gaussian bivariate kernel function, and H is a symmetric positive definite 2×2 diagonal matrix (bandwidth matrix).

The performance of a kernel density estimator is primarily determined by the choice of bandwidth, which controls the degree of smoothing, and secondarily by the choice of the kernel function, which in our case is Gaussian. A large body of literature [2, 16] exists on bandwidth selection for univariate and multivariate kernel density estimation. In this contribution, we follow a simple case scenario as described in [2]. With data observed from a bivariate normal density, the diagonal bandwidth matrix, denoted by

$$H = \begin{pmatrix} h_1 & 0 \\ 0 & h_2 \end{pmatrix} \tag{2}$$

can be well approximated by $h_b = \sigma_b \left(\frac{4}{(d+2)n}\right)^{\frac{1}{d+4}}$ for $b \in \{1, 2\}$, where σ_b is the standard deviation of the i-th variate and d denotes the problem's dimensionality. This method is often used when no other practical bandwidth selection scheme is available, despite the fact that most interesting data are non-Gaussian.

Data of such a process is visualized in Figure 3 which illustrates the initial dataset for a spatial relationship in a two-dimensional space (distance [km] and orientation [degrees] as the x and y-axis respectively), the Gaussian kernel density estimate of the probability density function (PDF) of the initial dataset and the generated samples using the estimated PDF, respectively.

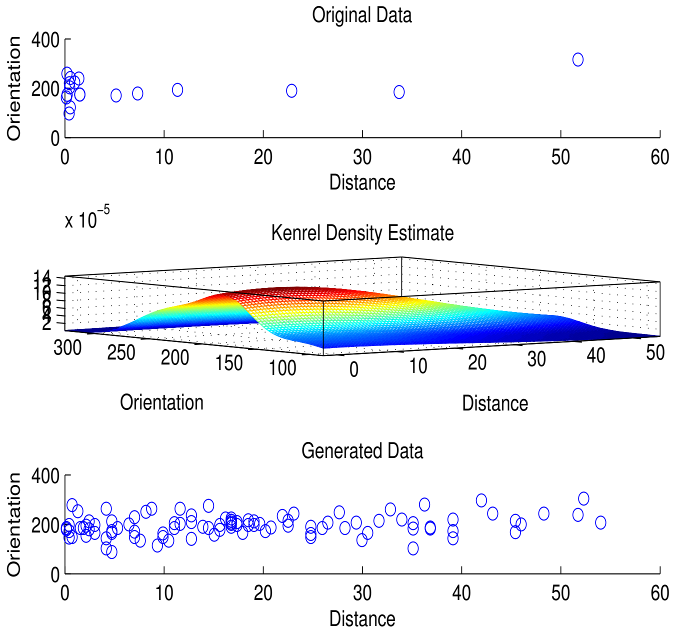


Figure 3: Spatial feature dataset population: (i) initial dataset (ii) estimated probability density function using KDE, and (iii) generated dataset.

Finally, as we explain in Section 5, we underline that we use the generated data only for training probabilistic models and not for testing, since the generated data exhibits a considerable bias.

4.2 Quantifying Qualitative Relations

The essential step in quantifying qualitative data is the mapping of the generated data to PDFs. Here we have to decide what kind of probabilistic model we desire to train. Using Gaussian kernel density estimation to populate our dataset, we naturally opted for Gaussian Mixture Models (GMMs). GMMs have been extensively used in many classification and general machine learning problems (cf. [1, 7]). They are very well known for (*i*) their formality, as they build on the formal probability theory, (*ii*) their practicality, as they have been implemented several times in practice, (*iii*) their generality, as they are capable of handling many different types of uncertainty, and (*iv*) their effectiveness because existing solutions that employ them known to be effective and scalable.

Generally speaking, a GMM is a weighted sum of M component Gaussian densities as

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x; \mu_i, \Sigma_i) \tag{3}$$

where x is a d-dimensional data vector (i.e., features - in our case $d = 2$), w_i , $1 \leq i \leq M$, are the mixture weights, and $g(x|\mu_i, \Sigma_i)$ is a Gaussian density function $\forall i$, with mean vector $\mu_i \in \mathbb{R}^d$ and covariance matrix $\Sigma_i \in \mathbb{R}^{d \times d}$ such that

$$\begin{aligned} &g(x; \mu_i, \Sigma_i) = \\ &(2\pi)^{-\frac{d}{2}} \det(\Sigma_i)^{-\frac{1}{2}} \exp(-\frac{1}{2}(x - \mu_i)^\intercal \Sigma_i^{-1}(x - \mu_i)) \end{aligned} \tag{4}$$

with mean vector $\mu_i \in \mathbb{R}^d$ and covariance matrix $\Sigma_i \in \mathbb{R}^{d \times d}$. The mixture weights satisfy the constraint that

$$\sum_{i=1}^{M} w_i = 1 \tag{5}$$

with $w_i \geq 0$.

The complete GMM is parameterized by the mean vectors, the covariance matrices and mixture weights $\forall i$. These parameters are collectively represented in Equation 3, by the notation $\lambda = \{w_i, \mu_i, \Sigma_i\}$ with $i = 1, \ldots, M$. In our setting, each spatial relation is modeled by a 2-dimensional GMM trained with each relation's spatial feature vectors, which are created as detailed in Sections 3.3 and 4.1.

For the parameter estimation of Gaussian component of each GMM, we use Expectation Maximization (EM) (cf. [5]). EM enables us to update the parameters of a given M-component mixture with respect to a feature vector set (generated spatial feature vector set in our case) $\mathcal{X} = \{X_1, \ldots, X_n\}$ with $1 \leq j \leq n$ and all $X_j \in \mathbb{R}^d$, such that the *log-likelihood* $\mathcal{L}$ of $\mathcal{X}$ calculated using Equation 6 increases with each re-etimation step. This means that we keep re-estimating model parameters until the log-likelihood $\mathcal{L}$ or the parameters converge.

$$\mathcal{L} = \sum_{j=1}^{n} \log(p(X_j|\lambda)) \tag{6}$$

The updates for the parameters of a GMM can be accomplished by iterative application of the following equations for all components $i \in \{1, ..., M\}$

$$P(i|X_j) = \frac{w_i g(X_j; \lambda_i)}{p(X_j|\lambda)} \tag{7}$$

$$w_i = \sum_{j=1}^{n} \frac{P(i|X_j)}{n} \tag{8}$$

$$\mu_i = \sum_{j=1}^{n} \frac{P(i|X_j)X_j}{n w_i} \tag{9}$$

$$\Sigma_i = \sum_{j=1}^{n} \frac{P(i|X_j)(X_j - \mu_i)(X_j - \mu_i)^\intercal}{n w_i} \tag{10}$$

The EM algorithm is not guaranteed to lead us to the solution yielding maximum log-likelihood on $\mathcal{X}$ among all maxima of the log-likelihood. Nevertheless, using the EM algorithm, if we are "close" to the global optimum (maximum) of the parameter space, then it is very likely we can obtain the globally optimal solution.

4.3 Model Optimization

A main issue in probabilistic modeling with GMMs is that a predefined number of components per Gaussian mixture is neither a dynamic nor an efficient and robust approach. The optimal number of Gaussian components should be decided based on each dataset. Hence, in this section we employ a greedy learning approach to dynamically estimate the number of components in a GMM. (cf. [29]). Typically a GMM is trained by starting with a random configuration of all components and improve upon this configuration with the EM algorithm. This greedy approach tries to build the mixture component in a more efficient way by starting from an one-component GMM, whose parameters are trivially computed by using EM (cf. Section 4.2), and then employing the following two steps until a stop criterion is met.

1. Insert a new component in the mixture
2. Apply EM until the log-likelihood $\mathcal{L}$ or the parameters of the GMM converge (cf. Section 4.2)

The stop criterion can either be a maximum pre-specified number of components, or it can be any other model selection criterion like Minimum Description Length [13] or Bayesian Information Criterion (BIC) [26]. In our case the algorithm stops if the maximum number of components is reached, or if the log-likelihood $\mathcal{L}$ after introducing a new component is lower than that of the previous model. For a more formal description let us consider a feature vector set (generated spatial feature vector set in our case) $\mathcal{X}$ under an M-component mixture $p^M(\mathcal{X}|\lambda)$. The greedy learning algorithm can be summarized in the following five steps:

1. Compute the one-component mixture $p^1(\mathcal{X}|\lambda)$ that yields maximum log-likelihood using the (EM) algorithm.
2. Find the optimal new component $g(\mathcal{X};\lambda^*)$ and the corresponding mixing weight w^*.
3. Set $p^{M+1}(\mathcal{X}|\lambda) = (1-w^*)p^M(\mathcal{X}|\lambda) + w^*g(\mathcal{X};\lambda^*)$ and $M = M+1$.
4. Update new model parameters using EM algorithm.
5. Terminate if (i) log-likelihood $\mathcal{L}$ of GMM starts to decrease, or (ii) max number of components is reached; else go to step 2.

The crucial step of the algorithm is the component insertion in Step 2. Several approaches exist here. One is to consider a number of candidates equal to the number of feature vectors but this would be rather expensive. The approach followed in this work is to pick an optimal number of candidate components as discussed in [29].

Figure 4 illustrates a converged 3-component GMM. In this case, the maximum number of Gaussian components was used as a stop criterion. Distance and orientation are used as uncorrelated random variables, which means that all Gaussian components in each GMM have diagonal covariance matrices. The x and the y-axes represent raw (not normalized) distance and orientation information in kilometers and degrees, respectively.

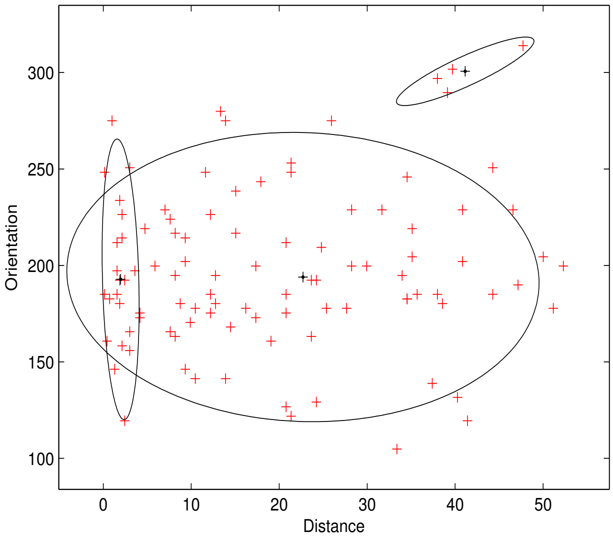


Figure 4: Converged max 3-component GMM.

4.4 Similarity Between Quantitative Spatial Relationships

In many probabilistic classification problems several metrics have been proposed to compute a distance measurement between different classes as a means to compare them. Measuring distance between converged PDFs which model different classes (spatial relationships in our case) is a measure of similarity between them. In our contribution, we use Kullback-Leibler (KL) divergence [21] as such a distance metric.

There are two main reasons for checking similarity between quantified spatial relations. Firstly, we want to observe the changes for each GMM as we increase the maximum number of Gaussian components during the training procedure. Secondly, we use KL divergence to measure the similarity between spatial relationships that tend to follow similar patterns, e.g., *Near* & *NextTo*, *In* & *On*.

KL divergence is a similarity measure between two probability distributions. So, let $\mathcal{F}_1(x)$ and $\mathcal{F}_2(x)$ be two probability distributions (GMMs in our case). By definition, the KL distance $\mathcal{D}(\mathcal{F}_1(x)||\mathcal{F}_2(x))$ between $\mathcal{F}_1(x)$ and $\mathcal{F}_2(x)$ is given as follows.

$$\mathcal{D}(\mathcal{F}_1(x)||\mathcal{F}_2(x)) = \int \mathcal{F}_1(x) \log \left\{ \tfrac{\mathcal{F}_1(x)}{\mathcal{F}_2(x)} \right\} dx \tag{11}$$

The KL divergence is always nonnegative and it is zero only when the two distributions are identical. Additionally KL divergence is not symmetric, i.e., $\mathcal{D}(\mathcal{F}_1(x)||\mathcal{F}_2(x)) \neq \mathcal{D}(\mathcal{F}_2(x)||\mathcal{F}_1(x))$.

It is common to encounter the symmetric version of the KL divergence between $\mathcal{F}_1(x)$ and $\mathcal{F}_2(x))$ as

$$\mathcal{D}_{sym}(\mathcal{F}_1(x)||\mathcal{F}_2(x)) = \frac{\mathcal{D}(\mathcal{F}_1(x)||\mathcal{F}_2(x)) + \mathcal{D}(\mathcal{F}_2(x)||\mathcal{F}_1(x))}{2} \tag{12}$$

In this work, we use the symmetric KL divergence in order to measure the similarity between GMMs.

5 Experimentation

The scope of this section is to assess the quantitative representation of qualitative geospatial data by means of probability distributions (GMMs). For this purpose, we investigate a set of spatial relationships for a specific geographic area (London). In terms of experiments, we compute probabilistic representations of spatial relationships by considering distance and orientation as dependent but, uncorrelated features (case one) and as correlated features (case two).

We visualize the results of the trained models and compare them to check if they intuitively perform well, e.g., they return visually reasonable results. In addition, we measure the KL divergence for spatial relationships between a baseline one-component model and the maximum number of Gaussian components model. Finally, based on visualization and KL divergence, we assess the informativeness and efficiency of distance and orientation features for quantitative modeling of spatial relations and observe how much different spatial relations may behave in a similar way.

5.1 Experimental Setup

The choice of an appropriate dataset is crucial in our experimentation. As mentioned is Section 3.2, the density of POIs is very high in urban regions. We decided to use data from such a dense region to find meaningful as well as consistent spatial relationships. We retrieved data for a bounding box that contains the greater area of London, UK. In this preprocessing step, we parsed our travel blog data (120k texts) set and retrieved sentences containing at least two POIs and whose coordinates are within the bounding box of Latitude $[51°, 52°]$ and longitude $[-1°, 1°]$. This resulted in 12k sentences. Using human annotation, we extracted instances of the eight most frequent spatial relations including *North, South, East, West, Near, In, On, NextTo.* This means also that in our travel blog dataset, people tend to use a mixture of directional, topology and vague metrical relations in order to describe POI locations. From this data, distance and orientation features where extracted as described in Section 3.

Given that only a small percentage of the collected data contains spatial relationship information, here to obtain a meaningful amount of useful data, we need to overall collect a large volume of texts. For example, considering the London case, approximately 10% of the 12000 sentences contained clear instances of spatial relationships. For the specific approach, this would have not been enough to train and test probabilistic models. We use KDE to create a semi-synthetic dataset based on the collected data (cf. Section 4.1). More specifically, we use KDE to estimate each spatial relationship's density function. This estimate is then used to generate more samples and so to train a probabilistic model. As explained is Section 4.1, we use the generated data only for training probabilistic models but not for testing because of the considerable bias.

Next, we employ the greedy EM algorithm to train bivariate GMMs based on the extracted distance and orientation features for each spatial relationship. The results are PDFs for each spatial relationship that, as the initial outset suggests, can be used to estimate the unknown position of spatial objects.

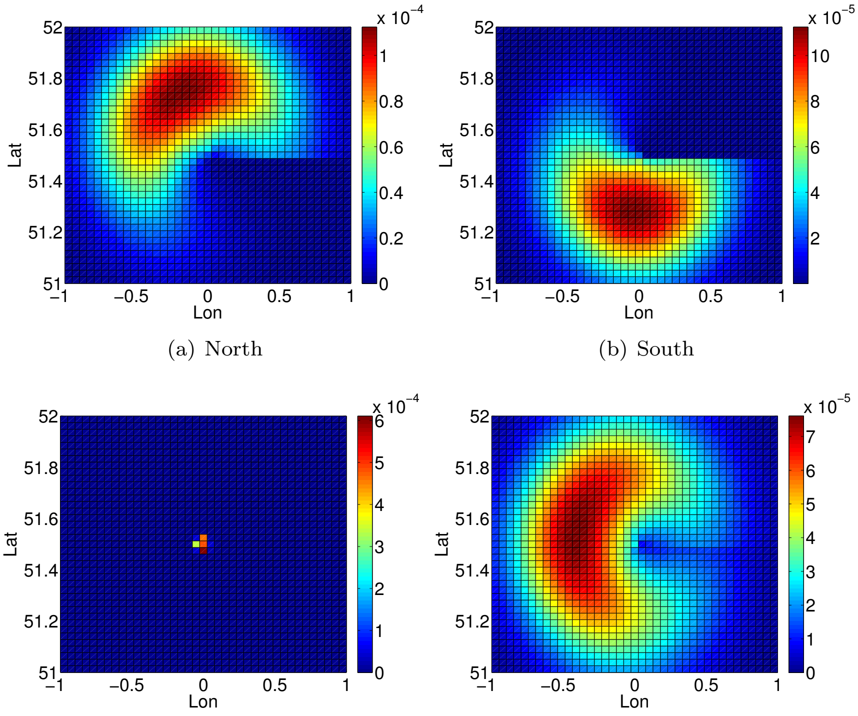
Our approach has been implemented in Matlab and all the experiments were conducted on an Intel(R) Core(TM) i5-2400 CPU at 3.10GHz with 8GB of RAM, running Ubuntu Linux 11.10.

5.2 Visualization of Quantitative Spatial Relations

The most important means of assessing the result is to visualize the quantified spatial relations. We divided the London bounding box to filter the input data by means of a 50×50 spatial grid. Each grid cell corresponds to a $4.4km \times 2.2km$ spatial extent (Longitude, Latitude). Given two spatial objects and the known location at the center of the grid, we plot for each grid cell the positional probability of the unknown location, i.e., how likely it would be for the unknown spatial object to be located in a specific grid cell. Using a heat map, warmer colors (red) indicate higher probabilities.

Figure 5 shows four spatial relationships modeled as one-component GMMs, with distance and orientation considered as uncorrelated random variables.

The proposed modeling based on distance and orientation features performs especially well in some of the cases. More specifically, for the cases of *North* (cf. Figure 5(a)), *South* (cf. Figure 5(b)) and *Near* (cf. Figure 5(c)) the proposed



(a) North

(b) South

(c) Near

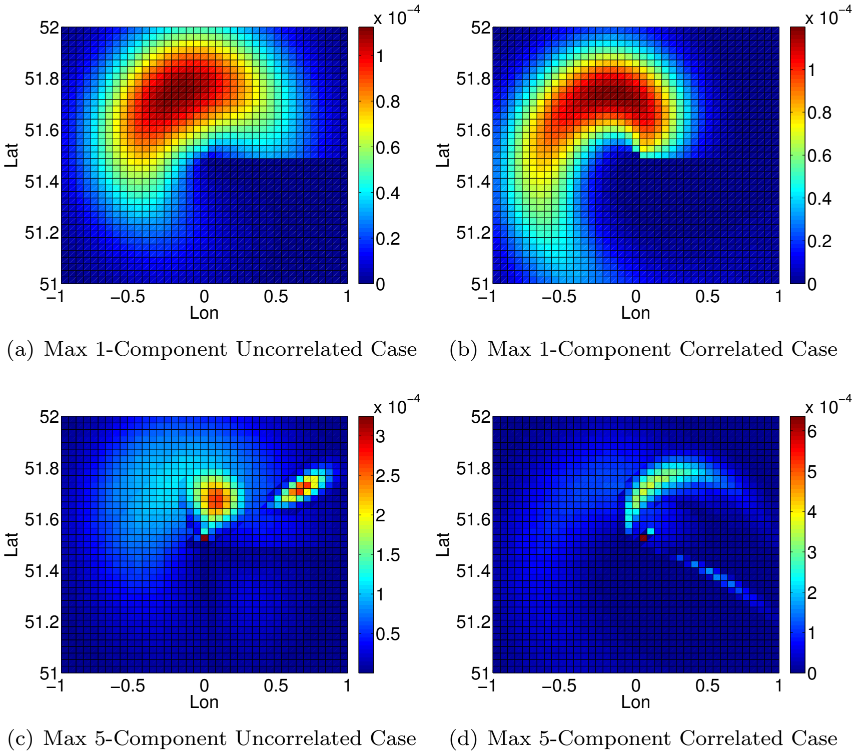
(d) In

Figure 5: Probabilistic heat maps for four basic spatial relationships: 1-Component Gaussian Mixture Models for the uncorrelated distance and orientation case. All figures illustrate the case where a POI is conntected with the center of the grid, with the respective spatial relation.

model returns high probabilities in the expected regions. On the other hand, the case of *In* (cf. Figure 5(d)) seems to include a lot of statistical noise due to the general uncertain nature of user generated content. For example, high distance and orientation variance values for the cases of *On* and *In* are caused by the the fact that most of the sentences that contain these spatial relations are of the form ***POI*** *in London* and ***POI*** *on river Thames*.

5.2.1 Optimal Number of Gaussian Components

An important parameter when generating GMMs is the maximum number of Gaussian components. Such a limit is simply a stop criterion in the GMM training process and does not mean that the final component will converge to this upper limit, e.g., it might already converge to a lower number of components. Figure 6 illustrates the case of spatial relation *North*. The heat maps of Figures 6(a) and 6(b) show the cases of a maximum of 1 Gaussian components per mixture when distance and orientation are considered as uncorrelated and correlated, respectively. The heat maps of Figures 6(c) and 6(d) show the cases of a maximum of 5 Gaussian components per mixture, when distance and



(a) Max 1-Component Uncorrelated Case (b) Max 1-Component Correlated Case

(c) Max 5-Component Uncorrelated Case (d) Max 5-Component Correlated Case

Figure 6: Probabilistic heat maps for *North*: (a), (b) show correlated and uncorrelated distance and orientation case for a max of 1 Gaussian component per GMM while (c), (d) show correlated and uncorrelated distance and orientation case for a max of 5 Gaussian components per GMM.

orientation are considered as uncorrelated and correlated, respectively.

For both uncorrelated and correlated cases, Figures 6(c) and 6(d) show that by stepwise increasing the maximum number of Gaussian components, high probabilities tend to accumulate in fragmented small regions. The reason is that higher number of mixtures per GMM leads to components that are converging on their parameters (mean, covariance, component weight) based on more dense regions of the dataset, e.g., regions with more data samples will become dominant components. The weight of such *dominant components* (it is denoted as w in Section 4.2) is higher than the weight of other components in the final GMM. This results in fragmented high probability regions in the final heat maps.

As a result of this phenomenon, a major question is to the best approach to decide about the number of components per GMM. From an intuitive point of view, a smaller number of Gaussian components performs better as it preserves spatial generality, i.e., trends. However, as high probability regions are larger, they might result in inefficient location prediction tasks. From a statistical point of view, a high number of mixture components results in more accurate probabilistic models and better classification performance in most of the cases.

Unfortunately, the latter approach leads to sparse and small, high probability regions, which could be characterized as biased to the specific characteristics of the geographic region from where the dataset is taken (London in our case).

In Figures 7(a) and 7(c), we depict the *average log-likelihood*, e.g., we estimated parameters and log-likelihoods for each spatial relation model running the greedy learning algorithm 100 times per maximum component step and stepwise increasing the maximum Gaussian components per GMM. Figures 7(a) and 7(c) show the cases of correlated and uncorrelated distance and orientation random variables, respectively. In both cases, most of the spatial relation models converge on a high number of components, i.e., 16-17. Only spatial relations *Near* and *NextTo* converge on a smaller number of components. This means that statistically, most of the spatial relationships should be modeled with an upper limit of Gaussian components close to 16 or 17. In practice, this will result in fragmented spatial probabilities (heat maps) as outlined above.

Concluding, we realize that based on the log-likelihood measurements, there are statistically correct and sometimes optimal solutions for deciding the number of components. The difficult part in our case is the balance between statistical and intuitive robustness.

Based on a user generated dataset, we believe that GMMs with a number of components between 1 and 10 are statistically correct (but not optimal) and intuitively efficient to model spatial relations.

5.2.2 Correlated vs. Uncorrelated Features

Correlation between distance and orientation is another important issue when training GMMs. Literature suggests that most of the classification approaches perform better when probabilistic models are trained taking into consideration the correlation between random variables. In our work, visualization shows that there is a high correlation between distance and orientation for some but not all cases. Figure 6 illustrates the case of *North*. For the heat maps shown in Figures 6(a) and 6(c), distance and orientation are considered uncorrelated, for Figures 6(b) and 6(d), they are considered as correlated random variables. Intuition suggests that we can not guarantee that the correlated case performs better, even if we are sure that distance and orientation are correlated. The *North* case should result in high probabilities for the top part of the grid as it should be the case for all directional relations. However, based on visual results and heat maps for all modeled spatial relations, we observe that distance and orientation seem to be less correlated for the cases of *In* (cf. Figure 5(d)) and *On*, and tend to have zero correlation, e.g., region around the center of the grid with almost equal probabilities, for the cases of *Near* (cf. Figure 5(c)) and *NextTo*. As expected, this leads as to the conclusion that some spatial relations are independent of orientation, e.g., only distance could model them efficiently. This also means that distance and orientation should be modeled as independent random variables.

Summing up, based on user-generated content, we believe that directional relations like *North* should be modeled taking correlation between distance and orientation into consideration. On the other hand, topological relations such as *In* and metric relations like *Near* tend to be independent of orientation, which means that correlation between distance and orientation should not be taken into consideration during modeling.

5.3 Similarity Between Quantified Spatial Relations

Besides a visual inspection, it is important to have a quality metric to assess the probabilistic spatial relation quantification. We use Kullback-Leibler (KL) divergence (i) to assess the similarity between converged GMMs of the same relation and (ii) to measure the similarity between some spatial relationships that tend to follow similar patterns.

Figures 7(b) and 7(d) illustrate the KL divergence between the baseline 1-component GMM and the final converged GMM after each step of increasing the maximum number of components for correlated and uncorrelated cases, respectively. Most of the models tend to diverge from the baseline model as we increase the maximum number of components. Only the models for *Near* and *NextTo* have low and zero distance from their baseline model. This matches the corresponding log-likelihoods illustrated in Figures 7(a) and 7(c), which remain almost stable. In these examples, with a small number of Gaussian components for *Near* and one Gaussian component for *NextTo* the log-likelihoods remains stable.

Finally, Figures 7(e) and 7(f) show that spatial relation pairs *Near-NextTo* and *On-In* exhibit similar characteristics. To assess this similarity, we measured the KL divergence for all cases of their models. The aforementioned figures show that the pair *On-In* seems to diverge as we increase the number of components for both correlated and uncorrelated cases. However, the pair *Near-NextTo* exhibits low values of KL divergence for all cases. This leads to the conclusion that people use more than one language expression to describe the same spatial relation. For our examples, this means that we could merge the cases of *Near* and *NextTo* into one probabilistic model.

6 Conclusions

The increase in available user-generated data provides a unique opportunity for the generation of rich datasets in geographical information science. In this work, we provide a quantitative approach for the representation of qualitative spatial relations extracted from such data based on training probabilistic models. The proposed scheme returns estimates of uncertain object locations based on distance and orientation features as provided by human reporters in relation to known object locations. To achieve these desiderata, we propose a greedy learning algorithm based on the Expectation Maximization (EM) framework to train probabilistic models over spatial relationships; here, we restrict our attention on GMM models. The proposed approach seems to be promising in terms of accurately capturing and representing spatial relationships. Distance and orientation features tend to describe all spatial relations that were extracted from user generated texts in an informative way. Moreover, our probabilistic approach seems to be robust in handling any uncertainties, which characterize observations in crowd-sourced text data. As a future research direction, we already have been investigating new NLP techniques for the automatic extraction of POIs and spatial relationship information from texts and we are very close to a practical and theoretically robust solution. This will enable us to evaluate additional probabilistic and deterministic modeling techniques and to develop efficient text-to-map applications.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme - Marie Curie Actions, Initial Training Network GEOCROWD (http://www.geocrowd.eu) under grant agreement No. FP7-PEOPLE-2010-ITN-264994.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] A. W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations (Oxford Statistical Science Series).* Oxford University Press, USA, Nov. 1997.

[3] R. Bunescu and R. Mooney. Subsequence Kernels for Relation Extraction. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 171–178. MIT Press, Cambridge, MA, 2006.

[4] H. Cunningham, Y. Wilks, and R. J. Gaizauskas. Gate - a general architecture for text engineering, 1996.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[6] E. Drymonas and D. Pfoser. Geospatial route extraction from texts. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Data Mining for Geoinformatics*, DMG '10, pages 29–37, New York, NY, USA, 2010. ACM.

[7] H. P. Duda, R. and D. Stork. *Pattern Classification.* John Wiley and Sons., 2001.

[8] M. Egenhofer. A formal definition of binary topological relationships. In W. Litwin and H.-J. Schek, editors, *Foundations of Data Organization and Algorithms*, volume 367 of *Lecture Notes in Computer Science*, pages 457–472. Springer Berlin / Heidelberg, 1989.

[9] M. J. Egenhofer and J. Herring. A mathematical framework for the definitions of topological relationships. In *Int'l Symp. on Spatial Data Handling*, 1990.

[10] M. J. Egenhofer and J. Sharma. Topological relations between regions in r^2 and z^2. In *SSD*, pages 316–336, 1993.

[11] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference of Empirical Methods in Natural Language Processing (EMNLP '11)*, Edinburgh, Scotland, UK, July 27-31 2011.

[12] A. U. Frank. Ontology for spatio-temporal databases. In *Spatio-Temporal Databases: The Chorochronos Approach*, pages 9–77, 2003.

[13] P. D. Grünwald. *The Minimum Description Length Principle (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[14] R. H. Güting. An introduction to spatial database systems. *The VLDB Journal*, 3(4):357–399, Oct. 1994.

[15] J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[16] R. J. Hyndman, X. Zhang, and M. L. King. Bandwidth selection for multivariate kernel density estimation using mcmc. Econometric Society 2004 Australasian Meetings 120, Econometric Society, 2004.

[17] W. Kainz, M. J. Egenhofer, and I. Greasley. Modeling spatial relations and operations with partially ordered sets. *International Journal of Geographical Information Systems*, 7:215–229, 1993.

[18] D. V. Kalashnikov, Y. Ma, S. Mehrotra, R. Hariharan, and C. Butts. Modeling and querying uncertain spatial information for situational awareness applications. In *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, GIS '06, pages 131–138, New York, NY, USA, 2006. ACM.

[19] P. Kordjamshidi, M. van Otterlo, and M.-F. Moens. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing*, 8(3):article 4, 36 p, 2011.

[20] M. Koubarakis, T. K. Sellis, A. U. Frank, S. Grumbach, R. H. Güting, C. S. Jensen, N. A. Lorentzos, Y. Manolopoulos, E. Nardelli, B. Pernici, H.-J. Schek, M. Scholl, B. Theodoulidis, and N. Tryfona, editors. *Spatio-Temporal Databases: The Chorochronos Approach*, volume 2520 of *Lecture Notes in Computer Science*. Springer, 2003.

[21] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951.

[22] Y. Ma, D. V. Kalashnikov, and S. Mehrotra. Toward managing uncertain spatial information for situational awareness applications. *IEEE Trans. on Knowl. and Data Eng.*, 20(10):1408–1423, Oct. 2008.

[23] D. Papadias and T. Sellis. Qualitative representation of spatial knowledge in two-dimensional space. *The VLDB Journal*, 3(4):479–516, Oct. 1994.

[24] D. Papadias, Y. Theodoridis, and T. Sellis. The retrieval of direction relations using r-trees, 1994.

[25] D. Papadias, Y. Theodoridis, T. Sellis, and M. J. Egenhofer. Topological relations in the world of minimum bounding rectangles: A study with r-trees. pages 92–103, 1995.

[26] G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

[27] R. Smith, M. Self, and P. Cheeseman. Autonomous robot vehicles. chapter Estimating uncertain spatial relationships in robotics, pages 167–193. Springer-Verlag New York, Inc., New York, NY, USA, 1990.

[28] S. Vasudevan and R. Siegwart. A bayesian approach to conceptualization and place classification: Incorporating spatial relationships (distances) to infer concepts, 2007.

[29] J. J. Verbeek, N. Vlassis, and B. Kröse. Efficient greedy learning of gaussian mixture models. *Neural Computation*, 15:469–485, 2003.

[30] Wanderlust. Extracting Semantic Relations from NaturalLanguage Text Using Dependency Grammar Patterns. *Proceedings of the Workshop on Semantic Search (SemSearch 2009) at the 18th Int. World Wide Web Conference (WWW 2009) , April 18, 2009, Madrid, Spain*, 2009.

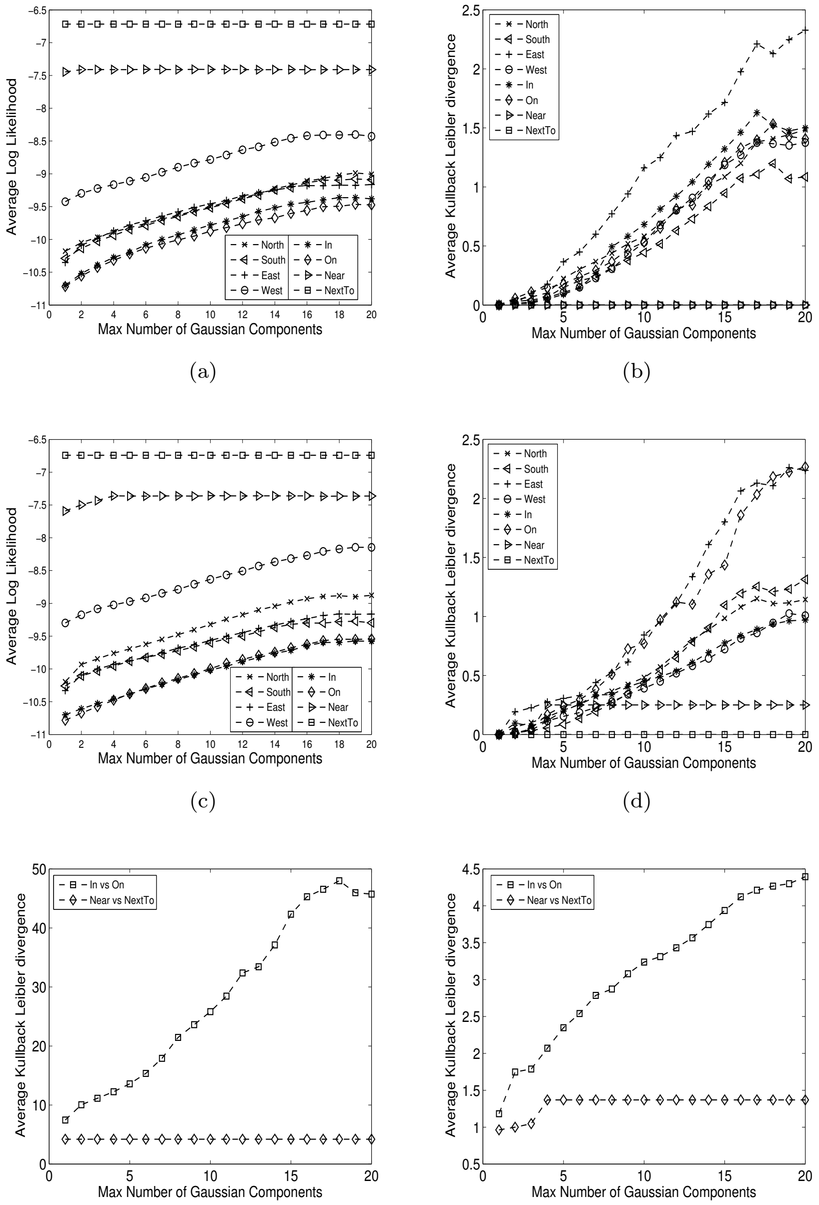
[31] Y. Wang and F. Makedon. R-histogram: quantitative representation of spatial relations for similarity-based image retrieval. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 323–326, New York, NY, USA, 2003. ACM.

[32] J. Xu and C. Yao. Formalizing natural-language spatial relations descriptions with fuzzy decision tree algorithm. *Proceedings of Spie the International Society for Optical Engineering*, 6420.

[33] Y. Yuan. Extracting spatial relations from document for geographic information retrieval. In *2011 19th International Conference on Geoinformatics*, pages 1 –5, june 2011.

[34] D. Zelenko, C. Aone, and A. Richardella. Kernel methods for relation extraction. *J. Mach. Learn. Res.*, 3:1083–1106, Mar. 2003.

[35] X. Zhang, C. Zhang, C. Du, and S. Zhu. Svm based extraction of spatial relations in text. In *2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)*, pages 529 –533, 29 2011-july 1 2011.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 7: (a), (c) Average log-likelihood vs maximum number of Gaussian components for correlated and uncorrelated distance and orientation case respectively. (b), (d) Average KL divergence between the baseline 1-component GMM and the final converged GMM after each step of increasing the maximum number of components for correlated and uncorrelated distance and orientation case respectively. (e), (f) Average KL diverge between spatial relationship pairs “In-On” and “Near-Nextto” for correlated and uncorrelated distance and orientation case respectively.