Spatial Data Mining in Practice: 
Principles and Case Studies

Christine KÖRNER a,1, Dirk HECKER a, Maike KRAUSE-TRAUDES a, Michael MAY a, Simon SCHEIDER a, Daniel SCHULZ a, Hendrik STANGE a, Stefan WROBEL a,b

a Fraunhofer IAIS, Sankt Augustin, Germany
b Department of Computer Science III, University of Bonn, Germany

Abstract. Almost any data can be referenced in geographic space. Such data permit advanced analyses that utilize the position and relationships of objects in space as well as geographic background information. Even though spatial data mining is still a young research discipline, in the past years research advances have shown that the particular challenges of spatial data can be mastered and that the technology is ready for practical application when spatial aspects are treated as an integrated part of data mining and model building. In this chapter in particular, we give a detailed description of several customer projects that we have carried out and which all involve customized data mining solutions for business relevant tasks. The applications range from customer segmentation to the prediction of traffic frequencies and the analysis of GPS trajectories. They have been selected to demonstrate key challenges, to provide advanced solutions and to arouse further research questions.

Keywords. spatial data mining, algorithms, case studies

Introduction

Over the past years the interest in spatial data has clearly been pushed by the wide availability of recording technologies such as the Global Positioning System (GPS), mobile phone data or radio frequency identification (RFID). Today, nearly all database systems support data types for the storage and processing of geographic data. However, knowledge discovery from geographic data is still a young research direction. In classic data mining many algorithms extend over multi-dimensional feature space and are thus inherently spatial. Yet, they are not necessarily adequate to model geographic space.

Spatial data mining combines statistics, machine learning, databases and visualization with geographic data. The task is to identify spatial patterns or objects that are potential generators of such patterns. This includes also the iden-
tification of information which is relevant to explain the spatial patterns and to present the results in an intuitive way that supports further analysis of the data.

Georeferenced data differ in a number of ways from traditional tabular data and therefore challenge the application of data mining methods to spatial problems [1]. First, autocorrelation is typical for data within a geographic context, yet it is unusual in traditional data mining. Second, spatial data types range from simple point data to complex objects and can further be combined to network structures and spatio-temporal data structures. These structures must be handled by the data mining algorithms and require sophisticated feature extraction methods. Furthermore, feature extraction is known to be the most time-consuming step during data mining, which is even more true for operations on spatial objects. A third challenge therefore is the development of specialized algorithms that interweave the feature extraction and the data mining step. Naturally, on-the-fly feature extraction in combination with early pruning of the search space can lead to substantial performance improvements. The following three paragraphs briefly introduce the nature of spatial data with respect to the presented challenges.

Spatial phenomena are characterized by autocorrelation. Tobler [2] formulates this basic principle as follows: “[…] everything is related to everything else, but near things are more related than distant things”. Autocorrelation is a powerful resource to improve inference, however it can cause poor performance for algorithms that ignore it [3]. Traditional data mining methods assume that data are independent and identically distributed (iid), thus they are not prepared to model autocorrelation. Spatial autocorrelation can be measured, for example, by Geary’s c or Moran’s I statistics [4].

References to geographic space can be modeled in several ways, resulting in different spatial data types. In general, continuous phenomena that spread in space, e.g. temperature or humidity, and geographically referenced objects, e.g. houses or rivers, are distinguished. The former type is modeled as field data and has been explored extensively in the area of geostatistics. One of the most well-known regression techniques for field data is probably Kriging [5]. The latter type is modeled using vectors in form of points, lines or polygons. Depending on the importance of their true spatial extent, objects may be generalized to point data by their centroid. Vector data can be combined to form networks or tessellations in space. The addition of temporal information, as found in trajectories or geographically referenced time series, introduces even more complexity to the data.

The relationships between spatial objects offer a rich source of information for data mining. Therefore, a number of spatial feature extraction and aggregation methods have been developed. One basic relational feature is the distance between two objects, which is usually measured using the Euclidean distance. For lines and polygons the shortest distance between any two points of the objects or their minimum bounding rectangles can be used. However, in practice distance is often calculated between centroids. More complex spatial features can be derived from the topological relationship between two objects as described by the 9-intersection model [6] or by the connectivity of a network. Another useful method for feature extraction is aggregation, which summarizes information in a particular neighborhood of an object. The neighborhood is commonly defined by buffers, drive-time zones or Voronoi polygons.
The following sections present recent industry projects at Fraunhofer IAIS that demonstrate the demand for and benefit of spatial data mining. The sections are organized to reflect the increasing complexity of spatial data types. Besides the application scenario, each project highlights one or more aspects of the above challenges for spatial data mining and shows a practical solution.

1. Spatial Data Mining for Marketing and Planning

This section presents three case studies in marketing and planning which utilize vector data for their analysis. The first study forecasts sales at potential new locations for a trading company and emphasizes the handling of large amounts of spatial features. The second and third study apply visual analytics and subgroup discovery for customer segmentation and the optimization of mobile networks respectively.

1.1. Sales Forecasting for Retail Location Planning

Choosing the appropriate site is crucial for the success of every retailing company. From a microeconomic point of view, the expected sales at a location are the most important decision criterion for the evaluation of potential new sites. However, sales forecasting is still a great challenge in retail location planning today. How can sales at potential new locations be predicted? And which factors influence sales the most?

Our project partner is one of Austria’s leading trading companies. In order to reduce the risk in location decisions while continuing growth, the company sought for an automated sales forecasting solution to evaluate possible new sites. In our project we identified and quantified the most important factors influencing sales at operating store locations for three different product lines and store formats (supermarket, hypermarket and drugstore). The main challenge of the project was to handle an abundance of attributes which possessed diverse levels of spatial resolution and for which the most appropriate resolution was not known beforehand. We applied support vector machines (SVM) for the regression task as they are robust in the face of high-dimensional data. SVMs are not spatial by themselves, therefore we conducted extensive feature extraction during which all spatial operations were performed.

The training set for model learning was made up of about 1,400 existing stores from all over Austria and a broad variety of socio-economic, demographic and market data on different administration levels as well as competitor information and points of interest (POI). Most of the socio-economic and market data were available on hierarchical spatial aggregation levels of states, districts resp. cities, municipalities and Zählsprengel as well as post code areas. Zählsprengel are subunits of municipalities at the lowest spatial aggregation level for which official statistics are available (around 1,000 inhabitants on average). They proved to be especially valuable for modeling purposes because they reflected most of the spatial variability.

In order to characterize the environment of individual shops, we first built trading areas for which socio-economic, demographic, competitor and POI infor-
mation was aggregated. The feature extraction process for each source of information is described in more detail in the following paragraphs. Generally, aggregation can be performed using buffers or drive time zones. They mark, for a fixed location, the area which lies within a given range or which can be reached within a given time respectively. However, location factors show different effects on different levels of spatial aggregation, and it had been unknown beforehand which levels would yield the highest impact. For instance, if attributes had been taken into account solely based on 5-minutes drive time zones, important positive shopping linkages which mostly appear within the range of a 3-minutes walk would have been lost. Therefore, we built several trading areas with varying spatial extent based on drive time zones for cars and pedestrians (1-5 and 1-3 minutes respectively) as well as buffers with a distance between 100 and 500 meters based on the street network. This resulted in a total of 13 trading areas per store.

Naturally, the trading areas did not correspond to the spatial units by which the socio-economic and demographic data were provided. Therefore, an assignment of attribute values in proportion to the intersecting area of a trading cell and other spatial units was made. Let $ta$ denote the trading area of interest and $u \in U$ the spatial units that carry some attribute $a()$. The assignment is specified by:

$$a(ta) = \sum_{u \in U} \frac{area(ta \cap u)}{area(u)} \ast a(u).$$

This approach assumes that socio-economic and demographic characteristics are equally distributed within a given spatial unit. However, especially rural areas with sparse population violate this assumption and skew the assignment. Therefore, we regarded only the proportion of built-up areas for the redistribution of attribute values.

The stress of competition was expressed by counting the number of competitor stores within the trading areas. In addition, we aggregated their membership to a competing retail chain, shop type and size, estimated turnover as well as opening hours. It is crucial to incorporate competitive effects in the forecasting model because competition is always a strong determinant of the amount of own sales. However, it is important to notice that competition must not necessarily be negative; it can also have positive impacts by raising the cumulative attraction of a site. We further took the distances to own shops of the company as competitive factor into account because they also draw off sales from a location (a phenomenon which is called retail cannibalization). Last, we included the geographical coordinates of the locations into our model to account for local and global trends.

We created location specific geographic features as, for example, population density, centrality, accessibility or a site’s shopping linkage potential. To assess the shopping linkage potential, we evaluated 3,000 different branches of POI with regard to their individual interception potential for the product lines and retail formats of our project partner. We again ascertained the number of relevant POI because it was expected that a high number of affine POI would increase the attraction of a site and thus lead to higher sales. The process of geographical
feature construction was supported by extensive visual analysis. The comparison of the spatial distribution of sales and touristic features (for instance the number of touristic POI as ski-lifts, hotels or holiday homes or the number of employees in the tourism sector) led to the discovery of a significant positive spatial correlation between sales and tourism.

At the end of the data preparation step, about 200 attributes for each trading area of each store were determined. Partly, these attributes had been considered as important a-priori. However, our analysis also revealed complex geographical features which had not been considered to be important sales determinants in certain parts of Austria before.

Taking the annual sales of each store of the year 2005 as response variable, we finally selected the most important features from the vast amount of independent variables and quantified their impact on sales. For this purpose we applied a wrapper approach with forward feature selection and calculated the regression coefficients for the selected attributes via support vector regression [7]. Due to the fact that the amount of sales for different merchandising lines and retail formats depends on different location factors, we developed specific models for new locations of supermarkets, hypermarkets and drugstores. The resulting regression models reduced the vast number of input variables to a set of 25 to 60 relevant features.

1.2. Customer Segmentation for Marketing Services

One of the leading German gas suppliers in the B-to-B market provides database marketing services for power authorities and local energy providers. A main challenge in this domain is to provide reliable knowledge about gas customers: What are the main factors which influence customer interest in natural gas? How can potential customers be reliably classified according to these characteristics? And how can this knowledge be used to automatically support the selection of addresses for direct marketing purposes?

Spatial data mining and knowledge discovery are considered to be a promising way to deal with the above challenges, as the application involves the development of models with geographically constrained validity, models using indirect and contingent relations on geographical objects as well as efficient methods for discovering this knowledge. The goals within our project were to a) find reliable candidate features for customer description and b) classify addresses according to the probability of customer interest in a sales representative visit.

The empirical basis of the study was a combined database of nationwide address data with description of buildings, a database of discrete geographical objects as rivers and elevation fields from a topographical map and a georeferenced sample of response data from about 500,000 nationwide interviews (see left plot in Figure 1).

In the data preparatory step the regional sample of response data was enriched with building data and geographic context. Thereby, the relation between the regional sample and the building data was realized by georeferencing the given addresses. The enriched sample served as training set for the analysis of interesting and statistically extraordinary subgroups and for the construction of a model for rule-based classification of addresses with high response probability.
In a first step we explored the data using techniques from visual analytics. Subsequently, the resulting hypotheses were tested for statistical significance using binomial tests and subgroup mining. Visual analytics is the “science of analytical reasoning facilitated by interactive visual interfaces” [8]. Especially in geographic context the visualization of information plays an important role to profit from background knowledge, flexible thinking and imagination of human analysts [9].

Subgroup discovery detects groups of objects with common characteristics that show a significant deviation in their target value with respect to the whole data set. In our application we searched for subgroups with a significantly larger response probability to marketing campaigns than in general. The quality of a subgroup depends on a quantitative and a qualitative term, which measure the size of a subgroup and the pureness of the target attribute within the subgroup respectively. More precisely, the quality \( q \) of a subgroup \( h \) is defined as

\[
q(h) = \frac{|p - p_0|}{\sqrt{p_0(1 - p_0)}} \sqrt{n}
\]

and accounts for the difference of target share between the subgroup \( p \) and the whole data set \( p_0 \), as well as the size \( n \) of the subgroup [10]. Spatial subgroups are formed if the subgroup definition involves operations on spatial components of the objects. However, spatial operations are expensive. They lead to a loss of performance during execution or require additional storage when computed in advance. Klösgen and May [11] developed a spatial subgroup mining system, which integrates spatial feature extraction into the mining process. They exploit the fact that it may not be necessary to compute all spatial relations due to early pruning in the mining process. The spatial joins are performed separately on each search level, which reduces the number of spatial operations and avoids redundant storage of features.

One major result of our study was that geographic relations, such as river distance and ground elevation, as well as the age of buildings can be used to improve the response probability of a sample of addresses. One example for an interesting subgroup of customers were people using heating oil instead of gas and living within 1 km distance from a larger river, which could be explained by
the specific flooding risk for oil tanks. Figure 1 (right) shows an example of this pattern as experienced during visual exploratory analysis.

1.3. Mobile Network Planning

The quality and coverage of a mobile network are key factors for the success of a mobile telecommunication company. In order to support decisions about the extension and optimization of such a network, we analyzed the capacity, quality and cost-effectiveness of the mobile network of one of the leading German mobile telecommunication companies. The goal of the project was to identify rural areas with a high demand in mobile network services and to relate the demand to demographic and geographic characteristics.

Mobile networks extend over geographic space and therefore make a strong claim on the inclusion of geographic data in the analysis process. A first explorative data analysis showed, for example, a decreasing network quality within cells in increasingly hilly areas. The overall input data of the project consisted of network usage, demographic and geographic information. In the data preparatory step we merged all three kinds of data and aggregated attribute values such as population and POI for radio network cells. In addition, we defined a target attribute which describes the demand of (future) network services:

\[
\text{cell potential} = \frac{\text{number of calls}}{\text{number of customers}} \times \text{population}.
\]

It weights the population of an area with the average number of calls of the present customers. Similar to the above project about customer segmentation, we applied subgroup discovery to detect variables that influence the demand for mobile network services. We used the SPIN! [12] spatial data mining platform, which has been developed within the EU project IST-1999-10536 SPIN!. It joins the power of data mining and geographic information systems by the implementation of spatial data mining algorithms and a dynamic visualization of spatial patterns. This allows for a mutual interaction between the system and the user, between automatically generated hypotheses and user defined hypotheses.

In our project we first analyzed cell potential according to network usage as, for example, call duration and transmitted data volume. A visualization of the best 30 subgroups suggested a spatial pattern along interstate highways. Figure 2 (left) shows results for the area of Stuttgart. Dark and light colored cells are not randomly distributed in space but form chain-like structures, dark colors indicating that cells participate in a high number of subgroup models. The map was then supplemented with various layers of geographic information, and a coincidence with interstate highways (dark lines) became visible. In the next step we added spatial information, including road network and public transportation data, to the analysis process. The resulting subgroups confirmed that cells along interstate highways and railways have an increased demand for mobile network services and should receive special attention during mobile network planning (see right plot in Figure 2).
2. Prediction of Traffic Frequencies and Detection of Customer Movement

In this section we present two case studies which involve network data and geographically referenced time series. The first study develops a traffic frequency map and emphasizes the tight integration of feature extraction and data mining algorithms for performance optimization. The second study extracts customer movements from tabloid sales data.

2.1. Frequency Map for German Cities

The German outdoor advertisement industry realizes a yearly turnover of about 780 million Euro. Its umbrella organization, which represents a joint market share of over 80 percent, provides performance indicators for poster sites on which the pricing of advertisement campaigns is based. The indicator consists of a quantitative and a qualitative measure. The quantitative term states the number of passing vehicles, pedestrians and public transport while the qualitative term specifies the average notice of passers-by. As part of an industrial project we developed a frequency map for German cities which today forms an essential part of price calculations in the German outdoor advertisement.

Essential for the prediction of traffic frequencies are the exploitation of geographic neighborhood, inclusion of background knowledge and performance optimization. We therefore applied a modified $k$-nearest neighbor (kNN) algorithm [13]. Nearest neighbor algorithms are generally able to incorporate spatial and non-spatial information based on the definition of appropriate distance functions. Thus, they are inherently spatial and exploit autocorrelation as a matter of principle. In order to gain background knowledge about the vicinity of a street, several geographically referenced attributes were aggregated. Furthermore, the large domain required a tight integration of spatial feature extraction and the algorithm in order to reduce expensive spatial operations.

The input data comprised several sources of different quality and resolution. The primary objects of interest were street segments, which generally denote a part of street between two intersections. Each segment possessed a geometry
object and had attached information about the type of street, direction, speed class etc. Germany contains in total 6.2 million street segments, for which about 100,000 traffic measurements were available. In addition, demographic and socio-economic data about the vicinity as well as nearby POI were known. Demographic and socio-economic data usually exist in aggregated form, for example, for official districts like post code areas. This information was likewise assigned to all street segments in an area. In contrast, POI are point data that mark attractive places like railway stations or restaurants. Clearly, areas with a high POI density are more frequented than areas with a low density. In order to obtain density information the POI data were aggregated. Two basic aggregation methods are buffers and drive time zones. As explained earlier, they mark, for a fixed location, the area which lies within a given range or which can be reached within a given time respectively. Drive time zones emphasize network constraints related to topology and allowed speed. Imagine, for example, two locations on opposite sides of a river. Their spatial distance is small, but the travel time between them depends on the location of the next bridge. For our application we created buffers around each street segment and calculated the number of relevant POI.

The central part of our traffic frequency prediction is a modified kNN algorithm, which models geographic space as a subcomponent of the general attribute space. The distance between two segments $x_a$ and $x_b$ is defined as the (normalized) sum of absolute distances of their attributes

$$d(x_a, x_b) = \sum_{i=1}^{m} |x_{ai} - x_{bi}|.$$

For fine tuning, the attributes were assigned domain dependent weights, which we will not discuss here further. The frequency $y_0$ of a street segment is calculated as the normalized weighted sum of frequencies from the $k$ nearest neighbors, each weight indirectly proportional to the distance between the two segments

$$y_0 = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i} \quad \text{with} \quad w_i = \frac{1}{d(x_0, x_i)}.$$

The kNN algorithm is known to use extensive resources as the distances between each street segment and all available measurements have to be calculated. For a city like Frankfurt this amounts to 43 million calculations (about 21,500 segments and 2,000 measurements). While differences in numerical attributes can be determined very fast, the geographic distance between line segments is computationally expensive. We therefore implemented the algorithm to perform a dynamic and selective calculation of distance from each street segment to the various measurement locations. First, at any time distances to only the top $k$ neighbors are stored, replacing them dynamically during the iteration over measurement sites. Second, a step-wise calculation of distance is applied. If the summarized distance of all non-spatial attributes already exceeds the maximal total distance of the current $k$ neighbors, the candidate neighbor can be safely discarded and no spatial calculation is necessary. Else, the distance between the minimum bounding rectangles (MBRs) of the line segments is calculated. The MBR distance is a

$$d(x_a, x_b) = \sum_{i=1}^{m} |x_{ai} - x_{bi}|.$$
lower bound for the actual distance between the line segments and less expensive to calculate. Again, if the distance of the non-spatial attributes plus the distance between the MBRs is greater or equal to the threshold, the instance can be discarded. Only if both tests are passed, the actual spatial distance is determined. For the city of Frankfurt, this integrated approach sped up calculations from nearly one day to about two hours. In addition, the dynamic calculations reduced the required disc space substantially.

2.2. Customer Movement Detection in Sales Data

In recent years, companies have spent great effort to systematically profile their customers in order to gain insights into target group characteristics and shopping behavior. Newspaper companies are especially interested in purchasing behavior as they face the challenge to supply each point of sale (POS) with the number of copies that are expected to be sold the next day. Existing forecasting systems employ complex time series models to predict the expected sales. However, they are bound to the temporal dimension and lack the understanding of local market situations and the customers' movement behavior at a particular selling point. Clearly, closures and sellouts influence shopping behavior and lead to one key business question: Where do readers buy a specific newspaper if their preferred shop is closed or has no more copies left? In an industry project with the publisher of Europe's leading tabloid newspaper, we developed a spatial model to detect and visualize local customer behavior.

The data basis for our model were approximately 110,000 POS, irregularly distributed over Germany. For each object a triannual history of sales figures was available. All objects were equipped with location information and could be mapped to a network of street segments. Information about the street network restrains vehicular as well as pedestrian movement and therefore simplifies the geographic space of possible movement. In addition, socio-demographic data about the vicinity of a POS as well as nearby POI were known. Both are needed to better understand, explain and learn the movement behavior of local target groups. For example, certain patterns or habits might correlate with certain demographic attributes or POI.

Shopping behavior is influenced by intrinsic as well as extrinsic factors [14,15]. This includes the individual destination, spatial barriers, mood (activation) and available selling points. In our model we assume that readers follow some routine. For example, the reader may buy the newspaper at his/her preferred selling point along his/her way to work. Such a routine can easily be interrupted by external factors as, for example, sellouts, vacation or openings of new shops requiring the customer to adapt his/her behavior. The challenge of the project was to detect, quantify and learn the behavior of customers after any such event and to predict the amount of copies that are additionally sold in alternative shops. Clearly, without personalized data customer movement can hardly be traced over a whole city. We therefore restricted our analysis to the local environment of a POS.

The first task in learning local movement patterns was to define a reasonable spatial unit for movement detection, which we call movement space (see left plot in Figure 3). If the unit is set too large, movement patterns will be lost in general
noise or overlaid by side effects due to events at other POS. Limiting the space too strongly, however, reduces the chance to detect reasonable movement patterns within. We employed two criteria to define the size of the unit, namely drive time zones and Voronoi neighbors. Drive time zones were used to set the initial (and maximal) extent of the movement space according to typical pedestrian walking speed. This area was further restricted based on the assumption that people who immediately seek an alternative POS will not pass by two alternative POS without buying. Of course, the individual choice depends on the knowledge of each customer about the set of selling points in his/her range (awareness set).

In order to limit the movement space, we calculated the convex hull of the second order POS Voronoi neighbor (see right plot in Figure 3). The resulting area was the space in which we looked for additional newspaper sales as an indicator for movement if the service at some POS had been unavailable. We call the set of all POS inside the movement space \textit{optional shops}.

The basic idea to detect local movement patterns in case of a changed local market situation (closures, sellouts, etc.) is to predict the sales of all optional shops assuming a typical shopping behavior and to compare the prediction with their actual sales. All shops showing an increased sale are likely to gain customers from the considered shop. In order to predict the expected number of copies at some POS, we calculated the sales based on shops with similar selling trends in the recent past. These shops are called \textit{reference shops}. The reference shops were dynamically determined by maximizing the similarity in selling trends applying a two week window before the registered event of the original POS. In this way, also seasonal or regional trends could be anticipated. Of course, all reference shops have to be located outside the movement space in order to be independent of any event-driven movement caused by the POS under consideration. If an optional shop sells a certain amount of copies above the expected number, it is likely that customers of the considered POS buy their newspaper alternatively at that point. Over time we gain robust knowledge about the movement behavior of the local customer base as well as a set of alternative shops inside the movement space.

With this knowledge newspaper companies can optimize the number of copies they deliver to each POS, taking into account not only time variant information but also the current local market situation. Moreover, the information about customer behavior provided by movement spaces allows to optimize location planning and to calculate the effect of opening or closing a POS in a specific area.
3. Mobility Mining in Outdoor Advertisement

Over the past five years GPS technology has steadily conquered its place in mass market and is on the threshold to become an every day companion. Besides the application in navigation systems, enterprises have also recognized the value of movement histories. The outdoor advertising industries of Germany and Switzerland commissioned nationwide GPS field studies to collect representative samples of mobile behavior. The data are used to calculate reach and gross contacts of poster campaigns for specific target populations.

This section describes a general approach for mobility mining in outdoor advertisement and highlights challenges of current industrial projects for Arbeitsgemeinschaft Media-Analyse e.V. (ag.ma) in Germany and Swiss Poster Research Plus (SPR+) in Switzerland.

3.1. Modeling of Poster Reach

The reach of a campaign states the percentage of population which has at least one contact with any poster of the campaign within a specified period of time. Poster reach allows to determine the optimal duration of some advertisement and to tune the configuration of poster networks as it expresses the publicity of some location and the spread of information within the population.

Given trajectories for a sample of the population and geographic coordinates of poster locations, the contacts with a given poster campaign can be extracted by spatial intersection and the reach can be determined. One challenge of calculating poster reach lies in the incompleteness of sample trajectories. For example, many trajectories are incomplete due to technical defects or because people forget (to switch on) their GPS devices. In addition, people tend to drop out of the study early, which leads to a decreasing number of participants with advancing time. What possibilities exist to handle incomplete data? In general, missing data can be treated in the data preparation step or within the modeling process. During data preparation, incomplete data objects can either be removed, ignored or tried to fill in by modeling. However, none of these possibilities are practicable in our application. First, if incomplete data objects are removed, the size of the data set decreases drastically because only a few test persons produce trajectories for the whole surveying period. Second, ignoring missing data leads to an underestimation of poster contacts and thus to an underestimation of poster reach as well. Finally, the reconstruction of missing trajectories is a fairly complex and ambitious task. We therefore treat missing data explicitly in the modeling step, applying a technique from the area of event history analysis.

Event history analysis (also survival analysis) [16] is a branch of statistics that investigates the probability of some event or the amount of time until a specific event occurs. It is usually applied in clinical studies and quality control where an event denotes, for example, the occurrence of some disease or the failure of a device. In our application an event denotes the first contact of a test person with a poster campaign. To calculate poster reach, we apply the Kaplan-Meier method which allows for censored data. This method adapts to differing sample sizes by calculating conditional probabilities between two consecutive events. If
no more data of a test person are available, the person is assumed to survive until the next event occurs and is censored afterwards. Thus, a gradual adjustment to the actual number of people in the sample is achieved.

3.2. Integration of Heterogeneous Mobility Data

The ag.ma, a joint industry committee of German advertising vendors and customers, commissioned a nationwide survey to collect mobility data using two different surveying technologies. From a total of about 30,000 test persons, one third was provided with GPS devices while the other test persons where queried about their movements in a Computer Assisted Telephone Interview (CATI). One task of the project was to analyze both data sets according to their content and structure-related differences, and to combine the data sets for modeling if possible.

Both surveying techniques bear the risk of incomplete and erroneous data. GPS devices may easily be forgotten or passed on to other family members while telephone interviews demand a precise and complete recollection of the activities of the previous day. We therefore compared the mobile behavior of both data sets. The analysis showed similar movement behavior as, for example, in the average number of trips per day or the average distance traveled.

The main structural difference of the data sets are the different surveying periods. While all GPS test persons collected data over a period of one week, CATI test persons were asked about their movements on the previous day of the interview only. However, a combination of both data sets with regard to their structure was possible due to the adaptive character of Kaplan-Meier. As Kaplan-Meier censors missing days, the modeling process is robust against varying lengths of surveying periods.

3.3. Extrapolation of Reach over Time and Space

In our project with SPR+ we investigate further research questions that concern the prediction of reach when only a limited number of measurements are available. The first task is to predict poster reach when the measurement period is shorter than the desired interval of time. The second challenge is to predict poster reach in a city where no measurements at all are available. In this case, the reach of a given campaign within one city has to be inferred from the mobility of another (similar) city.

For the extrapolation of reach beyond the surveying period, we combine two different extrapolation techniques. The first technique utilizes the reach of one week to fit a log-linear function and subsequently extrapolates values for longer periods. The second technique relies on the assumption of weekly periodic mobility patterns and replicates mobile behavior accordingly. Both techniques are interweaved according to the stability of available data.

The extrapolation for areas without GPS measurements is a great challenge. Neither GPS data nor other mobility information as, for example, traffic frequencies are available. In addition, individual poster characteristics which affect the intensity of a contact need to be taken into account for the calculation of reach. The extrapolation method therefore consists of three separate steps. First, the
traffic behavior at the poster locations of interest is inferred. Second, the passages are scaled according to individual poster characteristics. Finally, the reach of a campaign with a similar contact distribution is assigned to the campaign of interest. In the first step, various location attributes such as the type of street, type and number of nearby POI or the size of population define a similarity measure by which poster passages are extrapolated. In the next step, a scaling factor which transforms passages into poster contacts is applied. The factor depends on individual poster characteristics and is determined based on evaluations in GPS cities. The final assignment of poster reach depends again on a similarity measure which is defined on the contact distribution of the campaign of interest. The extrapolation method thus accounts for general traffic characteristics, yet allows for individual features of poster campaigns. In order to validate our extrapolation method, we applied the technique in a city with GPS measurements. The comparison of modeled and extrapolated values showed a high correlation.

4. Summary

In this chapter we present a collection of spatial data mining applications which have been carried out at Fraunhofer IAIS over the past years. The projects demonstrate the wide applicability of spatial data mining and the various facets of spatial data types, preprocessing methods and algorithms. We begin the chapter by case studies for marketing and planning, which involve spatial feature extraction on various levels of aggregation, extensive visual analysis and subgroup discovery. We then proceed to more advanced data types in form of street networks and geographically referenced time series. These case studies emphasize the benefit of specialized algorithms that allow for dynamic and selective computations and underline the necessity for application dependent definitions of neighborhood relationships. Finally, we introduce a case study using spatio-temporal trajectories which calls attention to the problem of missing data.

Without question, spatial data mining is an attractive research area with high impact on industry and many further challenges to meet.

Acknowledgment

The authors would like to thank all business partners for their close cooperation. The publication of this chapter would have been impossible without their interest and participation. Parts of this work have been inspired by research within the EU projects IST-1999-10536 SPIN! (Spatial Mining for Data of Public Interest) and IST-6FP-014915 GeoPKDD (Geographic Privacy-aware Knowledge Discovery and Delivery). Finally, the authors acknowledge and thank all members of the Department of Knowledge Discovery who have contributed by their research and constant work to the success of the presented projects.
References


