Pedestrian Flow Prediction in Extensive Road Networks using Biased Observational Data

Michael May
Fraunhofer IAIS
Schloss Birlinghoven
D-53754 Sankt Augustin
+49 2241 142039
michael.may@iais.fraunhofer.de

Simon Scheider
University of Münster (IfGI)
Weseler Straße 253
D-48151 Münster
+49 176 81028432
simonscheider@web.de

Roberto Rösler
Fraunhofer IAIS
Schloss Birlinghoven
D-53754 Sankt Augustin
+49 2241 143510
roberto.roesler@iais.fraunhofer.de

Daniel Schulz
Fraunhofer IAIS
Schloss Birlinghoven
D-53754 Sankt Augustin
+49 2241 142401
daniel.schulz@iais.fraunhofer.de

Dirk Hecker
Fraunhofer IAIS
Schloss Birlinghoven
D-53754 Sankt Augustin
+49 2241 141509
dirk.hecker@iais.fraunhofer.de

ABSTRACT

In this paper, we discuss an application of spatial data mining to predict pedestrian flow in extensive road networks using a large biased sample. Existing out-of-the-box techniques are not able to appropriately deal with its challenges and constraints, in particular with sample selection bias. For this purpose, we introduce *s-knn-apriori*, an efficient nearest neighbor based spatial mining algorithm that allows prior knowledge and deductive models to be included in a straightforward and easy way.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Induction, Knowledge acquisition; I.5.1 [**Models**]; I.5.2 [**Design Methodology**]

General Terms

Algorithms, Economics, Reliability, Human Factors, Verification

Keywords

Spatial data mining, pedestrian flow prediction, sample selection bias, prior knowledge, extensive road networks, large scale data

1. INTRODUCTION

The German outdoor advertising market has a yearly turnover of 1,200 million $. The value of each single transaction done in this business sector depends in an important part on a spatial data mining prediction. One of the three components of this prediction is a model for pedestrian flow. This is a very rare if not unique case where a spatial data mining model, like [11], is business critical for a whole branch of industry, comprising a large number of companies. As we will show, this application is far from requiring just a simple instantiation of existing tools for inductive learning or statistical modeling. The most distinct aspect of our task is to predict an extensive road network using a very large biased convenience sample. We describe a data mining algorithm that infers a frequency map, where for each street segment in every German city (approx. 6 Mio. segments), the model predicts the average number of pedestrians per hour.

ACM GIS '08, November 5-7, 2008. Irvine, CA, USA

(c) 2008 ACM ISBN 978-1-60558-323-5/08/11...$5.00

The Fachverband Außenwerbung e.V. (FAW) is the governing organization of German outdoor advertising. FAW provides performance indicators on which pricing of poster sites is based. The value of each site is characterized by a quantitative measure, the number of passing vehicles, pedestrians, and public transport, and a qualitative measure which specifies the expected notice of passers-by. The application has had a strong impact on the way business is done in outdoor advertising: It has become the basis for pricing of posters for that branch of industry; it can be used for selecting new poster sites and for planning campaigns. Outdoor advertising perceives the project as a milestone for their business. The reliability of the prediction result seems to be accepted by dozens of companies in this area.

In the past, for each poster site to be evaluated, video measurements have been made by GfK Germany. Measurements at a site have been taken manually at 4 different days and 4 different hours lasting 6 minutes. Over the years, a collection of more than 100.000 pedestrian flow measurements has been collected this way. The task was to use this data set along with readily available secondary data. Public transport is predicted using different data and a different approach, see [14].

In this paper, we discuss a general solution to account for these requirements by incorporating prior domain knowledge into an appropriately adapted inductive learning algorithm. Because of sample selection bias, we evaluate our algorithm in a two-fold way: by evaluating its error using the training data set and by checking the informal plausibility of its prediction results.

2. COMMON PRACTICAL CONSTRAINTS

As our task is to predict pedestrian flow in an extensive German wide road network which comprises 6.385.721 street segments, our method clearly must belong to the class of *large scale methods.* Furthermore, it is clear that in the near future it will economically not be feasible to devise a tailored experiment or observational study to cover such an extensive network. Therefore, we are in the classical situation that gave *knowledge discovery in large databases* its name. We are forced to use a *convenience sample* that was gathered for a distinct purpose.

For many methods of statistics and data mining, it is assumed that training and test data for an inductive model are sampled *in an independent manner from the same probability distribution*, so that data tuples are sampled with a probability equal to their prior probability in the population. However, this assumption is definitely violated in our prediction scenario. In general, a sample selection bias exists, if the probability of sampling a data tuple

(**x**,y), with target variable (e.g. pedestrian flow) y and feature vector **x**, statistically depends on its multidimensional value. We denote the sample selection procedure by the binary variable s (compare [6]). As outlined in [7], an accurate model trained from and evaluated on a biased sample may be highly inaccurate on test data taken as an iid sample. This is because the sampled conditional probability of a pedestrian flow value given a feature vector, P(y|**x**,s) misrepresents the actual probability P(y|**x**), as the target value is dependent both on s and x, and therefore it cannot be asymptotically approximated by an inductive learner, even if we assume that the "true" model was included in the learner's model space [7]. As we have a non-extensible convenience sample, it is very likely that not every possible feature vector is included in it. Therefore, its probability will unrealistically be estimated with 0 and there is at least a dependency on **x**. Moreover, there is also a class bias involved, because the market value of a poster site depends on traffic flow, and therefore these tend to be at places with comparatively high flow rates. *Therefore high flow rates are oversampled*.

There are some more reasons why the application is challenging. First, we have *measurement errors* and *outliers* present in the data set for a number of reasons. Second, the phenomenon modeled here is a prototypical example for *heterogeneity* and discontinuous non-smooth target value changes in the feature space [9]. Furthermore, in our scenario, it is essential that the method must be able to *correctly predict existing available measurements without estimation bias*, and that prediction must be *consistent with common sense expectation*. In the absence of empirical evidence in favor of the prediction, a model can still be trustworthy if it follows common sense.

3. APPROPRIATE PREDICTION METHODS

3.1 Discussion of Related Approaches

From section 2 it is clear that we should not use deductive prediction methods that strongly rely on *a priori doubtful theories*, like *spatial interaction models*, *entropy maximizing models* [4], or micro simulation models including *Monte Carlo simulation*, *Markov models* and *cellular automata* [2] [3]. This is because we need to correctly predict existing measurements without estimation bias, and because extensive detailed controlled studies are not available to calibrate the many parameters of such a model. Inductive regression approaches like in [5] and [13] are able to fit the data very closely and therefore more appropriate.

In order to account for the presence of outliers and heterogeneity, it is reasonable to use *kernel prediction methods*. Kernel prediction methods achieve the highest flexibility in estimating the regression function by fitting a local simple model separately at each point of the feature space [10]. As outliers in the data can affect only predictions in its neighborhood, using kernel methods is a way of restricting the influence of measurement errors. In general, kernel methods can account for high heterogeneity in the modeled domain because they do not rely on global functional assumptions [10]. Nevertheless, large scale data makes the usage of kernel methods a challenge, because the training set has to be queried at prediction time. Choosing a most scalable kernel method is therefore mandatory.

From the set of available remaining methods, like e.g. *structured local linear regression*, we chose the simplest one, k-Nearest Neighbor or *k-NN*. The most important reasons for this were the need for simple explanations and plausible values, and the importance to incorporate deductive domain prior knowledge in order to cope with sample selection bias.

On the one hand, the range of predicted values in k-NN can never exceed the empirically given value range of the training sample. Furthermore, as k-NN belongs to the class of *prototype methods* that predict a point from its nearest prototype, we always have an existing measured prototype available in order to explain a prediction by analogy reasoning. On the other hand, *prototyping* is a way of easily incorporating domain prior knowledge and to account for sample selection bias. This option will be described in the next section.

3.2 Dealing with Sample Selection Bias using s-kNN-apriori

As outlined in section 2, the predominant class bias in our sample exists due to an oversampling of high flow rates whose concrete mechanism is unknown. Prototype street segments with low pedestrian flow are under sampled and even *likely not to be present in our sample*. Therefore, sample correction methods [7] do not help us, since they rely on every point of the data space y, **x** to be sampled with a frequency > 0, and furthermore it is necessary to know their actual sample selection probability P(s|**x**,y). *Model averaging* (joint probability averaging) could perhaps be a useful way to make our prediction more accurate [7]. But this poses of course a problem of scalability in our large scale application.

The best way to deal with sampling bias in our context is therefore to incorporate domain prior knowledge in the form of *prototypes*. For example, we perfectly know how regions in the road network with low pedestrian flow look like. These are *housing areas* or *non-pedestrian areas* with low accessibility to important attractors, like *public transport stations*, *city halls and public buildings*, *tourist attractions and retail locations*. We will outline this idea in the following paragraphs by extending the basic k-NN formula.

The basic k-NN algorithm relies on a distance function between two n-dimensional data vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, in our case defined as the sum of the absolute distances among the n normalized attributes

$$\mathrm{d}(x_1, x_2) = (1-\alpha)\sum_{i=1}^{n}|x_{1i} - x_{2i}| + \alpha\, Gdist_{x1,x2}\,.$$

In the s-kNN algorithm, the minimum distance *Gdist* between the spatially extended geographic street segments of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is one weighted component in this distance measure, with α representing the tradeoff between spatial and non-spatial distance. Weights can fruitfully be assigned to other attributes as well.

The prediction y_0 at vector $\mathbf{x}_0$ is the normalized weighted sum of its k nearest neigbors,

$$y_0 = \frac{\sum_{j=1}^{k} w_j\, y_j}{\sum_{j=1}^{k} w_j}$$

where each weight w_j is a kernel $K(\boldsymbol{x}_0,x_j)$ with $w_j = 1/\,d(x_0, x_j)$.

For s-kNN-apriori, we constructed different prototypes with respect to the feature vector $\boldsymbol{x}_0$. We constructed a prototype based on the lowest empirically measured flow rate for all street

segments with zero or low attractor accessibilities. Likewise, we constructed an a-priori peak flow prototype for pedestrian areas. For other cases, we constructed a mixed model with additive and non-additive components based on number of public buildings, number of railway and public transport stations, number of restaurants and number of touristic POI. We assumed a plausible flow generation constant and thereby assigned a flow value.

We incorporated these prototypes into s-kNN as the $k+1^{th}$ neighbor and weighted all neighbors dependent on the *sparsity or confidence* of our kernel prediction, which is parameterized simply as the distance to the nearest measurement. Formally, in order to estimate pedestrian flow at feature vector x_0 we take the weighted sum of the respective a-priori value y_{ap} with respect to x_0 and the a posteriori values y_j derived from available measurements in the k-neighborhood of x_0

$$y_0 = \frac{\sum_{j=1}^{k} w_j y_j + (1 - w_1) y_{ap}(x_0)}{\sum_{j=1}^{k} w_j + (1 - w_1)}.$$

For example, if a segment is nearby crowded pedestrian areas, we use high values for y_{ap}, and if it is near a housing area, we will use low values for y_{ap}. The influence of this prototype is small if the nearest measurement y_j is near and has high weight. If the nearest neighbor has the maximum weight of 1, the a-priori term vanishes. If the measurements are all far away, the a-priori term dominates the average.

3.3 Predictive Features and Performance

The *spatial feature* component in s-K-NN is useful for the following reasons: Prediction is done between streets segments (fig. 1) carrying flow information. Segments are represented as poly-lines and can have complex shapes. To account for spatial dependence of flow, the spatial distance between segments needs to be calculated. In particular, two segments that meet in a point must have a distance of 0. Simple Euclidian distance between centroids can lead to highly misleading results, as shown in fig. 1: In the street network the segment with centroid A and the segment with centroid B meet each other. However, all points in the grey circle are closer to A than B with respect to Euclidian distance. B is not even among the 20 nearest neighbors of A!

Unfortunately, this preferable distance calculation heavily downgrades time complexity of our algorithm. Therefore we utilized the general idea of partial evaluation of the distance function (e.g.[8]) and applied it to s-kNN using Minimum Bounding Rectangles (MBR) [12]. Calculating the minimum distance between two MBRs is much cheaper than calculating the distance between the actual polylines. And if the distance of the non-spatial attributes plus the distances between the MBR is greater or equal to the k^{th} nearest neighbor distance, the instance can safely be discarded. For a city like Frankfurt a full computation would amount to 43 million spatial calculations (about 21.500 segments and 2.000 measurements). Using partial evaluation, calculations were sped up from nearly 1 day to about 2h (i.e. one order of magnitude).

We precalculated *non-spatial features* for every street segment from the available German wide Points of Interest (POI) and street network dataset including name of the road, presence of pedestrian area, number of railway stations, number of restaurants and number of public buildings in 250 m distance, and number of public transport stations in 20, 125 and 250 m distance. This is not supposed to be a final but a convenience list. Especially, we suppose that more sophisticated network features or detailed retail frontage figures, as in [5] may help improving the results.

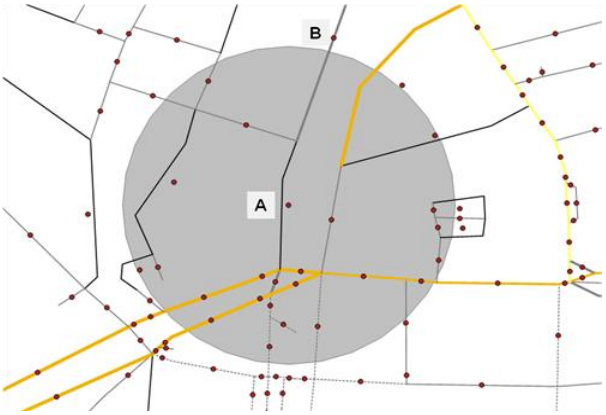

Figure 1. Distances between centroids does not well capture the distance among polylines.

4. EVALUATION

For reasons outlined in section 2, standard statistical evaluation techniques, like bootstrapping or cross validation, are not good indicators of the true error on the network. As a German wide controlled study is not available to us, we use instead a many-fold evaluation technique built on plausibility rather than on numerical performance measures.

We first studied how close an s-knn model fits a similarly biased sample by leave-one-out cross validation using the *coefficient of determination*, the *root mean squared error* and the *mean absolute error*. We compared this performance with available out-of-the-box techniques, like linear regression and Gaussian regression from the Weka toolkit.

Table 1. Leave-one-out cross validation on biased sample (n=943) in Dresden using fixed attribute set

Method	R^2	RMSE	MAE
Multiple linear regression	0,85	69,87	48,38
Gaussian Process (RBF & 10CV)	0,85	84,89	49,82
5-NN	0,92	51,22	26,58
s-5-NN	0,91	53,95	27,60
s-5-NN-apriori	0,92	51,71	25,08

In table 1, the results for the German city of Dresden illustrates that spatial k-NN is obviously among the closest predictors of the biased sample, and the a-priori version is at least equally good or even better. Note that this evaluation is *not appropriate to estimate overall predictive quality*. The need for it arises because in our scenario, a predictor must closely reproduce existing measurements.

Secondly, we evaluated the quality of the overall prediction results by visual analytics [1] and plausibility. Figure 2 illustrates an example of pedestrian peak flow in the city center of Dresden. In the absence of appropriate measurements, the famous central

tourist attractions *Semper Opera, Zwinger and Castle* in the upper left (compare circles in figure 2) are seriously underestimated by the purely inductive method s-knn. The presence of important POI on this location allows for a more plausible prediction with the s-knn-apriori version. Most similar measurements in the attribute space are located far away and are therefore very unlikely to correctly predict pedestrian flow.

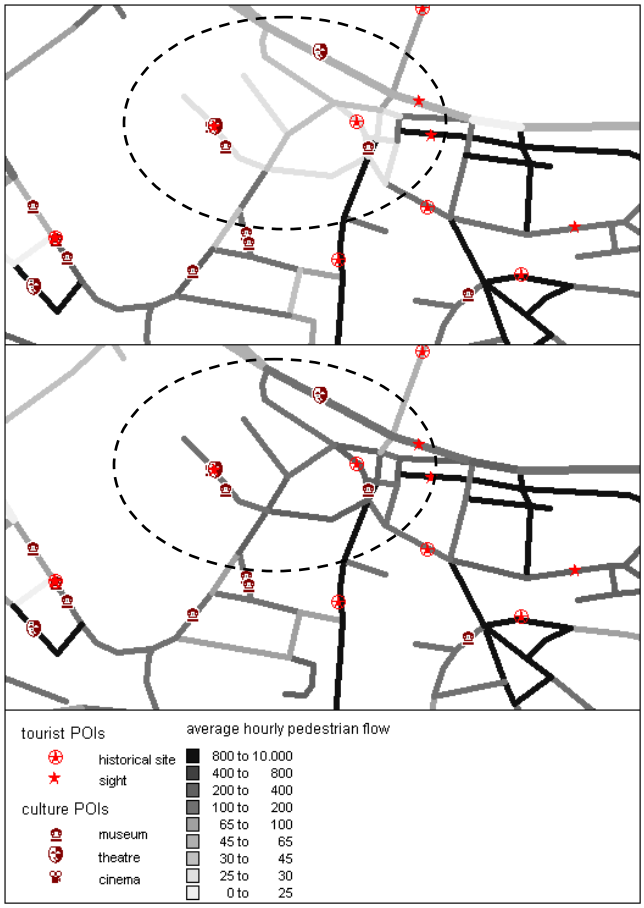

Figure 2. Historical city centre of Dresden predicted with basic s-5-NN (upper map) and s-5-NN-apriori.

5. CONCLUSION

We described a market relevant application of pedestrian flow prediction for large scale data under sample selection bias. We also discussed reliable solution methods and proposed a spatial data mining algorithm called s-knn-apriori, based on k-nearest-neighbor and prior domain knowledge in the form of synthetic prototypes. In this scenario, as in most applied scenarios of machine learning, a representative test or training sample is not available, and therefore a purely inductive prediction model cannot be reliably estimated. A reliable algorithm should instead incorporate prior domain knowledge and should be evaluated by plausibility checks using exploratory analysis tools.

For future work we plan to improve the available feature set and work on empirical validation studies in order to estimate and improve the numerical quality. We also try to incorporate more complex prior models based on Gaussian distributions.

6. ACKNOWLEDGMENTS

The project reported here has been supported over several years by the German Fachverband Außenwerbung e.V. (FAW). Many partners contributed to this project. GfK Germany is responsible for flow measurements and visibility conditions, MGE Data for various geographic data preparation tasks, DDS as a data provider. We gratefully acknowledge their contribution.

7. REFERENCES

[1] Andrienko, N. and Andrienko, G. 2005. Exploratory Analysis of Spatial and Temporal Data – A Systematic Approach, Springer New York 2005

[2] Blue, V.J. and Adler, J.L. 1998. Emergent fundamental pedestrian flows from cellular automata microsimulation. Transportation Research Record 1644, 29-36

[3] Borgers, A and Timmermans, H.J.P. 1986. City Centre Entry Points, Store Location Patterns and Pedestrian Route Choice Behaviour: A Microlevel Simulation Model. Socio-Econ. Plan. Sci. 20, 1 (1986), 25-31

[4] Butler, S. 1978. Modeling Pedestrian Movements in Central Liverpool, Working Paper 98, Institute of Transport Studies, University of Leeds

[5] Desyllas, J., Duxbury, E., Ward, J., Smith, A 2003. Pedestrian Demand Modeling of Large Cities: An Applied Example from London. Centre for Advanced Spatial Analysis Working Paper 62, University College London

[6] Fan W., Davidson I., Zadrozny B. and Yu P. 2005. An Improved Categorization of Classifier's Sensitivity on Sample Selection Bias. 5th IEEE International Conference on Data Mining, ICDM 2005, Louisiana

[7] Fan, W. and Davidson, I. 2007. On Sample Selection Bias and Its Efficient Correction via Model Averaging and Unlabeled Examples. Proc. SIAM Data Mining Conference 2007, Minneapolis

[8] Grother, P., Candela, G., Blue, J. 1997. Fast implementations of nearest neighbour classifiers, Pattern Recognition 30, 3 (1997), 459-465

[9] Harney, D 2002. Pedestrian modeling: current methods and future directions. Road & Transport Research 11, 4 (2002), 2–12

[10] Hastie, T., Tibshirani, R., Friedman, J. (2001).The elements of statistical learning: data mining, inference, and prediction (Springer series in statistics). Springer, New York, 2001.

[11] Klösgen, W., May, M. 2002. Spatial Subgroup Mining Integrated in an Object-Relational Spatial Database. Proc. PKDD 2002, 275-283

[12] Papadias, D., Sellis, T., Theodoridis, Y., Egenhofer. M. 1995. Topological relations in the world of minimum bounding rectangles: a study with R-trees, ACM Sigmod Record 24, 2 (1995), 92-103

[13] Pushkarev, B. and Zupan, J.M. 1971. Pedestrian travel demand. Highway Research Record 355, 37-53

[14] Scheider, S. and May, M. 2007. A Method for Inductive Estimation of Public Transport Traffic using Spatial Network Characteristics. 10th AGILE International Conference on Geographic Information Science 2007, Aalborg University