Pedestrian Quantity Estimation with Trajectory Patterns

Thomas Liebig, Zhao Xu, Michael May, and Stefan Wrobel

Fraunhofer IAIS
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
{Thomas.Liebig,Zhao.Xu,Michael.May,Stefan.Wrobel}@iais.fraunhofer.de

Abstract. In street-based mobility mining, traffic volume estimation receives increasing attention as it provides important applications such as emergency support systems, quality-of-service evaluation and billboard placement. In many real world scenarios, empirical measurements are usually sparse due to some constraints. On the other hand, pedestrians generally show some movement preferences, especially in closed environments, e.g., train stations. We propose a Gaussian process regression based method for traffic volume estimation, which incorporates topological information and prior knowledge on preferred trajectories with a trajectory pattern kernel. Our approach also enables effectively finding most informative sensor placements. We evaluate our method with synthetic German train station pedestrian data and real-world episodic movement data from the zoo of Duisburg. The empirical analysis demonstrates that incorporating trajectory patterns can largely improve the traffic prediction accuracy, especially when traffic networks are sparsely monitored.

Keywords: Pedestrian Quantity Estimation, Trajectory, Gaussian Process Regression, Graph Kernels

1 Introduction

Estimation of traffic volumes is a common task for street based traffic and the achieved values are highly interesting for risk analysis, quality of service evaluation, location ranking and mobility analysis applications. Particularly, for pedestrians traffic, knowledge on people's presence offers a vast chance for improvement of the signage and the infrastructure. Facilities provided to people depend on pedestrian movements and volumes. To give a few general examples: locations of information desks, shops or toilettes depend on the quantity of persons; pathwidths of the corridors in a stadium depend on people's quantity as well, mobile phone networks are planned according to the expected movements and even locations of advertisement billboards are placed such that they are potentially noticed by as many pedestrians as possible. Modelling the pedestrian quantities gives indispensable insights on visitor preferences and motivations at a particular public event or site and thus supports creation of intelligent environments.

In this work we focus on the estimation of traffic volumes for pedestrians within closed environments. These are sites or buildings which have in common, that no people reside inside but all present people leave after some time period. Thus, these closed environments have dedicated entrances and exits. Prominent examples are train stations, terminals, shops, shopping malls, parks as well as zoos. As shown in the previous examples, knowledge of pedestrian movement provides indispensable benefits to safety, marketing as well as infrastructural applications. Thus, over the past years many sensor technologies to fetch empirical measurements and record pedestrian volumes have been developed (most popular ones are video surveillance, laser beams and Bluetooth sensors). However, empirical measurements are usually rare due to some constraints, e.g., budget limitations. This arises the following questions.

- How can values on pedestrian quantities be estimated from few empirical measurements?
- At which places should a constrained number of quantity sensors be located?

Often, available data is limited to few measurements and some prior knowledge, e.g., floor plan sketches, knowledge on preferred routes by local domain experts. Incorporating prior knowledge is thus essential to address the above challenges. However there are few approaches taking into account the trajectory patterns, although pedestrians generally show some move preferences [1, 2], especially in closed environments, e.g., train stations. Consider for example an average daily traffic (ADT) prediction problem with traffic networks consisting of only one junction. As shown in Figure 1, a T-junction occurs in a wide corridor that goes straight. At the junction a small corridor intersects and an expert knows that it is most likely for persons to continue their walk straight ahead in the main corridor. Assume further to have a frequency sensor placed in the main corridor which measures a known amount of people within considered time interval. Under these circumstances, existing traffic volume estimation methods, e.g., k-nearest neighbour and standard Gaussian process regression, do not take into account the expert knowledge and thus may not effectively provide accurate estimations for the side corridor.

We propose a traffic volume estimation method based on Gaussian process regression, which incorporates topological information and the expert knowledge on preferred trajectories with a trajectory pattern kernel. By exploring trajectory patterns, our method can also effectively elicit most informative sensor placements. We demonstrate the advantages of our approach with two applications. The first one is pedestrian mobility analysis in German train stations. As the data available is some statistical analysis on the network characteristics. We draw synthetic (but realistic, see Section 4) traffic networks and pedestrian movement based on these analysis. Secondly, our approach is applied to the real world scenario at the zoo of Duisburg (Germany). The pedestrian mobility data of the visitors was collected with Bluetooth tracking technology [3]. The empirical analysis demonstrates that incorporating trajectory patterns can largely improve the traffic prediction accuracy, especially when traffic networks are sparsely monitored. Our work contributes an extensive approach to the pedestrian volume

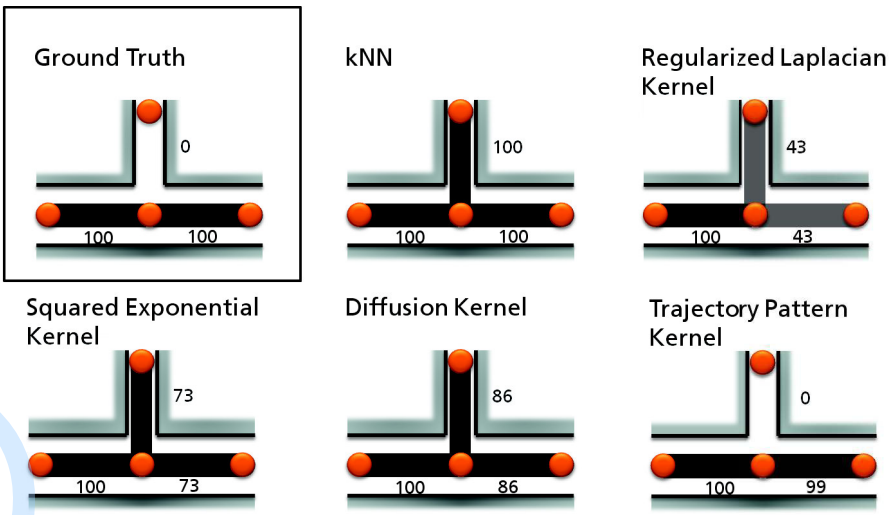

Fig. 1. T-junction example. Main corridor is horizontal. Expert knowledge which presumes that people walk straight ahead in the main corridor is given. Left corridor frequency measurement is given. Numbers denote relative frequencies in percent.

estimation problem and it provides an efficient, applicable solution to industrial real world scenarios.

The rest of the paper is organized as follows. Section 2 gives an overview of related work. Then we describe the proposed approach in Section 3. Empirical analysis and real world applications are presented in Section 4. Finally, we give our conclusions and discussions on future work in Section 5.

2 Related Work

Existing literature distinguishes between average daily traffic (ADT) estimation and average annual daily traffic (AADT) estimation. Whereas AADT focuses on estimation of a traffic volume depending on the day of the year, ADT estimation provides an average for a particular day. Naïve approach for AADT estimation is utilization of ordinary linear regression (OLR) [4]. Street segment attributes (e.g. number of lanes or function classes) are multiplied by weights which are subject for least squares regression. Improvements of this technique were achieved by respecting the geographical space by usage of geographical weighted regression (GWR) [4] and by application of k-nearest neighbor approaches (kNN) [5]. In [6] the AADT prediction of kNN for a particular location is improved by weighting measurements by their temporal distance to the prediction time. This approach showed better results than application of Gaussian maximum likelihood (GML) approaches for weighting of the historical data points. Recent improvements to kNN non parametric regression were made in [7]. Although performing the

k-nearest neighbor search in the attribute space of the street segments, this approach selects the spatially closest neighbours, as they have highest impact. In [7–9] the ADT estimation problem is addressed as business critical industrial data mining use case as the pricing in the outdoor advertisement sector in Germany and Switzerland relies on the estimated values [7]. Their proposed algorithm is a spatial k-nearest neighbour (S-kNN) approach that incorporates geometric distances for estimation of an unknown segment. The closer a measured segment is to an unmeasured one, the higher its impact. This is similar to the Kriging approach described in [10] but goes beyond it, as just the k-nearest neighbours were used for prediction.

The regression approaches in [11–13] are motivated by outdoor advertisement use cases. In contrast to the previously described methods their approaches operate in the space of the possible routes instead of the segment-attribute space. After an extensive path enumeration step, this work checks every possible path on plausibility and considers the resulting set of plausible paths for path frequency estimation using a least squares regression at the measurement locations. That approach contains a basic assumption on pedestrian route choice, namely pedestrians prefer the shortest path to travel from one location to another. But in some scenarios this assumption does not hold [14]. [15] applies Gaussian process regression(GPR) to the estimation of traffic frequencies within a public transport network. Their approach is not applicable to the problem of this work as pedestrian mobility patterns arise in the traffic flow due to the non-random but motivated individual behaviour, which was also result of the analysis of about 2'500 traces of train station visitors in [16]. More on challenges for pedestrian modelling can be found in [17]. [18] shows in a study of 210 infrastructure planning projects that the inaccuracy of traffic forecasts can be immense. In this paper we propose a new GPR based method to tackle the pedestrian quantity estimation problem which explores the prior knowledge of trajectory patterns.

3 GPR with Trajectory Patterns

In the paper we focus on the pedestrian quantity estimation in closed environments, e.g., train stations, shopping malls and zoos. Unlike the outdoor pedestrain quantity estimation, the continuous tracking technologies, e.g. global positioning system (GPS), are not feasible due to the lack of GPS signal in buildings and expensive deployment of the hardware. Recently developed technologies (lightbeams, video surveillance, Bluetooth meshes) record episodic movement data [19] or its location based aggregate, presence counts at low expenses. Episodic movement data is represented by tuples $< o, p, t >$ of moving object identifier o, discrete location identifier p and corresponding timestamp t. The location based aggregate, presence counts, for time interval Δt, as known as number of visits, quantity or traffic frequency, is defined as

$$NV(p, \Delta t) = | < o, p, t >, t \in \Delta t| \, . \tag{1}$$

To estimate the traffic volume at unmeasured locations, we propose a nonparametric Bayesian method, Gaussian process with a random-walk based trajec-

tory kernel. The method explores not only the commonly used information in the literature, e.g. traffic network structures (retrieved by tessellation from the floor plan sketch) and recorded (or aggregated) presence counts NV at some measurement locations, but also the move preferences of pedestrians (trajectory patterns) collected from the local experts.

Consider a traffic network $\tilde{\mathcal{G}}(\tilde{\mathbf{V}}, \tilde{\mathbf{E}})$ with N vertices and M edges. For some of the edges, we observe the pedestrian quantities, denoted as $\mathbf{y} = \{y_s := NV(\tilde{e}_s, \Delta t) : s = 1, \ldots, S\}$. Additionally, we have the information of the major pedestrian movement patterns $\mathcal{T} = \{T_1, T_2, \ldots\}$ over the traffic network, collected from the local experts or the tracking technology (e.g. Bluetooth tracking technology). Obviously, taking into account the trajectory patterns is beneficial to predict the unknown pedestrian quantities: The edges included in a trajectory pattern appear to have similar pedestrian quantities. To meet the challenge, we propose a nonparametric Bayesian regression model with trajectory based kernels.

The pedestrian quantity estimation over traffic networks can be viewed as a link prediction problem, where the predicted quantities associated with links (edges) are continuous variables. In the literature of statistical relational learning [20,21], commonly used GP relational methods are to introduce a latent variable to each vertex, and the values of edges is therefore modeled as a function of latent variables of the involved vertices, e.g. [22,23]. Although these methods have the advantage that the problem size remains linear with the size of the vertices, it is difficult to find appropriate functions to encode the relationship between the variables of vertices and edges for different applications.

In the scenario of pedestrian quantity estimation, we directly model the edge-oriented quantities [5, 6, 15] within a Gaussian process regression framework. First, we convert the original network $\tilde{\mathcal{G}}(\tilde{\mathbf{V}}, \tilde{\mathbf{E}})$ to an edge graph $\mathcal{G}(\mathbf{V}, \mathbf{E})$ that represents the adjacencies between edges of $\tilde{\mathcal{G}}$. In the edge graph $\mathcal{G}$, each vertex $v_i \in \mathbf{V}$ is an edge of $\tilde{\mathcal{G}}$; and two vertices of $\mathcal{G}$ are connected if and only if their corresponding edges share a common endpoint in $\tilde{\mathcal{G}}$. To each vertex $v_i \in \mathbf{V}$ in the edge graph, we introduce a latent variable f_i which represents the true pedestrian quantity at v_i. It is value of a function over the edge graph and the known trajectory pattens, as well as the possible information about the features of the vertex. The observed pedestrian quantities (within a time interval Δt) are conditioned on the latent function values with Gaussian noise ϵ_i

$$y_i = f_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \; . \tag{2}$$

As mathmatical form and parameters of the function are random and unknown, f_i is also unknown and random. For an infinite number of vertices, the function values $\{f_1, f_2, \ldots\}$ can be represented as an infinite dimensional vector. Within a nonparametric Bayesian framework, we assume that the infinite dimensional random vector follows a Gaussian process (GP) prior with mean function $m(x_i)$ and covariance function $k(x_i, x_j)$ [24]. In turn, any finite set of function values $\mathbf{f} = \{f_i : i = 1, ..., M\}$ has a multivariate Gaussian distribution with mean and covariances computed with the mean and covariance functions of the GP [24].

Without loss of generality, we assume zero mean so that the GP is completely specified by the covariance function. Formally, the multivariate Gaussian prior distribution of the function values $\mathbf{f}$ is written as

$$P(\mathbf{f}|\mathbf{X}) = \mathcal{N}(0, K),$$

where K denotes the $M \times M$ covariance matrix, whose ij-th entry is computed in terms of the covariance function. If there are vertex features $\mathbf{x} = \{x_1, ..., x_M\}$ available, e.g., the spatial representation of traffic edges, a typical choice for the covariance function is the squared exponential kernel with isotropic distance measure:

$$k(x_i, x_j) = \kappa^2 \exp\left(-\frac{\rho^2}{2}\sum_{d}^{D}(x_{i,d} - x_{j,d})^2\right), \tag{3}$$

where κ and ρ are hyperparameters.

Since the latent variables $\mathbf{f}$ are linked together into an edge graph $\mathcal{G}$, it is obvious that the covariances are closely related to the network structure: the variables are highly correlated if they are adjacent in $\mathcal{G}$, and vice versa. Therefore we can also employ graph kernels, e.g. the regularized Laplacian kernel, as the covariance functions:

$$K = \left[\beta(L + I/\alpha^2)\right]^{-1}, \tag{4}$$

where α and β are hyperparameters. L denotes the combinatorial Laplacian, which is computed as $L = D - A$, where A denotes the adjacency matrix of the graph $\mathcal{G}$. D is a diagonal matrix with entries $d_{i,i} = \sum_j A_{i,j}$.

Although graph kernels have some successful applications to public transportation networks [15], there are probably limitations when applying the network-based kernels to the scenario of closed environments: the pedestrians in a train station or a shopping mall have favorite or commonly used routes, they are not randomly distributed on the networks. In a train station, the pedestrian flow on the main corridor is most likely unrelated to that on the corridors leading to the offices, even if the corridors are adjacent. To incorporate the information of the move preferences (trajectory patterns, collected from the local experts or tracking technology) into the model, we explore a graph kernel inspired with the diffusion process [25].

Assume that a pedestrian randomly moves on the edge graph $\mathcal{G}$. From a vertex i he jumps to a vertex j with $n^k_{i,j}$ possible random walks of length k, where $n^k_{i,j}$ is equal to $[A^k]_{i,j}$. Intuitively, the similarity of two vertices is related to the number and the length of the random walks between them. Based on diffusion process, the similarity between vertices v_i and v_j is defined as

$$s(v_i, v_j) = \left[\sum_{k=1}^{\infty}\frac{\lambda^k}{k!}A^k\right]_{ij}, \tag{5}$$

where $0 \leq \lambda \leq 1$ is a hyperparameter. All possible random walks between v_i and v_j are taken into account in similarity computation, however the contributions of

longer walks are discounted with a coefficient $\lambda^k/k!$. The similarity matrix is not always positive semi-definite. To get a valid kernel, the combinatorial Laplacian is used and the covariance matrix is defined as [25]:

$$K = \left[\sum_{k=1}^{\infty} \frac{\lambda^k}{k!} L^k\right] = \exp(\lambda L) \ . \tag{6}$$

On a traffic network within closed environment, the pedestrian will move not randomly, but with respect to a set of trajectory patterns and subpatterns denoted as sequences of vertices [26], e.g.,

$$\left\{ \begin{array}{l} T_1 = v_1 \rightarrow v_3 \rightarrow v_5 \rightarrow v_6, \\ T_2 = v_2 \rightarrow v_3 \rightarrow v_4, \\ T_3 = v_4 \rightarrow v_5 \rightarrow v_1, \\ \quad \ldots \end{array} \right\}. \tag{7}$$

Each trajectory pattern T_ℓ can also be represented as an adjacency matrix in which $\hat{A}_{i,j} = 1$ iff $v_i \rightarrow v_j \in T_\ell$ or $v_i \leftarrow v_j \in T_\ell$. The subpatterns are subsequences of the trajectories. For example, the subpatterns of T_1 are $\{v_1 \rightarrow v_3, v_3 \rightarrow v_5, v_5 \rightarrow v_6, v_1 \rightarrow v_3 \rightarrow v_5, v_3 \rightarrow v_5 \rightarrow v_6\}$. Given a set of trajectory patterns $\mathcal{T} = \{T_1, T_2, \ldots\}$, a random walk is valid and can be counted in similarity computation, if and only if all steps in the walk belong to $\mathcal{T}$ and subpatterns of $\mathcal{T}$. Thus we have

$$\begin{aligned} \hat{s}(v_i, v_j) &= \left[\sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \hat{A}^k\right]_{ij}, & \hat{K} &= \left[\sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \hat{L}^k\right] = \exp(\lambda \hat{L}) \\ \hat{A} &= \sum_{\ell} \hat{A}_\ell, & \hat{L} &= \hat{D} - \hat{A}, \end{aligned} \tag{8}$$

where $\hat{D}$ is a diagonal matrix with entries $\hat{d}_{i,i} = \sum_j \hat{A}_{i,j}$.

For pedestrian quantities $\mathbf{f}_u$ at unmeasured locations u, the predictive distribution can be computed as follows. Based on the property of GP, the observed and unobserved quantities $(\mathbf{y}, \mathbf{f}_u)^T$ follows a Gaussian distribution

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_u \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \hat{K}_{\overline{u},\overline{u}} + \sigma^2 I & \hat{K}_{\overline{u},u} \\ \hat{K}_{u,\overline{u}} & \hat{K}_{u,u} \end{bmatrix}\right), \tag{9}$$

where $\hat{K}_{u,\overline{u}}$ is the corresponding entries of $\hat{K}$ between the unmeasured vertices u and measured ones $\overline{u}$. $\hat{K}_{\overline{u},\overline{u}}$, $\hat{K}_{u,u}$, and $\hat{K}_{\overline{u},u}$ are defined equivalently. I is an identity matrix of size $|\overline{u}|$. Finally the conditional distribution of the unobserved pedestrian quantities is still Gaussian with the mean m and the covariance matrix Σ:

$$\begin{aligned} m &= \hat{K}_{u,\overline{u}} (\hat{K}_{\overline{u},\overline{u}} + \sigma^2 I)^{-1} \, \mathbf{y} \\ \Sigma &= \hat{K}_{u,u} - \hat{K}_{u,\overline{u}} (\hat{K}_{\overline{u},\overline{u}} + \sigma^2 I)^{-1} \, \hat{K}_{\overline{u},u} \ . \end{aligned}$$

Besides pedestrian quantity estimation, incorporating trajectory patterns also enables effectively finding sensor placements that are most informative for traffic estimation on the whole network. To identify the most informative locations $\mathcal{I}$, we employ the exploration strategy, maximizing mutual information [27]

$$\arg\max_{\mathcal{I}\subset\mathbf{V}} H(\mathbf{V}\backslash\mathcal{I}) - H(\mathbf{V}\backslash\mathcal{I} \mid \mathcal{I}) \; . \tag{10}$$

It is equal to find a set of vertices $\mathcal{I}$, which maximally reduces the entropy of the traffic at the unmeasured locations $\mathbf{V}\backslash\mathcal{I}$. Since the entropy and the conditional entropy of Gaussian variables can be completely specified with covariances, the selection procedure is only based on covariances of vertices, not involves any pedestrian quantity observations. To solve the optimization problem, we employ a poly-time approximate method [27]. In particular, starting from an empty set $\mathcal{I} = \emptyset$, each vertex is selected with the criterion:

$$v_* \leftarrow \arg\max_{v\in\mathbf{V}\backslash\mathcal{I}} \; H_\epsilon(v \mid \mathcal{I}) - H_\epsilon(v \mid \overline{\mathcal{I}}) \; , \tag{11}$$

where $\overline{\mathcal{I}}$ denotes the vertex set $\mathbf{V}\backslash(\mathcal{I} \cup v)$. $H_\epsilon(x|Z) := H(x|Z')$ denotes an approximation of the entropy $H(x|Z)$, where any element z in $Z' \subset Z$ satisfies the constraint that the covariance between z and x is larger than a small value ϵ. Within the GP framework, the approximate entropy $H_\epsilon(x|Z)$ is computed as

$$\begin{aligned} H_\epsilon(x \mid Z) &= \frac{1}{2} \ln 2\pi e \sigma^2_{x|Z'} \\ \sigma^2_{x|Z'} &= \hat{K}_{x,x} - \hat{K}^T_{x,Z'} \hat{K}^{-1}_{Z',Z'} \hat{K}_{x,Z'} \; . \end{aligned} \tag{12}$$

The term $\hat{K}_{x,Z'}$ is the corresponding entries of $\hat{K}$ between the vertex x and a set of vertices Z'. $\hat{K}_{x,x}$ and $\hat{K}_{Z',Z'}$ are defined equivalently. Given the informative trajectory pattern kernel, the pedestrian quantity observations at the vertices selected with the criterion (11) can well estimate the situation of the whole network. Sec. 4.3 shows a successful application to the zoo of Duisburg data.

4 Experimental Analysis

Our intention here is to investigate the following questions: (Q1) Can the proposed method integrate with expert knowledge on preferred movement patterns in closed environments and thus improve the prediction accuracy on pedestrian quantity estimation? (Q2) Can the proposed method choose sensor locations to better monitor pedestrian quantities in an industrial scenario? To this aim, we evaluate the method on two datasets: synthetic German train station pedestrian data and real-world episodic movement data collected with Bluetooth tracking technology at the zoo of Duisburg (Germany). We compare our method with state-of-the-art traffic volume estimation approaches. As discussed in Section 2, kNN methods are extensively used for traffic volume estimation [5]. The latest version of this approach, the *Spatial kNN* [7], has many successful applications

and is considered as a baseline. In the experiments we use distance-weighted 5-Nearest Neighbors. Particularly this method detects for each unmeasured edge the 5 nearest neighbors among the measured ones. Their traffic frequencies are weighted by distance to achieve prediction for the unmeasured ones. As this is a geometric algorithm which requires spatial representation of the traffic network, we apply Fruchterman Reingold [28] algorithm to lay out the test networks in two-dimensional space and achieve spatial representations. Distances between edges are computed with Euclidian metric. Additionally, we compare our method to GPR with commonly used kernels, including regularized Laplacian (RL), squared exponential (SE) and diffusion kernel (Diff). The prediction performance of the methods is measured with mean absolute error (MAE) $MAE = n^{-1} \sum_{i=1}^{n} |y_i - f_i|$.

4.1 German Train Station Data

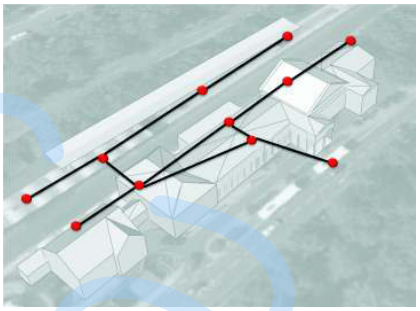

Fig. 2. Sketch of train station Hofheim (Germany) with traffic network overlay.

To approximate the true situation, we study traffic networks of 170 largest public train stations in Germany, an example shown as Fig. 2. The distributions of vertex-degree and vertex-number are visualized in Fig. 3. Given the collected information of the real-world train stations, the synthetic data is generated as follows. We apply the real vertex-degree distribution to the random network generator described in [29] and draw the train station like random graphs of order 10. In these graphs we generate pedestrian flows between dead ends (vertices of degree one), as no pedestrians permanently stay in a train station. The dead ends are selected pairwise and edge frequencies are sampled along the shortest connecting path with a random frequency of maximal 10,000 persons, which is a reasonable approximation for train station traffic networks. Afterwards we select a random set of edges (ranging from 10 to 50 percent of all edges) as monitored locations. Traffic frequencies at these edges are viewed as evidence to estimate frequencies at unmeasured ones. At each setting, we repeat the experiment 100 times and report the distributions of prediction performance for each method.

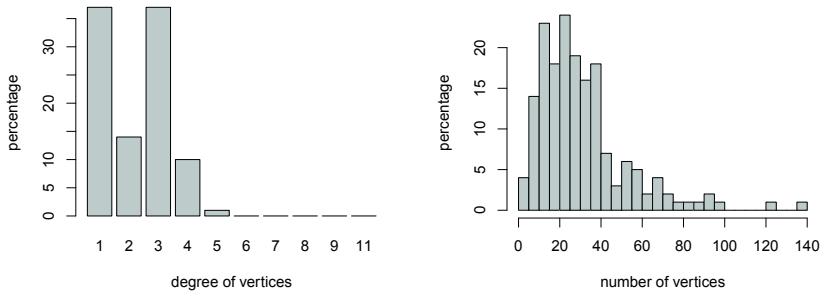

Fig. 3. Distributions of vertex-degree (right) and number of vertices (left) among 170 large German train stations.

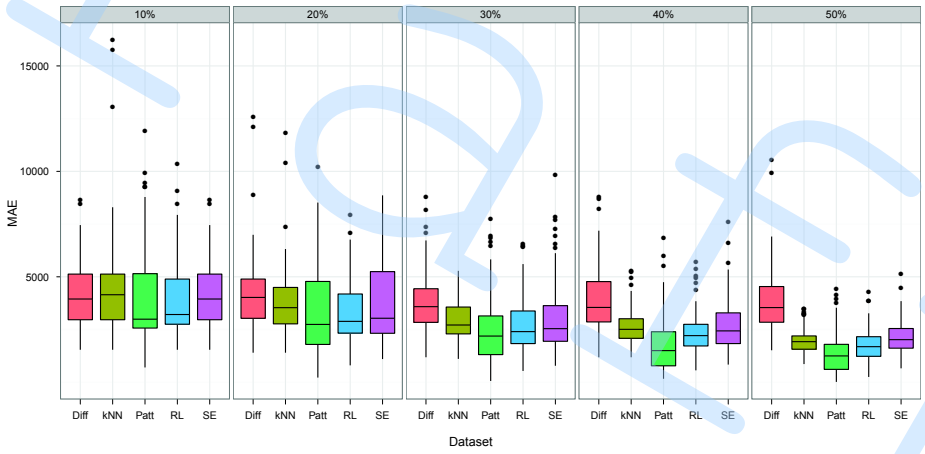

Fig. 4. Pedestrian quantity estimation on networks of train stations. Performance is measured by MAE at settings with different ratios of monitored edges (10 to 50 percent from left to right). The five methods: GPR with diffusion kernel (Diff), spatial k-nearest neighbor (kNN), GPR with trajectory pattern kernel (Patt), GPR with regularized Laplacian (RL) and GPR with squared exponential kernel (SE).

Experimental results are depicted in Fig. 4. Grouped in blocks are the different experiment configurations (different number of monitored edges). Statistics on the MAE distribution per method are depicted in the five boxplots. Throughout the tests, our method achieved minimal MAE and minimal average MAE, and therefore best results for the pedestrian quantity estimation problem. The proposed method outperformed commonly used kNN approach, especially when traffic networks are sparsely monitored. With increasing the number of monitored edges, all methods, except the GPR with diffusion kernel, provide better performance on pedestrian quantity estimation given that MAE decreased and did not scatter that much. Within the GP framework, the proposed trajectory pattern kernel achieved best performance compared to other kernels.

4.2 Zoo of Duisburg Data

Fig. 5. The network of the zoo in Duisburg (Germany) and the positions of the 15 Bluetooth scanners.

We apply the proposed method to a real world dataset of visitor movement in the zoo of Duisburg (Germany). The dataset consists of episodic movement data [19] and was collected with a mesh of 15 Bluetooth scanners [3,30] (see map in Fig. 5). Within a period of 7 days (07/26/11–08/02/11) all Bluetooth enabled devices (smartphones or intercoms) were scanned and log-entries attached to the log-file. Thus, the dataset consists of tuples (device identifier, timestamp, location). In order to perform the tests, the traffic network is build from the sensor positions. Each sensor becomes a vertex. To achieve ground truth for the traffic volume prediction, temporal aggregates of recorded transitions between sensors, as proposed in [19], become scaled by the Bluetooth representativity (in this case at the zoo approximately 6 percent). Due to the uncertainties in episodic movement data transitions in the dataset are not limited to neighbouring sensor positions, but occur between arbitrary pairs of sensors. In our case this results in a traffic network consisting of 102 edges and 15 vertices. The recorded trajectories of the zoo visitors become the trajectory pattern input to the *trajectory pattern kernel*. Similar to the previously synthetically generated data, the real

world experiments are conducted with different percentages of measured edges. Measurement edges are chosen uniformly at random 100 times for each dataset.

As shown in Figure 6 on the experimental results, the proposed method again achieved the best prediction performance for the pedestrian quantity estimation problem in comparison to other state-of-the-art methods. Incorporating expert knowledge on movement preferences allows for the model to well capture the dependencies of traffic at different edges and, in turn, to improve prediction accuracy.

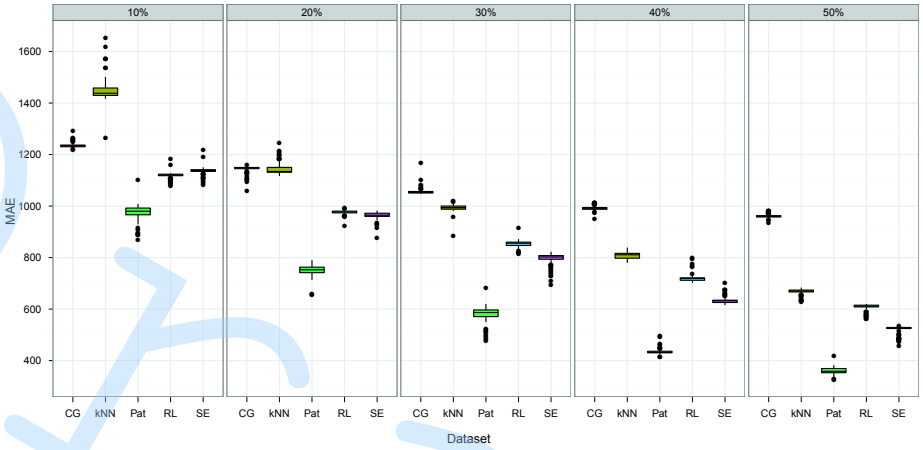

Fig. 6. Pedestrian quantity estimation on network of the zoo in Duisburg (Germany). Performance is measured by MAE at settings with different ratios of monitored edges (10 to 50 percent from left to right). The five methods: GPR with diffusion kernel (Diff), spatial k-nearest neighbor (kNN), GPR with trajectory pattern kernel (Patt), GPR with regularized Laplacian (RL) and GPR with squared exponential kernel (SE).

4.3 Sensor Placement with Trajectory Patterns

Besides the traffic volume estimation, another interesting task is to give a solution to the question where to place the sensors such that the traffic over the whole network can be well estimated. Based on the proposed trajectory pattern kernel, we perform the sensor placement procedure on the zoo of Duisburg data. Afterwards, pedestrian quantity estimation based on resulting sensor placement is carried out and performance is measured with MAE. The red horizontal line in Fig. 7 depicts the sensor placement performances in comparison to previous random placement. For sparse sensor distribution (low percentages of measurement edges), the sensor placement has a high positive impact on the prediction performance. However, for higher sensor numbers the random placement may outperform the mutual information based sensor placement. One reason is that

this placement is not optimal but near optimal. Another possible explanation is given by the data. Due to noise or other unexpected anomalies in the data which are not consistent to the prior knowledge on trajectory patterns.

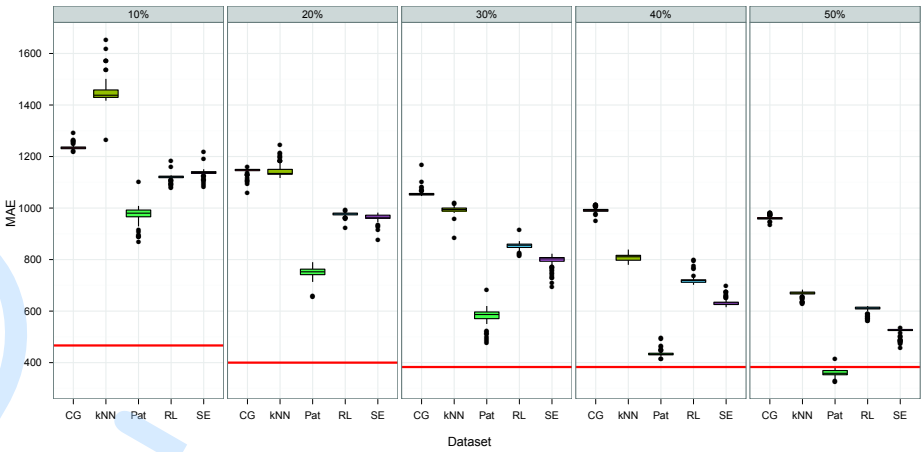

Fig. 7. Traffic Flow Estimation performance measured by MAE for 5 real world datasets with different ratios of known edges (10 to 50 percent) and five methods: GPR with Diffusion kernel (Diff), Spatial k-Nearest Neighbor (kNN), GPR with the proposed Trajectory Pattern kernel (Patt), GPR with Regularized Laplacian (RL) and GPR with Squared Exponential (SE) in comparison to (Patt) with mutual information based sensor placement (horizontal line).

5 Conclusions and Future Work

Pedestrian volume estimation is an important problem for mobility analysis in many application scenarios such as emergency support systems, quality-of-service evaluation and billboard placement, risk analysis and location ranking. This work proposed a nonparametric Bayesian method to tackle the pedestrian quantity estimation problem which explores the expert knowledge of trajectory patterns. We validated our proposed method on two datasets: synthetic German train station pedestrian data and real-world dataset collected with of Bluetooth tracking technology at the zoo of Duisburg. Furthermore, we addressed the question for sensor placement in an industrial scenario with the trajectory based graph kernel. The empirical analysis demonstrated that incorporating trajectory patterns can largely improve the traffic prediction accuracy in comparison to other state-of-the-art methods. Our work also provides an efficient and applicable solution to pedestrian volume estimation in industrial real world scenarios.

This work focussed on pedestrian volume estimation in closed environments (zoo, train station, terminal, etc.) because in closed environments different meth-

ods can be studied and compared under controlled circumstances. For instance, movements in these closed environments are not influenced by residing persons or unexpected pedestrian sinks or sources like tram stops, living houses, etc. Nevertheless, our proposed approach was not based on this assumption and future work should validate performance on arbitrary traffic networks. Another future research direction is to focus on temporal aspects of pedestrian movement and the creation of time dynamic models using at once dynamic expert knowledge and dynamic measurements. Also combination of measurements and expert knowledge at heterogeneous spatial granularities is promising for industrial applications (e.g. combination of (1) movement patterns among dedicated points of interest retrieved from social media and (2) pedestrian counts from video surveillance on (3) a city center traffic network). This question is of high interest in near future, as valuable (episodic) data on people's movement is expected to become widely available e.g. by billing data, logfiles on social media usage or wireless communication networks (GSM, WLAN, Bluetooth) [19].

Acknowledgements

This work was supported by the European Project Emergency Support System (ESS 217951) and the Fraunhofer ATTRACT Fellowship STREAM. We gratefully thank the zoo of Duisburg for supporting data collection.

References

1. Liebig, T., Körner, C., May, M.: Scalable sparse bayesian network learning for spatial applications. In: ICDM Workshops, IEEE Computer Society (2008) 420–425
2. Liebig, T., Körner, C., May, M.: Fast visual trajectory analysis using spatial bayesian networks. In: ICDM Workshops, IEEE Computer Society (2009) 668–673
3. Bruno, R., Delmastro, F.: Design and Analysis of a Bluetooth-based Indoor Localization System. In: 8th International Conference on Personal Wireless Communications. (2003) 711–725
4. Zhao, F., Park, N.: Using geographically weighted regression models to estimate annual average daily traffic. Journal of the Transportation Research Board **1879**(12) (2004) 99–107
5. Gong, X., Wang, F.: Three improvements on knn-npr for traffic flow forecasting. In: Proceedings of the 5th International Conference on Intelligent Transportation Systems, IEEE Press (2002) 736–740
6. Lam, W.H.K., Tang, Y.F., Tam, M.: Comparison of two non-parametric models for daily traffic forecasting in hong kong. Journal of Forecasting **25**(3) (2006) 173–192
7. May, M., Hecker, D., Körner, C., Scheider, S., Schulz, D.: A vector-geometry based spatial knn-algorithm for traffic frequency predictions. In: Data Mining Workshops, International Conference on Data Mining, Los Alamitos, CA, USA, IEEE Computer Society (2008) 442–447

8. Scheider, S., May, M.: A method for inductive estimation of public transport traffic using spatial network characteristics. In: Proceedings of the 10th AGILE International Conference on Geographic Information Sciences, Aalborg University (2007) 1–8
9. May, M., Scheider, S., Rösler, R., Schulz, D., Hecker, D.: Pedestrian flow prediction in extensive road networks using biased observational data. In: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems. GIS '08, New York, NY, USA, ACM (2008) 67
10. Wang, X., Kockelmann, K.M.: Forecasting network data: Spatial interpolation of traffic counts from texas data. Journal of the Transportation Research Board **2105**(13) (2009) 100–108
11. Liebig, T., Xu, Z.: Pedestrian monitoring system for indoor billboard evaluation. Journal of Applied Operational Research **4**(1) (2012) 28–36
12. Liebig, T.: A general pedestrian movement model for the evaluation of mixed indoor-outdoor poster campaigns. In: Proc. of the Third International Conference on Applied Operation Research - ICAOR'11, Tadbir Operational Research Group Ltd. (2011) 289–300
13. Liebig, T., Stange, H., Hecker, D., May, M., Körner, C., Hofmann, U.: A general pedestrian movement model for the evaluation of mixed indoor-outdoor poster campaigns. In: Proc. of the Third International Workshop on Pervasive Advertising and Shopping. (2010)
14. Kretz, T.: Pedestrian traffic: on the quickest path. Journal of Statistical Mechanics: Theory and Experiment **P03012** (mar 2009)
15. Neumann, M., Kersting, K., Xu, Z., Schulz, D.: Stacked gaussian process learning. In: Proceeding of the 9th IEEE International Conference on Data Mining (ICDM 2009), IEEE Computer Society (2009) 387–396
16. Li, M., Konomi, S., Sezaki, K.: Understanding and modeling pedestrian mobility of train-station scenarios. In Sabharwal, A., Karrer, R., Zhong, L., eds.: WINTECH, ACM (2008) 95–96
17. Harney, D.: Pedestrian modelling: current methods and future directions. Road Transport Research **11**(4) (2002) 38–48
18. Flyvbjerg, B., Skamris H., M.K., Buhl, S.L.: Inaccuracy in traffic forecasts. Transport Reviews **26**(1) (2006) 1–24
19. Andrienko, N., Andrienko, G., Stange, H., Liebig, T., Hecker, D.: Visual Analytics for Understanding Spatial Situations from Episodic Movement Data. KI - Künstliche Intelligenz (2012) 1–11
20. De Raedt, L.: Logical and Relational Learning. Springer (2008)
21. Getoor, L., Taskar, B., eds.: Introduction to Statistical Relational Learning. The MIT Press (2007)
22. Yu, K., Chu, W., Yu, S., Tresp, V., Xu, Z.: Stochastic relational models for discriminative link prediction. In: Neural Information Processing Systems. (2006)
23. Chu, W., Sindhwani, V., Ghahramani, Z., Keerthi, S.: Relational learning with gaussian processes. In: Neural Information Processing Systems. (2006)
24. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press (2006)
25. Kondor, R.I., Lafferty, J.D.: Diffusion kernels on graphs and other discrete input spaces. In: Proceeding of the International Conference on Machine Learning. (2002) 315–322
26. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proceedings of the Eleventh International Conference on Data Engineering, Washington, DC, USA, IEEE Computer Society (1995) 3–14

27. Krause, A., Guestrin, C., Gupta, A., Kleinberg, J.: Near-optimal sensor placements: maximizing information while minimizing communication cost. In: Proceedings of the 5th international conference on Information processing in sensor networks. IPSN '06, New York, NY, USA, ACM (2006) 2–10
28. Fruchterman, T.M.J., Reingold, E.M.: Graph Drawing by Force-directed Placement. Software: Practice and Experience **21**(11) (1991) 1129–1164
29. Viger, F., Latapy, M.: Efficient and Simple Generation of Random Simple Connected Graphs with Prescribed Degree Sequence. In Wang, L., ed.: Computing and Combinatorics. Volume 3595 of Lecture Notes in Computer Science. Springer Berlin/Heidelberg (2005) 440–449
30. Hallberg, J., Nilsson, M., Synnes, K.: Positioning with Bluetooth. In: 10th International Conference on Telecommunications. Volume 2. (2003) 954–958